SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 1–23

MAXIMUM PRINCIPLE OF OPTIMAL CONTROL FOR DEGENERATE QUASI-LINEAR ELLIPTIC EQUATIONS*

HONGWEI LOU†

Abstract. Optimal control problems governed by degenerate quasi-linear partial differential equations of elliptic type are considered. The optimal control systems considered may lack Cesari-type conditions, and therefore the corresponding approximate optimal control problem may have no solution. To yield the maximum principle of optimal pairs, relaxed controls are used to overcome the difficulties occurring when considering approximate problems. The relaxed controls used are defined by finite additive measures so that the case of the control set being noncompact can be treated.

Key words. optimal control, degenerate quasi-linear equation, necessary condition, finite additive measure, relaxed control

AMS subject classifications. 35J70, 49K20

PII. S0363012902401664

1. Introduction.

Let $\Omega \subset \mathbb{R}^n$ be a bounded domain. Consider the degenerate elliptic partial differential equation,

$$\begin{cases} -\operatorname{div}(\varphi(x,|\nabla y|)\frac{\nabla y}{|\nabla y|}) = f(x,y(x),u(x)) & \text{in } \Omega, \\ y|_{\partial\Omega} = 0, \end{cases} \tag{1.1}$$

and the cost functional

$$J(u(\cdot)) = \int_\Omega f^0(x,y(x),u(x))dx.$$

We assume the following:

(S1) Let $1 < p < +\infty$, $\Omega \subset \mathbb{R}^n$ be a bounded domain with $C^{1,1}$ boundary $\partial\Omega$, U be a separable metric space, and $\mathcal{U}_{ad} \equiv \{v : \Omega \to U | v \text{ is measurable}\}$.

(S2) Let $\varphi \in C(\bar{\Omega}\times[0,+\infty)) \bigcap C^1(\bar{\Omega}\times(0,+\infty))$ satisfy $\varphi(x,0) = 0$ for all $x \in \Omega$. Moreover, there exist $\Lambda > \lambda > 0$ such that for almost all $x \in \Omega$,

$$\lambda s^{p-2} \le \varphi_s(x,s) \le \Lambda s^{p-2} \qquad \forall s \in (0,+\infty), \tag{1.2}$$

$$\sum_{i=1}^n |\varphi_{x_i}(x,s)| \le \Lambda s^{p-1} \qquad \forall s \in (0,+\infty). \tag{1.3}$$

(S3) The function $f : \Omega \times \mathbb{R} \times U \to \mathbb{R}$ has the following properties: $f(\cdot,y,u)$ is measurable in Ω, $f(x,\cdot,u)$ is in $C^1(\mathbb{R})$. $f_y(x,\cdot,\cdot)$ and $f(x,\cdot,\cdot)$ are continuous in $\mathbb{R}\times U$. Moreover,

$$f_y(x,y,u) \le 0 \qquad \forall (x,y,u) \in \Omega \times \mathbb{R} \times U, \tag{1.4}$$

and for any $R > 0$, there exists a constant $M_R > 0$ such that

$$|f(x,y,u)| + |f_y(x,y,u)| \le M_R \qquad \forall (x,u) \in \Omega \times U,\ |y| \le R. \tag{1.5}$$

*Received by the editors January 30, 2002; accepted for publication (in revised form) October 21, 2002; published electronically March 19, 2003.
http://www.siam.org/journals/sicon/42-1/40166.html
†Mathematical Department, Fudan University, Shanghai 200433, China (hwlou@fudan.edu.cn).

(S4) The function $f^0 : \Omega \times \mathbb{R} \times U \to \mathbb{R}$ satisfies (S3) except for (1.4).

In our assumptions, U is not necessarily compact, which enables us to treat unbounded controls. The separability of U is used in applying the Filippov lemma and getting maximum conditions from that in integral form.

Our optimal control problem is as follows.

Problem (C). Find a $\bar{u}(\cdot) \in \mathcal{U}_{ad}$ such that

$$J(\bar{u}(\cdot)) = \inf_{u(\cdot)\in\mathcal{U}_{ad}} J(u(\cdot)).$$

The purpose of this paper is to give a Pontryagin maximum principle of an optimal control $\bar{u}(\cdot)$ to Problem (C).

Similar problems were considered by Casas and Fernández [2], [3] (see also [4]).

Because of the degeneracy of (1.1), to get necessary conditions of optimal controls, usually, one needs to consider approximate problems. In our case, no Cesari condition is assumed. Thus, the first difficulty we meet is that even if Problem (C) has an optimal control, the corresponding approximate optimal control problem may have no solution. On the other hand, when the approximate problems have optimal solutions, difficulties still occur in yielding maximum principle for Problem (C) from that for approximate problems. In [2], because of the speciality of the problem treated there, the first difficulty does not occur. Moreover, by using the convexity, the authors of [2] got strong convergence of optimal controls for approximate problems, obtaining the final results. To overcome the difficulties we just mentioned, we will consider relaxed controls. But another difficulty occurs when relaxed control is introduced. To ensure a sequence of relaxed controls having a subsequence converging weakly to a relaxed control, usually, we need to suppose that the control set U is compact (see [11], [21], and [25]). In this paper, we follow the track of Fattorini, who considered finite additive relaxed control in [9]. Such "relaxed controls" are different from those considered in [11], [21], and [25]. Using this new concept of "relaxed control," one is able to treat the case of U being a noncompact set.

Our main theorems are as follows.

THEOREM 1.1. *Suppose that* (S1)–(S4) *hold,* $1 < p < 2$. *Let* $(\bar{y}(\cdot), \bar{u}(\cdot)) \in W_0^{1,p}(\Omega) \times \mathcal{U}_{ad}$ *be an optimal state-control pair to Problem* (C). *Then, there exists a* $\bar{\psi}(\cdot) \in W_0^{1,2}(\Omega)$ *such that*

$$-\operatorname{div}\left\{\left[\frac{\varphi(x,|\nabla\bar{y}|)}{|\nabla\bar{y}|}\left(I-\frac{\nabla\bar{y}(\nabla\bar{y})^T}{|\nabla\bar{y}|^2}\right)+\varphi_s(x,|\nabla\bar{y}|)\frac{\nabla\bar{y}(\nabla\bar{y})^T}{|\nabla\bar{y}|^2}\right]\nabla\bar{\psi}\right\} \tag{1.6}$$

$$= f_y(x,\bar{y},\bar{u}(x))\bar{\psi} - f_y^0(x,\bar{y},\bar{u}(x)) \qquad \text{in } \{\nabla\bar{y}\neq 0\},$$

$$\nabla\bar{\psi}(x) = 0, \qquad a.e.\ x \in \{\nabla\bar{y} = 0\}, \tag{1.7}$$

and for almost all $x \in \Omega$,

$$H(x,\bar{y}(x),\bar{u}(x),\bar{\psi}(x)) = \max_{v\in U} H(x,\bar{y}(x),v,\bar{\psi}(x)), \tag{1.8}$$

where I *denotes the* $n \times n$ *identity matrix and*

$$H(x,y,v,\psi) = f(x,y,v)\psi - f^0(x,y,v) \tag{1.9}$$
$$\forall\ (x,y,v,\psi) \in \Omega \times \mathbb{R} \times U \times \mathbb{R}.$$

THEOREM 1.2. *Suppose that* (S1)–(S4) *hold,* $2 < p < +\infty$. *Moreover, suppose that there exists a constant* $\theta_0 > 0$ *such that*

$$f_y(x, y, u) \le -\theta_0 \qquad \forall\, (x, y, u) \in \Omega \times \mathbb{R} \times U. \tag{1.10}$$

Let $(\bar{y}(\cdot), \bar{u}(\cdot)) \in W_0^{1,p}(\Omega) \times \mathcal{U}_{ad}$ *be an optimal pair to Problem* (C). *Then, there exists a* $\bar{\psi}(\cdot) \in W_{loc}^{1,2}(\Omega)$ *satisfying* (1.6) *such that, for almost all* $x \in \{\nabla \bar{y} \neq 0\}$, (1.8) *holds, where* $W_{loc}^{1,2}(\Omega)$ *denotes the set of functions belonging to* $W^{1,2}(\Omega_0)$ *for any* $\Omega_0 \subset\subset \Omega$.

Because of the degeneracy of the state equation (1.1), usually, the conditions of the adjoint state $\bar{\psi}(\cdot)$ that we have given in Theorems 1.1 and 1.2 are not enough to determine it. Thus, in general, we can hardly determine the optimal controls from these necessary conditions (see also the similar theorems established in [2], [3]). Nevertheless, we can really use such conditions to find some important information about the optimal control. In [15], similar results were used to estimate the optimal controls, getting the regularities of optimal pairs.

In the case that $n = 1$ and $1 < p < 2$, our necessary conditions can be written in the form of equations of first order. Thus, in this case we are able to avoid using singular sets and we are sure that it keeps much important information of the optimal control.

THEOREM 1.3. *Suppose that* (S1)–(S4) *hold,* $1 < p < 2$, $\Omega = (a, b)$ *for some* $-\infty < a < b < +\infty$. *Let* $(\bar{y}(\cdot), \bar{u}(\cdot)) \in W_0^{1,p}(a, b) \times \mathcal{U}_{ad}$ *be an optimal pair to Problem* (C). *Then, there exist a* $\bar{\Psi}(\cdot) \in W^{1,+\infty}(a, b)$ *and a* $\bar{\psi}(\cdot) \in W_0^{1,2}(a, b) \bigcap W^{1,+\infty}(a, b)$ *such that*

$$\begin{cases} \bar{\Psi}'(x) = f_y(x, \bar{y}, \bar{u})\bar{\psi} - f_y^0(x, \bar{y}, \bar{u}), & a < x < b, \\ \bar{\psi}'(x) = -\dfrac{1}{\varphi_s(x, |\bar{y}'(x)|)}\bar{\Psi}(x), & a < x < b, \\ \bar{\psi}(a) = \bar{\psi}(b) = 0, \end{cases} \tag{1.11}$$

and for almost all $x \in (a, b)$

$$H(x, \bar{y}(x), \bar{u}(x), \bar{\psi}(x)) = \max_{v \in U} H(x, \bar{y}(x), v, \bar{\psi}(x)). \tag{1.12}$$

2. Classical control and relaxed control. In this section, we recall the concept of relaxed control and the relation between classical controls and relaxed controls. The concept of relaxed control can be traced back to the work of Young and McShane [27], [28], [29], [18], [19], [20] and their generalized curves (see also [22], [30]). Due to the development of measure theory and theory of generalized functions, McShane [21], Warga [25], and Gamkrelidze [11] gave the concept a new expression called "relaxed control" that is easier to understand than its archetype. In [9], to treat the case of the control set U being noncompact, Fattorini gave a new definition of relaxed control based on finite additive measures on U.

Now, let us recall some basic notions. Let Φ be a family of subsets of U. Φ is called a field of U if it contains the empty set $\emptyset$, the complement of each of its members, and the union of its two elements. Let $\mathcal{F}$ be the field generated by the closed sets of U, that is, $\mathcal{F}$ is the smallest field containing all closed subsets of U. A set function μ defined on $\mathcal{F}$ is called a finitely additive measure on U if $\mu(\emptyset) = 0$ and $\mu(A \bigcup B) = \mu(A) + \mu(B)$ for disjoint $A, B \in \mathcal{F}$ (see [8, p. 96]), where $\mu(A)$ can be a real number, $-\infty$ or $+\infty$. The total variation $|\mu|$ is defined by

$$|\mu|(A) \equiv \sup \sum_{j=1}^{m} |\mu(A_j)|,$$

where the supremum is taken over all finite sequences $\{A_j\}$ of disjoint sets in $\mathcal{F}$ with $A_j \subseteq A$. If

$$\|\mu\| \equiv |\mu|(U) < +\infty,$$

then μ is said to be of bounded variation. A finitely additive measure is called regular if for any $A \in \mathcal{F}$ and $\varepsilon > 0$, there exists two sets $B, D \in \mathcal{F}$ such that $\mathrm{cl}(B) \subseteq A \subseteq \mathrm{int}(D)$ (here $\mathrm{cl}(B)$ and $\mathrm{int}(D)$ denote the closure of B and the interior of D, respectively), and $|\mu|(E) < \varepsilon$ for any set $E \in \mathcal{F}$ with $E \subseteq D \setminus B$. Let $\mathcal{M}(U)$ be the space of all regular bounded finitely additive measures on U (μ is bounded if $\sup_{A\in\mathcal{F}} |\mu(A)| < +\infty$, or equivalently $\|\mu\| < +\infty$). Then $\mathcal{M}(U)$ is a Banach space under norm $\|\cdot\|$. Let $C(U)$ be the space of all continuous bounded functions defined on U with its supremum norm $\|\cdot\|_{C(U)}$. Then we have (see [8, Theorem IV.6.2]) the following lemma.

LEMMA 2.1. *The dual space $C(U)^*$ of $C(U)$ is isometrically isomorphic to $\mathcal{M}(U)$. An element $\mu \in \mathcal{M}(U)$ acts on elements $g(\cdot) \in C(U)$ in the form*

$$\langle \mu, g \rangle \equiv \int_U g(v)\mu(dv). \tag{2.1}$$

For an integration theory based on finitely additive measure, see [8, III.2]. An important difference between an integral based on finitely additive measures and that based on countably additive measures is that for a finitely additive measure μ and an μ-integrable nonnegative function $g(\cdot)$ on U,

$$\int_U g(v)\ \mu(dv) = 0$$

does not means $g = 0$, μ a.e. U.

Let $L^\infty(\Omega; \mathcal{M}(U))_w$ be the space of all $\mathcal{M}(U)$-valued $C(U)$-weakly measurable functions which are bounded almost everywhere. That is, $\sigma(\cdot) \in L^\infty(\Omega; \mathcal{M}(U))_w$ if

$$\sigma(x) \in \mathcal{M}(U) \qquad \forall\ x \in \Omega,$$

$$x \mapsto \langle \sigma(x), g \rangle \qquad \text{is measurable} \qquad \forall\ g \in C(U),$$

and there exists a constant $C > 0$ such that for almost all $x \in \Omega$,

$$\Big|\langle \sigma(x), g \rangle\Big| \le C\|g\|_{C(U)} \qquad \forall\ g \in C(U).$$

An element $\mu \in \mathcal{M}(U)$ is said to be a probability measure on U if μ is nonnegative and $\mu(U) = 1$. Let $\mathcal{M}^1_+(U)$ be the set of all probability measures on U, and let $\mathcal{R}(\Omega; U)$ be the set of all $\mathcal{M}^1_+(U)$-valued $C(U)$-weakly measurable functions in Ω, that is, $\sigma(\cdot) \in \mathcal{R}$ if

$$\sigma(x) \in \mathcal{M}^1_+(U) \qquad \forall\ x \in \Omega,$$

and

$$x \mapsto \int_U g(v)\ \sigma(x)(dv) \qquad \text{is measurable} \qquad \forall\ g \in C(U).$$

Any member of $\mathcal{R}(\Omega; U)$ will be called a relaxed control. Respectively, an element of $\mathcal{U}_{ad}$ is called a classical control.

By Lemma 2.1 and a discussion similar to that in the proof of Theorems 2.1 and 2.3 in [9] (see also [6] and [7]), we can get the following.

LEMMA 2.2. *The dual space $L^1(\Omega; C(U))^*$ of $L^1(\Omega; C(U))$ is isometrically isomorphic to $L^\infty(\Omega; \mathcal{M}(U))_w$.*

Thus, $\mathcal{R}(\Omega; U) \subset L^\infty(\Omega; \mathcal{M}(U))_w$ can be looked at as a subset of $L^1(\Omega; C(U))^*$ by setting

$$\langle \sigma, h\rangle \equiv \int_\Omega dx \int_U h(x,v)\, \sigma(x)(dv) \qquad \forall\, h \in L^1(\Omega, C(U)),\ \sigma \in \mathcal{R}(\Omega; U).$$

Moreover, using the Banach–Alaoglu theorem, we have the following.

LEMMA 2.3. *Suppose $\Omega \subset \mathbb{R}^n$ is a bounded domain and U is a metric space. Then for any sequence $\sigma_k(\cdot) \in \mathcal{R}(\Omega; U)$, there exists a $\sigma(\cdot) \in \mathcal{R}(\Omega; U)$ such that for any $h \in L^1(\Omega; C(U))$, there exists a subsequence $\sigma_{k_j}(\cdot)$ of $\sigma_k(\cdot)$ satisfying*

$$\langle \sigma_{k_j}, h\rangle = \int_\Omega dx \int_U h(x,v)\sigma_{k_j}(x)(dv) \to \langle \sigma, h\rangle.$$

For convenience, in this paper, the above property will be denoted as

$$\sigma_k(\cdot) \xrightarrow{\mathcal{N}} \sigma(\cdot) \qquad \text{in } \mathcal{R}(\Omega; U).$$

We note that in Lemma 2.3, the choice of the subsequence $\sigma_{k_j}(\cdot)$ is dependent of h. On the other hand, when U is compact, $L^1(\Omega; C(U))$ is separable. Consequently, $\sigma_{k_j}(\cdot)$ can be chosen independently of h in this case (see [11, Chapter 8] or [25, Chapter 4]). The following lemma shows that relaxed control can be approximated by classical control.

LEMMA 2.4. *Let* (S1)–(S4) *hold, and let $\sigma(\cdot) \in \mathcal{R}(\Omega; U)$. Then for every finite set $\{h_i\} \subset L^\infty(\Omega; C(U))$, $i = 1, \dots, N$, and any $\delta > 0$, there exist $u_1(\cdot), \dots, u_{N+1}(\cdot) \in \mathcal{U}_{ad}$ and nonnegative measurable functions $\alpha_1(\cdot), \dots, \alpha_{N+1}(\cdot)$ such that*

$$\sum_{i=1}^{N+1} \alpha_i(\cdot) = 1, \tag{2.2}$$

and for any $k = 1, \dots, N$,

$$\left\| \int_U h_k(\cdot, v)\sigma(\cdot)(dv) - \sum_{i=1}^{N+1} \alpha_i(\cdot) h_k(\cdot, u_i(\cdot)) \right\|_{L^\infty(\Omega)} \le \delta. \tag{2.3}$$

Moreover, there exists $\{u^l(\cdot)\} \subset \mathcal{U}_{ad}$ such that for any $k = 1, \dots, N$ and $h^0(\cdot) \in C(\bar\Omega)$,

$$\int_\Omega h^0(x) h_k(x, u^l(x))\, dx \to \int_\Omega dx \int_U h^0(x) h_k(x,v)\sigma(x)(dv) \quad \text{as } l \to +\infty. \tag{2.4}$$

Proof. For any $x \in \Omega$, we set

$$\tilde h_i(x) = \int_U h_i(x,v)\sigma(x)(dv), \qquad i = 1, \dots, N,$$

and

$$\mathcal{E}(x) = \{(h_1(x,v), \dots, h_N(x,v)) | v \in U\}.$$

Then $\mathcal{E}(x)$ is bounded. Moreover, it is easy to see that

$$(\tilde{h}_1(x), \dots, \tilde{h}_N(x)) \in \bar{co}\, \mathcal{E}(x),$$

since $\sum_{i=1}^N a_i \tilde{h}_i(x) + c \le 0$ for any $(a_1, \dots, a_N, c) \in \mathbb{R}^{N+1}$ such that

$$\sum_{i=1}^N a_i z_i + c \le 0 \qquad \forall\, (z_1, \dots, z_N) \in \mathcal{E}(x),$$

where $\bar{co}(A)$ denotes the closed convex hull of set $A \subset \mathbb{R}^N$. Therefore, by the Carathéodory theorem (see [25, p. 139], for example), for any $x \in \Omega$, we have $\hat{\alpha}_i(x) \ge 0$ and $z^i = (z_{i1}, \dots, z_{iN}) \in \bar{co}\, \mathcal{E}(x)$ $(i = 1, \dots, N+1)$ such that

$$\sum_{i=1}^{N+1} \hat{\alpha}_i(x) = 1,$$

and

$$(\tilde{h}_1(x), \dots, \tilde{h}_N(x)) = \sum_{i=1}^{N+1} \hat{\alpha}_i(x) z^i.$$

Consequently, there exist $\hat{u}_i(x) \in U$ $(i = 1, \dots, N+1)$ such that

$$\left| \tilde{h}_k(x) - \sum_{i=1}^{N+1} \hat{\alpha}_i(x) h_k(x, \hat{u}_i(x) \right| \le \delta, \qquad k = 1, \dots, N.$$

Let

$$X = \left\{ (\alpha_1, \dots, \alpha_{N+1}) \in \mathbb{R}^{N+1} | \alpha_i \ge 0, 1 \le i \le N+1, \sum_{k=1}^{N+1} \alpha_k = 1 \right\}.$$

Consider the multifunction $\Gamma : \Omega \to 2^{X \times U^{N+1}}$ defined by

$$\Gamma(x) = \Bigg\{ (\alpha_1, \dots, \alpha_{N+1}, v_1, \dots, v_{N+1}) \in X \times U^{N+1} | \\ |\tilde{h}_k(x) - \sum_{i=1}^{N+1} \alpha_i h_k(x, v_i)| \le \delta, k = 1, \dots, N \Bigg\}.$$

Then $\Gamma(\cdot)$ is measurable and takes nonempty closed set values. By Theorem 2.23 in [13, Chapter 3] (see also [10]), we have measurable functions $\alpha_i(\cdot) : \Omega \to [0, 1]$ and $u_i(\cdot) \in \mathcal{U}_{ad}$ $(i = 1, \dots, N+1)$ such that

$$\sum_{i=1}^{N+1} \alpha_i(x) = 1,$$

and

$$\left| \tilde{h}_k(x) - \sum_{i=1}^{N+1} \alpha_i(x) h_k(x, u_i(x)) \right| \le \delta, \qquad k = 1, \dots, N. \tag{2.5}$$

That is, (2.3) holds.

Now, we turn to prove (2.4). By (2.3), it suffices to prove that there exists a sequence $u^l(\cdot) \in \mathcal{U}_{ad}$ such that for any $h(\cdot,\cdot) \in L^1(\Omega; C(U))$,

$$\lim_{l\to+\infty} \int_\Omega h(x, u^l(x))\, dx = \int_\Omega \sum_{i=1}^{N+1} \alpha_i(x) h(x, \hat{u}_i(x))\, dx. \tag{2.6}$$

We will prove that $u^l(\cdot)$ can be chosen as

$$u^l(x) = \hat{u}_i(x), \qquad x \in \Omega_i^l,\ i = 1, \dots, N+1,$$

where

$$\begin{aligned} &\Omega_i^l \text{ is measurable}, \qquad i = 1, \dots, N+1, \\ &\Omega_i^l \bigcap \Omega_j^l = \emptyset \qquad \text{if } i \neq j, \end{aligned}$$

and

$$\bigcup_{i=1}^{N+1} \Omega_i^l = \Omega.$$

That is, we want to prove that

$$\lim_{l\to+\infty} \sum_{i=1}^{N+1} \int_{\Omega_i^l} h(x, \hat{u}_i(x))\, dx = \int_\Omega \sum_{i=1}^{N+1} \alpha_i(x) h(x, \hat{u}_i(x))\, dx. \tag{2.7}$$

To prove this, it is enough to prove that for any $g_i(\cdot) \in C(\bar{\Omega})$ $(i = 1, \dots, N+1)$,

$$\sum_{i=1}^{N+1} \int_{\Omega_i^l} g_i(x)\, dx \to \int_\Omega \sum_{i=1}^{N+1} g_i(x)\alpha_i(x)\, dx. \tag{2.8}$$

Now, we construct the sets Ω_i^l.

For each $l \geq 1$, we have $\Omega = \sum_{j=1}^{N_l} Q_j^l$, where

$$\begin{aligned} &Q_j^l \text{ is measurable}, \qquad j = 1, 2, \dots, N_l,\ l = 1, 2, 3, \dots, \\ &Q_j^l \bigcap Q_m^l = \emptyset \qquad \text{if } j \neq m, \end{aligned}$$

and

$$0 < diam\,(Q_j^l) \leq \frac{1}{l}, \qquad j = 1, 2, \dots, N_l,\ l = 1, 2, 3, \dots.$$

Furthermore, we can decompose Q_j^l as

$$Q_j^l = Q_{1,j}^l \bigcup \cdots \bigcup Q_{N+1,j}^l,$$

with $Q_{i,j}^l$ being measurable,

$$Q_{i,j}^l \bigcap Q_{m,j}^l = \emptyset \qquad \text{if } i \neq m,$$

and

$$|Q_{i,j}^l| = \int_{Q_j^l} \alpha_i(x)\, dx.$$

Let $\Omega_i^l = \bigcup_{j=1}^{N_l} Q_{i,j}^l$. Then, for any $g_i(\cdot) \in C(\bar{\Omega})$ $(i = 1, \dots, N+1)$, we get (2.8) from the following:

$$\begin{aligned}
&\int_\Omega \sum_{i=1}^{N+1} g_i(x)\alpha_i(x)\, dx = \sum_{i=1}^{N+1}\sum_{j=1}^{N_l} \int_{Q_j^l} g_i(x)\alpha_i(x)\, dx \\
&= \sum_{i=1}^{N+1}\sum_{j=1}^{N_l} g_i(x_{l,j}) \int_{Q_j^l} \alpha_i(x)\, dx + O\left(\omega\left(\frac{1}{l}\right)\right) \qquad \text{(where } x_{l,j} \in Q_j^l\text{)} \\
&= \sum_{i=1}^{N+1}\sum_{j=1}^{N_l} g_i(x_{l,j})|Q_{i,j}^l| + O\left(\omega\left(\frac{1}{l}\right)\right) = \sum_{i=1}^{N+1}\sum_{j=1}^{N_l} \int_{Q_{i,j}^l} g_i(x)\, dx + O\left(\omega\left(\frac{1}{l}\right)\right) \\
&= \sum_{i=1}^{N+1} \int_{\Omega_i^l} g_i(x)\, dx + O\left(\omega\left(\frac{1}{l}\right)\right),
\end{aligned}$$

where

$$\omega(r) \equiv \max_{|x-\tilde{x}|\le r} \sum_i |g_i(x) - g_i(\tilde{x})| \to 0 \qquad \text{as } r \to 0^+.$$

Consequently, we get (2.8). By this, we can obtain (2.4) easily and complete the proof. □

The relaxed problem corresponding to Problem (C) is the following.

Problem (R). Find a $\bar{\sigma}(\cdot) \in \mathcal{R}(\Omega; U)$ such that

$$J(\bar{\sigma}(\cdot)) = \inf_{\sigma(\cdot)\in\mathcal{R}(\Omega;U)} J(\sigma(\cdot)),$$

where

$$J(\sigma(\cdot)) = \int_\Omega dx \int_U f^0(x, y(x), v)\sigma(x)(dv), \tag{2.9}$$

and $y(\cdot) \equiv y(\cdot; \sigma(\cdot))$ is the solution of the following equation:

$$\begin{cases} -\operatorname{div}\left(\varphi(x, |\nabla y|)\dfrac{\nabla y}{|\nabla y|}\right) = \displaystyle\int_U f(x, y(x), v)\sigma(x)(dv) \quad \text{in } \Omega, \\ y|_{\partial\Omega} = 0. \end{cases} \tag{2.10}$$

It is easy to verify that the maps $(x, y, \sigma) \to \int_U f(x, y, v)\sigma(dv)$ and $(x, y, \sigma) \to \int_U f^0(x, y, v)\sigma(dv)$ have the analogous (S3) and (S4) properties. Moreover, by identifying $u(\cdot) \in \mathcal{U}_{ad}$ with Dirac measure-valued function $\delta_{u(\cdot)} \in \mathcal{R}(\Omega; U)$, $\mathcal{U}_{ad}$ can be looked upon as a subset of $\mathcal{R}(\Omega; U)$. On the other hand, if for some $\sigma(\cdot) \in \mathcal{R}(\Omega; U)$, $\sigma(x) = \delta_{u(x)}$ for almost all $x \in \Omega$, then $u(\cdot)$ must be measurable, that is, $u(\cdot) \in \mathcal{U}_{ad}$. It is easy to see that $J(\delta_{u(\cdot)}), y(\cdot; \delta_{u(\cdot)})$ defined by (1.2), (1.1) are equal to $J(u(\cdot)), y(\cdot; u(\cdot))$ defined by (2.9), (2.10), respectively. Therefore, the notation $J(\sigma(\cdot))$ would not cause any confusion. If an optimal relaxed control $\bar{\sigma}(\cdot)$ corresponding to

Problem (R) has the form $\delta_{\bar{u}(\cdot)}$, then $\bar{u}(\cdot)$ must be an optimal classical control corresponding to Problem (C). This idea can be used to get the existence theorem of an optimal classical control (see [16], [17] for such results in semilinear cases).

When using relaxed control to get the Pontryagin maximum principle for an optimal classical control, one needs to prove that

$$\inf_{u(\cdot)\in\mathcal{U}_{ad}} J(u(\cdot)) = \inf_{\sigma(\cdot)\in\mathcal{R}(\Omega;U)} J(\sigma(\cdot)). \tag{2.11}$$

In the next section, we will prove the above relationship and give some basic properties of states.

3. Approximation lemma. In this section, we want to prove (2.11). First, let us state a lemma which shows that (2.10) (as well as (1.2)) is well posed for any $\sigma(\cdot) \in \mathcal{R}(\Omega; U)$.

LEMMA 3.1. *Let* (S1)–(S3) *hold. Then for any* $\varepsilon \in [0,1]$, $\sigma(\cdot) \in \mathcal{R}(\Omega;U)$, *there exists a unique weak solution* $y_\varepsilon(\cdot) \equiv y_\varepsilon(\cdot;\sigma(\cdot)) \in W_0^{1,p}(\Omega)\bigcap L^\infty(\Omega)$ *of the following equation:*

$$\begin{cases} -\operatorname{div}\left(\varphi(x,\sqrt{\varepsilon^2+|\nabla y_\varepsilon|^2})\dfrac{\nabla y_\varepsilon}{\sqrt{\varepsilon^2+|\nabla y_\varepsilon|^2}}\right) \\ \qquad = \displaystyle\int_U f(x,y_\varepsilon(x),v)\sigma(x)(dv) \quad \text{in } \Omega, \\ y_\varepsilon|_{\partial\Omega} = 0. \end{cases} \tag{3.1}$$

Moreover, there exists a constant $C>0$, *independent of* $\varepsilon \in [0,1]$ *and* $\sigma(\cdot) \in \mathcal{R}(\Omega;U)$, *such that*

$$\|y\|_{L^\infty(\Omega)} \le C. \tag{3.2}$$

Hereafter, $y_\varepsilon(\cdot;\sigma(\cdot))$ denotes the solution of (3.1) corresponding to $\sigma(\cdot)$. The existence of a solution in the above lemma can be proved by the Schauder fixed point theorem, while (3.2) can be obtained by the De Giorgi–Moser estimate. We omit the proofs since it is similar to that of Theorem 6.11 in [13, p. 78].

The following lemma is a special case of Theorem 1 in [14], which shows the regularity of solutions to quasi-linear equations.

LEMMA 3.2. *Let* (S1)–(S2) *hold,* $\varepsilon \in [0,1]$ *and* $v(\cdot) \in L^\infty(\Omega)$. *Suppose that* $y_\varepsilon(\cdot) \in W_0^{1,p}(\Omega)$ *is the solution of the following equation:*

$$\begin{cases} -\operatorname{div}\left(\varphi(x,\sqrt{\varepsilon^2+|\nabla y_\varepsilon|^2}\,)\frac{\nabla y_\varepsilon}{\sqrt{\varepsilon^2+|\nabla y_\varepsilon|^2}}\right) = v \qquad \text{in } \Omega, \\ y_\varepsilon|_{\partial\Omega} = 0. \end{cases} \tag{3.3}$$

Then there exists a constant $C>0$ *and* $\alpha \in (0,1)$ *dependent only on* $p, \lambda, \Lambda, \Omega$ *and the upper bound of* $\|v\|_{L^\infty(\Omega)}$, *independent of* $\varepsilon \in [0,1]$, *such that* $y_\varepsilon(\cdot) \in C^{1,\alpha}(\bar{\Omega})$ *and*

$$\|y_\varepsilon\|_{C^{1,\alpha}(\bar{\Omega})} \le C. \tag{3.4}$$

By Lemma 3.1, for any $\sigma(\cdot) \in \mathcal{R}(\Omega;U)$, the corresponding solution $y_\varepsilon(\cdot;\sigma(\cdot))$ to (3.1) is bounded uniformly in $L^\infty(\Omega)$. Thus, by (S3), for any $\sigma(\cdot) \in \mathcal{R}(\Omega;U)$, the right-hand term in (3.1) is bounded uniformly in $L^\infty(\Omega)$. Then, using Lemma 3.2, we get the following proposition.

PROPOSITION 3.3. *Let* (S1)–(S3) *hold. Then there exists a constant* $C > 0$ *and* $\alpha \in (0,1)$*, independent of* $\varepsilon \in [0,1]$ *and* $\sigma(\cdot) \in \mathcal{R}(\Omega; U)$*, such that*

$$\|y_\varepsilon(\cdot;\sigma(\cdot))\|_{C^{1,\alpha}(\bar\Omega)} \leq C. \tag{3.5}$$

LEMMA 3.4. *Let* $1 < p < +\infty$*,* $\psi(\cdot) \in C[0,+\infty) \bigcap C^1(0,+\infty)$ *satisfy* $\psi(0) = 0$*. Moreover, there exist* $\Lambda > \lambda > 0$ *such that*

$$\lambda s^{p-2} \leq \psi'(s) \leq \Lambda s^{p-2} \qquad \forall\ s \in (0,+\infty). \tag{3.6}$$

Then, there exists a constant $C = C(p,\lambda,\Lambda) > 0$ *such that for any* $a, b \in \mathbb{R}^m$*,*

$$\left(\psi(|a|)\frac{a}{|a|} - \psi(|b|)\frac{b}{|b|}\right)\cdot(a-b) \geq \begin{cases} C|a-b|^p & \text{if } p \geq 2, \\ C\frac{|a-b|^2}{(|a|+|b|)^{2-p}} & \text{if } 1 < p < 2. \end{cases} \tag{3.7}$$

The above lemma can be obtained with the same argument to that of Lemma 1 in [24] by taking $a_j(x,\eta) = \psi(|\eta|)\frac{\eta_j}{|\eta|}$ and $\kappa = 0$. (This last fact implies that the inequality holds without the term 1, in the case $1 < p < 2$.)

Now, we give the main result of this section.

LEMMA 3.5. *Let* (S1)–(S4) *hold. Then for any* $\sigma(\cdot) \in \mathcal{R}(\Omega; U)$*, there exists* $\{u_k(\cdot)\} \subset \mathcal{U}_{ad}$ *such that*

$$J(u_k(\cdot)) \to J(\sigma(\cdot)) \qquad \text{as } k \to +\infty. \tag{3.8}$$

Consequently, (2.11) *holds.*

Proof. Denote $y(\cdot) = y(\cdot;\sigma(\cdot))$ and

$$(h_1(x,v), h_2(x,v)) = (f(x,y(x),v), f^0(x,y(x),v)).$$

Then $h_i \in L^\infty(\Omega; C(U))$ by Lemma 3.1 and (S3)–(S4). Thus, by Lemma 2.4, there exists $\{u_k(\cdot)\} \subset \mathcal{U}_{ad}$ such that

$$\lim_{k\to+\infty}\int_\Omega h^0(x) f(x,y(x),u_k(x))\,dx = \int_\Omega dx \int_U h^0(x) f(x,y(x),v)\sigma(x)(dv) \qquad \forall\ h^0(\cdot) \in C(\bar\Omega) \tag{3.9}$$

and

$$\lim_{k\to+\infty}\int_\Omega f^0(x,y(x),u_k(x))\,dx = \int_\Omega dx \int_U f^0(x,y(x),v)\sigma(x)(dv). \tag{3.10}$$

Then it follows from $y(\cdot) = y(\cdot;\sigma(\cdot))$ that

$$\begin{cases} -\operatorname{div}\left(\varphi(x,|\nabla y|)\dfrac{\nabla y}{|\nabla y|}\right) = f(x,y(x),u_k(x)) + r_k(x) & \text{in } \Omega, \\ y|_{\partial\Omega} = 0, \end{cases} \tag{3.11}$$

where

$$\lim_{k\to+\infty}\int_\Omega h^0(x) r_k(x)\,dx = 0 \qquad \forall\ h^0(\cdot) \in C(\bar\Omega). \tag{3.12}$$

Let $y_k(\cdot) = y(\cdot; u_k(\cdot))$, i.e.,

$$\begin{cases} -\operatorname{div}\left(\varphi(x, |\nabla y_k|)\dfrac{\nabla y_k}{|\nabla y_k|}\right) = f(x, y_k(x), u_k(x)) & \text{in } \Omega, \\ y_k|_{\partial\Omega} = 0. \end{cases} \tag{3.13}$$

By Proposition 3.3, $y_k(\cdot)$ is uniformly bounded in $C^{1+\alpha}(\bar{\Omega})$. Thus, at least in the sense of a subsequence, we can suppose that

$$y_k(\cdot) \to \tilde{y}(\cdot) \qquad \text{uniformly in } \ C^1(\bar{\Omega}).$$

Thus, by (3.11) and noting that

$$|f(x, y(x), u_k(x))| \le C,$$

we have

$$\begin{aligned} &\int_\Omega r_k(x)(y_k(x) - \tilde{y}(x))\, dx \\ &= \int_\Omega \varphi(x, |\nabla y|)\frac{\nabla y}{|\nabla y|} \cdot \nabla(y_k - \tilde{y})\, dx - \int_\Omega (y_k - \tilde{y}) f(x, y(x), u_k(x))\, dx \\ &\to 0 \qquad \text{as } k \to +\infty. \end{aligned}$$

Thus, by (3.11)–(3.13) and (1.4), we get

$$\begin{aligned} &\int_\Omega \left(\varphi(x, |\nabla y|)\frac{\nabla y}{|\nabla y|} - \varphi(x, |\nabla y_k|)\frac{\nabla y_k}{|\nabla y_k|}\right) \cdot \nabla(y - y_k)\, dx \\ &\le \int_\Omega r_k(x)(y(x) - y_k(x))\, dx \\ &\le \int_\Omega r_k(x)(y(x) - \tilde{y}(x))\, dx + \int_\Omega r_k(x)(\tilde{y}(x) - y_k(x))\, dx \\ &\to 0 \qquad \text{as } k \to +\infty. \end{aligned}$$

Then by Lemma 3.4, we get that

$$\tilde{y}(x) = y(x), \qquad \text{a.e. } \Omega.$$

Therefore,

$$y_k(\cdot) \to y(\cdot) \qquad \text{uniformly in } \ C^1(\bar{\Omega}),$$

not only in the sense of a subsequence. Finally, by (3.10),

$$\begin{aligned} J(u_k(\cdot)) &= \int_\Omega f^0(x, y_k(x), u_k(x))\, dx \\ &= \int_\Omega \Big(f^0(x, y_k(x), u_k(x)) - f^0(x, y(x), u_k(x))\Big)\, dx \\ &\quad + \int_\Omega f^0(x, y(x), u_k(x))\, dx \\ &\to J(\sigma(\cdot)), \end{aligned}$$

since

$$|f^0(x, y_k(x), u_k(x)) - f^0(x, y(x), u_k(x))| \le C|y_k(x) - y(x)|,$$

for some constant $C > 0$. Thus, we get the proof. ∎

4. Maximum principle for optimal pair to approximate problem. We will yield the maximum principle for Problem (C) from that for corresponding approximate problems in the next section. In this section, we consider the corresponding (relaxed) approximate problems.

Let $\bar{u}(\cdot) \in \mathcal{U}_{ad}$ be an optimal control to Problem (C), and let $\bar{y}(\cdot) \in W_0^{1,p}(\Omega)$ be the optimal state corresponding to $\bar{u}(\cdot)$. For $\varepsilon > 0$, $\tau \in (0,1)$, consider (3.1) and

$$J_{\tau,\varepsilon}(\sigma(\cdot)) = \int_\Omega dx \int_U f^{0,\tau}(x, y_\varepsilon(x), v)\, \sigma(x)(dv), \tag{4.1}$$

where $\sigma(\cdot) \in \mathcal{R}(\Omega; U)$, $y_\varepsilon(\cdot) = y_\varepsilon(\cdot; \sigma(\cdot))$,

$$f^{0,\tau}(x,y,v) = f^0(x,y,v) + \frac{\tau\rho(v, \bar{u}(x))}{1+\rho(v,\bar{u}(x))} \qquad \forall\, (x,y,v) \in \Omega \times \mathbb{R} \times U, \tag{4.2}$$

and $\rho(v,w)$ denotes the distance between $v, w \in U$. We want to find the necessary condition of a $\bar{\sigma}_{\tau,\varepsilon}(\cdot) \in \mathcal{R}(\Omega; U)$ satisfying

$$J_{\tau,\varepsilon}(\bar{\sigma}_{\tau,\varepsilon}(\cdot)) = \inf_{\sigma(\cdot)\in\mathcal{R}(\Omega;U)} J_{\tau,\varepsilon}(\sigma(\cdot)). \tag{4.3}$$

First, we give the following lemma, which shows the existence of such $\bar{\sigma}_{\tau,\varepsilon}(\cdot)$'s and the relation between $\bar{\sigma}_{\tau,\varepsilon}(\cdot)$ and $\bar{u}(\cdot)$.

LEMMA 4.1. *Suppose that* (S1)–(S4) *holds. Then for any* $\tau, \varepsilon \in (0,1)$, *there exists a* $\bar{\sigma}_{\tau,\varepsilon}(\cdot) \in \mathcal{R}(\Omega; U)$ *satisfying* (4.3). *Moreover* (*as* $\varepsilon \to 0^+$),

$$\bar{\sigma}_{\tau,\varepsilon}(\cdot) \xrightarrow{\mathcal{N}} \delta_{\bar{u}(\cdot)} \qquad \text{in } \mathcal{R}(\Omega; U).$$

Proof. By Proposition 3.3, there exist positive constants $C > 0$ and $\alpha \in (0,1)$, independent of $\varepsilon \in (0,1)$, such that for any $\sigma(\cdot) \in \mathcal{R}(\Omega; U)$,

$$\|y_\varepsilon(\cdot; \sigma(\cdot))\|_{C^{1,\alpha}(\bar{\Omega})} \le C. \tag{4.4}$$

Thus, it follows from (S4) that

$$\inf_{\sigma(\cdot)\in\mathcal{R}(\Omega;U)} J_{\tau,\varepsilon}(\sigma(\cdot)) > -\infty.$$

Let $\sigma_{\tau,\varepsilon,k}(\cdot) \in \mathcal{R}(\Omega; U)$ satisfy

$$\lim_{k\to+\infty} J_{\tau,\varepsilon}(\sigma_{\tau,\varepsilon,k}(\cdot)) = \inf_{\sigma(\cdot)\in\mathcal{R}(\Omega;U)} J_{\tau,\varepsilon}(\sigma(\cdot)).$$

Let $y_{\tau,\varepsilon,k}(\cdot) = y_\varepsilon(\cdot; \sigma_{\tau,\varepsilon,k}(\cdot))$. Then by (4.4), we can suppose that

$$y_{\tau,\varepsilon,k}(\cdot) \to \bar{y}_{\tau,\varepsilon}(\cdot) \qquad \text{uniformly in } \ C^1(\bar{\Omega}),$$

without losing generality. By Lemma 2.3, there exist $\bar{\sigma}_{\tau,\varepsilon}(\cdot) \in \mathcal{R}(\Omega; U)$ such that (as $k \to +\infty$),

$$\sigma_{\tau,\varepsilon,k}(\cdot) \xrightarrow{\mathcal{N}} \bar{\sigma}_{\tau,\varepsilon}(\cdot) \qquad \text{in } \mathcal{R}(\Omega; U).$$

Consequently, it follows from Lebesgue's dominated convergence theorem that for any $\psi(\cdot) \in C_0^\infty(\Omega)$,

$$\begin{aligned}
&\lim_{k\to+\infty} \int_\Omega \varphi\left(x, \sqrt{\varepsilon^2 + |\nabla y_{\tau,\varepsilon,k}|^2}\right) \frac{\nabla y_{\tau,\varepsilon,k}}{\sqrt{\varepsilon^2 + |\nabla y_{\tau,\varepsilon,k}|^2}} \cdot \nabla\psi \, dx \\
&\quad = \int_\Omega \varphi\left(x, \sqrt{\varepsilon^2 + |\nabla \bar{y}_{\tau,\varepsilon}|^2}\right) \frac{\nabla \bar{y}_{\tau,\varepsilon}}{\sqrt{\varepsilon^2 + |\nabla \bar{y}_{\tau,\varepsilon}|^2}} \cdot \nabla\psi \, dx.
\end{aligned}$$

Moreover, there exists a subsequence $\sigma_{\tau,\varepsilon,k_j}(\cdot)$ (which may be different for different $\psi(\cdot)$) of $\sigma_{\tau,\varepsilon,k}(\cdot)$ such that

$$\begin{aligned}\lim_{k\to+\infty}\int_\Omega dx\int_U \psi(x)f(x,y_{\tau,\varepsilon,k_j}(x),v)\,\sigma_{\tau,\varepsilon,k_j}(x)(dv)\\ =\int_\Omega dx\int_U \psi(x)f(x,\bar y_{\tau,\varepsilon}(x),v)\,\bar\sigma_{\tau,\varepsilon}(x)(dv).\end{aligned}\tag{4.5}$$

This means that $\bar y_{\tau,\varepsilon}(\cdot)=y_\varepsilon(\cdot;\bar\sigma_{\tau,\varepsilon}(\cdot))$. Consequently, as (4.5), we can get

$$\lim_{k\to+\infty} J_{\tau,\varepsilon}(\sigma_{\tau,\varepsilon,k}(\cdot))=J_{\tau,\varepsilon}(\bar\sigma_{\tau,\varepsilon}(\cdot)).$$

That is, $\bar\sigma_{\tau,\varepsilon}(\cdot)$ satisfies (4.3).

Now, we turn to see the relation between $\bar\sigma_{\tau,\varepsilon}(\cdot)$ and $\bar u(\cdot)$. Let

$$J_\tau(\sigma(\cdot))=J(\sigma(\cdot))+\int_\Omega dx\int_U\frac{\tau\rho(v,\bar u(x))}{1+\rho(v,\bar u(x))}\,\sigma(x)(dv).\tag{4.6}$$

By a similar discussion to that above, we can get a $\bar\sigma_\tau(\cdot)\in\mathcal{R}(\Omega;U)$ such that (as $\varepsilon\to0^+$)

$$\bar\sigma_{\tau,\varepsilon}(\cdot)\xrightarrow{\mathcal{N}}\bar\sigma_\tau(\cdot)\qquad\text{in }\mathcal{R}(\Omega;U)\tag{4.7}$$

and

$$J_\tau(\bar\sigma_\tau(\cdot))=\inf_{\sigma(\cdot)\in\mathcal{R}(\Omega;U)}J_\tau(\sigma(\cdot)).$$

Thus,

$$J_\tau(\bar\sigma_\tau(\cdot))\le J_\tau(\delta_{\bar u(\cdot)})=J(\delta_{\bar u(\cdot)}).$$

On the other hand, by Lemma 3.5,

$$J(\delta_{\bar u(\cdot)})\le J(\bar\sigma_\tau(\cdot)).$$

Combining the above two inequalities, we have

$$\int_\Omega dx\int_U\frac{\tau\rho(v,\bar u(x))}{1+\rho(v,\bar u(x))}\,\bar\sigma_\tau(x)(dv)=0.$$

Therefore,

$$\int_U\frac{\rho(v,\bar u(x))}{1+\rho(v,\bar u(x))}\,\bar\sigma_\tau(x)(dv)=0,\qquad\text{a.e. }x\in\Omega.$$

By Lemma 4.2, which we will prove later, we have

$$\bar\sigma_\tau(x)=\delta_{\bar u(x)},\qquad\text{a.e. }x\in\Omega.$$

Thus, it follows from (4.7) that (as $\varepsilon\to0^+$)

$$\bar\sigma_{\tau,\varepsilon}(\cdot)\xrightarrow{\mathcal{N}}\delta_{\bar u(\cdot)}.$$

The proof is completed. □

LEMMA 4.2. *Suppose that* $\mu \in \mathcal{M}_+^1(U)$ *and*

$$\int_U \frac{\rho(v,w)}{1+\rho(v,w)}\,\mu(dv) = 0, \tag{4.8}$$

for some $w \in U$. *Then* $\mu = \delta_w$.

Proof. If μ is an infinitive additive probability measure, then it is natural to get $\mu = \delta_w$ from (4.8). Yet, generally, for a nonnegative finite additive measure μ and a nonnegative function $h(\cdot)$ on U, one cannot get $h = 0$, μ a.e. U, from $\int_U h(v)\,\mu(dv) = 0$ (see [8]).

To prove the lemma, we need to prove that

$$\mu\{v \in U | \rho(v,w) > 0\} = 0.$$

By (4.8), for all $\alpha > 0$,

$$\mu\left\{v \in U | \frac{\rho(v,w)}{1+\rho(v,w)} \geq \frac{\alpha}{1+\alpha}\right\} = 0.$$

That is,

$$\mu\{v \in U | \rho(v,w) \geq \alpha\} = 0 \qquad \forall\, \alpha > 0.$$

Let $A = \{v \in U | \rho(v,w) > 0\}$, $A_\alpha = \{v \in U | 0 < \rho(v,w) < \alpha\}$ $(\alpha > 0)$. Then, A, A_α are open. Consequently, $A, A_\alpha \in \mathcal{F}$, and

$$\mu(A) = \mu(A_\alpha) \qquad \forall\, \alpha > 0.$$

Noting that μ is regular for all $\delta > 0$, there exists $B, D \in \mathcal{F}$ such that

$$\mathrm{cl}(B) \subseteq A \subseteq \mathrm{int}(D), \quad \mu(D \setminus B) < \delta.$$

Since $w \notin A$, $w \notin \mathrm{cl}(B)$. Therefore there exists a $\beta > 0$ such that

$$\{v \in U | \rho(v,w) < \beta\} \cap \mathrm{cl}(B) = \emptyset.$$

On the other hand, $A \subseteq \mathrm{int}(D) \subseteq D$. Thus,

$$(D \setminus B) \supseteq (A \setminus B) \supseteq A_\beta.$$

Therefore,

$$\delta > \mu(D \setminus B) = \mu((D \setminus B) \setminus A_\beta) + \mu(A_\beta) \geq \mu(A_\beta) = \mu(A).$$

Consequently, $\mu(A) = 0$. We get the proof. □

Now, we give the necessary condition of optimal pair $(\bar{y}_{\tau,\varepsilon}(\cdot), \bar{\sigma}_{\tau,\varepsilon}(\cdot))$ in the following lemma.

LEMMA 4.3. *Let* $\tau, \varepsilon \in (0,1)$. *Suppose that* (S1)–(S4) *hold,* $\bar{\sigma}_{\tau,\varepsilon}(\cdot) \in \mathcal{R}(\Omega; U)$ *satisfies* (4.3), $\bar{y}_{\tau,\varepsilon}(\cdot) = y_\varepsilon(\cdot; \bar{\sigma}_{\tau,\varepsilon}(\cdot))$. *Then there exists a* $\bar{\psi}_{\tau,\varepsilon}(\cdot) \in W_0^{1,2}(\Omega)$ *such that*

$$\left\{\begin{array}{l} -\mathrm{div}\left\{\left[\dfrac{\varphi(x, \sqrt{\varepsilon^2 + |\nabla \bar{y}_{\tau,\varepsilon}|^2}\,)}{\sqrt{\varepsilon^2 + |\nabla \bar{y}_{\tau,\varepsilon}|^2}}\left(I - \dfrac{\nabla \bar{y}_{\tau,\varepsilon}(\nabla \bar{y}_{\tau,\varepsilon})^T}{\varepsilon^2 + |\nabla \bar{y}_{\tau,\varepsilon}|^2}\right)\right.\right. \\ \qquad\qquad \left.\left. + \varphi_s\left(x, \sqrt{\varepsilon^2 + |\nabla \bar{y}_{\tau,\varepsilon}|^2}\,\right)\,\dfrac{\nabla \bar{y}_{\tau,\varepsilon}(\nabla \bar{y}_{\tau,\varepsilon})^T}{\varepsilon^2 + |\nabla \bar{y}_{\tau,\varepsilon}|^2}\right]\nabla \bar{\psi}_{\tau,\varepsilon}\right\} \\ \qquad = \displaystyle\int_U \left[f_y(x, \bar{y}_{\tau,\varepsilon}, v)\bar{\psi}_{\tau,\varepsilon} - f_y^0(x, \bar{y}_{\tau,\varepsilon}, v)\right]\,\bar{\sigma}_{\tau,\varepsilon}(x)(dv) \text{ in } \Omega, \\ \bar{\psi}_{\tau,\varepsilon}|_{\partial\Omega} = 0, \end{array}\right. \tag{4.9}$$

and

$$\int_\Omega dx \int_U \left[f(x, \bar{y}_{\tau,\varepsilon}(x), v)\bar{\psi}_{\tau,\varepsilon}(x) - f^0(x, \bar{y}_{\tau,\varepsilon}(x), v) - \frac{\tau\rho(v, \bar{u}(x))}{1+\rho(v, \bar{u}(x))} \right] (\sigma(x) - \bar{\sigma}_{\tau,\varepsilon}(x))(dv) \le 0 \qquad \forall\, \sigma(\cdot) \in \mathcal{R}(\Omega; U). \tag{4.10}$$

Proof. The proof of the above lemma is quite standard, relatively easy, and similar to the proof of Theorem 3.1 in [3]. It can be got quite directly from

$$0 \le J_{\tau,\varepsilon}(\bar{\sigma}_{\tau,\varepsilon}(\cdot) + \delta(\sigma(\cdot) - \bar{\sigma}_{\tau,\varepsilon}(\cdot))) - J_{\tau,\varepsilon}(\bar{\sigma}_{\tau,\varepsilon}(\cdot)) \qquad \forall\, \sigma(\cdot) \in \mathcal{R}(\Omega; U),\ \delta \in (0,1).$$

We give a sketch of the proof. Let $\sigma(\cdot) \in \mathcal{R}(\Omega; U)$, $\sigma^\delta_{\tau,\varepsilon}(\cdot) = \bar{\sigma}_{\tau,\varepsilon}(\cdot) + \delta(\sigma(\cdot) - \bar{\sigma}_{\tau,\varepsilon}(\cdot))$. Then $\sigma^\delta_{\tau,\varepsilon}(\cdot) \in \mathcal{R}(\Omega; U)$, and we have

$$0 \le J_{\tau,\varepsilon}(\sigma^\delta_{\tau,\varepsilon}(\cdot)) - J_{\tau,\varepsilon}(\bar{\sigma}(\cdot)) \qquad \forall\, \sigma(\cdot) \in \mathcal{R}(\Omega; U),\ \delta \in (0,1).$$

Let $y^\delta_{\tau,\varepsilon}(\cdot) = y_\varepsilon(\cdot; \sigma^\delta_{\tau,\varepsilon}(\cdot))$. Denoting

$$\eta^{\delta,t}_{\tau,\varepsilon}(\cdot) = \nabla \bar{y}_{\tau,\varepsilon}(\cdot) + t(\nabla y^\delta_{\tau,\varepsilon}(\cdot) - \nabla \bar{y}_{\tau,\varepsilon}(\cdot)), \qquad t \in [0,1],$$

$$A^\delta_{\tau,\varepsilon} = \int_0^1 \left[\frac{\varphi\Big(x, \sqrt{\varepsilon^2 + |\eta^{\delta,t}_{\tau,\varepsilon}|^2}\Big)}{\sqrt{\varepsilon^2 + |\eta^{\delta,t}_{\tau,\varepsilon}|^2}} \left(I - \frac{\eta^{\delta,t}_{\tau,\varepsilon} \left(\eta^{\delta,t}_{\tau,\varepsilon}\right)^T}{\varepsilon^2 + |\eta^{\delta,t}_{\tau,\varepsilon}|^2} \right) + \varphi_s\left(x, \sqrt{\varepsilon^2 + |\eta^{\delta,t}_{\tau,\varepsilon}|^2}\right) \frac{\eta^{\delta,t}_{\tau,\varepsilon} \left(\eta^{\delta,t}_{\tau,\varepsilon}\right)^T}{\varepsilon^2 + |\eta^{\delta,t}_{\tau,\varepsilon}|^2} \right] dt,$$

$$b^\delta_{\tau,\varepsilon} = \int_U \left[\int_0^1 f_y(x, \bar{y}_{\tau,\varepsilon} + t(y^\delta_{\tau,\varepsilon} - \bar{y}_{\tau,\varepsilon}), v)\, dt \right] \bar{\sigma}_{\tau,\varepsilon}(x)(dv),$$

$$c^\delta_{\tau,\varepsilon} = \int_U f(x, y^\delta_{\tau,\varepsilon}, v)\, (\sigma(x) - \bar{\sigma}_{\tau,\varepsilon}(x))(dv),$$

and

$$Y^\delta_{\tau,\varepsilon} = \frac{y^\delta_{\tau,\varepsilon} - \bar{y}_{\tau,\varepsilon}}{\delta},$$

we have

$$\begin{cases} -\operatorname{div}(A^\delta_{\tau,\varepsilon}\, \nabla Y^\delta_{\tau,\varepsilon}) = b^\delta_{\tau,\varepsilon} Y^\delta_{\tau,\varepsilon} + c^\delta_{\tau,\varepsilon}, & \text{in } \Omega, \\ Y^\delta_{\tau,\varepsilon}\big|_{\partial\Omega} = 0. \end{cases} \tag{4.11}$$

On the other hand, by Proposition 3.3,

$$\|y^\delta_{\tau,\varepsilon}(\cdot)\|_{C^{1,\alpha}(\bar{\Omega})}, \|\bar{y}_{\tau,\varepsilon}(\cdot)\|_{C^{1,\alpha}(\bar{\Omega})} \le C \tag{4.12}$$

for some $C > 0$ independent of $\delta \in (0,1)$. Consequently

$$\|\eta^{\delta,t}_{\tau,\varepsilon}\|_{C^\alpha(\bar\Omega)} \le C \qquad \forall\, t \in [0,1].$$

Therefore, for some $\beta > 0$

$$\|A^\delta_{\tau,\varepsilon}\|_{C^\beta(\bar\Omega;\mathbb{R}^{n\times n})} \le C, \tag{4.13}$$

and by (S3)

$$\|b^\delta_{\tau,\varepsilon}\|_{L^\infty(\Omega)} \le C,$$

$$\|c^\delta_{\tau,\varepsilon}\|_{L^\infty(\Omega)} \le C.$$

Then it follows easily from (4.11) and (S2) that

$$\|Y^\delta_{\tau,\varepsilon}\|_{W^{1,2}_0(\Omega)} \le C_\varepsilon$$

for some positive constant C_ε, independent of δ. Thus (as $\delta \to 0^+$),

$$y^\delta_{\tau,\varepsilon} = \bar y_{\tau,\varepsilon} + \delta Y^\delta_{\tau,\varepsilon} \to \bar y_{\tau,\varepsilon} \qquad \text{strongly in } W^{1,2}_0(\Omega). \tag{4.14}$$

Combining (4.14) with (4.12), we get that

$$y^\delta_{\tau,\varepsilon} \to \bar y_{\tau,\varepsilon} \qquad \text{uniformly in } C^1(\bar\Omega).$$

Therefore, at least in the sense of a subsequence,

$$\begin{aligned} A^\delta_{\tau,\varepsilon} \to A_{\tau,\varepsilon} \equiv{}& \frac{\varphi(x, \sqrt{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2}\,)}{\sqrt{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2}} \left(I - \frac{\nabla \bar y_{\tau,\varepsilon} \left(\nabla \bar y_{\tau,\varepsilon}\right)^T}{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2} \right) \\ &+ \varphi_s \left(x, \sqrt{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2}\, \right) \frac{\nabla \bar y_{\tau,\varepsilon} \left(\nabla \bar y_{\tau,\varepsilon}\right)^T}{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2} \\ &\text{uniformly in } C(\bar\Omega; \mathbb{R}^{n\times n}), \end{aligned}$$

$$\begin{aligned} b^\delta_{\tau,\varepsilon} \to b_{\tau,\varepsilon} \equiv{}& \int_U f_y(x, \bar y_{\tau,\varepsilon}, v) \bar\sigma_{\tau,\varepsilon}(x)(dv) \\ &\text{strongly in } L^2(\Omega), \\ c^\delta_{\tau,\varepsilon} \to c_{\tau,\varepsilon} \equiv{}& \int_U f(x, \bar y_{\tau,\varepsilon}, v)\, (\sigma(x) - \bar\sigma_{\tau,\varepsilon}(x))(dv) \\ &\text{strongly in } L^2(\Omega). \end{aligned}$$

Consequently,

$$Y^\delta_{\tau,\varepsilon}(\cdot) \to Y(\cdot) \qquad \text{strongly in } W^{1,2}_0(\Omega), \tag{4.15}$$

with $Y(\cdot)$ being the solution of the following equation:

$$\begin{cases} -\operatorname{div}\left(A_{\tau,\varepsilon} \nabla Y\right) = b_{\tau,\varepsilon} Y + c_{\tau,\varepsilon} & \text{in } \Omega, \\ Y|_{\partial\Omega} = 0. \end{cases} \tag{4.16}$$

Since $Y(\cdot) \in W_0^{1,2}(\Omega)$ is the unique solution of (4.16), we really get that (4.15) holds not only in the sense of a subsequence, though the convergence of a subsequence is enough in application. Now, it follows from the following inequality,

$$\begin{aligned}
0 &\le \frac{1}{\delta}\left(J_{\tau,\varepsilon}(\sigma^{\delta}_{\tau,\varepsilon}(\cdot)) - J_{\tau,\varepsilon}(\bar\sigma_{\tau,\varepsilon}(\cdot))\right)\\
&= \int_\Omega \left\{ Y^{\delta}_{\tau,\varepsilon} \int_U \left[\int_0^1 f^0_y(x, \bar y_{\tau,\varepsilon} + t(y^{\delta}_{\tau,\varepsilon} - \bar y_{\tau,\varepsilon}), v)\, dt\right] \bar\sigma_{\tau,\varepsilon}(x)(dv)\right\} dx\\
&\qquad + \int_\Omega dx \int_U f^{0,\tau}(x, y^{\delta}_{\tau,\varepsilon}, v)\, (\sigma(x) - \bar\sigma_{\tau,\varepsilon}(x))(dv),
\end{aligned}$$

that

$$\begin{aligned}
0 \le \int_\Omega \Big\{ Y(x) \int_U f^0_y(x, \bar y_{\tau,\varepsilon}(x), v)\, \bar\sigma_{\tau,\varepsilon}(x)(dv)\\
+ \int_U f^{0,\tau}(x, \bar y_{\tau,\varepsilon}(x), v)\, (\sigma(x) - \bar\sigma_{\tau,\varepsilon}(x))(dv)\Big\}\, dx.
\end{aligned} \tag{4.17}$$

Let $\bar\psi_{\tau,\varepsilon}(\cdot) \in W_0^{1,2}(\Omega)$ satisfy (4.9). We get (4.10) from (4.17). □

5. Proofs of the main theorems. Now, we will give the proofs of Theorems 1.1–1.3. Let $\bar y_{\tau,\varepsilon}(\cdot), \bar\sigma_{\tau,\varepsilon}(\cdot)$, and $\bar\psi_{\tau,\varepsilon}(\cdot)$ be given in Lemma 4.3.

Proof of Theorem 1.1. By (S2), we have

$$\varphi(x,s) \ge \frac{\lambda}{p-1} s^{p-1} \qquad \forall\, (x,s) \in \bar\Omega \times (0, +\infty).$$

For $x \in \Omega$, $\tau, \varepsilon \in (0,1)$, if

$$h_{\tau,\varepsilon}(x) \equiv \sqrt{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2}\, \varphi_s\left(x, \sqrt{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2}\,\right) - \varphi\left(x, \sqrt{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2}\,\right) \ge 0,$$

then

$$\begin{aligned}
&\varphi\left(x, \sqrt{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2}\,\right) \frac{|\nabla \bar\psi_{\tau,\varepsilon}|^2}{\sqrt{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2}} + h_{\tau,\varepsilon}(x) \frac{(\nabla \bar\psi_{\tau,\varepsilon})^T \nabla \bar y_{\tau,\varepsilon} \nabla \bar y^T_{\tau,\varepsilon} \nabla \bar\psi_{\tau,\varepsilon}}{(\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2)^{\frac{3}{2}}}\\
&\ge \varphi\left(x, \sqrt{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2}\,\right) \frac{|\nabla \bar\psi_{\tau,\varepsilon}|^2}{\sqrt{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2}} \ge \frac{\lambda}{p-1} (\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2)^{\frac{p-2}{2}} |\nabla \bar\psi_{\tau,\varepsilon}|^2.
\end{aligned}$$

Otherwise,

$$\begin{aligned}
&\varphi\left(x, \sqrt{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2}\,\right) \frac{|\nabla \bar\psi_{\tau,\varepsilon}|^2}{\sqrt{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2}} + h_{\tau,\varepsilon}(x) \frac{(\nabla \bar\psi_{\tau,\varepsilon})^T \nabla \bar y_{\tau,\varepsilon} \nabla \bar y^T_{\tau,\varepsilon} \nabla \bar\psi_{\tau,\varepsilon}}{(\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2)^{\frac{3}{2}}}\\
&\ge \varphi_s\left(x, \sqrt{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2}\,\right) |\nabla \bar\psi_{\tau,\varepsilon}|^2 \ge \lambda (\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2)^{\frac{p-2}{2}} |\nabla \bar\psi_{\tau,\varepsilon}|^2.
\end{aligned}$$

Thus, there exists a constant λ_0 (independent of $\tau, \varepsilon \in (0,1)$, $x \in \bar\Omega$) such that

$$\begin{aligned}
&\varphi\left(x, \sqrt{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2}\,\right) \frac{|\nabla \bar\psi_{\tau,\varepsilon}|^2}{\sqrt{\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2}} + h_{\tau,\varepsilon}(x) \frac{(\nabla \bar\psi_{\tau,\varepsilon})^T \nabla \bar y_{\tau,\varepsilon} \nabla \bar y^T_{\tau,\varepsilon} \nabla \bar\psi_{\tau,\varepsilon}}{(\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2)^{\frac{3}{2}}}\\
&\ge \lambda_0 (\varepsilon^2 + |\nabla \bar y_{\tau,\varepsilon}|^2)^{\frac{p-2}{2}} |\nabla \bar\psi_{\tau,\varepsilon}|^2.
\end{aligned} \tag{5.1}$$

Thus, it follows from (4.9), Proposition 3.3, and (S3)–(S4) that

$$
\begin{aligned}
(5.2)\qquad & \lambda_0 \int_\Omega (\varepsilon^2 + |\nabla \bar{y}_{\tau,\varepsilon}|^2)^{\frac{p-2}{2}} |\nabla \bar{\psi}_{\tau,\varepsilon}|^2 \, dx \\
&\le \int_\Omega dx \int_U \left[f_y(x, \bar{y}_{\tau,\varepsilon}, v) \bar{\psi}^2_{\tau,\varepsilon} - f^0_y(x, \bar{y}_{\tau,\varepsilon}, v) \bar{\psi}_{\tau,\varepsilon} \right] \bar{\sigma}(x)(dv) \\
&\le - \int_\Omega dx \int_U f^0_y(x, \bar{y}_{\tau,\varepsilon}, v) \bar{\psi}_{\tau,\varepsilon} \, \bar{\sigma}(x)(dv) \\
&\le C \int_\Omega |\bar{\psi}_{\tau,\varepsilon}(x)| \, dx.
\end{aligned}
$$

Since $1 < p < 2$, we get

$$
\begin{aligned}
&\lambda_0 \int_\Omega |\nabla \bar{\psi}_{\tau,\varepsilon}|^2 \, dx \\
&\quad \le \lambda_0 \| \varepsilon^2 + |\nabla \bar{y}_{\tau,\varepsilon}|^2 \|^{\frac{2-p}{2}}_{L^\infty(\Omega)} \int_\Omega (\varepsilon^2 + |\nabla \bar{y}_{\tau,\varepsilon}|^2)^{\frac{p-2}{2}} |\nabla \bar{\psi}_{\tau,\varepsilon}|^2 \, dx \\
&\quad \le C \int_\Omega |\bar{\psi}_{\tau,\varepsilon}(x)| \, dx.
\end{aligned}
$$

Consequently,

$$
\|\bar{\psi}_{\tau,\varepsilon}\|_{W^{1,2}_0(\Omega)} \le C. \tag{5.3}
$$

Thus (see [1] and [26]), we can suppose that (at least in the sense of subsequence), as $\varepsilon \to 0^+$,

$$
\bar{\psi}_{\tau,\varepsilon}(\cdot) \to \bar{\psi}_\tau(\cdot) \qquad \text{weakly in } \ W^{1,2}_0(\Omega), \text{ strongly in } \ L^2(\Omega). \tag{5.4}
$$

Moreover,

$$
\|\bar{\psi}_\tau\|_{W^{1,2}_0(\Omega)} \le C. \tag{5.5}
$$

On the other hand, by Proposition 3.3, we can suppose that, as $\varepsilon \to 0^+$,

$$
\bar{y}_{\tau,\varepsilon}(\cdot) \to \tilde{y}_\tau(\cdot) \qquad \text{uniformly in } C^1(\bar{\Omega}). \tag{5.6}
$$

By Lemma 4.1,

$$
\bar{\sigma}_{\tau,\varepsilon}(\cdot) \xrightarrow{\mathcal{N}} \delta_{\bar{u}(\cdot)} \qquad \text{in } \mathcal{R}(\Omega; U). \tag{5.7}
$$

Then, it is easy to see that $\tilde{y}_\tau(\cdot) = \bar{y}(\cdot)$. Denoting

$$
H^\tau(x, y, v, \psi) = f(x, y, v)\psi - f^{0,\tau}(x, y, v) \qquad \forall\, (x, y, v, \psi) \in \mathbb{R}^n \times \mathbb{R} \times U \times \mathbb{R}^n,
$$

by (4.10), for any $\sigma(\cdot) \in \mathcal{R}(\Omega; U)$, we have

(5.8)

$$
\begin{aligned}
0 &\ge \int_\Omega dx \int_U H^\tau(x, \bar{y}_{\tau,\varepsilon}(x), v, \bar{\psi}_{\tau,\varepsilon}(x))(\sigma(x) - \bar{\sigma}_{\tau,\varepsilon}(x))(dv) \\
&= \int_\Omega dx \int_U \Big(H^\tau(x, \bar{y}_{\tau,\varepsilon}(x), v, \bar{\psi}_{\tau,\varepsilon}(x)) - H^\tau(x, \bar{y}(x), v, \bar{\psi}_{\tau,\varepsilon}(x)) \Big)(\sigma(x) - \bar{\sigma}_{\tau,\varepsilon}(x))(dv) \\
&\quad + \int_\Omega dx \int_U \Big(H^\tau(x, \bar{y}(x), v, \bar{\psi}_{\tau,\varepsilon}(x)) - H^\tau(x, \bar{y}(x), v, \bar{\psi}_\tau(x)) \Big)(\sigma(x) - \bar{\sigma}_{\tau,\varepsilon}(x))(dv) \\
&\quad + \int_\Omega dx \int_U H^\tau(x, \bar{y}(x), v, \bar{\psi}_\tau(x))(\sigma(x) - \bar{\sigma}_{\tau,\varepsilon}(x))(dv) \\
&\equiv I_1 + I_2 + I_3.
\end{aligned}
$$

By (5.3), (5.6), and (S3)–(S4), we have (as $\varepsilon \to 0^+$)

$$
\begin{aligned}
(5.9)\qquad |I_1| &\le C \int_\Omega dx \int_U \|\bar{y}_{\tau,\varepsilon}(x) - \bar{y}\|_{C(\bar{\Omega})}(|\bar{\psi}_{\tau,\varepsilon}(x)| + 1)(\sigma(x) + \bar{\sigma}_{\tau,\varepsilon}(x))(dv) \\
&= 2C\|\bar{y}_{\tau,\varepsilon}(x) - \bar{y}\|_{C(\bar{\Omega})} \int_\Omega (|\bar{\psi}_{\tau,\varepsilon}(x)| + 1)\, dx \to 0.
\end{aligned}
$$

Similarly,

$$
(5.10)\qquad \lim_{\varepsilon \to 0^+} I_2 = 0.
$$

On the other hand, by (5.7), we have (choosing a subsequence if necessary)

$$
(5.11)\qquad I_3 \to \int_\Omega dx \int_U H^\tau(x, \bar{y}(x), v, \bar{\psi}_\tau(x))(\sigma(x) - \bar{\sigma}_\tau(x))(dv).
$$

Combining (5.8)–(5.11), we get

$$
\begin{aligned}
(5.12)\qquad &\int_\Omega dx \int_U \left[f(x, \bar{y}(x), v)\bar{\psi}_\tau(x) - f^{0,\tau}(x, \bar{y}(x), v)\right](\sigma(x) - \delta_{\bar{u}(x)})(dv) \\
&\le 0 \qquad\qquad \forall\, \sigma(\cdot) \in \mathcal{R}(\Omega; U).
\end{aligned}
$$

By (5.5), we can suppose that (at least in the sense of a subsequence), as $\tau \to 0^+$,

$$
(5.13)\qquad \bar{\psi}_\tau(\cdot) \to \bar{\psi}(\cdot) \qquad \text{weakly in } W_0^{1,2}(\Omega), \text{ strongly in } L^2(\Omega).
$$

Then we have

$$
\begin{aligned}
&\int_\Omega dx \int_U H(x, \bar{y}(x), v, \bar{\psi}(x))\, (\sigma(x) - \delta_{\bar{u}(x)})(dv) \\
&= \int_\Omega dx \int_U \left[f(x, \bar{y}(x), v)\bar{\psi}(x) - f^0(x, \bar{y}(x), v)\right](\sigma(x) - \delta_{\bar{u}(x)})(dv) \\
&= \lim_{\tau \to 0^+} \int_\Omega dx \int_U \left[f(x, \bar{y}(x), v)\bar{\psi}_\tau(x) - f^{0,\tau}(x, \bar{y}(x), v)\right](\sigma(x) - \delta_{\bar{u}(x)})(dv) \\
&\le 0 \qquad\qquad \forall\, \sigma(\cdot) \in \mathcal{R}(\Omega; U).
\end{aligned}
$$

In particular, for any $v(\cdot) \in \mathcal{U}_{ad}$,

$$
(5.14)\qquad \int_\Omega \left[H(x, \bar{y}(x), v(x), \bar{\psi}(x)) - H(x, \bar{y}(x), \bar{u}(x), \bar{\psi}(x))\right]\, dx \le 0.
$$

Then it is easily to get that for any $v \in U$,

$$
H(x, \bar{y}(x), v, \bar{\psi}(x)) \le H(x, \bar{y}(x), \bar{u}(x), \bar{\psi}(x)), \qquad \text{a.e. } x \in \Omega.
$$

Since U is separable and $H(x, \bar{y}(x), \cdot, \bar{\psi}(x))$ is continuous in U, we get (1.8).

On the other hand, by (4.9), (5.4), (5.6), (5.13), and $\tilde{y}_\tau(\cdot) = \bar{y}(\cdot)$, we have

$$
\begin{aligned}
&\int_\Omega \left\{ \left[\frac{\varphi(x, |\nabla \bar{y}|\,)}{|\nabla \bar{y}|} \left(I - \frac{\nabla \bar{y}(\nabla \bar{y})^T}{|\nabla \bar{y}|^2} \right) + \varphi_s(x, |\nabla \bar{y}|\,) \frac{\nabla \bar{y}(\nabla \bar{y})^T}{|\nabla \bar{y}|^2} \right] \nabla \bar{\psi} \right\} \cdot \nabla \xi\, dx \\
&\quad = \int_\Omega [f_y(x, \bar{y}, \bar{u}(x))\bar{\psi} - f_y^0(x, \bar{y}, \bar{u}(x))]\xi(x)\, dx \qquad \forall\, \xi(\cdot) \in C_c^\infty(\{\nabla \bar{y} \ne 0\}),
\end{aligned}
$$

where $C_c^\infty(\{\nabla\bar{y}\neq 0\})$ is the set of $C^\infty(\{\nabla\bar{y}\neq 0\})$ functions having compact support in $\{\nabla\bar{y}\neq 0\}$. That is, (1.6) holds. By (5.6), for any $\gamma>0$, there exists an $\varepsilon_\tau\in(0,1)$ such that

$$|\nabla\bar{y}_{\tau,\varepsilon}(x)|\leq\gamma\qquad\forall\, x\in\{\nabla\bar{y}=0\},\ \varepsilon\in(0,\varepsilon_\tau).$$

Thus, by (5.2)–(5.3), we get

$$\frac{\lambda_0}{(\varepsilon^2+\gamma^2)^{\frac{2-p}{2}}}\int_{\{\nabla\bar{y}=0\}}|\nabla\bar{\psi}_{\tau,\varepsilon}|^2\,dx\leq C\qquad\forall\,\varepsilon\in(0,\varepsilon_\tau).$$

Therefore

$$\int_{\{\nabla\bar{y}=0\}}|\nabla\bar{\psi}_\tau|^2\,dx\leq\frac{C\gamma^{2-p}}{\lambda_0}.$$

Consequently,

$$\int_{\{\nabla\bar{y}=0\}}|\nabla\bar{\psi}_\tau|^2\,dx=0,$$

$$\int_{\{\nabla\bar{y}=0\}}|\nabla\bar{\psi}|^2\,dx=0,$$

and we get (1.7). □

Proof of Theorem 1.2. Similarly, in this case, we have (5.2) and (5.6). By (1.10) and the first inequality in (5.2), we have

$$\begin{aligned}\theta_0\int_\Omega&\bar{\psi}^2_{\tau,\varepsilon}(x)\,dx\\&\leq-\int_\Omega dx\int_U f_y(x,\bar{y}_{\tau,\varepsilon},v)\bar{\psi}^2_{\tau,\varepsilon}(x)\bar{\sigma}(x)(dv)\\&\leq-\int_\Omega dx\int_U f^0_y(x,\bar{y}_{\tau,\varepsilon},v)\bar{\psi}_{\tau,\varepsilon}\,\bar{\sigma}(x)(dv)\\&\leq C\int_\Omega|\bar{\psi}_{\tau,\varepsilon}(x)|\,dx.\end{aligned}$$

Thus,

$$\int_\Omega\bar{\psi}^2_{\tau,\varepsilon}(x)\,dx\leq C.\tag{5.15}$$

Then, by (5.2), (5.6), and (5.15), we get that for any $\Omega_0\subset\subset\{\nabla\bar{y}\neq 0\}$, there exists a constant $C(\Omega_0)>0$, independent of $\tau,\varepsilon\in(0,1)$, such that

$$\|\bar{\psi}_{\tau,\varepsilon}\|_{W^{1,2}(\Omega_0)}\leq C(\Omega_0)\qquad\forall\,\varepsilon\in(0,1).\tag{5.16}$$

Thus, we have $\bar{\psi}_\tau(\cdot),\bar{\psi}(\cdot)\in W^{1,2}_{loc}(\{\nabla\bar{y}\neq 0\})$ such that

$$\bar{\psi}_{\tau,\varepsilon}(\cdot)\to\bar{\psi}_\tau(\cdot)\qquad\text{weakly in }\ W^{1,2}(\Omega_0)\ \forall\,\Omega_0\subset\subset\{\nabla\bar{y}\neq 0\},$$

as $\varepsilon\to 0^+$, and

$$\bar{\psi}_\tau(\cdot)\to\bar{\psi}(\cdot)\qquad\text{weakly in }\ W^{1,2}_{loc}(\Omega_0)\ \forall\,\Omega_0\subset\subset\{\nabla\bar{y}\neq 0\},$$

as $\tau \to 0^+$. Then, we get (1.6). Moreover, similar to (5.14), it follows that for any $v(\cdot) \in \mathcal{U}_{ad}$,

$$\int_{\Omega_0} \left[H(x, \bar{y}(x), v(x), \bar{\psi}(x)) - H(x, \bar{y}(x), \bar{u}(x), \bar{\psi}(x))\right] \, dx \le 0. \tag{5.17}$$

And consequently, (1.8) holds for almost all $x \in \{\nabla \bar{y} \neq 0\}$. □

Proof of Theorem 1.3. In the case that $n = 1$ and $1 < p < 2$, (4.9) becomes

$$\begin{cases} -\left[\left(\dfrac{\varepsilon^2 \varphi\left(x, \sqrt{\varepsilon^2 + |\bar{y}'_{\tau,\varepsilon}|^2}\right)}{(\varepsilon^2 + |\bar{y}'_{\tau,\varepsilon}|^2)^{\frac{3}{2}}} + \dfrac{|\bar{y}'_{\tau,\varepsilon}|^2 \varphi_s\left(x, \sqrt{\varepsilon^2 + |\bar{y}'_{\tau,\varepsilon}|^2}\right)}{\varepsilon^2 + |\bar{y}'_{\tau,\varepsilon}|^2}\right) \bar{\psi}'_{\tau,\varepsilon}\right]' \\ \qquad = \displaystyle\int_U \left[f_y(x, \bar{y}_{\tau,\varepsilon}, v)\bar{\psi}_{\tau,\varepsilon} - f^0_y(x, \bar{y}_{\tau,\varepsilon}, v)\right] \, \bar{\sigma}_{\tau,\varepsilon}(x)(dv) \text{ in } (a,b), \\ \bar{\psi}_{\tau,\varepsilon}(a) = \bar{\psi}_{\tau,\varepsilon}(b) = 0. \end{cases}$$

Let

$$\bar{\Psi}_{\tau,\varepsilon} = -\left(\frac{\varepsilon^2 \varphi\left(x, \sqrt{\varepsilon^2 + |\bar{y}'_{\tau,\varepsilon}|^2}\right)}{(\varepsilon^2 + |\bar{y}'_{\tau,\varepsilon}|^2)^{\frac{3}{2}}} + \frac{|\bar{y}'_{\tau,\varepsilon}|^2 \varphi_s\left(x, \sqrt{\varepsilon^2 + |\bar{y}'_{\tau,\varepsilon}|^2}\right)}{\varepsilon^2 + |\bar{y}'_{\tau,\varepsilon}|^2}\right) \bar{\psi}'_{\tau,\varepsilon}.$$

Then

$$\begin{cases} \bar{\Psi}'_{\tau,\varepsilon}(x) = \displaystyle\int_U \left[f_y(x, \bar{y}_{\tau,\varepsilon}, v)\bar{\psi}_{\tau,\varepsilon} - f^0_y(x, \bar{y}_{\tau,\varepsilon}, v)\right] \, \bar{\sigma}_{\tau,\varepsilon}(x)(dv), & a < x < b, \\ \bar{\psi}'_{\tau,\varepsilon}(x) = h_{\tau,\varepsilon}(x) \, \bar{\Psi}_{\tau,\varepsilon}(x), & a < x < b, \\ \bar{\psi}_{\tau,\varepsilon}(a) = \bar{\psi}_{\tau,\varepsilon}(b) = 0, \end{cases} \tag{5.18}$$

where

$$h_{\tau,\varepsilon} = -\frac{(\varepsilon^2 + |\bar{y}'_{\tau,\varepsilon}|^2)^{\frac{3}{2}}}{\varepsilon^2 \varphi\left(x, \sqrt{\varepsilon^2 + |\bar{y}'_{\tau,\varepsilon}|^2}\right) + |\bar{y}'_{\tau,\varepsilon}|^2 \sqrt{\varepsilon^2 + |\bar{y}'_{\tau,\varepsilon}|^2} \varphi_s\left(x, \sqrt{\varepsilon^2 + |\bar{y}'_{\tau,\varepsilon}|^2}\right)}.$$

By (5.3), we have

$$\|\bar{\psi}_{\tau,\varepsilon}\|_{C[a,b]} \le C,$$

and $\bar{\psi}'_{\tau,\varepsilon}(\cdot) \in L^2(a,b)$. By (S2),

$$|\bar{\Psi}_{\tau,\varepsilon}(x)| \le C\varepsilon^{p-2} |\bar{\psi}'_{\tau,\varepsilon}(x)| \qquad \forall \, x \in (a,b).$$

Thus, $\bar{\Psi}_{\tau,\varepsilon}(\cdot) \in L^2(a,b)$. Then, by (5.18), we have $\bar{\Psi}_{\tau,\varepsilon}(\cdot) \in W^{1,2}(a,b) \hookrightarrow C[a,b]$. Since $h_{\tau,\varepsilon}(\cdot) \in C(a,b)$, the second equation in (5.18) shows that $\bar{\psi}_{\tau,\varepsilon}(\cdot) \in C^1(a,b)$. Thus, since $\bar{\psi}_{\tau,\varepsilon}(a) = \bar{\psi}_{\tau,\varepsilon}(b) = 0$, there exists a $c_{\tau,\varepsilon} \in (a,b)$ such that $\bar{\psi}'_{\tau,\varepsilon}(c_{\tau,\varepsilon}) = 0$. Therefore, $\bar{\Psi}_{\tau,\varepsilon}(c_{\tau,\varepsilon}) = 0$. Consequently, noting that

$$|h_{\tau,\varepsilon}(x)| \le C \, \left(\varepsilon^2 + |\bar{y}'_{\tau,\varepsilon}|^2\right)^{\frac{2-p}{2}},$$

we get easily from (5.18) that

$$\|\bar{\Psi}_{\tau,\varepsilon}(\cdot)\|_{W^{1,\infty}(a,b)} \le C$$

and

$$\|\bar{\psi}_{\tau,\varepsilon}(\cdot)\|_{W^{1,\infty}(a,b)} \leq C.$$

Thus, at least in the sense of a subsequence, we have $\bar{\Psi}_\tau(\cdot)$, $\bar{\Psi}(\cdot) \in W^{1,q}(a,b)$, $\bar{\psi}_\tau(\cdot)$, $\bar{\psi}(\cdot) \in W_0^{1,q}(a,b)$ (for all $1 < q < +\infty$) such that

$$\begin{aligned} \bar{\Psi}_{\tau,\varepsilon}(\cdot) &\to \bar{\Psi}_\tau(\cdot) \qquad \text{weakly in } \ W^{1,q}(a,b), \\ \bar{\psi}_{\tau,\varepsilon}(\cdot) &\to \bar{\psi}_\tau(\cdot) \qquad \text{weakly in } \ W_0^{1,q}(a,b), \end{aligned}$$

as $\varepsilon \to 0^+$, and

$$\begin{aligned} \bar{\Psi}_\tau(\cdot) &\to \bar{\Psi}(\cdot) \qquad \text{weakly in } \ W^{1,q}(a,b), \\ \bar{\psi}_\tau(\cdot) &\to \bar{\psi}(\cdot) \qquad \text{weakly in } \ W_0^{1,q}(a,b), \end{aligned}$$

as $\tau \to 0^+$. Then, (1.11) follows easily from (5.18) and it follows from (1.11) that $\bar{\Psi}(\cdot)$, $\bar{\psi}(\cdot) \in W^{1,\infty}(a,b)$. Finally, (1.12) is just (1.8). We get the proof. □

REFERENCES

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

[2] E. Casas and L. A. Fernández, *Optimal control of quasilinear elliptic equations with non-differentiable coefficients at the origin*, Rev. Mat. Univ. Complut. Madrid, 4 (1991), pp. 227–250.

[3] E. Casas and L. A. Fernández, *Distributed control of systems governed by a general class of quasilinear elliptic equations*, J. Differential Equations, 104 (1993), pp. 20–47.

[4] E. Casas and J. Yong, *Maximum principle for state-constrained optimal control problems governed by quasilinear elliptic equations*, Differential Integral Equations, 8 (1995), pp. 1–18.

[5] E. DiBenedetto, *$C^{1,\alpha}$ local regularity of weak solutions of degenerate elliptic equations*, Nonlinear Anal., 7 (1983), pp. 827–850.

[6] J. Dieudonné, *Sur le théoréme de Lebesgue-Nikodym.* III., Ann. Univ. Grenoble Sect. Sci. Math. Phys. (N.S.), 23 (1948), pp. 25–53.

[7] J. Dieudonné, *Sur le théoréme de Lebesgue-Nikodym.* IV., J. Indian Math. Soc. (N.S.), 15 (1951), pp. 77–86.

[8] N. Dunford and J. T. Schwartz, *Linear Operators. Part* 1, Interscience, New York, 1958.

[9] H. O. Fattorini, *Relaxed controls in infinite dimensional systems*, Internat. Ser. Numer. Math., 100 (1991), pp. 115–128.

[10] A. F. Filippov, *On certain questions in the theory of optimal control*, SIAM J. Control Ser. A, 1 (1962), pp. 76–84.

[11] R. V. Gamkrelidze, *Principle of Optimal Control Theory*, Plenum Press, New York, 1978.

[12] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer–Verlag, Berlin, 1983.

[13] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser Boston, Cambridge, MA, 1995.

[14] G. M. Lieberman, *Boundary regularity for solutions of degenerate elliptic equations*, Nonlinear Anal., 12 (1988), pp. 1203–1219.

[15] H. Lou, *An optimal control problem governed by quasi-linear variational inequalities*, SIAM J. Control Optim., 41 (2002), pp. 1229–1253.

[16] H. Lou, *Existence of optimal controls for semilinear elliptic equations without Cesari type conditions*, ANZIAM J., to appear.

[17] H. Lou, *Existence of optimal controls for semilinear parabolic equations without Cesari type conditions*, J. Appl. Math. Optim., to appear.

[18] E. J. McShane, *Generalized curves*, Duke Math. J., 6 (1940), pp. 513–536.

[19] E. J. McShane, *Necessary conditions for generalized–curve problems of the calculus of variation*, Duke Math. J., 7 (1940), pp. 1–27.

[20] E. J. McShane, *Existence theorem for Bolza problems of the calculus of variations*, Duke Math. J., 7 (1940), pp. 28–61.

[21] E. J. McShane, *Relaxed controls and variational problems*, SIAM J. Control, 5 (1967), pp. 438–485.
[22] E. J. McShane, *The calculus of variations from the beginning through optimal control theory*, SIAM J. Control Optim., 27 (1989), pp. 916–939.
[23] C. B. Morrey, Jr., *Multiple Integrals in the Calculus of Variations*, Springer–Verlag, Berlin, 1966.
[24] P. Tolksdorf, *Regularity for a more general case of quasilinear elliptic equations*, J. Differential Equations, 51 (1984), pp. 126–150.
[25] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[26] K. Yosida, *Functional Analysis*, 6th ed., Springer–Verlag, Berlin, 1980.
[27] L. C. Young, *On approximation by polygon's in the calculus of variations*, Proc. Roy. Soc. A, 14 (1933), pp. 325–341.
[28] L. C. Young, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, C. R. Sci. Lettres Varsovie C. III, 30 (1937), pp. 212–234.
[29] L. C. Young, *Necessary conditions in the calculus of variations*, Acta Math., 69 (1938), pp. 239–258.
[30] L. C. Young, *Lectures on the Calculus of Variational and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 24–52

A NEW APPROACH OF STABILIZATION OF NONDISSIPATIVE DISTRIBUTED SYSTEMS*

AISSA GUESMIA†

Abstract. In this paper we propose a new approach to prove the nonlinear (internal or boundary) stabilization of certain nondissipative distributed systems (the usual energy is not decreasing). This approach leads to decay estimates (known in the dissipative case) when the integral inequalities method due to Komornik [*Exact Controllability and Stabilization. The Multiplier Method*, Masson, Paris, John Wiley, Chichester, UK, 1994] cannot be applied due to the lack of dissipativity.

First we study the stability of a semilinear wave equation with a nonlinear damping based on the equation

$$u'' - \Delta u + h(\nabla u) + f(u) + g(u') = 0.$$

We consider the general case with a function h satisfying a smallness condition, and we obtain uniform decay of strong and weak solutions under weak growth assumptions on the feedback function and without any control of the sign of the derivative of the energy related with the above equation.

In the second part we consider the case $h(\nabla u) = -\nabla\phi \cdot \nabla u$ with $\phi \in W^{1,\infty}(\Omega)$. We prove some precise decay estimates (exponential or polynomial) of equivalent energy without any restriction on ϕ.

The same results will be proved in the case of boundary feedback.

Finally, we comment on some applications of our approach to certain nondissipative distributed systems.

Some results of this paper were announced without proof in [A. Guesmia, *C. R. Acad. Sci. Paris Sér.* I *Math.*, 332 (2001), pp. 633–636].

Key words. stabilizability by a nonlinear feedback, partial differential equation, wave equation, Petrovsky system, elasticity, integral inequalities

AMS subject classifications. 35B40, 35L70, 35B37

PII. S0363012901394978

1. Introduction.

Consider the semilinear wave equation with a nonlinear internal dissipative term,

$$\text{(P)} \qquad \begin{cases} u'' - \Delta u + h(\nabla u) + f(u) + g(u') = 0 & \text{in} \quad \Omega \times \mathbb{R}^+, \\ u = 0 & \text{on} \quad \Gamma \times \mathbb{R}^+, \\ u(x,0) = u_0(x) \quad \text{and} \quad u'(x,0) = u_1(x) & \text{in} \quad \Omega, \end{cases}$$

and the nonlinear boundary feedback,

$$\text{(P}') \qquad \begin{cases} u'' - \Delta u + h(\nabla u) + f(u) = 0 & \text{in} \quad \Omega \times \mathbb{R}^+, \\ u = 0 & \text{on} \quad \Gamma_0 \times \mathbb{R}^+, \\ \partial_\nu u + g(u') = 0 & \text{on} \quad \Gamma_1 \times \mathbb{R}^+, \\ u(x,0) = u_0(x) \quad \text{and} \quad u'(x,0) = u_1(x) & \text{in} \quad \Omega, \end{cases}$$

where $\Omega \subset \mathbb{R}^n$ $(n \in \mathbb{N}^*)$ is an open bounded domain with smooth boundary Γ and $f, g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$ are continuous nonlinear functions satisfying some

*Received by the editors September 12, 2001; accepted for publication (in revised form) September 11, 2002; published electronically March 19, 2003.
http://www.siam.org/journals/sicon/42-1/39497.html

†UFR de Mathématiques, Informatique et Mécanique, Université de Metz, ISGMP, Bat. A, Ile du Saulcy, 57045 Metz cédex 01, France (guesmia@poncelet.sciences.univ-metz.fr).

general properties (see Assumptions 2.1–2.5 below). In (P′), ν represents the outward unit normal to $\Gamma = \Gamma_0 \cup \Gamma_1$, where Γ_0 and Γ_1 are closed and disjoint. In this paper Δ and ∇ stand, respectively, for the Laplacian and the gradian with respect to the spatial variables, $'$ denotes the derivative with respect to time t, and $\mathbb{R}^+ = [0, \infty[$.

The main goal of this paper is to show that strong and weak solutions to problems (P) and (P′) decay to zero when $t \to \infty$ and give some precise decay properties.

When $h \equiv 0$ the bibliography of works in this direction is truly long. We can cite, for instance, the works of Nakao [18, 21, 22], Kawashima, Nakao, and Ono [11], Nakao and Narazaki [19], Nakao and Ono [20], Haraux and Zuazua [10], Pucci and Serrin [23], and Zuazua [27], among others.

In [21], Nakao considered the following initial boundary value problem:

$$\text{(P1)} \qquad \begin{cases} u'' - \Delta u + \rho(u') + f(u) = 0 & \text{in} \quad \Omega \times \mathbb{R}^+, \\ u = 0 & \text{on} \quad \Gamma \times \mathbb{R}^+, \\ u(x,0) = u_0(x) \quad \text{and} \quad u'(x,0) = u_1(x) & \text{in} \quad \Omega, \end{cases}$$

where $\rho(v) = |v|^\beta v$, $\beta > -1$, $f(u) = bu|u|^\alpha$, α, $b > 0$ (in this paper $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}$ and $\mathbb{R}^n$), and Ω is a bounded domain of $\mathbb{R}^n$ $(n \geq 1)$, with a smooth boundary $\Gamma := \partial\Omega$. He showed that (P1) has a unique global weak solution if $0 \leq \alpha \leq 2/(n-2)$, $n \geq 3$, and a global unique strong solution if $\alpha > 2/(n-2)$, $n \geq 3$ (of course if $n = 1$ or 2, then there is no restriction on α). In addition to global existence the issue of the decay rate was addressed. In both cases, it has been shown that the energy of the solution decays algebraically if $\beta > 0$ and it decays exponentially if $\beta = 0$. This improves an earlier result obtained by the author in [22], where he studied the problem in an abstract setting and established a theorem concerning the decay of the solution energy only for the case $\alpha \leq 2/(n-2)$, $n \geq 3$. Later on, in a joint work with Ono [20], this result has been extended to the Cauchy problem for the equation

$$u'' - \Delta u + \lambda^2(x)u + \rho(u') + f(u) = 0, \qquad (x,t) \in \mathbb{R}^n \times \mathbb{R}^+,$$

where $\rho(u')$ behaves like $|u'|^\beta u'$ and $f(u)$ behaves like $-bu|u|^\alpha$. In this case the authors required that the initial data be small enough in $H^1 \times L^2$ norm and of compact support.

Pucci and Serrin [23] discussed the stability of the problem

$$\text{(P2)} \qquad \begin{cases} u'' - \Delta u + Q(x,t,u,u') + f(x,u) = 0 & \text{in} \quad \Omega \times \mathbb{R}^+, \\ u = 0 & \text{on} \quad \Gamma \times \mathbb{R}^+, \\ u(x,0) = u_0(x) \quad \text{and} \quad u'(x,0) = u_1(x) & \text{in} \quad \Omega \end{cases}$$

and proved that the energy of the solution is a Liapunov function. Although they did not discuss the issue of the decay rate, they did show that in general the energy goes to zero as t approaches infinity. They also considered an important special case of (P2), which occurs when $Q(x,t,u,u') = a(t)t^\alpha u'$ and $f(x,u) = V(x)u$, and showed that the behavior of the solutions depends crucially on the parameter α. If $|\alpha| \leq 1$, then the rest field is asymptotically stable. On the other hand, when $\alpha < -1$ or $\alpha > 1$ there are solutions that do not approach zero or approach nonzero functions $\phi(x)$ as $t \to \infty$.

Messaoudi [16] discussed an initial boundary value problem related to the equation

$$u'' - \Delta u + a(1 + |u'|^{m-2})u' + bu|u|^{p-2} = 0 \quad \text{in} \quad \Omega \times \mathbb{R}^+,$$

where a, $b > 0$, $m \geq 2$, $p > 2$, and proved that the energy of the solution decays exponentially. The proof of this result is based on a direct method used in [3] and [5].

Concerning the boundary feedback case, problem (P′) with $h \equiv 0$ has attracted considerable attention in the literature and, in recent years, important progress has been obtained in this context. New techniques were developed which allow us to stabilize a system through its boundary or control it from an initial to a final state (controllability). There is a large body of literature regarding boundary stabilization with linear feedback; we refer the reader to the following works: Lagnese [13], Russell [24], Triggiani [25], and You [26]. Now when the boundary feedback is nonlinear we can cite the works of Zuazua [28], Lasiecka and Tataru [14], Komornik [12], and Guesmia [5], among others. For such cases, the main purpose is to obtain the same stabilization results when a boundary feedback of the form

$$\partial_\nu u + a(x)u + b(x)g(u') = 0 \quad \text{on} \quad \Gamma_1 \times \mathbb{R}^+$$

is applied on a part Γ_1 of the boundary Γ of Ω which satisfies certain geometric conditions and a, b, and g are given functions, whereas no feedback is applied on the other part of the boundary, i.e.,

$$u = 0 \quad \text{on} \quad (\Gamma \setminus \Gamma_1) \times \mathbb{R}^+.$$

However, when $h \not\equiv 0$ very little is known in the literature; more general and recent results in this direction were obtained in [2]. In this paper the authors established well-posedness of the following large class of hyperbolic equations:

$$K(x,t)u'' - \Delta u + F(x,t,u,u',\nabla u) = f(x)$$

with boundary conditions and initial data as in (P′), where K, F, and f are given functions satisfying some hypotheses.

However, to obtain exponential stability of solutions using classical multipliers and integral inequalities, they assumed some additional hypotheses on F which require, in particular, that F is global Lipschitz with respect to its last variable, where the Lipschitz constant is a function on t and converges exponentially to 0 at ∞. This is a strong hypothesis which is not satisfied if, for example, the function F does not depend on time t, as in our case.

Hyperbolic-parabolic equations are interesting from the point of view of not only the general theory of PDEs but also to applications in mechanics. For instance, the transonic Karman equation

$$u'u'' - \Delta u = 0$$

models flows of compressible gas in the transonic region where the velocity of gas varies from subsonic values to supersonic ones (see [2] and the references therein).

We note that stability of problems with the nonlinear term $h(\nabla u)$ requires careful treatment because we have any information neither about the influence of the integral $\int_\Omega h(\nabla u)u'\,dx$ on the norm

$$\|(u,u')\|^2_{H_0^1(\Omega)\times L^2(\Omega)} = \int_\Omega \left(|u'(x,t)|^2 + |\nabla u(x,t)|^2\right) dx$$

nor about the sign of its derivative; that is, the energy E defined by (2.7) is not necessary decreasing (see identities (3.2) and (5.1)). Decrease of energy plays a crucial

role in studying the asymptotic stability of the solution, as it was considered in the prior literature, in particular, in the works cited above.

We also observe that our problem deals with nonlinearity, which involves the gradient combined with a nonlinear feedback. This situation was not previously considered and leads to new difficulties. In order to overcome these difficulties and obtain energy decay estimates, we give a new and direct approach based on a combination of some ideas given by Guesmia in [3, 4] and the multiplier technique.

In the case where h is linear we introduce a nonincreasing equivalent energy (see (2.14)) and then, by the use of appropriate multipliers and a well-known lemma due to Haraux–Komornik (see [12, Theorem 9.1]), the exponential and polynomial decay estimates are proved. In the case where h is nonlinear, the introduction of a such equivalent energy seems to be not possible. In this case, the main ingredient for proving the exponential stability is to obtain a generalized integral inequalities of the form

$$(*) \quad \begin{cases} \int_S^T E(t)dt \le a_1(E(S)+E(T)) + a_2(E(S)-E(T)) & \forall 0 \le S \le T < \infty, \\ E'(t) \le a_3 E(t) & \forall t \ge 0, \end{cases}$$

where a_i, $i = 1, 2, 3$, are nonnegative constants and where E stands for the classical energy (2.7). Then we show that if, in addition, $2a_1a_3 < 1$ or $a_1 \le a_2$, E must converge exponentially to 0 at ∞.

Notice that a positive function satisfying $(*)$ does not necessarily converge to 0 at ∞; if $a_1a_3 \ge 1 + a_2a_3$, then the function $E(t) = e^{a_3 t}$ satisfies $(*)$. As an open question, it would be interesting to know what happens if $a_1a_3 \in [\frac{1}{2}, 1 + a_2a_3[$ and $a_1 > a_2$.

The integral result $(*)$ gives a generalization to the Haraux–Komornik lemma, which concerns nonincreasing functions (that is, $a_3 = 0$).

The rest of this paper is organized as follows. In section 2 we establish assumptions and state our main results. In section 3 we obtain the uniform stability of (P). In section 4 we consider the case $h(\nabla u) = -\nabla\phi \cdot \nabla u$, where $\phi \in W^{1,\infty}(\Omega)$ and $\cdot$ denotes the scalar product in $\mathbb{R}^n$, and we prove some decay estimates of equivalent energy of (P). In sections 5 and 6 we prove the same results for (P′). Finally, in the last section we give some applications of our approach to Petrovsky, coupled, and elasticity systems.

2. Assumptions and main results. We begin this section stating the general hypotheses.

Assumption 2.1 (assumptions on f). $f : \mathbb{R} \to \mathbb{R}$ is a C^1 function such that $f(0) = 0$ and, deriving from a potential F, that is

$$F(s) = \int_0^s f(\sigma)\, d\sigma \qquad \forall s \in \mathbb{R},$$

$$F(s) \ge -as^2 \qquad \forall s \in \mathbb{R}, \tag{2.1}$$

with $0 \le a < \frac{1}{2c_0}$, where c_0 is the smallest positive constant (depending only on Ω) such that (Poincaré's inequality)

$$\int_\Omega |v|^2 dx \le c_0 \int_\Omega |\nabla v|^2 dx \qquad \forall v \in H_0^1(\Omega). \tag{2.2}$$

Also, there exists $b > 0$ such that

$$2bF(s) \le sf(s) \qquad \forall s \in \mathbb{R}. \tag{2.3}$$

Assumption 2.2 (assumptions on g). $g : \mathbb{R} \to \mathbb{R}$ is a C^1 function, nondecreasing, $g(0) = 0$, such that

$$g(s)s > 0 \qquad \forall s \neq 0. \tag{2.4}$$

Also, there exist two positive constants c_1 and c_2 such that

$$c_1|s| \leq |g(s)| \leq c_2|s| \qquad \forall s \in \mathbb{R}. \tag{2.5}$$

Assumption 2.3 (assumptions on h). $h : \mathbb{R}^n \to \mathbb{R}$ is a C^1 function such that ∇h is bounded and there exists $\beta > 0$ such that

$$|h(\zeta)| \leq \beta|\zeta| \qquad \forall \zeta \in \mathbb{R}^n. \tag{2.6}$$

We define the energy of the solution of (P) by the formula

$$E(t) = \int_\Omega \left(\left|u'\right|^2 + \left|\nabla u\right|^2 + 2F(u) \right) dx, \qquad t \in \mathbb{R}^+. \tag{2.7}$$

Remarks. 1. If the function f is increasing and $f(0) = 0$, then (2.1) and (2.3) are satisfied with $a = 0$ and $b = \frac{1}{2}$.

2. Condition (2.1) assures the following inequality:

$$\|(u, u')\|^2_{H_0^1(\Omega)\times L^2(\Omega)} \leq kE(t) \qquad \forall t \in \mathbb{R}^+, \tag{2.8}$$

where $k = \frac{1}{1-2ac_0} > 0$. Indeed, (2.1) and (2.2) imply that

$$\begin{aligned} E(t) &\geq \int_\Omega \left(\left|u'\right|^2 + \left|\nabla u\right|^2 - 2a\left|u\right|^2 \right) dx \\ &\geq \int_\Omega \left(\left|u'\right|^2 + (1 - 2ac_0)\left|\nabla u\right|^2 \right) dx \\ &\geq (1 - 2ac_0) \int_\Omega \left(\left|u'\right|^2 + \left|\nabla u\right|^2 \right) dx = (1 - 2ac_0)\|(u, u')\|^2_{H_0^1(\Omega)\times L^2(\Omega)}, \end{aligned}$$

which gives (2.8).

3. Under Assumptions 2.1, 2.2, 2.3 and using analogous considerations like the ones used in [2] (we omit the details), we can use Galerkin's method (semigroup theory is not suitable to treat degenerate problems) and prove that problem (P) possesses a unique strong solution, $u :]0, \infty[\to \mathbb{R}$, such that

$$u \in L^\infty(]0, \infty[; H_0^1(\Omega) \cap H^2(\Omega)), \qquad u' \in L^\infty(]0, \infty[; H_0^1(\Omega)), \tag{2.9}$$

and

$$u'' \in L^\infty(]0, \infty[; L^2(\Omega)).$$

Moreover, supposing that $\{u_0, u_1\}$ is in $H_0^1(\Omega) \times L^2(\Omega)$ and using density arguments, we can show that (P) has a unique weak solution $u : \Omega\times]0, \infty[\to \mathbb{R}$ in the space

$$C(]0, \infty]; H_0^1(\Omega)) \cap C^1(]0, \infty[; L^2(\Omega)). \tag{2.10}$$

Now we are in position to state our first main result.

THEOREM 2.1. *Assume that Assumptions* 2.1, 2.2, 2.3 *hold such that* $b < 1$ *and* β *satisfies the following smallness hypotheses:*

$$\frac{\beta}{2}\left(\sqrt{c_0+\left(\frac{2}{c_1}\right)^2}+\sqrt{c_0}\right)+\sqrt{\frac{c_0c_2\beta}{2\sqrt{2}}}\le 1-b,$$

$$\beta<\frac{b}{k^2\sqrt{c_0}},\quad \text{or}\quad \frac{k\sqrt{c_0}}{2}\le\frac{1}{c_1}+\frac{1}{2}\sqrt{\frac{c_0c_2}{\sqrt{2}\beta}}.$$

Then the energy determined by the strong solution u *decays exponentially. That is, to say for some positive constants* c,ω, *one has*

$$E(t)\le cE(0)e^{-\omega t}\qquad \forall t\in\mathbb{R}^+. \tag{2.11}$$

Furthermore, (2.11) *holds for the weak solution* u.

Remark. If F is positive (for example, $sf(s)\ge 0$ for all $s\in\mathbb{R}$), then β and b can be taken such that $b>0$ and

$$\frac{\beta}{2}\left(\sqrt{c_0+\left(\frac{2}{c_1}\right)^2}+(1+2k^2)\sqrt{c_0}\right)+\sqrt{\frac{c_0c_2\beta}{2\sqrt{2}}}<1,\qquad \beta<\frac{b}{k^2\sqrt{c_0}}$$

or

$$\frac{\beta}{2}\left(\sqrt{c_0+\left(\frac{2}{c_1}\right)^2}+\sqrt{c_0}\right)+\sqrt{\frac{c_0c_2\beta}{2\sqrt{2}}}<1,\qquad \frac{k\sqrt{c_0}}{2}\le\frac{1}{c_1}+\frac{1}{2}\sqrt{\frac{c_0c_2}{\sqrt{2}\beta}}.$$

We consider now the case $h(\nabla u)=-\nabla\phi\cdot\nabla u$, where $\phi\in W^{1,\infty}(\Omega)$ and g satisfies a hypothesis weaker than (2.5).

Assumption 2.4 (assumptions on g). $g:\mathbb{R}\to\mathbb{R}$ is a C^1 function, nondecreasing, $g(0)=0$, such that (2.4) holds and there exist four constants r, $p\ge 1$ and c_1, $c_2>0$ such that

$$c_1\min\{|s|,|s|^r\}\le|g(s)|\le c_2\max\{|s|^{\frac{1}{r}},|s|^p\}\qquad\forall s\in\mathbb{R}, \tag{2.12}$$

$$(n-2)p\le n+2. \tag{2.13}$$

We have the following stabilization result.

THEOREM 2.2. *Let* u *be a solution of* (P) *in the class* (2.10). *Under Assumptions* 2.1 *and* 2.4, *there exist two positive constants* ω, c *such that the equivalent energy of* (P), *defined by*

$$E(t)=\int_\Omega e^{\phi(x)}\left(|u'|^2+|\nabla u|^2+2F(u)\right)dx,\qquad t\in\mathbb{R}^+, \tag{2.14}$$

satisfies (2.11) *if* $r=1$, *and*

$$E(t)\le c(1+t)^{\frac{-2}{r-1}}\qquad\forall t\in\mathbb{R}^+ \tag{2.15}$$

if $r>1$.

Remarks. 1. If we take $g(s) = \alpha s$ for all $s \in \mathbb{R}$ with $\alpha > 0$ (that is, $r = p = 1$), then we find the results obtained in [15]. On the other hand, the case of $g(s) = \alpha(1 + |s|^{m-2})s$ for all $s \in \mathbb{R}$ with $m > 2$ (that is, $p = m - 1$ and $r = 1$) gives the results obtained in [16].

2. In Theorem 2.1 we can weaken assumption (2.6) by taking β as the Lipschitz constant of only the nonlinear part of h; that is, we assume that there exists $\bar{\zeta} \in \mathbb{R}^n$ such that

$$|h(\zeta) + \bar{\zeta} \cdot \zeta| \leq \beta|\zeta| \qquad \forall \zeta \in \mathbb{R}^n.$$

To prove this we have only to consider the equivalent energy defined by (2.14) where $\phi(x) = \bar{\zeta} \cdot x$.

3. It is possible to weaken the growth assumption (2.12) as was done for the study of elasticity systems in [3, 7] and the Petrovsky system in [6]. In order to simplify we shall only consider in this paper the case of assumption (2.12).

Now we are concerned by the stability of (P′). In order to obtain the estimates (2.11) and (2.15), the following assumptions are made on Γ and f. Let x^0 be a fixed point in $\mathbb{R}^n$. Then put

$$m = m(x) = x - x^0, \qquad R = \max_{x \in \bar{\Omega}} |m(x)|$$

and partition the boundary Γ into two nonempty sets:

$$\Gamma_0 = \{x \in \Gamma : \ m(x) \cdot \nu(x) \leq 0\}, \qquad \Gamma_1 = \{x \in \Gamma : \ m(x) \cdot \nu(x) \geq \delta > 0\}.$$

Examples. Concerning the existence of such a partition of Γ, we can take Ω as follows:

1. If $n = 1$, then Ω is a bounded open interval, say $\Omega =]x_1, x_2[\subset \mathbb{R}$, and our geometric hypotheses are satisfied in each of the following two cases:

(i) $\Gamma_0 = \{x_1\}$, $\Gamma_1 = \{x_2\}$, and $x^0 \leq x_1$,

(ii) $\Gamma_0 = \{x_2\}$, $\Gamma_1 = \{x_1\}$, and $x^0 \geq x_2$.

2. If $n \geq 2$ and $\Omega = \Omega_1 \setminus \bar{\Omega}_0$, where Ω_1 and Ω_0 are two open domains with boundary Γ_1, and Γ_0, respectively, $\bar{\Omega}_0 \subset \Omega_1$, and Ω_1 and Ω_0 are star-shaped with respect to some point $x^0 \in \Omega_0$ (a domain Ω is called star-shaped with respect to x^0 if $m \cdot \nu > 0$ on $\partial\Omega$), then our geometric hypotheses are satisfied.

3. If $n \geq 2$ and Ω is not of the form mentioned in the preceding example, then in general there is no point x^0 satisfying simultaneously the geometric hypotheses assumed on Γ_1 and Γ_0. By applying an approximational method, one could considerably weaken these geometric hypotheses, at least in dimensions $n = 2$, 3, by adapting an analogous argument given by Komornik–Zuazua for the wave equation (see [12] and the references therein).

Assumption 2.5 (assumptions on f). $f : \mathbb{R} \to \mathbb{R}$ is a C^1 function such that (2.3) and

$$F(s) \geq 0 \qquad \forall s \in \mathbb{R}. \tag{2.16}$$

The well-posedness of the problem (P′) can be established by standard Galerkin's method (see [15]); we do not discuss this point here. We use the notations

$$V = \{v \in H^1(\Omega) : \ v = 0 \ \text{ on } \ \Gamma_0\} \quad \text{and} \quad W = H^2(\Omega) \cap V;$$

we have the following:

1. For all $(u_0, u_1) \in W \times V$ such that $\partial_\nu u_0 + g(u_1) = 0$ on Γ_1, problem (P′) has a unique strong solution, $u :]0, \infty[\to \mathbb{R}$, such that

$$u \in L^\infty(]0, \infty[; W), \quad u' \in L^\infty(]0, \infty[; V), \text{ and } u'' \in L^\infty(]0, \infty[; L^2(\Omega)).$$

2. If $\{u_0, u_1\}$ is in $V \times L^2(\Omega)$, then (using density arguments) the solution is weak: $u : \Omega \times]0, \infty[\to \mathbb{R}$ in the space

$$C(]0, \infty]; V) \cap C^1(]0, \infty[; L^2(\Omega)). \tag{2.17}$$

THEOREM 2.3. *Let u be a solution of* (P′) *in the class* (2.17). *Assume, moreover, that Assumptions* 2.2, 2.3, 2.5 *hold with β small enough and $b > 1$ or f is linear. Then the energy of u, defined by* (2.7), *decays exponentially to zero in the sense of* (2.11).

We consider now the case $h(\nabla u) = -\nabla\phi \cdot \nabla u$, where $\phi \in W^{1,\infty}(\Omega)$.

We have the following stabilization result for (P′).

THEOREM 2.4. *Let u be a solution of* (P′) *in the class* (2.17). *Under Assumptions* 2.5, 2.4 *with $p = 1$, $R\|\nabla\phi\|_\infty < \min\{2, n\}$, and $b > \frac{n+R\|\nabla\phi\|_\infty}{n-R\|\nabla\phi\|_\infty}$ or f is linear and $\|\nabla\phi\|_\infty$ is small enough, where $\|\nabla\phi\|_\infty = \max_{x\in\bar{\Omega}} |\nabla\phi(x)|$, the results of Theorem* 2.2 *hold true.*

Remarks. 1. As an example of a function f satisfying Assumption 2.5, we can take $f(s) = \gamma s\,|s|^{q-1}$ with $\gamma \geq 0$ and $q \geq 1$. Condition (2.3) is satisfied for all $b \leq \frac{q+1}{2}$.

2. We have many possibilities to take the function g such that condtions (2.12) and (2.13) are satisfied, for example, $g(s) = \gamma|s|^{r-1}s$ if $|s| \leq 1$, and $g(s) = \gamma s$ if $|s| \geq 1$, where $\gamma > 0$.

3. Thanks to (2.16), the function F is positive, and then the usual energy (2.7) satisfies

$$\int_\Omega \left(\left|u'\right|^2 + \left|\nabla u\right|^2 \right) dx \leq E(t). \tag{2.18}$$

The quantity $(\int_\Omega \left|\nabla u\right|^2 dx)^{\frac{1}{2}}$ defines a norm on V equivalent to the usual norm induced by $H^1(\Omega)$; consequently, V is a Hilbert space with this norm.

4. If h is nonlinear and $r > 1$, we do not know if the energy of (P) and (P′) decays polynomially to zero.

5. In the case of uniform stability (Theorem 2.1 and Theorem 2.3), our proof allows us to obtain explicit constants c and ω in (2.11).

6. Theorem 2.1, Theorem 2.3, and Theorem 2.4 probably remain valid without the smallness conditions assumed on β, but we could not prove them.

3. Uniform decay: Proof of Theorem 2.1. To justify all the computations that follow, we assume first that the solution is strong, and by a standard density argument we deduce the result for weak solutions.

We are going to prove that the energy defined by (2.7) satisfies the estimate

$$E(S + T_0) \leq dE(S) \qquad \forall S \in \mathbb{R}^+ \tag{3.1}$$

with $0 < d < 1$ and $T_0 > 0$. (This will be fixed later in the course of the proof.) Using (3.1), inequality (3.9) below gives (2.11).

We start this section by giving an explicit formula for the derivative of the energy. A simple computation shows that

$$E'(t) = -2\int_\Omega u' g(u')\,dx - 2\int_\Omega u' h\left(\nabla u\right)\,dx. \tag{3.2}$$

Multiplying the first equation in (P) by u and integrating the obtained result over $\Omega \times [S,T]$, we obtain

$$\begin{aligned} 0 &= \int_S^T \int_\Omega u\left(u'' - \Delta u + h(\nabla u) + f(u) + g(u')\right) dx\,dt \\ &= \left[\int_\Omega u u' dx\right]_S^T + \int_S^T \int_\Omega \left(-\left|u'\right|^2 + \left|\nabla u\right|^2 + u f(u)\right) dx\,dt \\ &\quad + \int_S^T \int_\Omega u g(u')\,dx dt + \int_S^T \int_\Omega u h\left(\nabla u\right) dx\,dt. \end{aligned} \tag{3.3}$$

Hence, from (3.3), making use of the Cauchy–Schwarz inequality and taking assumption (2.6) and property (2.2) into account, we infer

$$\begin{aligned} &\int_S^T \int_\Omega \left(\left|u'\right|^2 + \left|\nabla u\right|^2 + u f(u)\right) dx\,dt \\ &\leq -\left[\int_\Omega u u' dx\right]_S^T + \int_S^T \int_\Omega \left(2\left|u'\right|^2 - u g(u')\right) dx\,dt \\ &\quad + \frac{\beta}{2\sqrt{c_0}} \int_S^T \int_\Omega \left|u\right|^2 dx\,dt + \frac{\sqrt{c_0}}{2\beta} \int_S^T \int_\Omega \left|h\left(\nabla u\right)\right|^2 dx\,dt \\ &\leq -\left[\int_\Omega u u' dx\right]_S^T + \int_S^T \int_\Omega \left(2\left|u'\right|^2 - u g(u')\right) dx\,dt \\ &\quad + \frac{\beta\sqrt{c_0}}{2} \int_S^T \int_\Omega \left|\nabla u\right|^2 dx\,dt + \frac{\beta\sqrt{c_0}}{2} \int_S^T \int_\Omega \left|\nabla u\right|^2 dx\,dt. \end{aligned}$$

Then, taking assumption (2.3) into account, from this inequality we deduce

$$\begin{aligned} &\int_S^T \int_\Omega \left(\left|u'\right|^2 + (1 - \beta\sqrt{c_0})\left|\nabla u\right|^2 + 2bF(u)\right) dx\,dt \\ &\leq -\left[\int_\Omega u u' dx\right]_S^T + \int_S^T \int_\Omega \left(2\left|u'\right|^2 - u g(u')\right) dx\,dt. \end{aligned} \tag{3.4}$$

Using (2.2), (2.8), and the Cauchy–Schwarz inequality, we can easily get

$$\begin{aligned} \left|\int_\Omega u u' dx\right| &\leq \frac{1}{2}\int_\Omega \left(\sqrt{c_0}\left|u'\right|^2 + \frac{1}{\sqrt{c_0}}\left|u\right|^2\right) dx \\ &\leq \frac{\sqrt{c_0}}{2}\int_\Omega \left(\left|u'\right|^2 + \left|\nabla u\right|^2\right) dx \leq \frac{k\sqrt{c_0}}{2} E(t); \end{aligned}$$

then

$$-\left[\int_\Omega uu'dx\right]_S^T \le \frac{k\sqrt{c_0}}{2}\left(E(S)+E(T)\right).$$

Next, we insert this inequality into (3.4); it follows that

$$\int_S^T\int_\Omega \left(|u'|^2+(1-\beta\sqrt{c_0})\,|\nabla u|^2+2bF(u)\right)dx\,dt \tag{3.5}$$

$$\le \frac{k\sqrt{c_0}}{2}\left(E(S)+E(T)\right)+\int_S^T\int_\Omega\left(2\,|u'|^2-ug(u')\right)dx\,dt.$$

Next, we want to majorize the last term in the right-hand side of (3.5).

Estimate for $\int_S^T\int_\Omega \left(2|u'|^2-ug(u')\right)dx\,dt$. Using (3.2) and the Cauchy–Schwarz inequality and taking the assumptions (2.4), (2.5), and (2.6) into account, it holds that

$$2\int_S^T\int_\Omega |u'|^2dx\,dt \le \frac{2}{c_1}\int_S^T\int_\Omega u'g(u')dx\,dt$$

$$=\frac{1}{c_1}\int_S^T\left(-E'(t)-2\int_\Omega u'h(\nabla u)dx\right)dt$$

$$\le \frac{1}{c_1}\left(E(S)-E(T)\right)+\frac{1}{c_1}\int_S^T\int_\Omega\left(\epsilon|u'|^2+\frac{\beta^2}{\epsilon}|\nabla u|^2\right)dx\,dt;$$

we choose $\epsilon>0$ such that $\frac{\beta^2}{\epsilon c_1}=\frac{\epsilon}{c_1}-\beta\sqrt{c_0}$, that is, $\epsilon=\frac{\beta}{2}(\sqrt{{c_1}^2c_0+4}+c_1\sqrt{c_0})$; then we deduce

$$2\int_S^T\int_\Omega |u'|^2dx\,dt \le \frac{1}{c_1}\left(E(S)-E(T)\right) \tag{3.6}$$

$$+\beta\int_S^T\int_\Omega\left(\frac{1}{2}\left(\sqrt{c_0+\left(\frac{2}{c_1}\right)^2}+\sqrt{c_0}\right)|u'|^2+\frac{1}{2}\left(\sqrt{c_0+\left(\frac{2}{c_1}\right)^2}-\sqrt{c_0}\right)|\nabla u|^2\right)dx\,dt.$$

Similary we have

$$-\int_S^T\int_\Omega ug(u')dx\,dt \le \frac{1}{2}\int_S^T\int_\Omega\left(\frac{1}{\epsilon}g^2(u')+\epsilon|u|^2\right)dx\,dt$$

$$\le \frac{1}{2}\int_S^T\int_\Omega\left(\frac{c_2}{\epsilon}u'g(u')+\epsilon c_0|\nabla u|^2\right)dx\,dt$$

$$=\frac{c_2}{2\epsilon}\int_S^T\left(-\frac{1}{2}E'(t)-\int_\Omega u'h(\nabla u)dx\right)dt+\frac{\epsilon c_0}{2}\int_S^T\int_\Omega|\nabla u|^2dx\,dt$$

$$\leq \frac{c_2}{4\epsilon}\left(E(S)-E(T)\right)+\frac{\epsilon c_0}{2}\int_S^T\int_\Omega |\nabla u|^2 dx\,dt$$

$$+\frac{c_2}{2\epsilon}\int_S^T\int_\Omega\left(\frac{\epsilon'\beta^2}{2}|\nabla u|^2+\frac{1}{2\epsilon'}|u'|^2\right)dx\,dt;$$

we choose $\epsilon=\beta\sqrt{\frac{c_2\epsilon'}{2c_0}}$ and $\epsilon'=\frac{1}{\sqrt{2}\beta}$. It follows that

$$-\int_S^T\int_\Omega ug(u')dx\,dt\leq\sqrt{\frac{c_0c_2\beta}{2\sqrt{2}}}\int_S^T\int_\Omega\left(|u'|^2+|\nabla u|^2\right)dx\,dt + \frac{1}{2}\sqrt{\frac{c_0c_2}{\sqrt{2}\beta}}\left(E(S)-E(T)\right). \tag{3.7}$$

Combining (3.5), (3.6), and (3.7), we conclude that

$$\begin{aligned}&\left(1-\frac{\beta}{2}\left(\sqrt{c_0+\left(\frac{2}{c_1}\right)^2}+\sqrt{c_0}\right)-\sqrt{\frac{c_0c_2\beta}{2\sqrt{2}}}\right)\int_S^T\int_\Omega\left(|u'|^2+|\nabla u|^2\right)dx\,dt\\&+b\int_S^T\int_\Omega 2F(u)dx\,dt\\&\leq\left(\frac{k\sqrt{c_0}}{2}+\frac{1}{c_1}+\frac{1}{2}\frac{\sqrt{c_0c_2}}{\sqrt{2}\beta}\right)E(S)+\left(\frac{k\sqrt{c_0}}{2}-\frac{1}{c_1}-\frac{1}{2}\frac{\sqrt{c_0c_2}}{\sqrt{2}\beta}\right)E(T).\end{aligned} \tag{3.8}$$

Hence, if we take β small enough so that $\frac{\beta}{2}(\sqrt{c_0+(\frac{2}{c_1})^2}+\sqrt{c_0})+\sqrt{\frac{c_0c_2\beta}{2\sqrt{2}}}\leq 1-b$ as it is assumed in Theorem 2.1, then, from (3.8) and making use of definition (2.7) of energy, we arrive at

$$\int_S^T E(t)dt \tag{3.9}$$

$$\leq\frac{k\sqrt{c_0}}{2b}\left(E(S)+E(T)\right)+\frac{1}{b}\left(\frac{1}{c_1}+\frac{1}{2}\sqrt{\frac{c_0c_2}{\sqrt{2}\beta}}\right)\left(E(S)-E(T)\right).$$

If F is positive, then we assume that $\frac{\beta}{2}(\sqrt{c_0+(\frac{2}{c_1})^2}+\sqrt{c_0})+\sqrt{\frac{c_0c_2\beta}{2\sqrt{2}}}<1$ and we obtain (3.9) with b replaced by

$$\bar{b}=\min\left\{b,1-\frac{\beta}{2}\left(\sqrt{c_0+\left(\frac{2}{c_1}\right)^2}+\sqrt{c_0}\right)-\sqrt{\frac{c_0c_2\beta}{2\sqrt{2}}}\right\}.$$

Now we return to equality (3.2). Using (2.4), (2.6), (2.8), and the Cauchy–Schwarz inequality, we infer

$$E'(t)\leq -2\int_\Omega u'h(\nabla u)dx\leq\int_\Omega\left(\beta|u'|^2+\frac{1}{\beta}|h(\nabla u)|^2\right)dx$$

$$\leq \beta \int_\Omega \left(|u'|^2 + |\nabla u|^2\right) dx \leq \beta k E(t);$$

then

$$E'(t) \leq \beta k E(t). \tag{3.10}$$

We may assume in the rest of this section that $E(t) > 0$ for all $t \geq 0$. Otherwise if $E(t_0) = 0$ for some $t_0 \geq 0$, then from (2.8) we have $u(t_0, x) = u'(t_0, x) = 0$ in Ω; hence $v(t, x) := u(t + t_0, x)$ solves (P) with $(0, 0)$ as initial data. By the uniqueness of solution we conclude that $v = v' = 0$; hence $E(t) = 0$ for all $t \geq t_0$ and then we have nothing to prove.

Now by Gronwall's lemma, we conclude from (3.10) that

$$E(t) \leq e^{\beta k(t-\tau)} E(\tau) \qquad \forall\, 0 \leq \tau \leq t < \infty. \tag{3.11}$$

On the other hand, (3.10) implies that

$$E(t) \geq \frac{1}{\beta k} \frac{\partial}{\partial t} \left((1 - e^{-\beta k(t-\tau)}) E(t) \right) \qquad \forall\, 0 \leq \tau \leq t < \infty. \tag{3.12}$$

Now we distinguish two cases (corresponding to the hypothesis assumed on β in Theorem 2.1).

Case 1. $\beta < \frac{b}{k^2 \sqrt{c_0}}$. We fix

$$T_0 > \frac{-1}{\beta k} \ln \left(1 - \frac{\beta k^2 \sqrt{c_0}}{b} \right). \tag{3.13}$$

From (3.12) with $\tau = S$ we have

$$\int_S^{S+T_0} E(t) dt \geq \frac{1}{\beta k} (1 - e^{-\beta k T_0}) E(S + T_0).$$

Combining this inequality and (3.9) with $T = S + T_0$, we arrive at

$$\left(\frac{1}{\beta k} (1 - e^{-\beta k T_0}) + \frac{1}{b} \left(\frac{1}{c_1} + \frac{1}{2} \sqrt{\frac{c_0 c_2}{\sqrt{2}\beta}} \right) - \frac{k\sqrt{c_0}}{2b} \right) E(S + T_0)$$

$$\leq \left(\frac{k\sqrt{c_0}}{2b} + \frac{1}{b} \left(\frac{1}{c_1} + \frac{1}{2} \sqrt{\frac{c_0 c_2}{\sqrt{2}\beta}} \right) \right) E(S).$$

Thanks to our choice (3.13) of T_0, we have

$$\frac{1}{\beta k} (1 - e^{-\beta k T_0}) > \frac{k\sqrt{c_0}}{b};$$

then we obtain (3.1) with

$$d = \frac{\frac{1}{b} \left(\frac{1}{c_1} + \frac{1}{2} \sqrt{\frac{c_0 c_2}{\sqrt{2}\beta}} \right) + \frac{k\sqrt{c_0}}{2b}}{\frac{1}{\beta k} (1 - e^{-\beta k T_0}) + \frac{1}{b} \left(\frac{1}{c_1} + \frac{1}{2} \sqrt{\frac{c_0 c_2}{\sqrt{2}\beta}} \right) - \frac{k\sqrt{c_0}}{2b}} \in]0, 1[.$$

We note that if a nonnegative function $E:\mathbb{R}^+\to\mathbb{R}^+$ satisfies the estimate (3.1), then it also satisfies (2.11). Indeed, let $t\in\mathbb{R}^+$; then $t=mT_0+t_0$ with $0\le t_0<T_0$ and $m\in\mathbb{N}$. From (3.1) and taking (3.11) with $t=t_0$ and $\tau=0$ into account, it holds that

$$E(t)\le dE((m-1)T_0+t_0)\le\cdots\le d^mE(t_0)$$

$$\le d^{\frac{1}{T_0}(t-t_0)}e^{\beta kt_0}E(0)\le\frac{e^{\beta kT_0}}{d}E(0)e^{\frac{\ln d}{T_0}t};$$

then we deduce (2.11), where $c=\frac{e^{\beta kT_0}}{d}$ and $\omega=-\frac{\ln d}{T_0}$.

Case 2. $\frac{k\sqrt{c_0}}{2}\le\frac{1}{c_1}+\frac{1}{2}\sqrt{\frac{c_0c_2}{\sqrt{2}\beta}}$. Inequality (3.9) implies that

$$\int_S^T E(t)dt\le a_0E(S)\qquad\forall 0\le S\le T<\infty,$$

where $a_0=\frac{k\sqrt{c_0}}{2b}+\frac{1}{b}(\frac{1}{c_1}+\frac{1}{2}\sqrt{\frac{c_0c_2}{\sqrt{2}\beta}})$. Let T go to ∞; we deduce

$$\int_S^\infty E(t)dt\le a_0E(S)\qquad\forall S\ge 0. \tag{3.14}$$

Introduce the function

$$\psi(S)=\int_S^\infty E(t)dt,\qquad S\ge 0.$$

It is positive and nonincreasing. Differentiating and using (3.14), we find that

$$\psi'(S)\le-\frac{1}{a_0}\psi(S),$$

hence $(\ln(\psi(S)))'\le-\frac{1}{a_0}$. Integrating in $[0,S]$ and using (3.14) again, we obtain that

$$\psi(S)\le a_0E(0)e^{-\frac{1}{a_0}S}\qquad\forall S\ge 0. \tag{3.15}$$

On the other hand, E being nonnegative and satisfying (3.12) (with $\tau=S$), $\psi(S)$ may be estimated as follows: let $T_0>0$,

$$\psi(S)\ge\int_S^{S+T_0}E(t)dt\ge\int_S^{S+T_0}\frac{1}{\beta k}\frac{\partial}{\partial t}\left((1-e^{-\beta k(t-S)})E(t)\right)dt$$

$$=\frac{1-e^{-\beta kT_0}}{\beta k}E(S+T_0).$$

Therefore, taking $t=S+T_0$ and choosing $T_0=\frac{1}{\beta k}\ln(1+\beta ka_0)$ (for which the quantity $\frac{e^{T_0/a_0}}{1-e^{-\beta kT_0}}$ reachs its minimum), hence we deduce from (3.15) the estimate

$$E(t)\le(1+\beta ka_0)^{1+\frac{1}{\beta ka_0}}E(0)e^{-\frac{1}{a_0}t}\qquad\forall t\ge T_0. \tag{3.16}$$

This inequality holds, in fact, also for $t\in[0,T_0]$. Indeed, by (3.11) with $\tau=0$, we have

$$E(t)\le e^{\beta kt}E(0)\le e^{(\beta k+\frac{1}{a_0})T_0}E(0)e^{-\frac{1}{a_0}t}=(1+\beta ka_0)^{1+\frac{1}{\beta ka_0}}E(0)e^{-\frac{1}{a_0}t}.$$

Then (3.16) holds true for all $t\ge 0$ and hence the inequality (2.11) follows with $c=(1+\beta ka_0)^{1+\frac{1}{\beta ka_0}}$ and $\omega=\frac{1}{a_0}$.

This concludes the proof of Theorem 2.1.

4. Energy decay estimates: Proof of Theorem 2.2. For the proof of Theorem 2.2 which concerns the stability of (P) in the particular case $h(\nabla u) = -\nabla\phi \cdot \nabla u$, with $\phi \in W^{1,\infty}(\Omega)$, we are going to prove that the equivalent energy E defined by (2.14) satisfies, for any $0 \le S < \infty$,

$$\int_S^\infty E^{\frac{r+1}{2}}(t)dt \le cE(S). \tag{4.1}$$

Here and in what follows we shall denote by c diverse positive constants, by ϵ diverse positive constants small enough, and by c_ϵ diverse positive constants depending on ϵ. (All these constants do not depend on S.) The inequality (4.1) gives (2.11) and (2.15) (see [12, Theorem 9.1]).

Using the first equation of (P) and the boundary condition, we can easily prove that the equivalent energy E satisfies

$$E'(t) = -2\int_\Omega e^{\phi(x)}u'g(u')dx, \qquad t \in \mathbb{R}^+. \tag{4.2}$$

Assumption (2.4) implies that the equivalent energy is nonincreasing. Given $0 \le S \le T < \infty$ arbitrarily, integrate (4.2) between S and T to get

$$\int_S^T \int_\Omega e^{\phi(x)}u'g(u')dx = \frac{1}{2}\left(E(S) - E(T)\right). \tag{4.3}$$

We multiply the first equation of (P) by $E^{\frac{r-1}{2}}(t)e^{\phi(x)}u$ and integrate over $\Omega\times[S,T]$ to get

$$\int_S^T \int_\Omega E^{\frac{r-1}{2}}(t)e^{\phi(x)}\left(\left|u'\right|^2 + \left|\nabla u\right|^2 + uf(u)\right)dxdt \tag{4.4}$$

$$= \int_S^T \int_\Omega E^{\frac{r-1}{2}}(t)e^{\phi(x)}\left(2\left|u'\right|^2 - ug(u')\right)dxdt$$

$$+ \frac{r-1}{2}\int_S^T \int_\Omega E^{\frac{r-3}{2}}(t)E'(t)e^{\phi(x)}uu'dxdt - \left[\int_\Omega E^{\frac{r-1}{2}}(t)e^{\phi(x)}uu'dxdt\right]_S^T.$$

The last two terms of (4.4) can be easily majorized by $cE^{\frac{r+1}{2}}(S)$ (see [3] and [5]). We follow now the proof given in [5]. We note $q = p + 1$,

$$\Omega^+ = \{x \in \Omega : \left|u'\right| > 1\}, \quad \text{and} \quad \Omega^- = \Omega \setminus \Omega^+.$$

We exploit the Cauchy–Schwarz, Hölder, and Young inequalities and the Sobolev imbedding $H_0^1(\Omega) \subset L^q(\Omega)$ to get

$$-\int_S^T \int_{\Omega^+} E^{\frac{r-1}{2}}(t)e^{\phi(x)}ug(u')dxdt$$

$$\le \int_S^T E^{\frac{r-1}{2}}(t)e^{\phi(x)}\left(\int_{\Omega^+}\left|u\right|^q dx\right)^{\frac{1}{q}}\left(\int_{\Omega^+}\left|g(u')\right|^{1+\frac{1}{p}}dx\right)^{\frac{p}{p+1}}dt$$

$$\leq \int_S^T E^{\frac{r-1}{2}}(t)e^{\phi(x)}\left(\epsilon\int_{\Omega^+}\left|u\right|^q dx + c_\epsilon\int_{\Omega^+}\left|g(u')\right|^{1+\frac{1}{p}}dx\right)dt$$

$$\leq \epsilon\int_S^T E^{\frac{r+q-1}{2}}(t)dt + c_\epsilon E^{\frac{r-1}{2}}(S)\int_S^T\int_{\Omega^+} e^{\phi(x)}u'g(u')dxdt$$

$$\leq \epsilon\int_S^T E^{\frac{r+1}{2}}(t)dt + c_\epsilon\left(E^{\frac{r+1}{2}}(S) - E^{\frac{r+1}{2}}(T)\right).$$

On the other hand, using the growth assumption (2.12) and Poincaré's inequality, we have

$$-\int_S^T\int_{\Omega^-} E^{\frac{r-1}{2}}(t)e^{\phi(x)}ug(u')dxdt$$

$$\leq \int_S^T E^{\frac{r-1}{2}}(t)e^{\phi(x)}\left(\epsilon\int_{\Omega^-}\left|u\right|^2 dx + c_\epsilon\int_{\Omega^-} g^2(u')dx\right)dt$$

$$\leq \epsilon\int_S^T E^{\frac{r-1}{2}}(t)\int_{\Omega^-} e^{\phi(x)}\left|\nabla u\right|^2 dxdt + c_\epsilon\int_S^T\int_{\Omega^-} E^{\frac{r-1}{2}}(t)\left(e^{\phi(x)}u'g(u')\right)^{\frac{2}{r+1}}dxdt$$

$$\leq \epsilon\int_S^T E^{\frac{r+1}{2}}(t)dt + c_\epsilon\left(E(S) - E(T)\right).$$

Taking the sum of the last two inequalities and substituting it into the right-hand side of (4.4), using (2.3), and choosing $\epsilon \in]0, b[$, we obtain that

$$\int_S^T E^{\frac{r+1}{2}}(t)dt \leq c\left(E^{\frac{r+1}{2}}(S) + E(S)\right) + c\int_S^T\int_\Omega E^{\frac{r-1}{2}}(t)e^{\phi(x)}\left|u'\right|^2 dxdt. \tag{4.5}$$

Using another time (2.12) and (4.3), we have

$$\int_S^T\int_{\Omega^+} E^{\frac{r-1}{2}}(t)e^{\phi(x)}\left|u'\right|^2 dxdt \leq cE^{\frac{r-1}{2}}(S)\int_S^T\int_{\Omega^+} e^{\phi(x)}u'g(u')dxdt$$

$$\leq c\left(E^{\frac{r+1}{2}}(S) - E^{\frac{r+1}{2}}(T)\right).$$

In the same way, using Young's inequality, we get

$$\int_S^T\int_{\Omega^-} E^{\frac{r-1}{2}}(t)e^{\phi(x)}\left|u'\right|^2 dxdt \leq c\int_S^T\int_{\Omega^-} E^{\frac{r-1}{2}}(t)\left(e^{\phi(x)}u'g(u')\right)^{\frac{2}{r+1}}dxdt$$

$$\leq \epsilon\int_S^T E^{\frac{r+1}{2}}(t)dt + c_\epsilon\int_S^T\int_{\Omega^-} e^{\phi(x)}u'g(u')dxdt$$

$$\leq \epsilon\int_S^T E^{\frac{r+1}{2}}(t)dt + c_\epsilon\left(E(S) - E(T)\right).$$

Substituting the sum of these two estimates into the right-hand side of (4.5), choosing ϵ small enough, and letting T go to ∞, we obtain

$$\int_S^\infty E^{\frac{r+1}{2}}(t)dt \leq c\left(1 + E^{\frac{r-1}{2}}(0)\right)E(S) \leq cE(S);$$

then (4.1) follows, which gives (2.11) and (2.15) and finishes the proof of Theorem 2.2.

5. Uniform decay: Proof of Theorem 2.3. In this section we prove the exponential decay of energy (2.7) for strong solutions of (P′), and by a density argument we obtain the same results for weak solutions.

The proof is similar to the one given in section 3.

Using the first equation in (P′) and the boundary conditions, we can easily prove that

$$E'(t) = -2\int_{\Gamma_1} u' g(u')\, d\Gamma - 2\int_\Omega u' h(\nabla u)\, dx. \tag{5.1}$$

Using Assumptions 2.2, 2.3, and 2.5, from (5.1) it holds that (see section 3)

$$E'(t) \le -2\int_\Omega u' h(\nabla u)\, dx \le \beta \int_\Omega \Big(|u'|^2 + |\nabla u|^2\Big)\, dx \le \beta E(t);$$

then E satisfies (3.11) and (3.12) with $k = 1$ (see (2.18)). Following the proof given in section 3, it is sufficient to prove that, for all $0 \le S \le T < \infty$,

$$\int_S^T E(t)dt \le \bar{a}(E(S) + E(T)) + \hat{a}(E(S) - E(T)) \tag{5.2}$$

with $\bar{a}$, $\hat{a} > 0$ and $2\beta\bar{a} < 1$ or $\bar{a} \le \hat{a}$. Then the proof can be completed as in section 3.

To prove (5.2), let $\epsilon_0 \in]0, 1[$ (will be chosen later in the course of the proof); we multiply the first equation in (P′) by

$$2m \cdot \nabla u + (n - \epsilon_0)u,$$

integrating the obtained result over $\Omega \times [S, T]$ and using the boundary conditions. We are going to estimate the terms of the result formula. We have

$$I_1 := \int_S^T \int_\Omega u''\Big(2m \cdot \nabla u + (n - \epsilon_0)u\Big)dxdt$$

$$= \Big[\int_\Omega u'\Big(2m \cdot \nabla u + (n - \epsilon_0)u\Big)dx\Big]_S^T - \int_S^T \int_\Omega \Big(m \cdot \nabla (u')^2 + (n - \epsilon_0)\,|u'|^2\Big)dxdt$$

$$= \epsilon_0 \int_S^T \int_\Omega |u'|^2\, dxdt - \int_S^T \int_{\Gamma_1} (m \cdot \nu)\,|u'|^2\, d\Gamma dt$$

$$+ \Big[\int_\Omega u'\Big(2m \cdot \nabla u + (n - \epsilon_0)u\Big)dx\Big]_S^T.$$

We estimate the last term in this inequality; we have

$$\int_\Omega \Big(2m \cdot \nabla u + (n - \epsilon_0)u\Big)^2 dx - \int_\Omega (2m \cdot \nabla u)^2 dx$$

$$= \int_\Omega \Big((n - \epsilon_0)^2\,|u|^2 + 2(n - \epsilon_0)m \cdot \nabla (u)^2\Big)\, dx$$

$$= \int_\Omega \Big((n-\epsilon_0)^2 |u|^2 - 2(n-\epsilon_0) n |u|^2 \Big) dx + 2(n-\epsilon_0) \int_{\Gamma_1} (m\cdot\nu) |u|^2 d\Gamma$$

$$= (\epsilon_0+n)(\epsilon_0-n) \int_\Omega |u|^2 dx + 2(n-\epsilon_0) \int_{\Gamma_1} (m\cdot\nu) |u|^2 d\Gamma$$

$$\leq 2(n-\epsilon_0) R \int_{\Gamma_1} |u|^2 d\Gamma;$$

then

$$\int_\Omega \Big(2m\cdot\nabla u + (n-\epsilon_0)u \Big)^2 dx \leq \int_\Omega (2m\cdot\nabla u)^2 dx + 2(n-\epsilon_0) R \int_{\Gamma_1} |u|^2 d\Gamma. \tag{5.3}$$

Since, for all $\epsilon > 0$,

$$\Big| \int_\Omega \Big(2m\cdot\nabla u + (n-\epsilon_0)u \Big) u' dx \Big|$$

$$\leq \frac{\epsilon}{2} \int_\Omega |u'|^2 dx + \frac{1}{2\epsilon} \Big(\int_\Omega (2m\cdot\nabla u)^2 dx + 2(n-\epsilon_0) R \int_{\Gamma_1} |u|^2 d\Gamma \Big)$$

$$\leq \int_\Omega \left(\frac{\epsilon}{2} |u'|^2 + \frac{2R^2}{\epsilon} |\nabla u|^2 dx \right) + \frac{R}{\epsilon} (n-\epsilon_0) \bar{c} \int_\Omega |\nabla u|^2 dx,$$

where $\bar{c}$ is the positive constant satisfying (Poincaré's inequality)

$$\int_{\Gamma_1} |v|^2 d\Gamma \leq \bar{c} \int_\Omega |\nabla v|^2 dx \qquad \forall v \in V.$$

Choosing $\epsilon = 2\sqrt{R(R + \frac{\bar{c}}{2}(n-\epsilon_0))}$, we obtain

$$\Big| \int_\Omega \Big(2m\cdot\nabla u + (n-\epsilon_0)u \Big) u' dx \Big| \leq \sqrt{R\Big(R + \frac{\bar{c}}{2}(n-\epsilon_0) \Big)} E(t) := a_1 E(t).$$

Then we deduce

$$I_1 \geq -a_1 (E(S) + E(T)) - R \int_S^T \int_{\Gamma_1} |u'|^2 d\Gamma dt + \epsilon_0 \int_S^T \int_\Omega |u'|^2 dx dt. \tag{5.4}$$

On the other hand, taking the generalized Green formula and recalling the identity

$$2\nabla u \cdot \nabla (m\cdot\nabla u) = 2|\nabla u|^2 + m\cdot\nabla(|\nabla u|^2)$$

(note also that on Γ_0 we have $\nabla u = \partial_\nu u \nu$), we infer

$$I_2 := \int_S^T \int_\Omega (-\Delta u) \Big(2m\cdot\nabla u + (n-\epsilon_0)u \Big) dx dt$$

$$= (2-\epsilon_0) \int_S^T \int_\Omega |\nabla u|^2 dx dt - \int_S^T \int_{\Gamma_0} (m\cdot\nu) |\nabla u|^2 d\Gamma dt$$

$$+ \int_S^T \int_{\Gamma_1} \Big((m\cdot\nu) |\nabla u|^2 - (n-\epsilon_0) u \partial_\nu u - 2(m\cdot\nabla u)\partial_\nu u \Big) d\Gamma dt.$$

Using the definition of Γ_0 and Γ_1, we deduce

$$I_2 \geq (2-\epsilon_0)\int_S^T\int_\Omega |\nabla u|^2\,dxdt$$

$$+\int_S^T\int_{\Gamma_1}\Big(\delta\,|\nabla u|^2-(n-\epsilon_0)u\partial_\nu u-\delta\,|\nabla u|^2-\frac{R^2}{\delta}(\partial_\nu u)^2\Big)d\Gamma dt;$$

then

$$I_2 \geq (2-\epsilon_0)\int_S^T\int_\Omega |\nabla u|^2\,dxdt-\int_S^T\int_{\Gamma_1}\Big((n-\epsilon_0)u\partial_\nu u+\frac{R^2}{\delta}(\partial_\nu u)^2\Big)d\Gamma dt. \tag{5.5}$$

Similarly, using (2.6), (5.3), and the Cauchy–Schwarz inequality, we have

$$I_3 := \int_S^T\int_\Omega h(\nabla u)\Big(2m\cdot\nabla u+(n-\epsilon_0)u\Big)dxdt$$

$$\geq -\frac{R}{\beta}\int_S^T\int_\Omega h^2(\nabla u)dxdt-\frac{\beta}{4R}\int_S^T\Big(4R^2\int_\Omega |\nabla u|^2\,dx+2(n-\epsilon_0)R\int_{\Gamma_1}|u|^2\,d\Gamma\Big)dt;$$

we conclude that

$$I_3 \geq -2\beta R\int_S^T\int_\Omega |\nabla u|^2\,dxdt-\frac{\beta}{2}(n-\epsilon_0)\int_S^T\int_{\Gamma_1}|u|^2\,d\Gamma dt. \tag{5.6}$$

Using (2.3) and the fact that F is nonnegative and $F(0)=0$, we obtain

$$\begin{aligned} I_4 &:= \int_S^T\int_\Omega f(u)\Big(2m\cdot\nabla u+(n-\epsilon_0)u\Big)dxdt\\ &\geq (n-\epsilon_0)b\int_S^T\int_\Omega 2F(u)dxdt+\int_S^T\int_\Omega 2m\cdot\nabla(F(u))dxdt\\ &\geq ((n-\epsilon_0)b-n)\int_S^T\int_\Omega 2F(u)dxdt+\int_S^T\int_{\Gamma_1}2(m\cdot\nu)F(u)d\Gamma dt; \end{aligned}$$

then we deduce

$$I_4 \geq ((n-\epsilon_0)b-n)\int_S^T\int_\Omega 2F(u)dxdt. \tag{5.7}$$

Now we distinguish two cases.

Case 3. If $b>1$, then assuming that $\beta R<1$ and choosing $\epsilon_0=\min\{1-\beta R,\frac{b-1}{b+1}n\}$, we deduce that $\min\{\epsilon_0,2-\epsilon_0-2\beta R,(n-\epsilon_0)b-n\}=\epsilon_0$. Combining (5.4)–(5.7), taking the fact that $I_1+I_2+I_3+I_4=0$ in account, we obtain

$$\epsilon_0\int_S^T\int_\Omega\Big(|u'|^2+|\nabla u|^2+2F(u)\Big)dxdt\leq a_1(E(S)+E(T))$$

$$+\int_S^T\int_{\Gamma_1}\left(R\left|u'\right|^2+\frac{\beta}{2}(n-\epsilon_0)\left|u\right|^2+(n-\epsilon_0)u\partial_\nu u+\frac{R^2}{\delta}(\partial_\nu u)^2\right)d\Gamma dt.$$

Case 4. If f is linear, $f(s)=\alpha s$ for some positive constant α, then $b=1$ and we conclude from (5.7) that

$$\begin{aligned}I_4&\geq-\epsilon_0\int_S^T\int_\Omega 2F(u)dxdt=\epsilon_0\int_S^T\int_\Omega 2F(u)dxdt-2\epsilon_0\int_S^T\int_\Omega 2F(u)dxdt\\&=\epsilon_0\int_S^T\int_\Omega 2F(u)dxdt-2\epsilon_0\alpha\int_S^T\int_\Omega\left|u\right|^2dxdt\\&\geq\epsilon_0\int_S^T\int_\Omega 2F(u)dxdt-2\epsilon_0\alpha\hat{c}\int_S^T\int_\Omega\left|\nabla u\right|^2dxdt,\end{aligned}$$

where $\hat{c}$ is the smallest imbedding positive constant satisfying

$$\int_\Omega\left|v\right|^2dx\leq\hat{c}\int_\Omega\left|\nabla v\right|^2dx\qquad\forall v\in V. \tag{5.8}$$

Assuming that $\beta R<1$ and choosing $\epsilon_0=\frac{1-\beta R}{1+\alpha\hat{c}}$, then $\min\{\epsilon_0,2-\epsilon_0-2\beta R-2\epsilon_0\alpha\hat{c}\}=\epsilon_0$ and the same inequality obtained in Case 3 holds true.

We now use the boundary condition on Γ_1; we have in both previous cases

$$\epsilon_0\int_S^T E(t)dt\leq a_1(E(S)+E(T)) \tag{5.9}$$

$$+\int_S^T\int_{\Gamma_1}\left(R\left|u'\right|^2+\frac{R^2}{\delta}g^2(u')+\frac{\beta}{2}(n-\epsilon_0)\left|u\right|^2-(n-\epsilon_0)ug(u')\right)d\Gamma dt.$$

Using (5.1), the Cauchy–Schwarz inequality and taking the assumptions (2.4), (2.5), and (2.6) into account, it holds that

$$\int_S^T\int_{\Gamma_1}\left(R|u'|^2+\frac{R^2}{\delta}g^2(u')\right)dx\,dt\leq\left(\frac{R}{c_1}+\frac{R^2}{\delta}c_2\right)\int_S^T\int_{\Gamma_1}u'g(u')dx\,dt$$

$$=\frac{1}{2}\left(\frac{R}{c_1}+\frac{R^2}{\delta}c_2\right)\int_S^T\left(-E'(t)-2\int_\Omega u'h(\nabla u)dx\right)dt$$

$$\leq\frac{1}{2}\left(\frac{R}{c_1}+\frac{R^2}{\delta}c_2\right)(E(S)-E(T))+\frac{1}{2}\left(\frac{R}{c_1}+\frac{R^2}{\delta}c_2\right)\beta\int_S^T\int_\Omega\left(|u'|^2+|\nabla u|^2\right)dx\,dt;$$

we note $a_2:=\frac{1}{2}(\frac{R}{c_1}+\frac{R^2}{\delta}c_2)$ and deduce

$$\int_S^T\int_{\Gamma_1}\left(R|u'|^2+\frac{R^2}{\delta}g^2(u')\right)dx\,dt\leq a_2\left(E(S)-E(T)\right)+\beta a_2\int_S^T E(t)dt. \tag{5.10}$$

Similarly, we have

$$
\begin{aligned}
&(n-\epsilon_0)\int_S^T\int_{\Gamma_1}\left(\frac{\beta}{2}|u|^2-ug(u')\right)dx\,dt\\
&\le\frac{1}{2}(n-\epsilon_0)\int_S^T\int_{\Gamma_1}\left(\frac{1}{\epsilon}g^2(u')+(\beta+\epsilon)|u|^2\right)dx\,dt\\
&\le\frac{1}{2}(n-\epsilon_0)\int_S^T\int_{\Gamma_1}\left(\frac{c_2}{\epsilon}u'g(u')+(\beta+\epsilon)|u|^2\right)dx\,dt\\
&=\frac{c_2}{2\epsilon}(n-\epsilon_0)\int_S^T\left(-\frac{1}{2}E'(t)-\int_\Omega u'h(\nabla u)dx\right)dt\\
&\quad+\frac{1}{2}(\beta+\epsilon)(n-\epsilon_0)\bar{c}\int_S^T\int_\Omega|\nabla u|^2dx\,dt\\
&\le\frac{c_2}{4\epsilon}(n-\epsilon_0)(E(S)-E(T))+\frac{1}{2}(\beta+\epsilon)(n-\epsilon_0)\bar{c}\int_S^T\int_\Omega|\nabla u|^2dx\,dt\\
&\quad+\frac{c_2}{2\epsilon}(n-\epsilon_0)\int_S^T\int_\Omega\left(\frac{\epsilon'\beta^2}{2}|\nabla u|^2+\frac{1}{2\epsilon'}|u'|^2\right)dx\,dt,
\end{aligned}
$$

we choose $\epsilon=\beta\sqrt{\frac{c_2\epsilon'}{2\bar{c}}}$, $\epsilon'=\frac{1}{\beta\sqrt{2}}$, and we note $a_3:=\frac{1}{2}(n-\epsilon_0)\sqrt{\frac{\bar{c}c_2}{\sqrt{2}\beta}}$, $a_4:=(n-\epsilon_0)(\frac{\beta\bar{c}}{2}+\sqrt{\frac{\bar{c}c_2\beta}{2\sqrt{2}}})$. It follows that

$$
\begin{aligned}
&(n-\epsilon_0)\int_S^T\int_{\Gamma_1}\left(\frac{\beta}{2}|u'|^2-ug(u')\right)dx\,dt\\
&\le a_4\int_S^T E(t)dt+a_3\left(E(S)-E(T)\right).
\end{aligned}
\tag{5.11}
$$

Combining (5.9), (5.10), and (5.11), we have

$$
\begin{aligned}
&(\epsilon_0-\beta a_2-a_4)\int_S^T E(t)dt\\
&\le a_1\left(E(S)+E(T)\right)+(a_2+a_3)\left(E(S)-E(T)\right).
\end{aligned}
\tag{5.12}
$$

If β is small enough so that $2\beta a_1<a_5:=\epsilon_0-\beta a_2-a_4$, that is,

$$
\beta(2a_1+a_2)+a_4<\epsilon_0=\begin{cases}\min\{1-\beta R,\frac{b-1}{b+1}n\} & \text{if } b>1,\\ \frac{1-\beta R}{1+\alpha\hat{c}} & \text{if } f \text{ is linear}\end{cases}
$$

(note that $\beta(2a_1+a_2)+a_4$ goes to 0 when β goes to 0), we conclude (5.2) with $\bar{a}=\frac{a_1}{a_5}$ and $\hat{a}=\frac{a_2+a_3}{a_5}$. We fix then $T_0>\frac{-1}{\beta}\ln(1-2\beta\bar{a})$. Using (3.12) with $\tau=S$, we have

$$
\int_S^{S+T_0}E(t)dt\ge\frac{1}{\beta}(1-e^{-\beta T_0})E(S+T_0).
$$

We insert this inequality into (5.2) with $T = S + T_0$ and obtain

$$\left(\frac{1}{\beta}(1 - e^{-\beta T_0}) + \hat{a} - \bar{a}\right) E(S + T_0) \leq (\hat{a} + \bar{a}) E(S).$$

Thanks to the hypothesis on T_0, we have $\frac{1}{\beta}(1 - e^{-\beta T_0}) > 2\bar{a}$, which implies (3.1) with $d = \frac{\hat{a}+\bar{a}}{\frac{1}{\beta}(1-e^{-\beta T_0})+\hat{a}-\bar{a}}$.

If $\beta a_2 + a_4 < \epsilon_0$ and $a_1 \leq a_2 + a_3$ (that is, $\sqrt{R(R + \frac{\bar{c}}{2}(n - \epsilon_0))} \leq \frac{1}{2}(\frac{R}{c_1} + \frac{R^2}{\delta} c_2) + \frac{1}{2}(n - \epsilon_0)\sqrt{\frac{\bar{c} c_2}{\sqrt{2}\beta}}$), we conclude from (5.12) that (3.14) follows with $a_0 = \frac{a_1+a_2+a_3}{a_5}$.

Then in both cases the proof of Theorem 2.3 can be completed as in section 3.

6. Decay estimates: Proof of Theorem 2.4. To prove Theorem 2.4, which concerns the stability of (P′) in the particular case $h(\nabla u) = -\nabla\phi \cdot \nabla u$, with $\phi \in W^{1,\infty}(\Omega)$, it is sufficient to prove that the equivalent energy E defined by (2.14) satisfies (4.1) (see section 4).

In this section, we shall denote by c diverse positive constants, by ϵ diverse positive constants small enough (which can be changed from a line to another), and by c_ϵ diverse positive constants depending on ϵ.

A simple computation shows that

$$E'(t) = -2 \int_{\Gamma_1} e^{\phi(x)} u' g(u') dx, \qquad t \in \mathbb{R}^+. \tag{6.1}$$

Assumption (2.4) implies that the equivalent energy is nonincreasing.

We fix $\epsilon_0 > 0$ and we multiply the first equation in (P′) by

$$E^{\frac{r-1}{2}}(t) e^{\phi(x)} \Big(2m \cdot \nabla u + (n - \epsilon_0) u\Big),$$

integrating the obtained result over $\Omega \times [S, T]$ and using the boundary conditions. We have

$$\begin{aligned}
I_1 &:= \int_S^T \int_\Omega E^{\frac{r-1}{2}}(t) e^{\phi(x)} u'' \Big(2m \cdot \nabla u + (n - \epsilon_0) u\Big) dx dt \\
&= \left[\int_\Omega E^{\frac{r-1}{2}}(t) e^{\phi(x)} u' \Big(2m \cdot \nabla u + (n - \epsilon_0) u\Big) dx\right]_S^T \\
&\quad - \frac{r-1}{2} \int_S^T \int_{\Gamma_1} E^{\frac{r-3}{2}}(t) E'(t) e^{\phi(x)} \Big(2m \cdot \nabla u + (n - \epsilon_0) u\Big) dx dt \\
&\quad - \int_S^T \int_\Omega E^{\frac{r-1}{2}}(t) e^{\phi(x)} \int_\Omega \Big(m \cdot \nabla (u')^2 + (n - \epsilon_0) |u'|^2\Big) dx dt \\
&= \int_S^T \int_\Omega E^{\frac{r-1}{2}}(t) e^{\phi(x)} (\epsilon_0 + m \cdot \nabla\phi) |u'|^2 dx dt - \int_S^T \int_{\Gamma_1} E^{\frac{r-1}{2}}(t) e^{\phi(x)} (m \cdot \nu) |u'|^2 d\Gamma dt \\
&\quad + \Big[\int_\Omega E^{\frac{r-1}{2}}(t) e^{\phi(x)} u' \Big(2m \cdot \nabla u + (n - \epsilon_0) u\Big) dx\Big]_S^T \\
&\quad - \frac{r-1}{2} \int_S^T \int_{\Gamma_1} E^{\frac{r-3}{2}}(t) E'(t) e^{\phi(x)} \Big(2m \cdot \nabla u + (n - \epsilon_0) u\Big) dx dt.
\end{aligned}$$

The last two terms of this equality can be easily majorized by $cE^{\frac{r+1}{2}}(S)$; then we deduce

$$I_1 \geq -cE^{\frac{r+1}{2}}(S) - R\int_S^T \int_{\Gamma_1} E^{\frac{r-1}{2}}(t)e^{\phi(x)} |u'|^2 \, d\Gamma dt \tag{6.2}$$

$$+(\epsilon_0 - R\|\nabla\phi\|_\infty)\int_S^T \int_\Omega E^{\frac{r-1}{2}}(t)e^{\phi(x)} |u'|^2 \, dxdt.$$

On the other hand, taking the generalized Green formula (see section 5), we infer

$$I_2 := \int_S^T \int_\Omega E^{\frac{r-1}{2}}(t)e^{\phi(x)}(-\Delta u - \nabla\phi\cdot\nabla u)\Big(2m\cdot\nabla u + (n-\epsilon_0)u\Big)dxdt$$

$$= (2-\epsilon_0)\int_S^T \int_\Omega E^{\frac{r-1}{2}}(t)e^{\phi(x)} |\nabla u|^2 \, dxdt - \int_S^T \int_{\Gamma_0} E^{\frac{r-1}{2}}(t)e^{\phi(x)}(m\cdot\nu) |\nabla u|^2 \, d\Gamma dt$$

$$+\int_S^T \int_{\Gamma_1} E^{\frac{r-1}{2}}(t)e^{\phi(x)}\Big((m\cdot\nu) |\nabla u|^2 - (n-\epsilon_0)u\partial_\nu u - 2(m\cdot\nabla u)\partial_\nu u\Big)d\Gamma dt.$$

Using the definition of Γ_0 and Γ_1, we deduce

$$I_2 \geq (2-\epsilon_0)\int_S^T \int_\Omega E^{\frac{r-1}{2}}(t)e^{\phi(x)} |\nabla u|^2 \, dxdt \tag{6.3}$$

$$-\int_S^T \int_{\Gamma_1} E^{\frac{r-1}{2}}(t)e^{\phi(x)}\left((n-\epsilon_0)u\partial_\nu u + \frac{R^2}{\delta}(\partial_\nu u)^2\right)d\Gamma dt.$$

Using (2.3) and the fact that F is nonnegative, we obtain

$$I_3 := \int_S^T \int_\Omega E^{\frac{r-1}{2}}(t)e^{\phi(x)} f(u)\Big(2m\cdot\nabla u + (n-\epsilon_0)u\Big)dxdt$$

$$\geq (n-\epsilon_0)b\int_S^T \int_\Omega 2E^{\frac{r-1}{2}}(t)e^{\phi(x)}F(u)dxdt + \int_S^T \int_\Omega 2E^{\frac{r-1}{2}}(t)e^{\phi(x)}m\cdot\nabla(F(u))dxdt$$

$$\geq \int_S^T \int_\Omega ((n-\epsilon_0)b - n - m\cdot\nabla\phi)2E^{\frac{r-1}{2}}(t)e^{\phi(x)}F(u)dxdt$$

$$+\int_S^T \int_{\Gamma_1} 2E^{\frac{r-1}{2}}(t)e^{\phi(x)}(m\cdot\nu)F(u)d\Gamma dt;$$

then we conclude that

$$I_3 \geq ((n-\epsilon_0)b - n - R\|\nabla\phi\|_\infty)\int_S^T \int_\Omega 2F(u)dxdt. \tag{6.4}$$

Thanks to the assumptions in Theorem 2.4, we have the following.

Case 5. If $R\|\nabla\phi\|_\infty < \min\{2,n\}$ and $b > \frac{n+R\|\nabla\phi\|_\infty}{n-R\|\nabla\phi\|_\infty}$, we can choose $\epsilon_0 \in]R\|\nabla\phi\|_\infty, \min\{2, n-\frac{n+R\|\nabla\phi\|_\infty}{b}\}[$ and then

$$\min\{\epsilon_0 - R\|\nabla\phi\|_\infty, 2-\epsilon_0, (n-\epsilon_0)b - n - R\|\nabla\phi\|_\infty\} > 0.$$

Case 6. If f is linear, $f(s) = \alpha s$ for some positive constant α, then $b = 1$ and we conclude from (6.4) that

$$\begin{aligned} I_3 &\geq (-\epsilon_0 - R\|\nabla\phi\|_\infty)\int_S^T\int_\Omega 2F(u)dxdt \\ &= (\epsilon_0 - R\|\nabla\phi\|_\infty)\int_S^T\int_\Omega 2F(u)dxdt - 2\epsilon_0\int_S^T\int_\Omega 2F(u)dxdt \\ &= (\epsilon_0 - R\|\nabla\phi\|_\infty)\int_S^T\int_\Omega 2F(u)dxdt - 2\epsilon_0\alpha\int_S^T\int_\Omega |u|^2\,dxdt \\ &\geq (\epsilon_0 - R\|\nabla\phi\|_\infty)\int_S^T\int_\Omega 2F(u)dxdt - 2\epsilon_0\alpha\hat{c}\int_S^T\int_\Omega |\nabla u|^2\,dxdt, \end{aligned}$$

where $\hat{c}$ is the positive constant defined by (5.8). Then, assuming that $R\|\nabla\phi\|_\infty < \frac{2}{1+2\alpha\hat{c}}$ and taking $\epsilon_0 \in]R\|\nabla\phi\|_\infty, \frac{2}{1+2\alpha\hat{c}}[$, the quantity $\min\{\epsilon_0 - R\|\nabla\phi\|_\infty, 2-(1+2\alpha\hat{c})\epsilon_0\}$ is positive.

Combining (6.2)–(6.4), taking the fact that $I_1 + I_2 + I_3 = 0$ into account, and using the boundary condition on Γ_1, we obtain in both previous cases

$$\begin{aligned} (6.5)\quad \int_S^T\int_\Omega E^{\frac{r+1}{2}}(t)dt &\leq cE^{\frac{r+1}{2}}(S) \\ &\quad + c\int_S^T\int_{\Gamma_1} E^{\frac{r-1}{2}}(t)e^{\phi(x)}\Big(|u'|^2 + g^2(u') + |ug(u')|\Big)d\Gamma dt. \end{aligned}$$

We now estimate the last term of (6.5). We exploit the Cauchy–Schwarz inequality and the Sobolev imbedding $V \subset L^2(\Gamma_1)$ to get

$$\int_{\Gamma_1}|ug(u')|d\Gamma \leq \epsilon\int_{\Gamma_1}|u|^2d\Gamma + c_\epsilon\int_{\Gamma_1}g^2(u')d\Gamma \leq \epsilon E(t) + c_\epsilon\int_{\Gamma_1}g^2(u')d\Gamma.$$

Substituting this inequality into the right-hand side of (6.5) and choosing $\epsilon > 0$ small enough, we obtain that

$$\begin{aligned} (6.6)\qquad \int_S^T\int_\Omega E^{\frac{r+1}{2}}(t)dt &\leq cE^{\frac{r+1}{2}}(S) \\ &\quad + c\int_S^T\int_{\Gamma_1} E^{\frac{r-1}{2}}(t)e^{\phi(x)}\Big(|u'|^2 + g^2(u')\Big)d\Gamma dt. \end{aligned}$$

We follow now the proof given in section 4. We note

$$\Gamma^+ = \{x \in \Gamma_1 : \big|u'\big| > 1\} \quad \text{and} \quad \Gamma^- = \Gamma_1 \setminus \Gamma^+.$$

By (2.12) and (6.1) we have

$$\int_S^T \int_{\Gamma^+} E^{\frac{r-1}{2}}(t) e^{\phi(x)} \Big(\big|u'\big|^2 + g^2(u') \Big) dxdt \le cE^{\frac{r-1}{2}}(S) \int_S^T \int_{\Gamma^+} e^{\phi(x)} u' g(u') dxdt$$
$$\le c \left(E^{\frac{r+1}{2}}(S) - E^{\frac{r+1}{2}}(T) \right).$$

In the same way (using Young's inequality), we get

$$\int_S^T \int_{\Gamma^-} E^{\frac{r-1}{2}}(t) e^{\phi(x)} \Big(\big|u'\big|^2 + g^2(u') \Big) dxdt \le c \int_S^T \int_{\Gamma^-} E^{\frac{r-1}{2}}(t) \Big(e^{\phi(x)} u' g(u') \Big)^{\frac{2}{r+1}} dxdt$$
$$\le \epsilon \int_S^T E^{\frac{r+1}{2}}(t) dt + c_\epsilon \int_S^T \int_{\Gamma^-} e^{\phi(x)} u' g(u') dxdt$$
$$\le \epsilon \int_S^T E^{\frac{r+1}{2}}(t) dt + c_\epsilon \left(E(S) - E(T) \right).$$

Substituting the sum of these two estimates into the right-hand side of (6.6), choosing ϵ small enough, and letting T go to ∞, we obtain (4.1). This finishes the proof of Theorem 2.4.

Remark. Using the method developed above, the same results can be easily obtained if we replace the first equation in (P) by

$$u'' - \Delta u + q_1(x) h(\nabla u) + q_2(x) f(u) + q_3(x) g(u') = 0 \quad \text{in} \quad \Omega \times \mathbb{R}^+,$$

and the first equation and the boundary condition on Γ_1 in (P′) by

$$\begin{cases} u'' - \Delta u + q_1(x) h(\nabla u) + q_2(x) f(u) = 0 & \text{in} \quad \Omega \times \mathbb{R}^+, \\ \partial_\nu u + q_4(x) u + q_3(x) g(u') = 0 & \text{on} \quad \Gamma_1 \times \mathbb{R}^+, \end{cases}$$

where $q_i : \Omega \to \mathbb{R}$ are bounded functions such that $q_2(x) \ge 0$, $q_4(x) \ge 0$, $q_3(x) \ge a_0 > 0$. If $q_4(x) \ge b_0 > 0$, we may take $\Gamma_0 = \emptyset$.

We define the equivalent energy of (P) and (P′), respectively, by

$$E(t) = \int_\Omega e^{\varphi(x)} \left(\big|u'\big|^2 + \big|\nabla u\big|^2 + 2q_2(x) F(u) \right) dx, \tag{6.7}$$

$$E(t) = \int_\Omega e^{\varphi(x)} \left(\big|u'\big|^2 + \big|\nabla u\big|^2 + 2q_2(x) F(u) \right) dx + \int_{\Gamma_1} e^{\varphi(x)} \big|u\big|^2 d\Gamma \tag{6.8}$$

if $h(\nabla u) = -\nabla\phi \cdot \nabla u$ with $\phi \in W^{1,\infty}(\Omega)$, where $\varphi \in W^{1,\infty}(\Omega)$ satisfying $\nabla\varphi = q_1(x)\nabla\phi$.

In the general case, we assume that $\beta \|q_1\|_\infty$ is small enough as in Theorem 2.1 and Theorem 2.3, where β, c_1, and c_2 are replaced by $\beta\|q_1\|_\infty$, $a_0 c_1$, and $c_2 \|q_3\|_\infty$, respectively, and we define the energy of (P) and (P′), respectively, by (6.7) and (6.8) with $\varphi \equiv 0$. In order to get ride of the lower-order term, which is $\int_{\Gamma_1} \big|u\big|^2 d\Gamma$, we use the solution of an auxiliary elliptic problem as an additional multiplier (see [4, Lemma 4.2]).

7. Some applications of our method. In [6], we considered the following Petrovsky system:

$$\begin{cases} u'' + \Delta^2 u + q(x)u + g(u') = 0 & \text{in} \quad \Omega \times \mathbb{R}^+, \\ u = \partial_\nu u = 0 & \text{on} \quad \Gamma \times \mathbb{R}^+, \\ u(x,0) = u_0(x) \quad \text{and} \quad u'(x,0) = u_1(x) & \text{in} \quad \Omega, \end{cases} \tag{7.1}$$

where Ω is a bounded domain in $\mathbb{R}^n$ $(n \geq 1)$ with a smooth boundary Γ and ν is the outward unit normal vector to Γ. For g continuous, increasing, satisfying $g(0) = 0$, and $q : \Omega \to \mathbb{R}^+$ a bounded function, we proved a global existence and a regularity result. We also established, under suitable growth conditions on g, decay results for weak, as well as strong, solutions. Precisely, we showed that the solution decays exponentially if g behaves like a linear function, whereas the decay is of a polynomial order otherwise. Similar results to the above system, coupled with a semilinear wave equation, have been established by Guesmia in [5]. In [17], Messaoudi studied the problem

$$\begin{cases} u'' + \Delta^2 u + au'\big|u'\big|^{m-2} - bu\big|u\big|^{p-2} = 0 & \text{in} \quad \Omega \times \mathbb{R}^+, \\ u = \partial_\nu u = 0 & \text{on} \quad \Gamma \times \mathbb{R}^+, \\ u(x,0) = u_0(x) \quad \text{and} \quad u'(x,0) = u_1(x) & \text{in} \quad \Omega, \end{cases}$$

where a, $b > 0$ and p, $m > 2$. This is a similar problem to (7.1), which contains a nonlinear source term competing with the damping factor. He established an existence result and showed that the solution continues to exist globally if $m \geq p$; however, it blows up in finite time if $m < p$. In this paper no result of stability was announced.

In [7], we obtained some stabilization results of the following elasticity system:

$$\begin{cases} u_i'' - \sigma_{ij,j} + g_i(u_i') = 0 & \text{in} \quad \Omega \times \mathbb{R}^+, \\ u_i = 0 & \text{on} \quad \Gamma \times \mathbb{R}^+, \\ u_i(x,0) = u_i^0(x) \quad \text{and} \quad u_i'(x,0) = u_i^1(x) & \text{in} \quad \Omega, \\ i = 1, \dots, n, \end{cases} \tag{7.2}$$

where the unknown $u = (u_1, \dots, u_n) : \Omega \to \mathbb{R}^n$. Here, $\sigma_{ij,j} = \sum_{j=1}^{j=n} \frac{\partial \sigma_{ij}}{\partial x_j}$, $\sigma_{ij} = \sum_{k,l=1}^{k,l=n} a_{ijkl}\varepsilon_{ij}$, $\varepsilon_{ij} = \frac{1}{2}(u_{i,j} + u_{j,i})$, $u_{i,j} = \frac{\partial u_i}{\partial x_j}$, $u_{j,i} = \frac{\partial u_j}{\partial x_i}$, and $a_{ijkl} \in W^{1,\infty}(\Omega)$. We proved some decay estimates which are crucially dependent on the behavior of the damping g_i at the origin and infinity. In [8], we extended these results to the case of localized dissipations; that is, the damping is effective only in a neighborhood of a suitable subset of the boundary.

In [4], we considered the problem of exact controllability and boundary stabilization of elasticity systems with coefficients a_{ijkl} depending also on time t. The stabilization results obtained in [4] were generalized in [3] to the nonlinear feedback case. The results obtained in [3] and [4] improve and generalize some ones obtained earlier by Alabau and Komornik [1] in the case where g_i is linear and $a_{ijkl} = \text{const}$.

The decrease of energy plays a crucial role in studying the asymptotic stability of the systems cited above. The situation of nondissipative systems (that is, the energy is not decreasing) was not previously considered.

Using the method developed in previous sections, we can extend Theorems 2.1–2.4 to the following more general nondissipative problems.

7.1. Petrovsky system.

$$\begin{cases} u'' + \Delta^2 u + q_1(x)h(\Delta u) + q_2(x)f(u) + q_3(x)g(u') = 0 & \text{in} \quad \Omega \times \mathbb{R}^+, \\ u = \partial_\nu u = 0 & \text{on} \quad \Gamma \times \mathbb{R}^+, \\ u(x,0) = u_0(x) \quad \text{and} \quad u'(x,0) = u_1(x) & \text{in} \quad \Omega, \end{cases}$$

where h, f, $g : \mathbb{R} \to \mathbb{R}$ are three given functions satisfying Assumptions 2.1–2.5 and q_i are three given functions defined as in the remark above. Here $c_0 > 0$ is the smallest imbedding positive constant (depending only on Ω) satisfying

$$\int_\Omega \big|v\big|^2 dx \le c_0 \int_\Omega \big|\Delta v\big|^2 dx \qquad \forall v \in H_0^2(\Omega).$$

The energy and the equivalent energy are, respectively, defined by

$$E(t) = \int_\Omega \Big(\big|u'\big|^2 + \big|\Delta u\big|^2 + 2q_2(x)F(u) \Big) dx, \qquad t \in \mathbb{R}^+,$$

in the general case, and

$$E(t) = \int_\Omega e^{\varphi(x)} \Big(\big|u'\big|^2 + \big|\Delta u\big|^2 + 2q_2(x)F(u) \Big) dx, \qquad t \in \mathbb{R}^+$$

if $h(\Delta u) = -\phi(x)\Delta u$, with $\phi \in L^\infty(\Omega)$, where $\varphi \in W^{2,\infty}(\Omega)$ satisfying $\Delta\varphi = q_1(x)\phi(x)$.

7.2. Coupled system. We consider the nonlinear coupled wave equation and Petrovsky system:

$$\begin{cases} u_1'' + \Delta^2 u_1 + q_1(x)h_1(\Delta u_1) + q_2(x)f_1(u_1) \\ \qquad\qquad + q_3(x)g_1(u_1') + a_1(x)u_2 = 0 & \text{in} \quad \Omega \times \mathbb{R}^+, \\ u_2'' - \Delta u_1 + l_1(x)h_2(\nabla u_2) + l_2(x)f_2(u_2) \\ \qquad\qquad + l_3(x)g_2(u_2') + a_2(x)u_1 = 0 & \text{in} \quad \Omega \times \mathbb{R}^+, \\ u_2 = u_1 = \partial_\nu u_1 = 0 & \text{on} \quad \Gamma \times \mathbb{R}^+, \\ u_i(x,0) = u_i^0(x) \quad \text{and} \quad u_i'(x,0) = u_i^1(x), \ i = 1,2 & \text{in} \quad \Omega, \end{cases}$$

where a_1, a_2 are two bounded functions with norms small enough (see [5]) and the l_i, h_i, f_i, and g_i are given functions defined as q_i, h, f, and g, respectively.

If $h_1(\Delta u_1) = -\phi_1(x)\Delta u_1$ and $h_2(\nabla u_2) = -\nabla\phi_2 \cdot \nabla u_2$ with $\phi_1 \in L^\infty(\Omega)$ and $\phi_2 \in W^{1,\infty}(\Omega)$, then we assume that $a_1(x)e^{\varphi_1(x)} = a_2(x)e^{\varphi_2(x)}$, where $\varphi_1 \in W^{2,\infty}(\Omega)$ and $\varphi_2 \in W^{1,\infty}(\Omega)$ satisfying $\Delta\varphi_1 = q_1(x)\phi_1(x)$ and $\nabla\varphi_2 = l_1(x)\nabla\phi_2$; we define the equivalent energy by

$$E(t) = \int_\Omega e^{\varphi_1(x)} \Big(\big|u_1'\big|^2 + \big|\Delta u_1\big|^2 + 2q_2(x)F_1(u_1) \Big) dx \tag{7.3}$$

$$+ \int_\Omega e^{\varphi_2(x)} \Big(\big|u_2'\big|^2 + \big|\nabla u_2\big|^2 + 2l_2(x)F_2(u_2) \Big) dx + 2\int_\Omega e^{\varphi_1(x)} a_1(x)u_1u_2 dx,$$

which is nonincreasing,

$$E'(t) = -2\int_\Omega \Big(e^{\varphi_1(x)} q_3(x)u_1'g_1(u_1') + e^{\varphi_2(x)} l_3(x)u_2'g_2(u_2') \Big) \le 0.$$

In the general case, we assume that $a_1(x) = a_2(x)$ and we define the energy by (7.3) with $\varphi_1 \equiv \varphi_2 \equiv 0$.

7.3. Elasticity systems. We are interested in the precise decay property of the solution for elasticity systems:

$$
\begin{cases}
u_i'' - \sigma_{ij,j} + q_{1,i}(x)h_i(\sigma_{i1},\dots,\sigma_{in}) \\
\qquad\qquad + q_{2,i}(x)f_i(u_i) + q_{3,i}(x)g_i(u_i') = 0 & \text{in} \quad \Omega\times\mathbb{R}^+, \\
u_i = 0 & \text{on} \quad \Gamma\times\mathbb{R}^+, \\
u_i(x,0) = u_i^0(x) \quad \text{and} \quad u_i'(x,0) = u_i^1(x) & \text{in} \quad \Omega, \\
i = 1,\dots,n,
\end{cases} \tag{7.4}
$$

with the same notations as before. Here for $i = 1,\dots,n$, h_i, f_i, and g_i satisfy the same hypothesis as h, f, and g in section 2, respectively, and $q_{1,i}$, $q_{2,i}$, and $q_{3,i}$ are defined as q_1, q_2, and q_3 in section 7.1, respectively.

We define the equivalent energy of (7.4) by the formula

$$
E(t) = \int_\Omega \sum_{i=1}^{i=n} e^{\varphi_i(x)} \left(\left|u_i'\right|^2 + \sum_{j=1}^{j=n} \sigma_{ij}\varepsilon_{ij} + 2q_{2,i}(x)F_i(u_i) \right) dx,
$$

where $\varphi_i \equiv 0$ if h_i is nonlinear, and if h_i is linear, $h_i(\zeta) = -\nabla\phi_i\cdot\zeta$ for all $\zeta\in\mathbb{R}^n$ with $\phi_i \in W^{1,\infty}(\Omega)$, then we take $\varphi_i \in W^{1,\infty}(\Omega)$ such that $\nabla\varphi_i = q_{1,i}(x)\nabla\phi_i$. In the case where all the functions h_i are linear, our system is dissipative:

$$
E'(t) = -2\int_\Omega \sum_{i=1}^{i=n} e^{\varphi_i(x)} q_{3,i}(x)u_i' g_i(u_i')dx \le 0.
$$

We obtain the results of Theorem 2.1 and Theorem 2.2.

Under some geometric condition as in [3], the results of Theorem 2.3 and Theorem 2.4 can be easily proved in the case of boundary feedback; that is, we consider the homogenous Dirichlet condition on Γ_0, and we consider the following one on Γ_1 (see [3]):

$$
\sum_{j=1}^{j=n} \sigma_{ij}\nu_j + q_{4,i}(x)u_i + q_{3,i}(x)g_i(u_i') = 0.
$$

Remark. The method developed in this paper is direct and very flexible; it can be applied to various nondissipative problems (elasticity, thermoelasticity, Kirchoff, von Karman, coupled systems, ...) with an internal or a boundary feedback, and it can generalize the decay estimates (known in the dissipative case) to the nondissipative one.

Open questions. The main restrictive assumptions under which the stability results are valid are the smallness conditions on β (defined by (2.6)) assumed in Theorems 2.1, 2.3, and 2.4. In the case of nonlinear function h, these assumptions are required to obtain the inequalities $(*)$ (given in the introduction). In Theorem 2.4 (stability of (P') with $h(\nabla u) = -\nabla\phi\cdot\nabla u$), the smallness assumption on β is required to absorb some terms caused by the use of the second multiplier $m\cdot\nabla u$. It would be interesting to know if the stability estimates still hold true under weaker assumption on β, using more sophisticated tools, for example, general multipliers. And if it is not the case, it would be interesting to know if other weaker stability estimates can be obtained.

Another important aspect of the case of nonlinear function h is assumption (2.5) imposed on the damping g. It would be interesting to prove the same polynomial

stability (obtained in the case of linear function h) under the weaker assumption (2.12). With this perspective, it would be interesting to look at what we can conclude at ∞ on a positive function satisfying the following inequalities more general than $(*)$:

$$\begin{cases} \int_S^T E^{a_0}(t)dt \le a_1(E(S)+E(T)) + a_2(E(S)-E(T)) & \forall 0 \le S \le T < \infty, \\ E'(t) \le a_3 E(t) & \forall t \ge 0, \end{cases}$$

where a_i, $i = 0, 1, 2, 3$, are nonnegative constants.

It would also be very interesting (particularly from the point of view of applications) to explore a more general class of hyperbolic equations based on the equation

$$K(x,t)u'' - Au + F(x,t,u,u',\nabla u) = 0,$$

where K and F are given functions and $Au = \sum_{i,j=1}^{n} \partial_{x_i}(a_{ij}(x,t)\partial_{x_j}u)$ is a second-order elliptic differential operator with smooth coefficients a_{ij}.

Acknowledgments. The author is very grateful to the referees and the associate editor for their valuable comments and suggestions.

REFERENCES

[1] F. Alabau and V. Komornik, *Boundary observability, controllability, and stabilization of linear elastodynamic systems*, SIAM J. Control Optim., 37 (1998), pp. 521–542.

[2] M. M. Cavalcanti, N. A. Lar'kin, and J. A. Soriano, *On solvability and stability of solutions of nonlinear degenerate hyperbolic equations with boundary damping*, Funkcial. Ekvac., 41 (1998), pp. 271–289.

[3] A. Guesmia, *Existence globale et stabilisation frontière non linéaire d'un système d'élasticité*, Portugal. Math., 56 (1999), pp. 361–379.

[4] A. Guesmia, *On linear elasticity systems with variable coefficients*, Kyushu J. Math., 52 (1998), pp. 227–248.

[5] A. Guesmia, *Energy decay for a damped nonlinear coupled system*, J. Math. Anal. Appl., 239 (1999), pp. 38–48.

[6] A. Guesmia, *Existence globale et stabilisation interne non linéaire d'un système de Petrovsk*, Bull. Belg. Math. Soc. Simon Stevin, 5 (1998), pp. 583–594.

[7] A. Guesmia, *Existence globale et stabilisation interne non linéaire d'un système d'élasticité*, Portugal. Math., 55 (1998), pp. 333–347.

[8] A. Guesmia, *On the decay estimates for elasticity systems with some localized dissipations*, Asymptot. Anal., 22 (2000), pp. 1–13.

[9] A. Guesmia, *Une nouvelle approche pour la stabilisation des systèmes distribués non dissipatifs*, C. R. Acad. Sci. Paris Sér. I. Math., 332 (2001), pp. 633–636.

[10] A. Haraux and E. Zuazua, *Decay estimates for some semilinear damped hyperbolic problems*, Arch. Ration. Mech. Anal., 100 (1988), pp. 191–206.

[11] S. Kawashima, M. Nakao, and K. Ono, *On decay property of solutions to the Cauchy problem of the semilinear wave equation with a dissipative term*, J. Math. Soc. Japan, 47 (1995), pp. 617–653.

[12] V. Komornik, *Exact Controllability and Stabilization. The Multiplier Method*, Masson, Paris, John Wiley, Chichester, UK, 1994.

[13] J. E. Lagnese, *Note on boundary stabilization of wave equations*, SIAM J. Control Optim., 26 (1988), pp. 1250–1256.

[14] I. Lasiecka and D. Tataru, *Uniform boundary stabilization of semilinear wave equations with nonlinear boundary damping*, Differential Integral Equations, 6 (1993), pp. 507–533.

[15] S. Messaoudi, *Energy decay of solutions of a semilinear wave equation*, Int. J. Appl. Math., 9 (2000), pp. 1037–1048.

[16] S. A. Messaoudi, *Decay of the solution energy for a nonlinearly damped wave equation*, Arabian. J. Sci. Eng. Sect. A Sci., 26 (2001), pp. 63–68.

[17] S. Messaoudi, *Global existence and nonexistence in a system of Petrovsky*, J. Math. Anal. Appl., 265 (2002), pp. 296–308.

[18] M. Nakao, *Convergence of solutions of the wave equation with a nonlinear dissipative term to the steady state*, Mem. Fac. Sci. Kyushu Univ., 30 (1976), pp. 257–265.

[19] M. Nakao and T. Narazaki, *Existence and decay of solutions of some nonlinear wave equations in noncylindrical domains*, Math. Rep., 11 (1978), pp. 117–125.
[20] M. Nakao and K. Ono, *Global existence to the cauchy problem of the semilinear wave equation with a nonlinear dissipation*, Funkcialaj Ekvacioj, 38 (1995), pp. 417–431.
[21] M. Nakao, *Remarks on the existence and uniqueness of global decaying solutions of the nonlinear dissipative wave equations*, Math Z., 206 (1991), pp. 265–275.
[22] M. Nakao, *Decay of solutions of some nonlinear evolution equations*, J. Math. Anal. Appl., 60 (1977), pp. 542–549.
[23] P. Pucci and J. Serrin, *Asymptotic stability for nonautonomous dissipative wave systems*, Comm. Pure Appl. Math., 49 (1996), pp. 177–216.
[24] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.
[25] R. Triggiani, *Wave equation on a bounded domain with boundary dissipation: An operator approach*, J. Math. Anal. Appl., 137 (1989), pp. 438–461.
[26] Y. You, *Energy decay and exact controllability for the Petrovsky equation in a bounded domain*, Adv. in Appl. Math., 11 (1990), pp. 372–388.
[27] E. Zuazua, *Stability and decay for a class of nonlinear hyperbolic problems*, Asymptot. Anal., 1 (1988), pp. 161–185.
[28] E. Zuazua, *Uniform stabilization of the wave equation by nonlinear boundary feedback*, SIAM J. Control Optim., 28 (1990), pp. 466–477.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 53–75

GENERAL LINEAR QUADRATIC OPTIMAL STOCHASTIC CONTROL PROBLEMS WITH RANDOM COEFFICIENTS: LINEAR STOCHASTIC HAMILTON SYSTEMS AND BACKWARD STOCHASTIC RICCATI EQUATIONS*

SHANJIAN TANG†

Abstract. Consider the minimization of the following quadratic cost functional:

$$J(u) := E\langle Mx_T, x_T\rangle + E\int_0^T (\langle Q_s x_s, x_s\rangle + \langle N_s u_s, u_s\rangle)\, ds,$$

where x is the solution of the following linear stochastic control system:

$$\begin{aligned} dx_t =& (A_t x_t + B_t u_t)\, dt + \sum_{i=1}^{d} (C_t^i x_t + D_t^i u_t)\, dW_t^i, \\ x_0 =& h \in \mathbb{R}^n, \qquad u_t \in \mathbb{R}^m; \end{aligned}$$

u is a square integrable adapted process. The problem is conventionally called the stochastic LQ (the abbreviation of "linear quadratic") problem. We are concerned with the following general case: the coefficients A, B, C^i, D^i, Q, N, and M are allowed to be adapted processes or random matrices. We prove the existence and uniqueness result for the associated Riccati equation, which in our general case is a backward stochastic differential equation with the generator (the drift term) being *highly nonlinear* in the two unknown variables. This solves Bismut and Peng's long-standing open problem (for the case of a Brownian filtration), which was initially proposed by the French mathematician J. M. Bismut [in *Séminaire de Probabilités* XII, Lecture Notes in Math. 649, C. Dellacherie, P. A. Meyer, and M. Weil, eds., Springer-Verlag, Berlin, 1978, pp. 180–264]. We also provide a rigorous derivation of the Riccati equation from the stochastic Hamilton system. This completes the interrelationship between the Riccati equation and the stochastic Hamilton system as two different but equivalent tools for the stochastic LQ problem.

There are two key points in our arguments. The first one is to connect the *existence* of the solution of the Riccati equation to the *homomorphism* of the stochastic flows derived from the optimally controlled system. Actually, we establish their equivalence. As a consequence, we can construct solutions to a sequence of suitably modified Riccati equations in terms of the associated stochastic Hamilton systems (and the optimal controls). The second key point is to establish a new type of a priori estimate for solutions of Riccati equations, with which we show that the sequence of constructed solutions has a limit which is a solution to the original Riccati equation.

Key words. linear quadratic optimal stochastic control, random coefficients, Riccati equation, backward stochastic differential equations, stochastic Hamilton flows, homomorphism of stochastic flows, optimality conditions

AMS subject classifications. 93E20, 49K45, 49N10, 60H10

PII. S0363012901387550

1. Formulation of the problem and basic assumptions.

Consider the following so-called linear quadratic (LQ in short form) optimal stochastic control problem: minimize over $u \in \mathcal{L}^2_{\mathcal{F}}(0, T; \mathbb{R}^m)$ the following quadratic cost functional:

$$J(u; 0, h) := E\langle Mx_T, x_T\rangle + E\int_0^T (\langle Q_s x_s, x_s\rangle + \langle N_s u_s, u_s\rangle)\, ds, \tag{1.1}$$

*Received by the editors April 9, 2001; accepted for publication (in revised form) September 23, 2002; published electronically March 19, 2003. This research was supported by a Research Fellowship from the Alexander von Humboldt Foundation and by the National Natural Science Foundation of China under project 79790130.

http://www.siam.org/journals/sicon/42-1/38755.html

†Department of Mathematics, Fudan University, Shanghai 200433, China (sjtangk@online.sh.cn).

where x is the solution of the following linear stochastic control system:

$$\begin{cases} dx_t = (A_t x_t + B_t u_t)\,dt + \sum_{i=1}^{d} (C_t^i x_t + D_t^i u_t)\,dW_t^i, \\ x_0 = h \in \mathbb{R}^n. \end{cases} \tag{1.2}$$

Here, $\{W_t := (W_t^1, \dots, W_t^d)', 0 \le t \le T\}$ is a d-dimensional standard Brownian motion defined on some probability space $(\Omega, \mathcal{F}, P)$. Denote by $\{\mathcal{F}_t, 0 \le t \le T\}$ the augmented natural filtration of the standard Brownian motion W. The control u belongs to the Banach space $\mathcal{L}^2_{\mathcal{F}}(0, T; \mathbb{R}^m)$, which consists of all $\mathbb{R}^m$-valued square integrable $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted processes.

Throughout this paper, we make the following three assumptions on the coefficients of the above problem.

(A1) Assume that the matrix processes $A : [0, T] \times \Omega \to \mathbb{R}^{n\times n}$, $B : [0, T] \times \Omega \to \mathbb{R}^{n\times m}$; $C^i : [0, T] \times \Omega \to \mathbb{R}^{n\times n}$, $D^i : [0, T] \times \Omega \to \mathbb{R}^{n\times m}$, $i = 1, \dots, d$; $Q : [0, T] \times \Omega \to \mathbb{R}^{n\times n}$, $N : [0, T] \times \Omega \to \mathbb{R}^{m\times m}$ and the random matrix $M : \Omega \to \mathbb{R}^{n\times n}$ are uniformly bounded and $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted or $\mathcal{F}_T$-measurable.

(A2) Assume that the state weighting matrix process Q and the control weighting matrix process N are almost surely and almost everywhere (abbreviated hereafter as a.s.a.e.) symmetric and nonnegative. Also assume that the terminal state weighting random matrix M is almost surely (hereafter abbreviated as a.s.) symmetric and nonnegative.

(A3) Assume that the control weighting matrix process N is uniformly positive.

In this paper, we shall demonstrate how to use (the optimal control processes of the stochastic LQ problem and) the solutions of the stochastic Hamilton system to construct a solution of the Riccati equation (see (3.1) in section 3), which in general is a highly nonlinear backward stochastic differential equation (BSDE in short form). We connect the existence of a solution of the Riccati equation to the homomorphism of the stochastic flows derived from the optimally controlled system and identify their equivalence. In this way, on one hand, we complete the interrelationship—partially existing in the literature—between the stochastic Hamilton system and the Riccati equation (see section 3). On the other hand, we solve the long-standing open problem which was initially proposed in 1978 by J. M. Bismut [4] (see section 4).

The rest of our paper is organized as follows. Section 2 describes the stochastic Hamilton system theory associated with the above LQ problem—most of which has been known in the literature. Section 3 recalls known connections of the Riccati equation to the stochastic LQ problem and to the associated stochastic Hamilton system. Section 4 reviews some previous results concerning the Riccati equation, which are known to the author. Section 5 sketches the main ideas and the main results of this paper. The next three sections (6–8) are devoted to the detailed proofs of the main results. Finally, in section 9, we give some concluding comments.

2. The stochastic Hamilton system. Let τ be a $\{\mathcal{F}_t, 0 \le t \le T\}$-stopping time such that $0 \le \tau \le T$. Consider the initial-data-parameterized stochastic LQ problem: minimize over $u \in \mathcal{L}^2_{\mathcal{F}}(0, T; \mathbb{R}^m)$ the quadratic cost functional

$$J(u; \tau, h) := E\langle M x_T^{\tau,h;u}, x_T^{\tau,h;u}\rangle + E\int_\tau^T (\langle Q_s x_s^{\tau,h;u}, x_s^{\tau,h;u}\rangle + \langle N_s u_s, u_s\rangle)\,ds, \tag{2.1}$$

where $x^{\tau,h;u}$ is the solution of the following linear stochastic control system:

$$\begin{cases} dx_t = (A_t x_t + B_t u_t)\,dt + \sum_{i=1}^{d} (C_t^i x_t + D_t^i u_t)\,dW_t^i, & \tau \le t \le T, \\ x_\tau = h \in L^2(\Omega, \mathcal{F}_\tau, P; \mathbb{R}^n). \end{cases} \tag{2.2}$$

Assumptions (A1), (A2), and (A3) imply that the above stochastic LQ problem has a unique optimal control. See Bismut [4] for the proof of this result. A further step is to characterize the optimal control.

THEOREM 2.1. *Let assumptions* (A1), (A2), *and* (A3) *be satisfied. Then, a necessary and sufficient condition for* u *to be an optimal control of the parameterized stochastic LQ problem is*

$$N_t u_t + B_t' y_t + \sum_{i=1}^{d} (D_t^i)' z_t^i = 0, \quad \tau \le t \le T, \quad a.s. \tag{2.3}$$

Here (y, z) (*called the adjoint processes with* $z := (z^1, \dots, z^d)$) *is the* (*unique*) *solution* (*see Pardoux and Peng* [16]) *of the BSDE* (*conventionally called the adjoint equation*):

$$\begin{cases} -dy_t = \left[A_t' y_t + \sum_{i=1}^{d} (C_t^i)' z_t^i + Q_t x_t^{\tau,h;u} \right] dt - \sum_{i=1}^{d} z_t^i\,dW_t^i, & \tau \le t \le T, \\ y_T = M x_T^{\tau,h;u}. \end{cases} \tag{2.4}$$

The proof is simple. In fact, the necessary part results from some simple variational calculus and some dual representation considerations. This part is conventionally called the stochastic maximum principle (see Bensoussan [2], Peng [20], and Tang and Li [22], for example). The sufficient part stems from the convexity of the cost functional $J(\cdot; \tau, h)$. All the details of the proof of Theorem 2.1 can be given similar to the work in Bismut [3, 4].

From (2.3), we get the optimal control

$$u_t = -N_t^{-1} \left[B_t' y_t + \sum_{i=1}^{d} (D_t^i)' z_t^i \right], \qquad \tau \le t \le T. \tag{2.5}$$

The so-called *stochastic Hamilton system* is given by

$$\begin{cases} dx_t = (A_t x_t + B_t u_t)\,dt + \sum_{i=1}^{d} (C_t^i x_t + D_t^i u_t)\,dW_t^i, & \tau \le t \le T, \\ u_t := -N_t^{-1} \left[B_t' y_t + \sum_{i=1}^{d} (D_t^i)' z_t^i \right], & \tau \le t \le T, \\ -dy_t = \left[A_t' y_t + \sum_{i=1}^{d} (C_t^i)' z_t^i + Q_t x_t \right] dt - \sum_{i=1}^{d} z_t^i\,dW_t^i, & \tau \le t \le T, \\ x_\tau = h \in L^2(\Omega, \mathcal{F}_\tau, P; \mathbb{R}^n), \quad y_T = M x_T, \quad z_t := (z_t^1, \dots, z_t^d). \end{cases} \tag{2.6}$$

It is a system of forward-backward stochastic differential equations (FBSDEs in short form). The solution consists of a triple (x, y, z).

THEOREM 2.2. *Let assumptions* (A1), (A2), *and* (A3) *be satisfied. Then, for each fixed pair* (τ, h) *with* $\tau \in [0,T]$ *a.s. and* $h \in L^2(\Omega, \mathcal{F}_s, P; \mathbb{R}^n)$, *the stochastic Hamilton system* (2.6) *has a unique adapted solution, which is a triple of stochastic processes parameterized by the initial data* (τ, h), *denoted by*

$$\{(\phi_{\tau,t}(h), \psi_{\tau,t}(h), \mu_{\tau,t}(h)); \tau \le t \le T, h \in L^2(\Omega, \mathcal{F}_\tau, P; \mathbb{R}^n)\}.$$

Moreover, we have for some deterministic positive constant β,

$$E \max_{\tau \le t \le T} |\phi_{\tau,t}(h)|^2 + E \max_{\tau \le t \le T} |\psi_{\tau,t}(h)|^2 + E \int_\tau^T |\mu_{\tau,t}(h)|^2 \, dt \le \beta E|h|^2. \tag{2.7}$$

LEMMA 2.1. *Let assumptions* (A1), (A2), *and* (A3) *be satisfied. If*

$$\{(\phi_{\tau,t}(h), \psi_{\tau,t}(h), \mu_{\tau,t}(h)); \tau \le t \le T, h \in L^2(\Omega, \mathcal{F}_\tau, P; \mathbb{R}^n)\}$$

is a solution to (2.6), *then for any* $\{\mathcal{F}_t, 0 \le t \le T\}$*-stopping time* $\gamma \in [\tau, T]$,

$$\begin{aligned} &\langle \psi_{\tau,\gamma}(h), \phi_{\tau,\gamma}(h) \rangle \\ &= E^{\mathcal{F}_\gamma} \left\{ \int_\gamma^T (\langle Q_r \phi_{\tau,r}(h), \phi_{\tau,r}(h) \rangle + \langle N_r u_r, u_r \rangle) \, dr + \langle M \phi_{\tau,T}(h), \phi_{\tau,T}(h) \rangle \right\}, \end{aligned} \tag{2.8}$$

where u *is the optimal control given by* (2.5).

Proof of Lemma 2.1. Using Itô's formula to compute the term $\langle \psi_{\tau,r}(h), \phi_{\tau,r}(h) \rangle$ and then taking the conditional expectation, we get the desired result. □

Proof of Theorem 2.2. Assumptions (A1), (A2), and (A3) imply the existence of an optimal control u. From Theorem 2.1, u should satisfy (2.5), and then (x, y, z) is a solution. The existence part is proved. The uniqueness assertion is obvious once (2.7) is true. Therefore, it remains to prove that (2.7) holds.

From Lemma 2.1, we have

$$\begin{aligned} &E \int_\tau^T \langle Q_r \phi_{\tau,r}(h), \phi_{\tau,r}(h) \rangle \, dr + E \langle M \phi_{\tau,T}(h), \phi_{\tau,T}(h) \rangle \le E \left(|\psi_{\tau,\tau}(h)| \cdot |h| \right), \\ &E \int_\tau^T \langle N_r u_r, u_r \rangle \, dr \le E \left(|\psi_{\tau,\tau}(h)| \cdot |h| \right). \end{aligned} \tag{2.9}$$

While we have (2.6), it follows from a classical a priori estimate for BSDEs (see El Karoui, Peng, and Quenez [6]) that

$$\begin{aligned} &E \max_{\tau \le t \le T} |\psi_{\tau,t}(h)|^2 + E \int_\tau^T |\mu_{\tau,t}(h)|^2 \, dt \\ &\le \beta E \int_\tau^T |Q_r \phi_{\tau,r}(h)|^2 \, dr + \beta E |M \phi_{\tau,T}(h)|^2 \\ &\le \beta E \int_\tau^T \langle Q_r \phi_{\tau,r}(h), \phi_{\tau,r}(h) \rangle \, dr + \beta E \langle M \phi_{\tau,T}(h), \phi_{\tau,T}(h) \rangle. \end{aligned}$$

Here and in the following, β stands for a universal deterministic positive constant, possibly changing from lines to lines. Therefore, we have from (2.9)

$$E \max_{\tau \le t \le T} |\psi_{\tau,t}(h)|^2 + E \int_\tau^T |\mu_{\tau,t}(h)|^2 \, dt \le \beta E \left(|\psi_{\tau,\tau}(h)| \cdot |h| \right). \tag{2.10}$$

In particular,

$$E|\psi_{\tau,\tau}(h)|^2 \le \beta E(|\psi_{\tau,\tau}(h)|\cdot|h|) \le \beta(E|\psi_{\tau,\tau}(h)|^2 E|h|^2)^{1/2},$$

which implies

$$E|\psi_{\tau,\tau}(h)|^2 \le \beta E|h|^2. \tag{2.11}$$

On the other hand, we derive from the forward stochastic differential equation in (2.6) the following estimate:

$$E\max_{\tau\le t\le T}|\phi_{\tau,t}(h)|^2 \le \beta\Big\{E|h|^2 + E\int_\tau^T |u_t|^2\,dt\Big\}.$$

From (2.9) and assumption (A3), we have

$$E\max_{\tau\le t\le T}|\phi_{\tau,t}(h)|^2 \le \beta\Big\{E|h|^2 + E\int_\tau^T \langle N_t u_t, u_t\rangle\,dt\Big\} \le \beta\Big\{E|h|^2 + E(|\psi_{\tau,\tau}(h)|\cdot|h|)\Big\}. \tag{2.12}$$

Combining (2.10), (2.11), and (2.12), it is easy to see that (2.7) holds. The proof is complete. □

An indirect proof of Theorem 2.2 can also be given similar to the proof in Bismut [4]. The above uniqueness proof is a direct one, which is an adaptation of relevant arguments of Peng and Wu [21].

In the above, the solution of the LQ problem is reduced in an equivalent way to the solution of the associated stochastic Hamilton system (2.6). However, the stochastic Hamilton system (2.6) is a system of fully coupled FBSDEs, which is not a satisfactory characterization to the optimal control. Some efforts have been made by Bismut [3, 4] to decouple the FBSDEs, along the lines of Lions [15, Chapter III, section 4]. In the following, we summarize his relevant results and refine his partial arguments in the more general case of random initial times.

Let e_i denote the unit vector of $\mathbb{R}^n$ whose ith component is one. Define, for $\tau \le t \le T$,

$$\begin{aligned} X_{\tau,t} &:= (\phi_{\tau,t}(e_1),\dots,\phi_{\tau,t}(e_n)),\\ Y_{\tau,t} &:= (\psi_{\tau,t}(e_1),\dots,\psi_{\tau,t}(e_n)),\\ Z_{\tau,t} &:= (\mu_{\tau,t}(e_1),\dots,\mu_{\tau,t}(e_n)). \end{aligned} \tag{2.13}$$

Then,

$$E\max_{\tau\le t\le T}|X_{\tau,t}|^2 + E\max_{\tau\le t\le T}|Y_{\tau,t}|^2 + E\int_\tau^T |Z_{\tau,t}|^2\,dt < \infty.$$

Since $Y_{\tau,t}$ is almost surely continuous in $t\in[\tau,T]$, the meaning of $Y_{\tau,\tau}$ is clear, and set $P_\tau := Y_{\tau,\tau}$. Note that it is not clear whether the matrix-valued process $\{P_t, 0\le t\le T\}$ is continuous.

THEOREM 2.3. *Let assumptions* (A1), (A2), *and* (A3) *be satisfied. Then, we have, for any* $h\in L^2(\Omega,\mathcal{F}_\tau,P;\mathbb{R}^n)$ *and any* $\{\mathcal{F}_t, 0\le t\le T\}$*-stopping time* $\gamma\in[\tau,T]$,

$$\phi_{\tau,\gamma}(h) = X_{\tau,\gamma}h, \quad \psi_{\tau,\gamma}(h) = Y_{\tau,\gamma}h, \quad \mu_{\tau,\gamma}(h) = Z_{\tau,\gamma}h. \tag{2.14}$$

In particular,

$$\psi_{\tau,\tau}(h) = Y_{\tau,\tau}h = P_\tau h, \qquad h \in L^2(\Omega, \mathcal{F}_\tau, P; \mathbb{R}^n), \tag{2.15}$$

and therefore, for every stopping $\tau \le T$,

$$\psi_{0,\tau}(h) = Y_{\tau,\tau}\phi_{0,\tau}(h) = P_\tau \phi_{0,\tau}(h). \tag{2.16}$$

Remark 2.1. From the first equality of (2.14), it follows that for $s \le t$, the random linear transformation $\phi_{s,t}(\cdot)$ is a.s. a homomorphism if and only if the transformation matrix $X_{s,t}$ *a.s.* has an inverse.

Proof. We can use Itô's formula to verify that $\{(X_{\tau,t}h, Y_{\tau,t}h, Z_{\tau,t}h), \tau \le t \le T\}$ satisfies (2.6). By the uniqueness assertion of Theorem 2.2, we get (2.14). Letting $\gamma = \tau$ in the second equality of (2.14), we get (2.15). Finally, since $\psi_{0,\tau}(h) = \psi_{\tau,\tau}(\phi_{0,\tau}(h))$, then (2.16) follows from (2.15). The proof is complete. □

We have the following feedback representation of the adjoint process $\{y_t, 0 \le t \le T\}$.

THEOREM 2.4. *Let assumptions* (A1), (A2), *and* (A3) *be satisfied. Let* (x, y, z) *be the solution of* (2.6) *as* $\tau = 0$. *Then for any stopping time* $\gamma \le T$, *we have*

$$y_\gamma = P_\gamma x_\gamma, \quad a.s. \tag{2.17}$$

Moreover, we have

$$\begin{aligned} P_\gamma &= E^{\mathcal{F}_\gamma}\left\{\int_\gamma^T (\langle Q_s X_{\gamma,s}, X_{\gamma,s}\rangle + \langle N_s U_{\gamma,s}, U_{\gamma,s}\rangle)\, ds + \langle M X_{\gamma,T}, X_{\gamma,T}\rangle\right\}, \\ P_T &= M, \end{aligned} \tag{2.18}$$

and P_γ *is symmetric, nonnegative, and uniformly bounded. Here,*

$$U_{\gamma,s} = -N_s^{-1}\left[B_s' Y_{\gamma,s} + \sum_{i=1}^d D_s^i Z_{\gamma,s}^i\right], \qquad \gamma \le s \le T. \tag{2.19}$$

Proof of Theorem 2.4. We have $x_\gamma = \phi_{0,\gamma}(h)$ and $y_\gamma = \psi_{0,\gamma}(h)$. Then (2.17) is identical to (2.16). The rest assertions of Theorem 2.4 can be proved in a way similar to the proof of Proposition II.4 of Bismut [4, p. 211]. □

3. The Riccati equation: Known connections to the LQ problem and the Hamilton system. In view of deterministic LQ theory, it is natural to connect the stochastic LQ problem with the Riccati equation. In fact, the Riccati equation results from decoupling the stochastic Hamilton system. However, the way how to go from the stochastic Hamilton system to the Riccati equation has not yet been—to the best of the author's knowledge—established in a rigorous manner. In the literature, a formal approach to derive the associated Riccati equation from the stochastic Hamilton system a priori assumes that there is a semimartingale K of the form

$$K_t = K_0 + \int_0^t K_1(s)\, ds + \int_0^t \sum_{i=1}^d L_s^i\, dW_s^i$$

such that

$$y_t = K_t x_t.$$

Then, use Itô's formula to compute $K_t x_t$, compare with y_t, and identify the integrands of the Lebesgue integral and Itô's integral, respectively. As a consequence, the following Riccati equation can be derived:

$$\begin{cases} dK_t = -G(A_t, B_t, C_t, D_t; Q_t, N_t; K_t, L_t)\,dt + \sum_{i=1}^{d} L_t^i\,dW_t^i, \\ K_T = M, \qquad L_t := (L_t^1, \dots, L_t^d), \end{cases} \tag{3.1}$$

where for any $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C = (C^1, \dots, C^d) \in (\mathbb{R}^{n\times n})^d$, $D = (D^1, \dots, D^d) \in (\mathbb{R}^{n\times m})^d$; $Q \in \mathbb{R}^{n\times n}$(being nonnegative), $N \in \mathbb{R}^{m\times m}$(being uniformly positive), $K \in \mathbb{R}^{n\times n}$, $L = (L^1, \dots, L^d) \in (\mathbb{R}^{n\times n})^d$ with K being symmetric and nonnegative and L^i $(i = 1, \dots, d)$ being symmetric, we have defined

$$\begin{aligned} G_1(B, C, D, N; K, L) := & -\left[KB + \sum_{i=1}^{d}(C^i)'KD^i + \sum_{i=1}^{d} L^i D^i\right]\left[N + \sum_{i=1}^{d}(D^i)'KD^i\right]^{-1} \\ & \times \left[KB + \sum_{i=1}^{d}(C^i)'KD^i + \sum_{i=1}^{d} L^i D^i\right]', \\ G(A, B, C, D; Q, N; K, L) := & A'K + KA + Q + \sum_{i=1}^{d}(C^i)'KC^i + \sum_{i=1}^{d}[(C^i)'L^i + L^i C^i] \\ & + G_1(B, C, D, N; K, L). \end{aligned} \tag{3.2}$$

It is a BSDE with the generator $G(A_t, B_t, C_t, D_t; Q_t, N_t; K, L)$ being nonlinear in K and L. For the full details on the above-mentioned formal derivation, we refer to Bismut [3, 4]. The above backward stochastic Riccati differential equation (3.1) will be hereafter abbreviated as BSRDE (3.1). Note that the semimartingale property of K is assumed rather than being proved.

DEFINITION 3.1. *A solution of BSRDE* (3.1) *is defined as a pair* (K, L) *of adapted matrix processes such that*

(i) $\int_0^T |L_s|^2\,ds < \infty$, *a.s.*;

(ii) $N + \sum_{i=1}^{d}(D^i)'KD^i$ *is a.s.a.e. positive; moreover,*

$$\int_0^T |G(A_s, B_s, C_s, D_s; Q_s, N_s; K_s, L_s)|\,ds < \infty, \ \text{a.s.};$$

and

(iii) $K_t = M + \int_t^T G(A_s, B_s, C_s, D_s; Q_s, N_s; K_s, L_s)\,ds - \int_t^T \sum_{i=1}^{d} L_s^i\,dW_s^i$ *for all* $t \in [0, T]$.

In the literature, we have the following rigorous connections of the Riccati equation to the stochastic Hamilton system and to the stochastic LQ problem.

THEOREM 3.1. *Let assumptions* (A1), (A2), *and* (A3) *be satisfied. Let* τ *be a* $\{\mathcal{F}_t, 0 \le t \le T\}$*-stopping time such that* $0 \le \tau \le T$, *and let* (x, y, z, u) *be the solution of* (2.6) *with* u *being the optimal control. Assume that* (K, L) *is a solution to BSRDE* (3.1) *such that* K *is nonnegative and uniformly bounded and* L *is square integrable. Then, we have, for* $t \in [\tau, T]$,

$$\begin{aligned} & y_t = K_t x_t, \\ & z_t^i = L_t^i x_t + K_t(C_t^i x_t + D_t^i u_t), \qquad i = 1, 2, \dots, d, \\ & z_t := (z_t^1, \cdots, z_t^d), \qquad L_t := (L_t^1, \dots, L_t^d). \end{aligned} \tag{3.3}$$

Proof of Theorem 3.1. Use Itô's formula to compute $K_t x_t$ and compare it with y_t. The identification of the integrands of Lebesgue integrals and Itô's integrals yields the desired connections (3.3). □

THEOREM 3.2. *Let assumptions* (A1), (A2), *and* (A3) *be satisfied.* τ *is a* $\{\mathcal{F}_t, 0 \le t \le T\}$*-stopping time such that* $0 \le \tau \le T$. *Let* u *be the optimal control of the parameterized stochastic LQ problem and* x *be the solution of* (2.2). *Assume that* (K, L) *is a solution to BSRDE* (3.1) *such that* K *is nonnegative and uniformly bounded and* $L := (L^1, \dots, L^d)$ *is square integrable. Then, the optimal control* u *has the following closed form: for* $t \in [\tau, T]$,

$$u_t = -\left[N_t + \sum_{i=1}^{d} (D_t^i)' K_t D_t^i\right]^{-1} \left[K_t B_t + \sum_{i=1}^{d} (C_t^i)' K_t D_t^i + \sum_{i=1}^{d} L_t^i D_t^i\right]' x_t. \tag{3.4}$$

Moreover, for every $h \in L^2(\Omega, \mathcal{F}_\tau, P; \mathbb{R}^n)$, *we have*

$$\begin{aligned} E\langle K_\tau h, h\rangle = \inf_{u \in \mathcal{L}^2_{\mathcal{F}}(0,T;\mathbb{R}^m)} E\Big\{ & \langle M\phi_{\tau,T}(h), \phi_{\tau,T}(h)\rangle \\ & + \int_\tau^T \left(\langle Q_r \phi_{\tau,r}(h), \phi_{\tau,r}(h)\rangle + \langle N_r u_r, u_r\rangle\right) dr \Big\}, \end{aligned} \tag{3.5}$$

where $\phi_{t,\cdot}$ *is the state process starting from* h *at time* t *under the control process* u.

Remark 3.1. Formula (3.4) provides a characterization of the optimal control in terms of the solution of BSRDE (3.1). BSRDE (3.1) is not a coupled equation, and this characterization is preferred to (2.3).

Remark 3.2. Putting (3.4) into the second equality of (3.3), we have

$$\begin{aligned} z_t^i = {} & (L_t^i + K_t C_t^i) x_t \\ & - K_t D_t^i \left[N_t + \sum_{i=1}^{d} (D_t^i)' K_t D_t^i\right]^{-1} \left[K_t B_t + \sum_{i=1}^{d} (C_t^i)' K_t D_t^i + \sum_{i=1}^{d} L_t^i D_t^i\right]' x_t, \\ & i = 1, \dots, d, \quad t \in [\tau, T]. \end{aligned} \tag{3.6}$$

Proof of Theorem 3.2. Putting (3.3) into (2.3), we get (3.4). Combining Lemma 2.1, the first relation in (3.3), and Theorem 2.1, we get (3.5). The proof is complete. □

Obviously, the above mathematically rigorous relationship between Riccati equation (3.1) and the stochastic Hamilton system (2.6) is *not complete*. There is *a gap* here. In this paper, we close this gap by providing a rigorous argument to derive Riccati equation (3.1) from the stochastic Hamilton system (see sections 5–8 below). In this way, we show the existence of a solution to Riccati equation (3.1)—which solves Bismut and Peng's long-standing open problem (for the case of a Brownian filtration) initially proposed in 1978 by J. M. Bismut [4] (see section 4 below for more information).

4. Previous results on the Riccati equation and some comments. The connection of the stochastic LQ problem to BSRDE (3.1) indicated in Theorem 3.2 directs the study of the former to the study of the latter. And some crucial results and useful methods have been developed for the latter.

When the coefficients A, B, C^i, D^i, Q, N, M are all deterministic, then $L^1 = \cdots = L^d = 0$ and BSRDE (3.1) is reduced to the following nonlinear matrix ordinary

differential equation:

$$\begin{cases} \dfrac{d}{dt}K_t = -G(A_t, B_t, C_t, D_t; Q_t, N_t; K_t, 0), & 0 \le t < T, \\ K_T = M, \end{cases} \tag{4.1}$$

which was essentially solved by Wonham [23] by applying Bellman's principle of quasi linearization (see Bellman [1]) and a monotone convergence result of symmetric matrices.

The attention to the randomness of the coefficients A, B, C, D, Q, N, M is dated back at least to Bismut [3, 4]. Bismut [3, 4] studied the case of random coefficients, but he could solve only some special simple cases at that time. Let the integer $0 \le d_0 \le d$, and denote by $\{\mathcal{G}_t, 0 \le t \le T\}$ the P-augmented natural filtration generated by the $(d - d_0)$-dimensional Brownian motion $(W^{d_0+1}, \dots, W^d)$. Bismut [3, 4] assumed that the randomness of the coefficients only comes from the smaller filtration $\{\mathcal{G}_t, 0 \le t \le T\}$, which leads to $L^1 = \cdots = L^{d_0} = 0$. He further assumed in the paper [3] that

$$C^{d_0+1} = \cdots = C^d = 0, \qquad D^{d_0+1} = \cdots = D^d = 0, \tag{4.2}$$

under which the generator G does not involve L at all. In the work [4], Bismut assumed only that

$$D^{d_0+1} = \cdots = D^d = 0, \tag{4.3}$$

under which the generator G depends on the second unknown variable L only in a linear way. Moreover, his method consists of constructing a contraction mapping and then using a fixed point theorem.

Later, Peng [18] gave a nice treatment on the proof of existence and uniqueness for BSRDE (4.5) by using Bellman's principle of quasi linearization and a monotone convergence result of symmetric matrices—a generalization of Wonham's approach to the random situation.

The solution of the general BSRDE (3.1), whose generator is allowed to contain a quadratic term of L, turns out to become a long-standing problem. As early as 1978, Bismut [4] commented on page 220 that "Nous ne pourrons pas démontrer l'existence de solution pour l'équation (2.49) dans le cas général." (We could not prove the existence of solution for equation (2.49) for the general case.) On page 238, he pointed out that the essential difficulty for the solution of the general BSRDE (3.1) lies in the integrand of the martingale term which appears in the generator in a quadratic way. Since then, the stochastic LQ problem seemed to have been silent until Peng [18]. Two decades later in 1998, Peng [19] formally included the above problem in his list of open problems on BSDEs.

Taking this opportunity, the author would like to acknowledge that as early as 1993, Professor S. Peng has introduced the open case of the general adapted possibly nonzero D in his private communication with the author.

To overcome the difficult feature of the quadratic growth in the martingale term of the generator G, Kobylanski [8] and Lepeltier and San Martin [13, 14] have developed a quite useful technique. Unfortunately, this technique is essentially of one-dimensional nature, and is difficult to be adapted to the underlying matrix-valued BSRDEs.

Recently, Kohlmann and Tang have made some progress towards solving the above open problem. See [9, 10, 11, 12] and the references therein. However, it is still very far from the complete solution.

In what follows, we shall give a complete solution with a new constructive method —which is totally different from the previous methods.

5. Main ideas and results of the paper. The traditional formal derivation of the Riccati equation (3.1) uses the following a priori hypothesis: assume that P in the adjoint-primal processes relation (see section 3)

$$y_t = P_t x_t, \qquad 0 \le t \le T,$$

or in the associated value function formula (see Peng [18, p. 299, eq. (5.4)])

$$V(t,x) := \langle P_t x, x\rangle, \qquad (t,x) \in [0,T] \times \mathbb{R}^n,$$

is a semimartingale. Note that P_t will be shown to be the first component K_t of the solution (K,L) of BSRDE (3.1).

The hypothesis has not yet been proved to be true, to the author's best knowledge. Therefore, the relevant arguments existing in the literature are formal, rather than being rigorous.

In this paper, we provide a rigorous derivation of Riccati equation (3.1) from the stochastic Hamilton system (2.6). We construct a solution of Riccati equation (3.1), using a basis of fundamental solutions of the stochastic Hamilton system (2.6).

Our *new observation* is the *connection* of the *existence* issue for BSRDE (3.1) to the *homomorphism* of the stochastic flows $\{\phi_{0,\cdot}(h), h \in \mathbb{R}^n\}$: if the stochastic flows $\{\phi_{0,\cdot}(h), h \in \mathbb{R}^n\}$ are a.s. homomorphism at each time t, i.e., if they a.s. have an inverse at each time t, then we can produce a BSDE from the first equality of (3.3), which turns out to be exactly BSRDE (3.1). Theorem 5.2 below states their equivalence.

The above-stated connection points out a promising new approach to the study of BSRDE (3.1). However, it is not directly known whether the stochastic flows $\{\phi_{0,\cdot}(h), h \in \mathbb{R}^n\}$ are a.s. homomorphism at each time t. This difficulty is overcome by considering the *first degenerate time* τ of the associated flows transformation (see the precise definition below), which is a stopping time.

The *first degenerate time* τ of the associated flows transformation will be shown to be a.s. $+\infty$, i.e., the stochastic flows $\{\phi_{0,\cdot}(h), h \in \mathbb{R}^n\}$ are a.s. homomorphism at each time $t \in [0,T]$. To prove this fact, we need to study the regularity of the approaching quantities $L(k)$ (see the definitions below) near the degenerate time τ, and then the following a priori estimate is needed.

THEOREM 5.1. *Let assumptions* (A1), (A2), *and* (A3) *be satisfied. Let* (K,L) *be a solution of BSRDE* (3.1) *with* K *being a.s.a.e. nonnegative and uniformly bounded. Then, there is a deterministic constant* β_0 *such that the following estimate holds:*

$$E\left(\int_0^T |L_s|^2\, ds\right)^p \le \beta_0, \qquad \forall p \ge 1. \tag{5.1}$$

Here, β_0 *depends only on* p, *the uniform upper bound of* K *and all the coefficients, and the uniformly lower bound of (positive) eigenvalues of* $\{N_t, 0 \le t \le T\}$.

In Theorem 5.1, K is assumed to be uniformly bounded. This distinguishes Theorem 5.1 from the known a priori estimates for BSDEs given by Pardoux and Peng [17] and El Karoui, Peng, and Quenez [6]. Note that the operator P_t which connects the adjoint process y_t and the primal state x_t "should" be the first component K_t of the solution (K,L) of BSRDE (3.1). In sections 2 and 3, through a flows analysis on the associated stochastic Hamilton system, we have obtained rich information on P_t: nonnegativity and uniform boundedness. However, we could say almost nothing about L: the difficulty lies in the fact that $L := (L^1, \dots, L^d)$ and $z := (z^1, \dots, z^d)$

in the second equality (3.3) or (3.6) is in general known only to be square integrable over time rather than continuous, which creates a difficulty in explaining the second equality of (3.3) or (3.6) at a time. Fortunately, Theorem 5.1 permits us to draw a useful property of L from those known properties on P obtained in sections 2 and 3.

Before going further, it is necessary to introduce the following notation.

Let $u(e_1), \dots, u(e_n)$ denote the optimal controls corresponding to the initial states $e_1, \dots, e_n$, respectively, at the initial time 0. Define

$$\begin{aligned} &U := (u(e_1), \dots, u(e_n)), \qquad X := (\phi_{0,\cdot}(e_1), \dots, \phi_{0,\cdot}(e_n)), \\ &Z^i := (\mu^i_{0,\cdot}(e_1), \dots, \mu^i_{0,\cdot}(e_n)), \qquad Y := (\psi_{0,\cdot}(e_1), \dots, \psi_{0,\cdot}(e_n)). \end{aligned} \tag{5.2}$$

Then, it is straightforward that X solves the matrix-valued stochastic differential equation (SDE in short form):

$$\begin{cases} dX_t = (A_t X_t + B_t U_t)\, dt + \displaystyle\sum_{i=1}^{d} (C^i_t X_t + D^i_t U_t)\, dW^i_t, \\ X_0 = I_{n\times n}, \text{the unit matrix of dimension } n \times n, \end{cases} \tag{5.3}$$

and the pair (Y, Z) of adapted processes solves the matrix-valued BSDE:

$$\begin{cases} -dY_t = \Big[A'_t Y_t + Q_t X_t + \displaystyle\sum_{i=1}^{d} (C^i_t)' Z^i_t \Big]\, dt - \sum_{i=1}^{d} Z^i_t\, dW^i_t, \\ Y_T = MX_T, \qquad Z_t := (Z^1_t, \dots, Z^d_t). \end{cases} \tag{5.4}$$

Moreover, it follows from Theorem 2.1 that the triple (U, Y, Z) of processes satisfies the following:

$$N_t U_t + B'_t Y_t + \sum_{i=1}^{d} (D^i_t)' Z^i_t = 0 \quad a.s.a.e.\ t \in [0, T]. \tag{5.5}$$

The following theorem states the *equivalence* between the *existence* of a solution to BSRDE (3.1) and the *homomorphism* of the stochastic flows $\{\phi_{0,\cdot}(h), h \in \mathbb{R}^n\}$.

THEOREM 5.2. *Let assumptions* (A1), (A2), *and* (A3) *be satisfied. Let* (X, Y, Z, U) *be defined by* (5.2). *Then, the existence of a solution to BSRDE* (3.1) *is equivalent to the homomorphism of the stochastic flows* $\{\phi_{0,\cdot}(h), h \in \mathbb{R}^n\}$, *i.e., to the fact that* X_t *a.s. has an inverse at each* $t \in [0, T]$. *More precisely, we have the following two assertions:*

(i) *If BSRDE has a solution* (K, L) *with* K *being a uniformly bounded nonnegative matrix valued process, then* X_t *a.s. has an inverse at each* $t \in [0, T]$.

(ii) *Conversely, if* X_t *a.s. has an inverse at each* $t \in [0, T]$, *then* (K, L) *defined by*

$$\begin{aligned} &K_t := Y_t X_t^{-1}, \\ &L^i_t := Z^i_t X_t^{-1} - Y_t X_t^{-1}(C^i_t + D^i_t U_t X_t^{-1}), \qquad i = 1, \dots, d, \\ &L_t := (L^1_t, \dots, L^d_t), \qquad 0 \le t \le T, \end{aligned} \tag{5.6}$$

solves BSRDE (3.1). *Moreover,* K *is nonnegative and uniformly bounded, and* L *is square integrable.*

Remark 5.1. By virtue of (5.5), we have

$$U_t = -N_t^{-1}\left[B_t' Y_t + \sum_{i=1}^{d}(D_t^i)' Z_t^i\right], \qquad t \in [0,T].$$

Therefore, L_t constructed in Theorem 5.2(ii) is expressible only in terms of the solutions (X, Y, Z) of stochastic Hamilton system (2.6).

If X_t a.s. has an inverse at each $t \in [0,T]$, then it follows from the above theorem that BSRDE (3.1) has a solution, and the uniqueness is easily derived from Theorem 3.2 (see (3.5)). This shows that BSRDE (3.1) has a unique solution. Unfortunately, as already pointed out in the beginning of this section, it is not obvious that X_t a.s. has an inverse at *every* $t \in [0,T]$. This seems to discourage us from going on. However, there are also the following two encouraging facts:

(i) X a.s. has an inverse on the subinterval $[0,\tau]$ when $\tau > 0$ is sufficiently small.

(ii) The a priori known uniform boundedness of the above-constructed $P = \{Y_t X_t^{-1}, 0 \le t \le T\}$ helps to improve the *regularity* of the constructed L (see Theorem 5.1). This property can be studied through their BSRDE (see (6.2) in section 6 below) using the classical techniques from the theory of BSDEs.

In the following, we develop the above two points. For this purpose, define

$$\tau := \inf\{t \in [0,T] : \det(X_t) \le 0\}. \tag{5.7}$$

Here and in the following, we use the convention that $\inf \emptyset = +\infty$. Since $X_0 = I_{n\times n}$, $\det(X_0) = 1$, and since X is a continuous process, we have $\tau > 0$, a.s. Define

$$\tau_k := \inf\left\{t \in [0,T] : \det(X_t) \le \frac{1}{k+1}\right\}, \qquad k = 1, \dots. \tag{5.8}$$

We have

$$\tau_k \uparrow \tau, \qquad \text{as } k \uparrow +\infty, \tag{5.9}$$

and

$$\det(X_{t\wedge\tau_k}) \ge \frac{1}{k+1} > 0 \qquad \forall t \in [0,T] \tag{5.10}$$

for each positive integer k, i.e., $X_{t\wedge\tau_k}$ a.s. has an inverse at every $t \in [0,T]$.

THEOREM 5.3. *Let assumptions* (A1), (A2), *and* (A3) *be satisfied. Then, we have*

$$\tau = +\infty, \qquad a.s.$$

Therefore, X_t a.s. has an inverse for each $t \in [0,T]$, and BSRDE (3.1) *has a unique adapted solution (K, L) with $K := \{Y_t X_t^{-1}, 0 \le t \le T\}$ being nonnegative and uniformly bounded and $L := \{\{Z_t^i X_t^{-1} - Y_t X_t^{-1}(C_t^i + D_t^i U_t X_t^{-1})\}_{i=1}^d, 0 \le t \le T\}$ being square integrable.*

6. A priori estimate for the Riccati equation: The proof of Theorem 5.1. This section is devoted to the proof of Theorem 5.1.

Define, for $k = 1, 2, \dots,$

$$\sigma_k := T \wedge \inf\left\{t \in [0,T] : \int_0^t |L_s|^2\, ds > k\right\}. \tag{6.1}$$

Note the convention $\inf \emptyset = +\infty$. Then, σ_k is a stopping time for each positive integer k. Moreover, as $k \to \infty$, we have $\sigma_k \uparrow T$, a.s.

$\{(K_{t\wedge\sigma_k}, \chi_{[0,\sigma_k]}L(t)), 0 \le t \le T\}$ satisfies the following:

$$\begin{aligned} dK_{t\wedge\sigma_k} = &-\chi_{[0,\sigma_k]}(t)\Big[A_t'K_t + K_tA_t + \sum_{i=1}^{d}(C_t^i)'K_tC_t^i + Q_t + \sum_{i=1}^{d}[(C_t^i)'L_t^i + L_t^iC_t^i] \\ &+ G_1(B_t, C_t, D_t, N_t; K_t, L_t)\Big]\,dt + \sum_{i=1}^{d}\chi_{[0,\sigma_k]}(t)L_t^i\,dW_t^i, \qquad 0 \le t < T, \end{aligned} \tag{6.2}$$

where G_1 is defined by (3.2).

Using Itô's formula, we get

$$\begin{aligned} d|K_{t\wedge\sigma_k}|^2 = &-\chi_{[0,\sigma_k]}(t)\Big[4\,\mathrm{tr}\,\left(K_t^2A_t\right) + \sum_{i=1}^{d}2\,\mathrm{tr}\,\left(K_t(C_t^i)'K_tC_t^i\right) + 2\,\mathrm{tr}\,\left(K_tQ_t\right) \\ &+ \sum_{i=1}^{d}4\,\mathrm{tr}\,\left(K_tL_t^iC_t^i\right) + 2\,\mathrm{tr}\,\left[K_tG_1(B_t, C_t, D_t, N_t; K_t, L_t)\right] - |L_t|^2\Big]\,dt \\ &+ \sum_{i=1}^{d}2\chi_{[0,\sigma_k]}(t)\,\mathrm{tr}\,\left(K_tL_t^i\right)\,dW_t^i, \qquad 0 \le t < T. \end{aligned} \tag{6.3}$$

We observe that since

$$G_1(B_t, C_t, D_t, N_t; K_t, L_t) \le 0, \qquad K_t \ge 0, \tag{6.4}$$

we have

$$\mathrm{tr}\,\left[K_tG_1(B_t, C_t, D_t; K_t, L_t)\right] = \mathrm{tr}\,\left[K_t^{\frac{1}{2}}G_1(B_t, C_t, D_t; K_t, L_t)K_t^{\frac{1}{2}}\right] \le 0. \tag{6.5}$$

Hence, from (6.3), it follows

$$\begin{aligned} &\int_0^T \chi_{[0,\sigma_k]}(s)|L_s|^2\,ds \\ &= |K_{T\wedge\sigma_k}|^2 - |K_0|^2 + \int_0^T 2\chi_{[0,\sigma_k]}(s)\,\mathrm{tr}\,\left[K_sG_1(B_s, C_s, D_s; K_s, L_s)\right]\,ds \\ &\quad + \int_0^T \chi_{[0,\sigma_k]}(s)\Big[4\,\mathrm{tr}\,\left(K_s^2A_s\right) + \sum_{i=1}^{d}2\,\mathrm{tr}\,\left(K_s(C_s^i)'K_sC_s^i\right) + 2\,\mathrm{tr}\,\left(K_sQ_s\right) \\ &\quad + \sum_{i=1}^{d}4\,\mathrm{tr}\,\left(K_sL_s^iC_s^i\right)\Big]\,ds - \int_0^T 2\chi_{[0,\sigma_k]}(s)\sum_{i=1}^{d}\mathrm{tr}\,\left(K_sL_s^i\right)\,dW_s^i \\ &\le \int_0^T \chi_{[0,\sigma_k]}(s)\Big[4\,\mathrm{tr}\,\left(K_s^2A_s\right) + \sum_{i=1}^{d}2\,\mathrm{tr}\,\left(K_s(C_s^i)'K_sC_s^i\right) + 2\,\mathrm{tr}\,\left(K_sQ_s\right) \\ &\quad + \sum_{i=1}^{d}4\,\mathrm{tr}\,\left(K_sL_s^iC_s^i\right)\Big]\,ds + |K_{T\wedge\sigma_k}|^2 - \int_0^T 2\chi_{[0,\sigma_k]}(s)\sum_{i=1}^{d}\mathrm{tr}\,\left(K_sL_s^i\right)\,dW_s^i. \end{aligned} \tag{6.6}$$

Therefore (using assumption (A1) and the uniform boundedness of K),

$$\left(\int_0^T \chi_{[0,\sigma_k]}|L|^2\,ds\right)^p \le \beta\left[1+\left(\int_0^T \chi_{[0,\sigma_k]}|L|\,ds\right)^p + \left|\int_0^T 2\chi_{[0,\sigma_k]}\sum_{i=1}^d \operatorname{tr}(KL^i)\,dW_s^i\right|^p\right]. \tag{6.7}$$

We have from the Burkholder–Davis–Gundy inequality the following:

$$E\left|\int_0^T \sum_{i=1}^d 2\chi_{[0,\sigma_k]}\operatorname{tr}\left(KL^i\right)\,dW_s^i\right|^p \le \beta E\left|\int_0^T \chi_{[0,\sigma_k]}|K|^2|L|^2\,ds\right|^{p/2}, \tag{6.8}$$

while from the Cauchy–Schwarz inequality, we have

$$E\left(\int_0^T \chi_{[0,\sigma_k]}|L|\,ds\right)^p \le T^{p/2}E\left(\int_0^T \chi_{[0,\sigma_k]}|L|^2\,ds\right)^{p/2}. \tag{6.9}$$

Finally, in view of (6.8) and (6.9), we get from (6.7)

$$E\left(\int_0^T \chi_{[0,\sigma_k]}|L|^2\,ds\right)^p \le \beta+\beta E\left(\int_0^T \chi_{[0,\sigma_k]}|L|^2\,ds\right)^{p/2},$$

which by the elementary inequality $2ab \le a^2+b^2$ implies the following:

$$E\left(\int_0^T \chi_{[0,\sigma_k]}|L|^2\,ds\right)^p \le \beta, \qquad k=1,2,\ldots.$$

Using Fatou's lemma, we have

$$\begin{aligned} E\left(\int_0^T |L|^2\,ds\right)^p &= E\lim_{k\to+\infty}\left(\int_0^{\sigma_k}|L|^2\,ds\right)^p = E\lim_{k\to+\infty}\left(\int_0^T \chi_{[0,\sigma_k]}|L|^2\,ds\right)^p \\ &\le \varliminf_{k\to+\infty} E\left(\int_0^T \chi_{[0,\sigma_k]}|L|^2\,ds\right)^p \le \beta, \qquad k=1,2,\ldots, \end{aligned}$$

which completes the proof. □

7. Equivalence between the existence of the solution to the Riccati equation and the homomorphism of the stochastic flows: The proof of Theorem 5.2. This section is devoted to the proof of Theorem 5.2—the equivalence between the existence of a solution to BSRDE (3.1) and the homomorphism of the stochastic flows $\{\phi_{0,\cdot}(h), h\in\mathbb{R}^n\}$.

For the reader's convenience, we begin with the following lemma, which is an immediate adaptation of Gal'chuk [7, basic theorem on pp. 756–757] to the underlying semimartingale

$$\{(\overbrace{1,\ldots,1}^{d})'t+W_t, 0\le t\le T\}.$$

LEMMA 7.1. *Assume that the vector functions* $f : \Omega \times [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ *and* $g : \Omega \times [0,T] \times \mathbb{R}^n \to \mathbb{R}^{n\times d}$ *satisfy the following two conditions:*

(i) *For each* $x \in \mathbb{R}^n$, $f(\cdot,x)$ *and* $g(\cdot,x)$ *are* $\{\mathcal{F}_t, 0 \le t \le T\}$*-adapted processes. Moreover,*

$$\int_0^T |f(t,0)|\,dt < \infty, \qquad \int_0^T |g(t,0)|^2\,dt < \infty, \quad a.s.$$

(ii) *There exist two positive functions* α_1 *and* α_2 *such that they are* $\{\mathcal{F}_t, 0 \le t \le T\}$*-adapted. Moreover,*

$$\int_0^T \alpha_1(t)\,dt < \infty, \qquad \int_0^T |\alpha_2(t)|^2\,dt < \infty, \quad a.s.$$

For any $x, y \in \mathbb{R}^n$,

$$|f(t,x) - f(t,y)| \le \alpha_1(t)|x-y|,$$
$$|g(t,x) - g(t,y)| \le \alpha_2(t)|x-y|.$$

Then, the SDE

$$dx_t = f(t,x_t)\,dt + g(t,x_t)\,dW_t,\ 0 < t \le T; \qquad x(0) = h$$

has a unique strong solution.

The proof of Theorem 5.2 is divided into two parts.

Proof of (i). If BSRDE has a solution $\{(P_t, L_t), 0 \le t \le T\}$, then it follows from Theorem 3.1 that X solves the following matrix-valued SDE:

$$\begin{cases} dX_t = \widehat{A}_t X_t\,dt + \sum_{i=1}^d \widehat{C}_t^i X_t\,dW_t^i, \\ X_0 = I_{n\times n}, \end{cases} \tag{7.1}$$

where $\widehat{A}$ and $\widehat{C}^i$ $(i = 1,\dots,d)$ are defined as follows:

$$\begin{aligned} \widehat{A}_t &:= A_t - B_t\Big[N_t + \sum_{i=1}^d (D_t^i)' P_t D_t^i\Big]^{-1}\Big[P_t B_t + \sum_{i=1}^d (C_t^i)' P_t D_t^i + \sum_{i=1}^d L_t^i D_t^i\Big]', \\ \widehat{C}_t^i &:= C_t^i - D_t^i\Big[N_t + \sum_{i=1}^d (D_t^i)' P_t D_t^i\Big]^{-1}\Big[P_t B_t + \sum_{i=1}^d (C_t^i)' P_t D_t^i + \sum_{i=1}^d L_t^i D_t^i\Big]'. \end{aligned} \tag{7.2}$$

In view of assumption (A1) and Definition 3.1, the coefficients of the above optimal closed system $\widehat{A}$ and $\widehat{C}_i$ $(i = 1,\dots,d)$ satisfy the following conditions:

$$\int_0^T |\widehat{A}_s|^2\,ds < \infty, \qquad \int_0^T |\widehat{C}_s^i|^2\,ds < \infty, \quad i = 1,\dots,d, \quad \text{a.s.}$$

It follows from Lemma 7.1 that the linear matrix-valued SDE

$$\begin{cases} d\mathcal{X}_t = -\mathcal{X}_t\Big(\widehat{A}_t - \sum_{i=1}^d (\widehat{C}_t^i)^2\Big)\,dt - \mathcal{X}_t \sum_{i=1}^d \widehat{C}_t^i\,dW_t^i, \\ \mathcal{X}_0 = I_{n\times n} \end{cases} \tag{7.3}$$

has a unique strong solution on $[0,T]$. Using Itô's formula, we can verify that $X\mathcal{X}$ solves the following SDE:

$$\begin{cases} dV_t = \Big(\widehat{A}_t V_t - V_t\widehat{A}_t + V_t\sum_{i=1}^d(\widehat{C}_t^i)^2 - \sum_{i=1}^d \widehat{C}_t^i V_t \widehat{C}_t^i\Big)\,dt + \sum_{i=1}^d(\widehat{C}_t^i V_t - V_t\widehat{C}_t^i)\,dW_t^i, \\ V_0 = I_{n\times n}. \end{cases} \tag{7.4}$$

Obviously, the constant matrix process $I_{n\times n}$ is its solution. Since this equation has a unique solution (by virtue of Lemma 7.1), we have $X_t\mathcal{X}_t = I_{n\times n}$ a.s. Therefore, X_t a.s. has an inverse at each $t\in[0,T]$. □

Proof of (ii). Applying Itô's formula to compute X_t^{-1}, we have from (5.3) the following:

$$\begin{aligned} dX_t^{-1} =& -X_t^{-1}(dX_t)X_t^{-1} + \sum_{i=1}^d X_t^{-1}(C_t^iX_t + D_t^iU_t)X_t^{-1}(C_t^iX_t + D_t^iU_t)X_t^{-1}\,dt \\ =& -X_t^{-1}\Big[A_t + B_tU_tX_t^{-1} - \sum_{i=1}^d(C_t^i + D_t^iU_tX_t^{-1})(C_t^i + D_t^iU_tX_t^{-1})\Big]\,dt \\ & - \sum_{i=1}^d X_t^{-1}(C_t^i + D_t^iU_tX_t^{-1})\,dW_t^i. \end{aligned} \tag{7.5}$$

Then, we have (in view of (5.4))

$$\begin{aligned} d(Y_tX_t^{-1}) =& -\sum_{i=1}^d Z_t^iX_t^{-1}(C_t^i + D_t^iU_tX_t^{-1})\,dt + Y_t\,d(X_t^{-1}) + (dY_t)X_t^{-1} \\ =& -\sum_{i=1}^d Y_tX_t^{-1}(C_t^i + D_t^iU_tX_t^{-1})\,dW_t^i - \sum_{i=1}^d Z_t^iX_t^{-1}(C_t^i + D_t^iU_tX_t^{-1})\,dt \\ & - Y_tX_t^{-1}\Big[A_t + B_tU_tX_t^{-1} - \sum_{i=1}^d(C_t^i + D_t^iU_tX_t^{-1})(C_t^i + D_t^iU_tX_t^{-1})\Big]\,dt \\ & - \Big[A_t'Y_tX_t^{-1} + Q_t + \sum_{i=1}^d(C_t^i)'Z_t^iX_t^{-1}\Big]\,dt + \sum_{i=1}^d Z_t^iX_t^{-1}\,dW_t^i \\ =& -\sum_{i=1}^d[Z_t^iX_t^{-1} - Y_tX_t^{-1}(C_t^i + D_t^iU_tX_t^{-1})](C_t^i + D_t^iU_tX_t^{-1})\,dt \\ & - \Big[A_t'Y_tX_t^{-1} + Q_t + \sum_{i=1}^d(C_t^i)'Z_t^iX_t^{-1}\Big]\,dt - Y_tX_t^{-1}(A_t + B_tU_tX_t^{-1})\,dt \\ & + \sum_{i=1}^d[Z_t^iX_t^{-1} - Y_tX_t^{-1}(C_t^i + D_t^iU_tX_t^{-1})]\,dW_t^i. \end{aligned} \tag{7.6}$$

Recall that

$$L_t^i := Z_t^iX_t^{-1} - Y_tX_t^{-1}(C_t^i + D_t^iU_tX_t^{-1}), \qquad i = 1,\dots,d. \tag{7.7}$$

Since X^{-1} is a.s. bounded on $[0,T]$, we have

$$\int_0^T |L_s|^2\,ds < \infty, \quad \text{a.s.}$$

However, it is not clear now that

$$E\int_0^T |L_s|^2\,ds < \infty.$$

From (7.7), it follows that

$$Z_t^i X_t^{-1} = L_t^i + Y_t X_t^{-1}(C_t^i + D_t^i U_t X_t^{-1}). \tag{7.8}$$

On the one hand, putting (7.8) into (7.6), we get

$$\begin{aligned} d(Y_tX_t^{-1}) =& -\left\{A_t'Y_tX_t^{-1} + Q_t + \sum_{i=1}^d (C_t^i)'[L_t^i + Y_tX_t^{-1}(C_t^i + D_t^iU_tX_t^{-1})]\right\}dt \\ & - Y_tX_t^{-1}(A_t + B_tU_tX_t^{-1})\,dt - \sum_{i=1}^d L_t^i(C_t^i + D_t^iU_tX_t^{-1})\,dt + \sum_{i=1}^d L_t^i\,dW_t^i \\ =& -\left\{A_t'Y_tX_t^{-1} + Y_tX_t^{-1}A_t + Q_t \right. \\ & \left. + \sum_{i=1}^d (C_t^i)'Y_tX_t^{-1}C_t^i + \sum_{i=1}^d [(C_t^i)'L_t^i + L_t^iC_t^i]\right\}dt \\ & - \left[Y_tX_t^{-1}B_t + \sum_{i=1}^d (C_t^i)'Y_tX_t^{-1}D_t^i + \sum_{i=1}^d L_t^iD_t^i\right]U_tX_t^{-1}\,dt + \sum_{i=1}^d L_t^i\,dW_t^i. \end{aligned} \tag{7.9}$$

On the other hand, in view of (5.5) and (7.8), we have

$$N_tU_tX_t^{-1} + B_t'Y_tX_t^{-1} + \sum_{i=1}^d (D_t^i)'[L_t^i + Y_tX_t^{-1}(C_t^i + D_t^iU_tX_t^{-1})] = 0; \tag{7.10}$$

that is,

$$\left(N_t + \sum_{i=1}^d (D_t^i)'Y_tX_t^{-1}D_t^i\right)U_tX_t^{-1} + B_t'Y_tX_t^{-1} + \sum_{i=1}^d (D_t^i)'Y_tX_t^{-1}C_t^i + \sum_{i=1}^d (D_t^i)'L_t^i = 0. \tag{7.11}$$

At this stage, we apply Theorem 2.4 and get

$$K_t = Y_tX_t^{-1} = P_t, \quad \text{a.s.} \tag{7.12}$$

From Theorem 2.4, it also follows that K is nonnegative and is uniformly bounded.

Putting (7.12) into (7.7) and (7.11), we have

$$L_t^i = Z_t^iX_t^{-1} - K_t(C_t^i + D_t^iU_tX_t^{-1}), \qquad i = 1,\dots,d, \tag{7.13}$$

and

$$(7.14)\quad \left(N_t+\sum_{i=1}^{d}(D_t^i)'K_tD_t^i\right)U_tX_t^{-1}+B_t'K_t+\sum_{i=1}^{d}(D_t^i)'K_tC_t^i+\sum_{i=1}^{d}(D_t^i)'L_t^i=0.$$

The last equality implies (noting K_t is symmetric and nonnegative) that

$$(7.15)\quad U_tX_t^{-1}=-\left(N_t+\sum_{i=1}^{d}(D_t^i)'K_tD^i\right)^{-1}\left[B_t'K_t+\sum_{i=1}^{d}(D_t^i)'K_tC_t^i+\sum_{i=1}^{d}(D_t^i)'L_t^i\right].$$

Putting (7.12) and (7.15) into (7.9), we see that (K,L) solves BSRDE (3.1). Moreover, we can apply Theorem 5.1 here and derive that L is square integrable. The proof is complete.

8. Existence and uniqueness of the solution to the BSRDE: The proof of Theorem 5.3. This section is devoted to the proof of Theorem 5.3.

The uniqueness easily follows from formula (3.5) in Theorem 3.2. In fact, (3.5) implies immediately the uniqueness of the first component of the adapted solution of BSRDE (3.1). Then, the uniqueness of the second component is proved by a direct application of Itô's formula to the difference of the two BSRDEs with the consideration that the first components are the same.

The following is concentrated on the proof of the existence part, by actually constructing a solution using the stochastic Hamilton system.

It is convenient to introduce the following notation:

$$(8.1)\quad \begin{aligned}&P_t:=Y_tX_t^{-1},\\ &L_t^i:=Z_t^iX_t^{-1}-P_t(C_t^i+D_t^iU_tX_t^{-1}),\qquad i=1,\dots,d,\\ &L_t:=(L_t^1,\cdots,L_t^d).\end{aligned}$$

Note that the two notations P_t and L_t are well defined only when X_t has an inverse. At the moment, we know that P and L are a.s. well defined on $[0,\tau)$. Please keep in mind that in what follows, P_t and L_t are assumed to be zero whenever they are not well defined in the context.

From (8.1), we see that both P_t and L_t might a.s. have a singularity at $t=\tau$.

The proof of the existence part consists of three steps. In the first step, we describe the counterparts of section 7 when the mathematical objects (processes, SDEs, BSDEs, BSRDE,...) are stopped or truncated by the sequence of stopping times $\{\tau_k\}_{k=1}^{\infty}$. In the second step, we perform a limit analysis to some concerned integrals and processes by giving five lemmas. In this step, Theorem 5.1 plays a crucial rule. Finally, in the third step, we pass to the limit in the sequence of stopped/truncated BSRDEs, based on the previous analysis at the limit.

Step 1. Define for $0\le t\le T$,

$$(8.2)\quad \begin{aligned}X_t^k:=&X_{t\wedge\tau_k},\qquad Y_t^k:=Y_{t\wedge\tau_k},\\ Z_t(k):=&\chi_{[0,\tau_k]}(t)Z_t,\qquad U_t^k:=\chi_{[0,\tau_k]}(t)U_t\end{aligned}$$

and

$$(8.3)\quad \begin{aligned}&A_t(k):=\chi_{[0,\tau_k]}(t)A_t,\qquad B_t(k):=\chi_{[0,\tau_k]}(t)B_t,\\ &C_t^i(k):=\chi_{[0,\tau_k]}(t)C_t^i,\qquad D_t^i(k):=\chi_{[0,\tau_k]}(t)D_t^i,\quad i=1,\dots,d,\\ &C_t(k):=(C_t^1(k),\dots,C_t^d(k)),\qquad D_t(k):=(D_t^1(k),\dots,D_t^d(k)),\\ &Q_t(k):=\chi_{[0,\tau_k]}(t)Q_t.\end{aligned}$$

Then, X^k a.s. has an inverse at each $t \in [0,T]$. Therefore, we can define, similar to section 7,

$$\begin{aligned} &P_t^k := Y_t^k (X_t^k)^{-1} = Y_{t\wedge\tau_k}(X_{t\wedge\tau_k})^{-1} = P_{t\wedge\tau_k}, \\ &L_t^i(k) := Z_t^i(k)(X_t^k)^{-1} - P_t^k[C_t^i(k) + D_t^i(k)U_t^k(X_t^k)^{-1}], \qquad i = 1,\dots,d, \\ &L_t(k) := (L_t^1(k),\dots,L_t^d(k)). \end{aligned} \tag{8.4}$$

From Theorem 2.4, it follows that P^k is uniformly bounded and $P_t^k \geq 0$.

From (5.3), it follows that X^k is the solution of the matrix-valued SDE:

$$\begin{cases} dX_t^k = [A_t(k)X_t^k + B_t(k)U_t^k]\,dt + \sum_{i=1}^d [C_t^i(k)X_t^k + D_t^i(k)U_t^k]\,dW_t^i, \\ X_0^k = I_{n\times n}. \end{cases} \tag{8.5}$$

From (5.4), it follows that the pair $(Y^k, Z(k))$ of processes is the solution of the matrix-valued BSDE:

$$\begin{cases} -dY_t^k = \left[A_t(k)'Y_t^k + Q_t(k)X_t^k + \sum_{i=1}^d C_t^i(k)'Z_t^i(k)\right]dt - \sum_{i=1}^d Z_t^i(k)\,dW_t^i, \\ Y_T^k = P_T^k X_T^k, \qquad Z_t(k) := (Z_t^1(k),\dots,Z_t^d(k)). \end{cases} \tag{8.6}$$

Moreover, from (5.5) we have

$$N_t U_t^k + B_t(k)'Y_t^k + \sum_{i=1}^d D_t^i(k)'Z_t^i(k) = 0. \tag{8.7}$$

Note that the terminal condition of BSDE (8.6) comes from the first relations in (8.4).

Proceeding identically as in the previous section, we can show that the pair $(P^k, L(k))$ of processes satisfies the following BSRDE:

$$\begin{cases} dP_t^k = -\Big\{A_t(k)'P_t^k + P_t^k A_t(k) + Q_t(k) + \sum_{i=1}^d C_t^i(k)'P_t^k C_t^i(k) \\ \qquad\qquad + \sum_{i=1}^d [C_t^i(k)'L_t^i(k) + L_t^i(k)C_t^i(k)]\Big\}\,dt \\ \qquad\qquad - G_1(B_t(k), C_t(k), D_t(k), N_t; P_t^k, L_t(k))\,dt + \sum_{i=1}^d L_t^i(k)\,dW_t^i, \quad 0 \leq t \leq T, \\ L_t(k) := (L_t^1(k),\dots,L_t^d(k)). \end{cases} \tag{8.8}$$

Step 2. We have the following five lemmas.

LEMMA 8.1. *Under the assumptions of Theorem* 5.3 *and the above notation, we have*

$$E\int_0^T |\chi_{[0,\tau)}(t)L_t|^2\,dt < \infty.$$

Before starting the proof, let us remark that according to the convention given at the beginning of this section, we have defined in the above

$$\chi_{[0,\tau)}(t)L_t = 0 \quad \text{as } t \geq \tau.$$

Similar other terms will also appear in the following expositions, and we will not repeat this kind of statement.

Proof of Lemma 8.1. From Theorem 2.4, it follows that P^k is uniformly bounded with respect to (t, ω, k) and is nonnegative. Then it follows from Theorem 5.1 that

$$E\int_0^T |L_t(k)|^2\,dt \leq \beta, \tag{8.9}$$

where β is a positive constant independent of k. Since $\tau_k \uparrow \tau$ a.s., we have

$$|L_t(k)|^2 = \chi_{[0,\tau_k]}(t)|L_t|^2 \to \chi_{[0,\tau)}(t)|L_t|^2, \quad \text{a.s.a.e.}$$

Using Fatou's lemma, we further obtain

$$E\int_0^T |\chi_{[0,\tau)}(t)L_t|^2\,dt \leq \varliminf_{k\to\infty} E\int_0^T |L_t(k)|^2\,dt \leq \beta. \tag{8.10}$$

This completes the proof. □

LEMMA 8.2. *Under the assumptions of Theorem* 5.3 *and the above notation, we have*

$$\lim_{k\to\infty}\int_0^T \sum_{i=1}^d L_t^i(k)\,dW_t^i = \int_0^T \sum_{i=1}^d \chi_{[0,\tau)}(t)L_t^i\,dW_t^i, \quad \textit{a.s.} \tag{8.11}$$

Proof of Lemma 8.2. It suffices to prove that the quadratic variation of the difference of the two stochastic integrals a.s. converges to zero, i.e., to show

$$\lim_{k\to\infty}\int_0^T |L_t(k) - \chi_{[0,\tau)}(t)L_t|^2\,dt = 0, \quad \text{a.s.}$$

Since $\tau_k \uparrow \tau$ a.s., we have

$$L_t(k) - \chi_{[0,\tau)}(t)L_t = -\chi_{[\tau_k,\tau)}(t)L_t \to 0, \quad \text{a.s.a.e.}$$

While

$$|L_t(k) - \chi_{[0,\tau)}(t)L_t|^2 = |\chi_{[\tau_k,\tau)}(t)L_t|^2 \leq |\chi_{[0,\tau)}(t)L_t|^2,$$

the last term of which is by Lemma 8.1 a.s. square integrable on $[0,T]$, i.e.,

$$\int_0^T |\chi_{[0,\tau)}(t)L_t|^2\,dt < \infty, \quad \text{a.s.}$$

The desired result then follows from Lebesgue's dominated convergence theorem. □

LEMMA 8.3. *Under the assumptions of Theorem* 5.3 *and the above notation, we have*

$$\begin{aligned}&\lim_{k\to\infty}\int_0^T G(A_t(k), B_t(k), C_t(k), D_t(k); Q_t(k), N_t; P_t^k, L_t(k))\,dt\\ &= \int_0^T \chi_{[0,\tau)}(t)G(A_t, B_t, C_t, D_t; Q_t, N_t; P_t, L_t)\,dt, \quad \textit{a.s.}\end{aligned} \tag{8.12}$$

Proof of Lemma 8.3. Note that

$$\begin{aligned}&\int_0^T G(A_t(k),B_t(k),C_t(k),D_t(k);Q_t(k),N_t;P_t^k,L_t(k))\,dt\\&=\int_0^{T\wedge\tau_k}\chi_{[0,\tau)}(t)G(A_t,B_t,C_t,D_t;Q_t,N_t;P_t,L_t)\,dt.\end{aligned}\tag{8.13}$$

Since P is uniformly bounded and $\chi_{[0,\tau)}L$ is square integrable with respect to (t,ω), the underlying integrand is integrable on $[0,T]\times\Omega$. Hence the desired result then follows.

Note that $P_0^k=Y_0$ for $k=1,2,\ldots$. From Lemmas 8.2 and 8.3, Theorem 2.4, and BSRDE (8.8), the following statement is immediate.

LEMMA 8.4. *Under the assumptions of Theorem* 5.3 *and the above notation,* P_T^k *a.s. converges. The limit* P_T^∞ *is uniformly bounded. Moreover,*

$$Y_{T\wedge\tau}=P_T^\infty X_{T\wedge\tau}.\tag{8.14}$$

We can write

$$\begin{aligned}P_t^k&=Y_0-\int_0^{t\wedge\tau_k}\chi_{[0,\tau)}(s)G(A_s,B_s,C_s,D_s;Q_s,N_s;P_s,L_s)\,ds\\&\quad+\int_0^{t\wedge\tau_k}\chi_{[0,\tau)}(s)\sum_{i=1}^d L_s^i\,dW_s^i.\end{aligned}\tag{8.15}$$

Proceeding identically as before, we can show the following lemma.

LEMMA 8.5. *Under the assumptions of Theorem* 5.3 *and the above notation,* P_t^k *strongly converges to* P_t^∞*, which is defined as follows:*

$$\begin{aligned}P_t^\infty&:=Y_0-\int_0^{t\wedge\tau}\chi_{[0,\tau)}(s)G(A_s,B_s,C_s,D_s;Q_s,N_s;P_s,L_s)\,ds\\&\quad+\int_0^{t\wedge\tau}\chi_{[0,\tau)}(s)\sum_{i=1}^d L_s^i\,dW_s^i\\&=Y_0-\int_0^t\chi_{[0,\tau)}(s)G(A_s,B_s,C_s,D_s;Q_s,N_s;P_s,L_s)\,ds\\&\quad+\int_0^t\chi_{[0,\tau)}(s)\sum_{i=1}^d L_s^i\,dW_s^i.\end{aligned}\tag{8.16}$$

Step 3. We can pass to the limit in BSRDE (8.8). This shows that $(P^\infty,\chi_{[0,\tau)}L)$ satisfies the following BSRDE:

$$\begin{cases}d\widetilde{P}_t=-\Big\{A_t(\infty)'\widetilde{P}_t+\widetilde{P}_tA_t(\infty)+Q_t(\infty)+\sum_{i=1}^d C_t^i(\infty)'\widetilde{P}_tC_t^i(\infty)\\\qquad+\sum_{i=1}^d[C_t^i(\infty)'\widetilde{L}_t^i+\widetilde{L}_t^iC_t^i(\infty)]+G_1(B_t(\infty),C_t(\infty),D_t(\infty),N_t;\widetilde{P}_t,\widetilde{L}_t)\Big\}\,dt\\\qquad+\sum_{i=1}^d\widetilde{L}_t^i\,dW_t^i,\qquad 0\le t\le T,\\\widetilde{P}_T=P_T^\infty,\qquad \widetilde{L}_t:=(\widetilde{L}_t^1,\ldots,\widetilde{L}_t^d).\end{cases}\tag{8.17}$$

Here,

$$\begin{aligned}
&A_t(\infty) := \chi_{[0,\tau]}(t)A_t, \qquad B_t(\infty) := \chi_{[0,\tau]}(t)B_t,\\
&C_t^i(\infty) := \chi_{[0,\tau]}(t)C_t^i, \qquad D_t^i(\infty) := \chi_{[0,\tau]}(t)D_t^i, \quad i = 1, \dots, d,\\
&C_t(\infty) := (C_t^1(\infty), \dots, C_t^d(\infty)), \qquad D_t(\infty) := (D_t^1(\infty), \dots, D_t^d(\infty)),\\
&Q_t(\infty) := \chi_{[0,\tau]}(t)Q_t, \qquad U_t(\infty) := \chi_{[0,\tau]}(t)U_t.
\end{aligned} \tag{8.18}$$

Note that $\{(X_{t\wedge\tau}, Y_{t\wedge\tau}, \chi_{[0,\tau)}(t)Z_t), 0 \le t \le T\}$ satisfies the following:

$$\begin{cases}
d\widetilde{X}_t = [A_t(\infty)\widetilde{X}_t + B_t(\infty)U_t(\infty)]\,dt + \sum_{i=1}^{d}[C_t^i(\infty)\widetilde{X}_t + D_t^i(\infty)U_t(\infty)]\,dW_t^i,\\
-d\widetilde{Y}_t = \left[A_t(\infty)'\widetilde{Y}_t + Q_t(\infty)\widetilde{X}_t + \sum_{i=1}^{d} C_t^i(\infty)'\widetilde{Z}_t^i\right]dt - \sum_{i=1}^{d}\widetilde{Z}_t^i\,dW_t^i,\\
\widetilde{X}_0 = I_{n\times n}, \qquad \widetilde{Y}_T = P_T^\infty \widetilde{X}_T,\\
0 = N_tU_t(\infty) + B_t(\infty)'\widetilde{Y}_t + \sum_{i=1}^{d} D_t^i(\infty)'\widetilde{Z}_t^i.
\end{cases} \tag{8.19}$$

It follows from Theorem 5.2(i) that $X_{T\wedge\tau}$ a.s. has an inverse. On the other hand, by definition of stopping time τ and the trajectory-continuity of process X, we see that $X_{T\wedge\tau}$ is degenerate on $\{\tau < \infty\}$. Therefore, to avoid a contradiction, it is necessary that

$$P(\{\tau < \infty\}) = 0.$$

Therefore, X a.s. has an inverse at each $t \in [0,T]$, $P^\infty = YX^{-1}$, $\chi_{[0,\tau)}L = L$, and BSRDE (8.17) coincides with BSRDE (3.1). The proof is complete. □

9. Concluding comments. The results of this paper can be adapted to the singular case (N is allowed to be only nonnegative) but with suitable additional conditions such as the following:

(A4) Assume that the matrix process $\sum_{i=1}^{d}(D^i)'D^i$ and the terminal state weighting random matrix M are uniformly positive.

This subject will be detailed elsewhere.

The singular case has received much recent interests because of its appearance in financial mean-variance problems. More generally, N can also be possibly negative—this is the so-called indefinite case. On these features, the interested reader is referred to Chen and Yong [5], Kohlmann and Tang [9, 12], Yong and Zhou [24], and the references therein.

Finally, the quadratic optimal control problem of linear stochastic evolution system with random coefficients can also be discussed. The details will be presented elsewhere.

Acknowledgments. The author would like to thank Prof. Michael Kohlmann for his help and relevant discussions as well as his hospitality during the author's stay in Konstanz. This work was reported in April 2001 at the School of Mathematics and System Science, University of Shandong, Jinan 250100. The author is grateful to Prof. Peng for his kind invitation and helpful comments. Last but not least, the author would like to thank the associate editor and the two referees for their helpful comments.

REFERENCES

[1] R. Bellman, *Functional equations in the theory of dynamic programming, positivity and quasilinearity*, Proc. Natl. Acad. Sci. USA, 41 (1955), pp. 743–746.

[2] A. Bensoussan, *Stochastic Control by Functional Analysis Methods*, North–Holland, Amsterdam, 1982.

[3] J.-M. Bismut, *Linear quadratic optimal stochastic control with random coefficients*, SIAM J. Control Optim., 14 (1976), pp. 419–444.

[4] J. M. Bismut, *Controle des systems lineares quadratiques: applications de l'integrale stochastique*, in Séminaire de Probabilités XII, Lecture Notes in Math. 649, C. Dellacherie, P. A. Meyer, and M. Weil, eds., Springer-Verlag, Berlin, 1978, pp. 180–264.

[5] S. Chen and J. Yong, *Stochastic linear quadratic optimal control problems*, Appl. Math. Optim., 43 (2001), pp. 21–45.

[6] N. El Karoui, S. Peng, and M. C. Quenez, *Backward stochastic differential equations in finance*, Math. Finance, 7 (1997), pp. 1–71.

[7] L. I. Gal'chuk, *Existence and uniqueness of a solution for stochastic equations with respect to semimartingales*, Theory Probab. Appl., 23 (1978), pp. 751–763.

[8] M. Kobylanski, *Résultats d'existence et d'unicité pour des équations différentielles stochastiques rétrogrades avec des générateurs á croissance quadratique*, C. R. Acad. Sci. Paris Sér I. Math., 324 (1997), pp. 81–86.

[9] M. Kohlmann and S. Tang, *Minimization of risk and LQ theory*, SIAM J. Control Optim., submitted.

[10] M. Kohlmann and S. Tang, *Global adapted solution of one-dimensional backward stochastic Riccati equations, with application to the mean-variance hedging*, Stochastic Process. Appl., 97 (2002), pp. 255–288.

[11] M. Kohlmann and S. Tang, *Multidimensional backward stochastic Riccati equations and applications*, SIAM J. Control Optim., 41 (2003), pp. 1696–1721.

[12] M. Kohlmann and S. Tang, *New developments in backward stochastic Riccati equations and their applications*, in Mathematical Finance (Konstanz, 2000), M. Kohlmann and S. Tang, eds., Birkhäuser, Basel, 2001, pp. 194–214.

[13] J. P. Lepeltier and J. San Martin, *Backward stochastic differential equations with continuous coefficient*, Statist. Probab. Lett., 32 (1997), pp. 425–430.

[14] J. P. Lepeltier and J. San Martin, *Existence for BSDE with superlinear-quadratic coefficient*, Stochastics Stochastics Rep., 63 (1998), pp. 227–240.

[15] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1971.

[16] E. Pardoux and S. Peng, *Adapted solution of backward stochastic equation*, Systems Control Lett., 14 (1990), pp. 55–61.

[17] E. Pardoux and S. Peng, *Backward stochastic differential equations and quasi-linear parabolic partial differential equations*, in Stochastic Partial Differential Equations and Their Applications, Lecture Notes in Control and Inform. Sci. 176, B. L. Rozovskii and R. S. Sowers, eds., Springer, Berlin, Heidelberg, New York 1992, pp. 200–217.

[18] S. Peng, *Stochastic Hamilton–Jacobi–Bellman equations*, SIAM J. Control Optim., 30 (1992), pp. 284–304.

[19] S. Peng, *Open problems on backward stochastic differential equations*, in Control of Distributed Parameter and Stochastic Systems (Hangzhou, 1998), S. Chen, et al., eds., Kluwer Academic Publishers, Boston, 1999, pp. 265–273.

[20] S. Peng, *A general stochastic maximum principle for optimal control problems*, SIAM J. Control Optim., 28 (1990), pp. 966–979.

[21] S. Peng and Z. Wu, *Fully coupled forward-backward stochastic differential equations and applications to optimal control*, SIAM J. Control Optim., 37 (1999), pp. 825–843.

[22] S. Tang and X. Li, *Necessary conditions for optimal control of stochastic systems with random jumps*, SIAM J. Control Optim., 32 (1994), pp. 1447–1475.

[23] W. M. Wonham, *On a matrix Riccati equation of stochastic control*, SIAM J. Control, 6 (1968), pp. 681–697.

[24] J. Yong and X. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer-Verlag, Berlin, New York, 1999.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 76–97

STABILIZATION OF LPV SYSTEMS: STATE FEEDBACK, STATE ESTIMATION, AND DUALITY*

FRANCO BLANCHINI† AND STEFANO MIANI‡

Abstract. In this paper we consider the problem of stabilizing linear parameter varying (LPV) systems by means of gain scheduling control. This technique amounts to designing a controller which is able to update its parameters on-line according to the variations of the plant parameters. We first consider the state feedback case and show a design procedure based on the construction of a Lyapunov function for discrete-time LPV systems in which the parameter variations are affine and occur in the state matrix only. This procedure produces a nonlinear static controller. We show that, different from the robust stabilization case, we can always derive a linear controller, that is, nonlinear controllers cannot outperform linear ones for the gain scheduling problem. Then we show that this procedure has a dual version which leads to the construction of a linear gain scheduling observer. The two procedures may be combined to derive an observer-based linear gain scheduling compensator.

Key words. LPV systems, gain scheduling, nonquadratic Lyapunov functions, robust stabilization

AMS subject classifications. 93D15, 93C55, 93D05

PII. S0363012900372283

1. Introduction.

The control of linear parameter varying (LPV) systems is a problem which is encountered in several applications in the industrial world. In many cases the parameter variations cannot be measured on-line and in this case the robust control approach is appropriate. The problem is different in the case in which the parameters can be measured on-line. In this case the compensator may take advantage of this knowledge and improve its performance [5].

The gain scheduling approach is classical and is often used in industrial applications. However, only quite recently these techniques, which have been heuristically applied, became a subject of mathematical investigation. In [29], [30] a rigorous stability analysis of some general gain-scheduled schemes is provided. A pole placement-like technique for gain scheduling synthesis is proposed in [28]. In [17] it is considered the problem of designing a nonlinear controller whose linearizations in several operating points match the linear controllers designed for these points. A linear matrix inequality (LMI) technique for the gain scheduled control design is proposed in [27]. A μ-analysis approach is proposed in [16]. A more recent technique based on a set-theoretic approach is presented in [31] for discrete-time LPV systems with bounded variations. The reader is referred to [26] for a tutorial exposition.

Although the knowledge of the plant parameters is an advantage for the compensator, an interesting exception has been investigated in the literature which concerns the state feedback case. Indeed, for the class of control-affine nonlinear continuous-time systems in which the term associated with the control is certain [22] or the so-called convex processes [10], the knowledge of the parameter is not an advantage

*Received by the editors May 15, 2000; accepted for publication (in revised form) August 8, 2002; published electronically March 19, 2003. This work was supported by MURST, Italy.
http://www.siam.org/journals/sicon/42-1/37228.html

†Dipartimento di Matematica ed Informatica, Università di Udine, via delle Scienze 208, 33100 Udine, Italy (blanchini@uniud.it).

‡Dipartimento di Ingegneria Elettrica, Gestionale e Meccanica, Università di Udine, via delle Scienze 208, 33100 Udine, Italy (miani.stefano@uniud.it).

for the compensator as far as it concerns its stabilization capability. As a particular case, gain scheduling design for state feedback LPV systems can be handled without restriction as a robust design, with the remarkable advantage that no parameter measurement is needed.

In the discrete-time case this property does not hold. As we will see, there are trivial examples of LPV discrete-time systems that can be stabilized via gain scheduling controllers, but no stabilizing controllers which ignore the parameters exist. This fact motivated us to derive a procedure to design a gain scheduling control in which the exploitation of the parameter measurement is allowed. In particular, we consider a procedure for the robust stabilization case [7], [12], which is in principle unsuitable for the gain scheduling design, and we show that it still helps if applied to a proper subsystem evidenced by a suitable transformation.

Furthermore, we show that in the discrete-time gain scheduling case nonlinear compensators cannot outperform linear ones. This is in contrast to the robust stabilization case in which nonlinear controllers can outperform linear ones [11] (see also [21] for further implications). Thus limiting the attention to linear compensators is not a restriction for gain scheduling state feedback stabilization. (This result holds for continuous-time systems as well [10].)

Then we consider the dual problem of designing a gain scheduling state observer based only on the knowledge of the current output and parameter measurements of the type considered in [20], [25], [32]. We show that necessary and sufficient conditions for the existence of such a linear observer are the dual of those for the existence of a gain scheduling state feedback compensator (linear or not). We show how this kind of duality does not hold for the robust state estimation problem, i.e., when the parameters are not available to the observer, and by means of a simple example we show that the class of linear gain scheduling observers proposed does not parameterize the class of gain scheduling observers.

Finally we show that the two procedures for state feedback stabilization and state detection can be combined together to solve the problem of stabilization by means of a linear observer-based compensator.

2. Problem statement and basic results. In this paper we consider LPV systems of the form

$$\begin{aligned} x(k+1) &= A(w(k))x(k) + Bu(k), \\ y(k) &= Cx(k), \end{aligned} \tag{1}$$

where $x(k) \in \mathbb{R}^n$ is the state, $u(k) \in \mathbb{R}^q$ is the control input, $y(k) \in \mathbb{R}^p$ is the control output, and $w(k)$ is a time-varying parameter. We assume that the state matrix $A(w(k))$ is constrained to belong to the matrix polytope

$$A(w(k)) = \sum_{h=1}^{m} A^{(h)} w_h(k), \tag{2}$$

where, for every k,

$$w(k) \in \mathcal{W} = \left\{ w = [w_1, \ldots, w_m]^T : w_h \geq 0, h = 1, \ldots, m, \sum_{h=1}^{m} w_h = 1 \right\}, \tag{3}$$

and $A^{(h)}$, $h = 1, 2, \ldots, m$, are assigned constant $n \times n$ matrices. A function $w(\cdot) : \mathbb{N}_o \to \mathbb{R}^m$ will be said *admissible* if $\mathrm{im}(w) \subseteq \mathcal{W}$.

In the following we will work under the assumption below.

Assumption 2.1. Matrices B and C have full column and row rank, respectively.

The basic problem considered in this paper is the stabilization of system (1) by means of a state observer and an estimated state feedback compensator which are scheduled on the parameter $w(k)$.

DEFINITION 2.1. *The system* (1) *is gain scheduling state feedback (GSSF) stabilizable if there exists a (possibly dynamic) continuous state feedback compensator whose equations are a function of the time-varying parameter* $w(k)$*,*

$$\begin{aligned} z(k+1) &= F(z(k), x(k), w(k)), \\ u(k) &= G(z(k), x(k), w(k)), \end{aligned} \tag{4}$$

such that the resulting closed-loop system is globally uniformly asymptotically stable for every admissible function $w(\cdot)$*.*

The next definition is essentially the dual.

DEFINITION 2.2. *The system* (1) *is gain scheduling detectable (GSD) if there exists a (possibly dynamic) system whose equations are a function of the time–varying parameter* $w(k)$*,*

$$\begin{aligned} z(k+1) &= F(z(k), y(k), u(k), w(k)), \\ \hat{x}(k) &= G(z(k), y(k), u(k), w(k)), \end{aligned} \tag{5}$$

and such that for all $x(0)$*,* $z(0)$*,* $w(\cdot)$ *the condition* $e(k) \doteq \hat{x}(k) - x(k) \to 0$ *as* $k \to \infty$ *is assured for every admissible function* $w(\cdot)$*.*

With obvious meaning, we will say that (1) is *robustly stabilizable* (RS) and *robustly detectable* (RD) if (4) and (5) do not depend on the parameter w. Furthermore, we will distinguish the case in which $F(\cdot)$ and $G(\cdot)$ in (4) and (5), for fixed $w(k) \in \mathcal{W}$, are linear with respect to $x(k)$ and $z(k)$ (respectively, $y(k)$, $u(k)$, and $z(k)$).

2.1. Robust stabilization and state detection. One of the main points of the paper is to show that there is no apparent duality between the problem of robust detection and robust state feedback stabilization while, in turn, a kind of duality relationship exists between the gain scheduling stabilization and detection problems.

Let us consider the following system:

$$\begin{aligned} x(k+1) &= A(w(k))x(k) + Bu(k), \\ y(k) &= x(k); \end{aligned} \tag{6}$$

let us define its dual as follows:

$$\begin{aligned} x(k+1) &= A^T(w(k))x(k) + u(k), \\ y(k) &= B^T x(k). \end{aligned} \tag{7}$$

As is well known, if A is a constant matrix, we have a duality relation which says that (6) is stabilizable if and only if (7) is detectable. In our case such a relation does not exist as long as the parameter w is unknown, as can be shown by the next counterexample. Consider the system

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 1+w(k) & 0 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix},$$
$$y(k) = x_2(k)$$

with $|w(k)| \le \bar{w}$. Such a system is such that there is no observer which can estimate asymptotically $x_1(k)$ ignoring $w(k)$. Indeed, in the second state equation $x_1(k)$ enters

in the term $\xi \doteq (1+w(k))x_1(k)$. Although $\xi(k) = x_2(k+1)$ can be determined with one step delay, it is impossible to estimate $x_1(k)$ from $\xi(k)$, if $\bar{w} > 0$, up to the fact it belongs to the uncertainty interval $\xi/(1+\bar{w}) \le x_1 \le \xi/(1-\bar{w})$, whose size grows arbitrarily if $x_1(0) \neq 0$. Conversely, its dual system

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \begin{bmatrix} 2 & 1+w(k) \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(k)$$

with the control $u(k) = -4x_1(k) - 2x_2(k)$ is stable provided that $\bar{w}$ is sufficiently small. Thus, roughly, robust stabilizability does not imply the robust detectability of the dual.

It can be easily shown that robust detectability does not imply the robust stabilizability of the dual as well. To this aim it is sufficient to consider the following simple example with $A(w(k)) = w(k)$, $B = 1$, and $C = 1$ (say the system and its dual coincide):

$$x(k+1) = w(k)x(k) + u(k), \tag{8}$$

$$y(k) = x(k) \tag{9}$$

with $|w(k)| \le 2$, which cannot be robustly stabilized by any state feedback, but is obviously detectable. Note in passing that the last example shows that robust state feedback stabilization and the gain scheduling state feedback stabilization are different problems for discrete-time systems. Indeed, the system can be gain scheduling stabilized (e.g., by $u(k) = -w(k)x(k)$), but not robustly stabilized. This fact was pointed out in [10] and will motivate the results of the next section in which we will present a procedure for the GSSF stabilization by means of the procedures already available for robust control synthesis [7], [12].

2.2. Some definitions and preliminary results. In this section we recall some basic results concerning the stability of LPV systems. We denote by C-set a convex and compact set containing the origin as an interior point. We say that a set $\mathcal{P}$ is 0-symmetric if $x \in \mathcal{P}$ implies $-x \in \mathcal{P}$. We denote by $\text{int}\{\mathcal{P}\}$ the interior of $\mathcal{P}$ and for any real $\lambda > 0$; we denote by $\lambda\mathcal{P}$ the scaled set $\lambda\mathcal{P} = \{x : \frac{x}{\lambda} \in \mathcal{P}\}$. Given $x \in \mathbb{R}^n$ the one norm and the infinity norm are defined as

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \qquad \|x\|_\infty = \max_i |x_i|,$$

respectively. The corresponding induced norms for matrices are

$$\|H\|_1 = \sup_{x\neq 0} \frac{\|Hx\|_1}{\|x\|_1} = \sup_j \sum_{i=1}^{n} |H_{ij}|, \qquad \|H\|_\infty = \sup_{x\neq 0} \frac{\|Hx\|_\infty}{\|x\|_\infty} = \sup_i \sum_{j=1}^{n} |H_{ij}|.$$

As it is known, the two norms above are dual in $\mathbb{R}^n$ [19].

Definition 2.3 (see [18]). *The set $\mathcal{P} \in \mathbb{R}^n$ obtained as the convex hull of finitely many vectors*

$$\mathcal{P} = conv\{x_1, \dots, x_r\} = \left\{ x : x = \sum_{i=1}^{r} \alpha_i x_i, \sum_{i=1}^{r} \alpha_i = 1,\ \alpha_i \ge 0, x_i \in \mathbb{R}^n, i = 1, \dots, r \right\}$$

is called a (convex) polytope. We will denote by $vert\{\mathcal{P}\}$ the minimal set of vectors $\{x_1, \dots, x_l\}$, $l \le r$, such that $conv\{x_1, \dots, x_l\} = conv\{x_1, \dots, x_r\}$. A vector $x \in vert\{\mathcal{P}\}$ is called a vertex of $\mathcal{P}$.

DEFINITION 2.4 (see [18]). *A set $\mathcal{P} \in \mathbb{R}^n$ is polyhedral if it is the intersection of a finite number of closed half-spaces:*

$$\mathcal{P} = \{x : f_i x \leq g_i,\ f_i^T \in \mathbb{R}^n,\ \ g_i \in \mathbb{R},\ i = 1, \dots, s\}. \tag{10}$$

If we restrict our attention to polyhedral 0-symmetric C-sets, the following proposition holds.

PROPOSITION 2.5. *Any 0-symmetric polyhedral C-set $\mathcal{P} \in \mathbb{R}^n$ is a polytope and can be represented in the form*

$$\mathcal{P} = \{x :\ \|Fx\|_\infty \leq 1\}, \tag{11}$$

where $F \in \mathbb{R}^{s \times n}$ is a full column rank matrix, or in the dual form

$$\mathcal{P} = \{x = X\alpha,\ \ \alpha \in \mathbb{R}^l,\ \|\alpha\|_1 \leq 1\}, \tag{12}$$

where $X \in \mathbb{R}^{n \times l}$ is a full row rank matrix.

Proof. The proof that a bounded polyhedral set is a polytope can be found in [18, Theorem 20.9]. As far as the representation (11), by Definition 2.4, in view of the 0-symmetry of $\mathcal{P}$, $\mathcal{P}$ can be represented by inequalities of the form

$$|f_i x| \leq g_i,\ \ g_i \geq 0,\ \ \text{for } i = 1, \dots, s. \tag{13}$$

Now since $\mathcal{P}$ is a C-set, $0 \in \text{int}\{\mathcal{P}\}$, thus $g_i > 0$ for every i. Dividing each of the (13) by g_i we get $|F_i x| \leq 1$, $i = 1, \dots, s$, with $F_i = \frac{f_i}{g_i}$, which is equivalent to (11). To conclude, if F were not full column rank, then for every arbitrarily large $k > 0$ there would exist x, $\|x\| > k$, such that $Fx = 0$, say $x \in \mathcal{P}$ and $\|x\| > k$, in contrast with the boundedness of $\mathcal{P}$.

Now note that, if $\mathcal{P}$ is 0-symmetric, the vertices of $\mathcal{P}$ can be stacked columnwise into a matrix $X_{symm} = [x_1, \dots x_l, -x_1, \dots, -x_l] = [X, -X] \subset \mathbb{R}^{n \times 2*l}$. Then the corresponding convex hull is formed by all the vectors

$$x = \sum_{i=1}^{l} \beta_i x_i - \gamma_i x_i = \sum_{i=1}^{l} (\beta_i - \gamma_i) x_i \tag{14}$$

with $\beta_i,\ \gamma_i \geq 0$, and $\sum_{i=1}^{l} (\beta_i + \gamma_i) = 1$. The vector α whose components are $\alpha_i = \beta_i - \gamma_i$ is such that $\|\alpha\|_1 \leq 1$, which leads to (12). The argument can be reversed. If x is as in (12), then define $\beta_i = \max\{\alpha_i, 0\}$ and $\gamma_i = \min\{\alpha_i, 0\}$; we get expression (14), with $\sum_{i=1}^{l} (\beta_i + \gamma_i) \leq 1$. This means that x is in the convex hull of the columns of X_{symm}. □

We will say that a function $\Psi(x) :\ \mathbb{R}^n \to \mathbb{R}^+$ is *polyhedral* if it is the Minkowski functional of a polyhedral 0-symmetric C-set $\mathcal{P}$. Such a function can be represented either by means of the "plane" representation (11),

$$\Psi(x) = \|Fx\|_\infty, \tag{15}$$

or by means of the dual "vertex" representation

$$\Psi(x) = \inf\{\|p\|_1,\ \ \text{s.t.}\ \ x = Xp\}. \tag{16}$$

In this paper we consider polyhedral functions as candidate Lyapunov functions for asserting the stability of the closed-loop system, according to the next definitions.

DEFINITION 2.6. *The polyhedral function $\Psi(x)$ is a Lyapunov function for the dynamic system $x(k+1) = f(x(k), w(k))$, $w(k) \in \mathcal{W}$, if for every admissible function $w(\cdot)$ the following condition holds:*

$$\Psi(x(k+1)) \leq \lambda \Psi(x(k)),$$

for some nonnegative $\lambda < 1$ and every solution $x(k)$.

DEFINITION 2.7. *A dynamic system $x(k+1) = f(x(k), w(k))$, $w(k) \in \mathcal{W}$, is robustly stable if it is globally uniformly asymptotically stable for every admissible function $w(\cdot)$.*

The main point in considering polyhedral functions is that they are described by a finite number of parameters and have been shown to be nonconservative for proving the robust stability for the class of systems considered here, say a system in the considered class is robustly stable if and only if it admits a polyhedral Lyapunov function. (Note that the same property does not hold for quadratic Lyapunov functions.) More precisely we have the following lemmas.

LEMMA 2.8 (see [12]). *The system*

$$x(k+1) = A(w(k))x(k), \tag{17}$$

where $A(w)$ is as in (2), is robustly stable if and only if there exists an integer $l \geq n$ and a full row rank matrix $X \in \mathbb{R}^{n \times l}$ and m matrices $P^{(h)} \in \mathbb{R}^{l \times l}$ such that, for every $h = 1, \ldots, m$, $\|P^{(h)}\|_1 \leq \lambda < 1$ and

$$A^{(h)} X = X P^{(h)}.$$

LEMMA 2.9 (see [2], [3], [4], [5]). *The system (17), where $A(w)$ is as in (2), is robustly stable if and only if there exists an integer $s \geq n$ and a full column rank matrix $F \in \mathbb{R}^{s \times n}$ and m matrices $H^{(h)} \in \mathbb{R}^{s \times s}$ such that, for every $h = 1, \ldots, m$, $\|H^{(h)}\|_\infty \leq \lambda < 1$ and*

$$F A^{(h)} = H^{(h)} F.$$

The coefficient λ above turns out to be the same index of the speed of convergence of Definition 2.6. Lemma 2.9 rephrases the results in [2], [3], [4], [5], and the proof follows immediately by the fact that if the system is stable, then it admits a polyhedral Lyapunov function (see also [14], [23], [24]). Note that the matrices F and X in the above lemmas have the meaning of (15) and (16), namely, they characterize a polyhedral function. Note also that the two lemmas are dual in the sense that by transposing all the matrices Lemma 2.8 reduces to Lemma 2.9 for the dual system

$$x(k+1) = A^T(w(k))x(k)$$

and vice versa. This also means that (17) is robustly stable if and only if the dual system is robustly stable.

3. Solution of the gain scheduling state feedback stabilization problem. In this section we consider the problem of determining a state feedback control of the gain scheduling type for the system (1). To this aim we introduce the following important definition of exponential stabilizability.

DEFINITION 3.1. *We say that the system (1) is* exponentially *GSSF stabilizable if it is GSSF stabilizable and there exists two constants $0 \leq \lambda < 1$ and $C \geq 0$ such that*

$$\|x_{cl}(k)\| \leq C \lambda^k \|x_{cl}(0)\|,$$

where x_{cl} is the overall closed-loop system state. The constant λ is referred to as the speed of convergence.

We will see that any stabilizable system of the considered class can indeed be exponentially stabilized.

3.1. Problem solvability. To solve our problem we consider the class of polyhedral functions as candidate Lyapunov functions. The next theorem motivates this choice by showing that the existence of one such Lyapunov function is not only sufficient for the gain scheduling stabilizability of system (1), but it is also necessary as well.

THEOREM 3.2. *The following statements are equivalent:*

(i) *System* (1) *is GSSF stabilizable.*

(ii) *System* (1) *is exponentially GSSF stabilizable.*

(iii) *There exists a static control law $u(k) = \Phi(x(k), w(k))$ and a polyhedral function* (16) *which is a Lyapunov function for system* (1).

(iv) *There exist a full row rank matrix $X \in \mathbb{R}^{n\times l}$, m matrices $P^{(h)} \in \mathbb{R}^{l\times l}$ and $U^{(h)} \in \mathbb{R}^{q\times l}$, such that for every $h = 1, \dots, m$*

$$A^{(h)}X + BU^{(h)} = XP^{(h)}, \quad \textit{with} \ \ \|P^{(h)}\|_1 < 1. \tag{18}$$

Proof. We skip the obvious implications (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i) and relegate to the appendix the proof that (i) implies (iv).

(iv) $\Rightarrow$ (iii) Assume that (18) holds and let

$$\lambda = \max_{h=1,\dots,m} \|P^{(h)}\|_1 < 1. \tag{19}$$

Consider the function (16) associated with X:

$$\Psi(x) = \inf\{\|p\|_1, \ \ \text{s.t.} \ \ x = Xp\}.$$

Since the infimum is actually a minimum, for every x there exists $\hat{p}$ such that $x = X\hat{p}$ and $\|\hat{p}\|_1 = \Psi(x)$. Let

$$u(x, w) = \sum_{h=1}^{m} \ w_h \ \ U^{(h)}\hat{p}. \tag{20}$$

For all possible $w = [w_1, \dots, w_m]^T \in \mathcal{W}$ we have that

$$\begin{aligned} A(w)x + Bu &= \sum_{h=1}^{m} \ w_h \ A^{(h)}x + \sum_{h=1}^{m} \ w_h \ BU^{(h)}\hat{p} = \sum_{h=1}^{m} \ w_h \ [A^{(h)}X + BU^{(h)}]\hat{p} \\ &= \sum_{h=1}^{m} \ w_h \ XP^{(h)}\hat{p} = X\sum_{h=1}^{m} \ w_h \ P^{(h)}\hat{p} = X\hat{p}'(w). \end{aligned}$$

The vector $\hat{p}'(w) = \sum_{h=1}^{m} \ w_h P^{(h)}\hat{p}$ is a convex combination of the vectors $P^{(h)}\hat{p}$ whose norms are such that $\|P^{(h)}\hat{p}\|_1 \le \lambda\|\hat{p}\|_1$; thus $\|\hat{p}'(w)\|_1 \le \lambda\|\hat{p}\|_1$. This means that for all $w \in \mathcal{W}$

$$\Psi(A(w)x + Bu) = \inf\{\|p\|_1, \ \ \text{s.t.} \ \ [A(w)x + Bu] = Xp\} \le \lambda\|\hat{p}\|_1,$$

and then

$$\Psi(A(w)x + Bu) \le \lambda\Psi(x).$$

Therefore Ψ is a Lyapunov function for the closed-loop system. □

The previous theorem states that (18) is crucial, since the existence of a solution in terms of X, $P^{(h)}$, and $U^{(h)}$ is necessary and sufficient for the GSSF stabilization problem to be solvable. The next theorem states that as long as the parameter w is known to the controller, we can always implement a linear controller for the system, namely, GSSF stabilizability implies GSSF stabilizability by means of a linear controller. This result extends that in [10] to the discrete-time case. To introduce the theorem we augment (18) as follows:

$$\begin{bmatrix} A^{(h)} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} X \\ Z \end{bmatrix} + \begin{bmatrix} B & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} U^{(h)} \\ V^{(h)} \end{bmatrix} = \begin{bmatrix} X \\ Z \end{bmatrix} P^{(h)} \tag{21}$$

(we remind the reader that $\|P^{(h)}\|_1 < 1$), where Z is an arbitrary matrix such that $\left[\begin{smallmatrix} X \\ Z \end{smallmatrix}\right]$ is square invertible, and $V^{(h)} \doteq ZP^{(h)}$. We are now able to state the following theorem, which states that gain scheduling stabilizability is equivalent to gain scheduling stabilizability via linear control.

THEOREM 3.3. *If system* (1) *is GSSF stabilizable, then it is GSSF stabilizable* via linear control. *A stabilizing control is given by*

$$\begin{aligned} u(k) &= K(w)x(k) + H(w)z(k), \\ z(k+1) &= G(w)x(k) + F(w)z(k), \end{aligned} \tag{22}$$

where

$$\begin{bmatrix} K(w) & H(w) \\ G(w) & F(w) \end{bmatrix} = \sum_{h=1}^{m} \begin{bmatrix} K^{(h)} & H^{(h)} \\ G^{(h)} & F^{(h)} \end{bmatrix} w_h \tag{23}$$

with

$$\begin{bmatrix} K^{(h)} & H^{(h)} \\ G^{(h)} & F^{(h)} \end{bmatrix} = \begin{bmatrix} U^{(h)} \\ V^{(h)} \end{bmatrix} \begin{bmatrix} X \\ Z \end{bmatrix}^{-1}. \tag{24}$$

Proof. Given system (1) with the control (22)–(24), the closed-loop system is

$$\begin{bmatrix} x(k+1) \\ z(k+1) \end{bmatrix} = A_{aug}(w) \begin{bmatrix} x(k) \\ z(k) \end{bmatrix}, \tag{25}$$

where

$$\begin{aligned} A_{aug}(w) &= \begin{bmatrix} A(w) + BK(w) & BH(w) \\ G(w) & F(w) \end{bmatrix} \\ &= \begin{bmatrix} \sum_{h=1}^m A^{(h)} w_h + B \sum_{h=1}^m K^{(h)} w_h & B \sum_{h=1}^m H^{(h)} w_h \\ \sum_{h=1}^m G^{(h)} w_h & \sum_{h=1}^m F^{(h)} w_h \end{bmatrix} \qquad (26) \\ &= \sum_{h=1}^{m} w_h \begin{bmatrix} A^{(h)} + BK^{(h)} & BH^{(h)} \\ G^{(h)} & F^{(h)} \end{bmatrix}, \text{ with } \sum_{h=1}^{m} w_h = 1, \quad w_h \geq 0. \qquad (27) \end{aligned}$$

By construction we have that

$$\begin{bmatrix} A^{(h)} + BK^{(h)} & BH^{(h)} \\ G^{(h)} & F^{(h)} \end{bmatrix} \begin{bmatrix} X \\ Z \end{bmatrix} = \begin{bmatrix} X \\ Z \end{bmatrix} P^{(h)}; \tag{28}$$

therefore, according to Lemma 2.8, the system in (25) is robustly stable and the proof is completed. □

3.2. Computation of the Lyapunov function. The results above are nonconstructive as long as we cannot provide algorithms to compute the matrices X, $P^{(h)}$, and $U^{(h)}$ in (18). As observed in [9], such types of equations are not convenient to solve the problem, since they are bilinear as long as X and $P^{(h)}$ are both unknown. Note also that the same problem holds in the robust stabilization case, where we have to cope with the same equation with the difference that [9], [12] the matrices $U^{(h)}$ are all equal according to the following proposition.

PROPOSITION 3.4 (see [7], [12], [13], [15]). *The system*

$$x(k+1) = A(w(k))x(k) + B(w(k))u(k), \tag{29}$$

where $A(w(k))$ *is as in* (2) *and*

$$B(w(k)) = \sum_{h=1}^{m} B^{(h)} w_h(k) \tag{30}$$

and $B^{(h)} \in \mathbb{R}^{n\times q}$, $h = 1, 2, \ldots, m$, *are assigned constant matrices, is robustly state feedback stabilizable if and only if there exists a full row rank matrix* $X \in \mathbb{R}^{n\times l}$, m *matrices* $P^{(h)}$, *and a* single *matrix* $U \in \mathbb{R}^{q\times l}$ *such that, for every* $h = 1, \ldots, m$,

$$A^{(h)}X + B^{(h)}U = XP^{(h)}, \quad \textit{with} \ \ \|P^{(h)}\|_1 < 1. \tag{31}$$

In [7], [12] an iterative procedure is proposed to compute X, $P^{(h)}$, and U. Clearly this procedure might be applied in a conservative way to our problem since if a system is robustly stabilizable, then it is also GSSF stabilizable. However, for discrete-time systems this is a conservative way of proceeding since, as we have seen, the knowledge of w can be an advantage for the compensator. The next result will allow us to exploit the existing procedures for the solution of the robust stabilization problem for GSSF stabilizability.

In view of Assumption 2.1, we consider the system in the following form:

$$\hat{A}(w) = \begin{bmatrix} \hat{A}_{11}(w) & \hat{A}_{12}(w) \\ \hat{A}_{21}(w) & \hat{A}_{22}(w) \end{bmatrix} = \sum_{h=1}^{m} w_i \begin{bmatrix} \hat{A}_{11}^{(h)} & \hat{A}_{12}^{(h)} \\ \hat{A}_{21}^{(h)} & \hat{A}_{22}^{(h)} \end{bmatrix} \tag{32}$$

and

$$\hat{B} = \begin{bmatrix} 0 \\ I \end{bmatrix}. \tag{33}$$

It is immediately seen that this form is achieved by applying to all the generating matrices $A^{(h)}$ the linear transformation

$$T = \begin{bmatrix} \tilde{B} & B \end{bmatrix}, \tag{34}$$

where $\tilde{B}$ is any matrix such that T is invertible, so that $\hat{A}^{(h)} = T^{-1}A^{(h)}T$ and $\hat{B} = T^{-1}B$.

The system is thus in the form

$$\hat{x}_1(k+1) = \hat{A}_{11}(w)\hat{x}_1(k) + \hat{A}_{12}(w)\hat{x}_2(k), \tag{35}$$

$$\hat{x}_2(k+1) = \hat{A}_{21}(w)\hat{x}_1(k) + \hat{A}_{22}(w)\hat{x}_2(k) + u(k). \tag{36}$$

By means of this form, we can recast the GSSF stabilization problem in a robust stabilization problem for the pair $(\hat{A}_{11}(w(k)), \hat{A}_{12}(w(k)))$ according to the following theorem.

THEOREM 3.5. *System* (1) *is GSSF stabilizable if and only if the system*

$$\hat{x}_1(k+1) = \hat{A}_{11}(w)\hat{x}_1(k) + \hat{A}_{12}(w)\hat{x}_2(k) \tag{37}$$

(where $\hat{x}_2$ now has to be thought of as an input signal) is robustly stabilizable.

Proof. If. Assume that (37) is robustly stabilizable. Then by Proposition 3.4 there exist a full row rank matrix $\hat{X}_1$, a matrix $\hat{X}_2$, and m matrices $\|P_1^{(h)}\|_1 < 1$ such that, for every $h = 1, \ldots, m$,

$$\hat{A}_{11}^{(h)}\hat{X}_1 + \hat{A}_{12}^{(h)}\hat{X}_2 = \hat{X}_1 P_1^{(h)}. \tag{38}$$

Now define the elements $U^{(h)}$ as the matrices that match the equalities

$$\hat{A}_{21}^{(h)}\hat{X}_1 + \hat{A}_{22}^{(h)}\hat{X}_2 + U^{(h)} = \hat{X}_2 P_1^{(h)}.$$

Combining the two equations we get

$$\begin{bmatrix} \hat{A}_{11}^{(h)} & \hat{A}_{12}^{(h)} \\ \hat{A}_{21}^{(h)} & \hat{A}_{22}^{(h)} \end{bmatrix} \begin{bmatrix} \hat{X}_1 \\ \hat{X}_2 \end{bmatrix} + \begin{bmatrix} 0 \\ I \end{bmatrix} U^{(h)} = \begin{bmatrix} \hat{X}_1 \\ \hat{X}_2 \end{bmatrix} P_1^{(h)}. \tag{39}$$

Assume now, for the moment being, that the matrix

$$\hat{X} \doteq \begin{bmatrix} \hat{X}_1 \\ \hat{X}_2 \end{bmatrix} \tag{40}$$

has full row rank. Then, by Theorem 3.2, (39) implies that the system (1) is GSSF stabilizable. If the matrix $\hat{X}$ in (40) has not full row rank, then we can augment the equation as follows:

$$\begin{aligned} \begin{bmatrix} \hat{A}_{11}^{(h)} & \hat{A}_{12}^{(h)} \\ \hat{A}_{21}^{(h)} & \hat{A}_{22}^{(h)} \end{bmatrix} \begin{bmatrix} \hat{X}_1 & 0 \\ \hat{X}_2 & \epsilon I \end{bmatrix} + \begin{bmatrix} 0 \\ I \end{bmatrix} \begin{bmatrix} U^{(h)} & \epsilon\tilde{U}^{(h)} \end{bmatrix} \\ = \begin{bmatrix} \hat{X}_1 & 0 \\ \hat{X}_2 & \epsilon I \end{bmatrix} \begin{bmatrix} P^{(h)} & \epsilon R^{(h)} \\ 0 & 0 \end{bmatrix}, \end{aligned} \tag{41}$$

where $R^{(h)}$, $h = 1, \ldots, m$, are matrices such that[1] $\hat{X}_1 R^{(h)} = \hat{A}_{12}^{(h)}$, $\tilde{U}^{(h)} \doteq \hat{X}_2 R^{(h)} - \hat{A}_{22}^{(h)}$, and $\epsilon > 0$. The matrix

$$\hat{X}_{aug} = \begin{bmatrix} \hat{X}_1 & 0 \\ \hat{X}_2 & \epsilon I \end{bmatrix}$$

now has full row rank and the matrices

$$P_{aug}^{(h)} = \begin{bmatrix} P^{(h)} & \epsilon R^{(h)} \\ 0 & 0 \end{bmatrix}$$

have 1-norm $\|P_{aug}^{(h)}\|_1 = \|P^{(h)}\|_1$, provided that ϵ is sufficiently small. Then (41) implies that (1) is GSSF stabilizable.

[1] We remind the reader that $\hat{X}_1$ has full row rank.

Only if. If (1) is GSSF stabilizable, then, by implication of (i)→ (iv) of Theorem 3.2, there exists a full row rank matrix X, m matrices $P^{(h)}$ and $U^{(h)}$ satisfying (18). By choosing the transformation matrix T as in (34), we get (39) with the matrix in (40) of full row rank. By selecting the first block, we get (38), with $\hat{X}_1$ of full row rank. In view of Proposition 3.4, this in turn implies that (37) is RS. □

Thus the algorithm to solve the gain scheduling stabilization problem turns out to be the following procedure.

PROCEDURE 3.1.

1. *Given the matrices $A^{(h)}$ and B, take $\tilde{B}$ such that $T = [\tilde{B}\ B]$ is invertible and compute the form* (35)–(36) *by means of the state transformation $T = [\tilde{B}\ B]$.*
2. *Compute the matrices $\hat{X}_1$ and $\hat{X}_2$ in* (38) *by computing a polyhedral Lyapunov function for the robust stabilization of the subsystem* (37) *and let $\hat{X}$ be defined as in* (40). *(In the event that $\hat{X}$ is not full row rank, apply the augmentation as in* (41).*)*
3. *Apply the reverse transformation to* (39) *and let $X \doteq T\hat{X}$ to achieve*
$$A^{(h)}X + BU^{(h)} = XP^{(h)}.$$
4. *Compute the linear dynamic control as in* (22)–(23).

4. Solution of the gain scheduling state estimation problem. In the previous section we have considered the problem of designing a state feedback gain scheduling compensator. Among the results, we have seen that if a system is gain scheduling stabilizable, then it is gain scheduling stabilizable via linear state feedback control. Now we cope with the dual problem of designing a linear state observer for the system

$$\begin{aligned} x(k+1) &= A(w(k))x(k) + v(k), \\ y(k) &= Cx(k). \end{aligned} \tag{42}$$

The candidate linear observer we consider is of the form

$$\begin{aligned} z(k+1) &= P(w(k))z(k) - L(w(k))y(k) + T_1(w(k))v(k), \\ \hat{x}(k) &= Q(w(k))z(k) + R(w(k))y(k), \end{aligned} \tag{43}$$

with $z(k) \in \mathbb{R}^s$.

DEFINITION 4.1. *The system* (43) *is a linear gain scheduled asymptotic observer for system* (42) *if it has the following properties:*

(i) Convergence. *For all $w(k)$, $v(k)$, and $x(0)$ and $z(0)$ we have $\hat{x}(k) - x(k) \to 0$ as $k \to \infty$.*

(ii) Initialization. *If $x(0) = 0$, $z(0) = 0$, then $\hat{x}(k) = x(k)$ for all $w(k) \in \mathcal{W}$ and $v(k)$.*

(iii) Internal stability. *If $x(k) \to 0$ as $k \to \infty$, then $z(k) \to 0$.*

It is known that, for a given *constant* $\bar{w}$, the system (43) represents the most general form of a linear observer [25]. Furthermore, for a given constant $\bar{w}$, for (43) to be an observer the following necessary and sufficient conditions must hold:

$$P(\bar{w})T_1(\bar{w}) - T_1(\bar{w})A(\bar{w}) = L(\bar{w})C, \tag{44}$$

$$Q(\bar{w})T_1(\bar{w}) + R(\bar{w})C = I, \tag{45}$$

and $P(\bar{w})$ is a stable matrix (i.e., its eigenvalues are inside the open unit disk).

The main problem now is to see what happens when $w(k)$ is time-varying as in our case. The next lemma leads to a simplification of the structure of (43), namely, that there is no restriction in considering observers of the form

$$\begin{aligned} z(k+1) &= P(w(k))z(k) - L(w(k))y(k) + T_1 v(k), \\ \hat{x}(k) &= Q(w(k))z(k) + R(w(k))y(k), \end{aligned} \tag{46}$$

with T_1 *constant and of full column rank.*

LEMMA 4.2. *Assume that there exists an observer of the form* (43). *Then there exists an observer with the same structure where the matrix* $T_1(w) = T_1$ *is constant and of full column rank.*

Proof. See the appendix. ∎

Define now the error variable $r(k)$ as

$$r(k) = z(k) - T_1 x(k), \tag{47}$$

whose associated error equation, derived from (44)–(46), is

$$\begin{array}{rcl} r(k+1) & = & P(w(k))r(k), \\ e(k) \doteq \hat{x}(k) - x(k) & = & Q(w(k))r(k). \end{array} \tag{48}$$

The previous equation leads us immediately to the following basic result.

LEMMA 4.3. *The system* (46), *with* T_1 *of full column rank, is an observer for* (42) *only if the time-varying system*

$$r(k+1) = P(w(k))r(k)$$

is robustly stable and T_1 *is such that, for all* $w \in \mathcal{W}$,

$$\begin{array}{rcl} P(w)T_1 - T_1 A(w) & = & L(w)C, \\ Q(w)T_1 + R(w)C & = & I. \end{array} \tag{49}$$

The proof of the above Lemma is immediate and thus it is omitted. We are in the position now to prove the main result of this section.

THEOREM 4.4. *The following statements are equivalent.*

(i) *There exists a linear gain scheduled observer for* (42) *of the form* (46) *with* T_1 *full column rank.*

(ii) *There exist a full column rank matrix* F, m *matrices* $H^{(h)}$, *and* m *matrices* $Y^{(h)}$ *such that, for every* $h = 1, \ldots, m$, *the dual equation of* (18) *holds:*

$$FA^{(h)} + Y^{(h)}C = H^{(h)}F, \quad \textit{with} \quad \|H^{(h)}\|_\infty < 1, \tag{50}$$

(iii) *The dual system*

$$x(k+1) = A^T(w(k))x(k) + C^T u(k)$$

is gain scheduling stabilizable.

Proof. (ii) $\Longleftrightarrow$ (iii) The equivalence comes from the duality between (50) and (18). Indeed, if (18) is satisfied, we achieve (50) by transposition by setting $F = X^T$, $Y^{(h)} = {U^{(h)}}^T$, $H^{(h)} = {P^{(h)}}^T$. Note that $\|H^{(h)}\|_\infty = \|P^{(h)}\|_1$, being the ∞-norm equal to the 1-norm of the transpose.

(ii) $\Rightarrow$ (i) Consider the dynamic observer (46) with

$$P(w(k)) = \sum_{h=1}^{m} w_h(k) H^{(h)}, \tag{51}$$

$$L(w(k)) = \sum_{h=1}^{m} w_h(k) Y^{(h)}, \tag{52}$$

$$T_1 = F, \tag{53}$$

$$Q(w(k)) = F^\dagger, \tag{54}$$

$$R(w(k)) = 0, \tag{55}$$

where $F^\dagger$ is any left inverse of F (which exists since F is full column rank). With the usual change of variable $r(k) = z(k) - Fx(k)$ the overall system dynamics (system + observer) becomes

$$r(k+1) = \sum_{h=1}^{m} w_h(k) H^{(h)} r(k), \tag{56}$$

$$e(k) = \hat{x}(k) - x(k) = F^\dagger r(k), \tag{57}$$

which is stable since $\| \sum_{h=1}^{m} w_h H^{(h)} \|_\infty < 1$ and satisfies the requirements in Definition 4.1.

(i) $\Rightarrow$ (ii) The first equation in (48) represents a stable system. A stable system always admits a polyhedral Lyapunov function. In particular, by Lemma 2.9 there exist m matrices $H^{(h)}$ such that $\|H^{(h)}\|_\infty \le \lambda < 1$ and a full column rank matrix $\hat{T}$ such that

$$\hat{T} P^{(h)} = H^{(h)} \hat{T}, \tag{58}$$

with $P^{(h)} = P(w^{(h)})$, where $w^{(h)}$ are the vertices of the polytope $\mathcal{W}$, namely,

$$w^{(h)} = [0 \ \dots 0 \ 1 \ 0 \dots 0]^T.$$

Define in a similar way $L^{(h)} = L(w^{(h)})$ to get

$$P^{(h)} T_1 - T_1 A^{(h)} = L^{(h)} C.$$

Then by multiplying by $\hat{T}$ on the left we get, in view of (58),

$$H^{(h)} \underbrace{\hat{T} T_1}_{F} = \underbrace{\hat{T} T_1}_{F} A^{(h)} + \underbrace{\hat{T} L^{(h)}}_{Y^{(h)}} C. \tag{59}$$

Note that $F = \hat{T} T_1$ is of full column rank since both $\hat{T}$ and T_1 are such; thus the last equation is exactly (50). □

The previous result is constructive, since, by duality, we can always apply the procedure to design a GSSF stabilizer to the dual system $(A^T(w), C^T)$ to design the observer.

It is important to remark that the proposed class of observers does not parameterize the whole class of observers as in Definition 2.2. For instance, it can be shown that the system

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \begin{bmatrix} w(k) & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix},$$

$$y(k) = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix}$$

with $|w(k)| \le 2$ is *not* gain scheduling observable by means of an observer of the form (43) though the following *nonlinear* system,

$$\begin{aligned} z(k+1) &= w(k), \\ \hat{x}_1(k) &= z(k) y(k), \\ \hat{x}_2(k) &= y(k), \end{aligned}$$

is an observer. The trouble is that, in general, computing this kind of nonlinear observer can be very hard. It could be done by means of the observability matrix $\mathcal{O}(k)$ associated with the last n steps (obtaining expressions which are very involved) or by means of standard Kalman filtering [1], [25]. Unfortunately, for our problem, these solutions have the following problem: one has to impose some *detectability conditions* [1] *for all possible sequences* $w(k) \in \mathcal{W}$. To the best of the authors' knowledge, this problem has not been solved and seems to be very difficult.

To recap, we report the following table which shows the relations between the GSSF stabilization (via linear control), = GSSF(L), and the gain scheduling estimation (via linear observer), = GSE(L):

$$\begin{array}{ccc} \text{GSSF} & \Longleftrightarrow & \text{GSSFL} \\ & & \Uparrow \\ & & \text{dual conditions.} \\ & & \Downarrow \\ \text{GSE} & \Longleftarrow & \text{GSEL} \end{array}$$

5. A separation principle for design. We briefly describe a way to stabilize system (1) by means of a linear observer and a gain scheduling estimated state feedback. The next theorem is a simple consequence of the results of the previous sections. It shows that one can always synthesize a stabilizing compensator by separately designing an observer and a state feedback control if the system is GSSF stabilizable and GSD via linear observer.

THEOREM 5.1. *Assume that* $(A(w), B)$ *is GSSF stabilizable and* $(C, A(w))$ *is GSD via linear observer. Then the dynamic controller (we dropped the time-dependence in* $w(k)$ *for clarity)*

$$\begin{array}{rcl} z_c(k+1) &=& G(w)\hat{x}(k) + F(w)z_c(k), \\ z_o(k+1) &=& P(w)z_o(k) - L(w)y(k) + T_1 B u(k), \\ \hat{x}(k) &=& Q(w)z_o(k) + R(w)y(k), \\ u(k) &=& K(w)\hat{x}(k) + H(w)z_c(k), \end{array} \tag{60}$$

where the matrices in the above equations are as in (22)–(24) *in Theorem* 3.3 *and* (51)–(55) *in Theorem* 4.4*, asymptotically stabilizes the plant.*

Proof. By combining (60) with the system dynamics, setting $r(k) = z_o(k) - T_1 x(k)$, and recalling (48), after some algebra the closed-loop system can be written as

$$\begin{bmatrix} x(k+1) \\ z_c(k+1) \\ r(k+1) \end{bmatrix} = \begin{bmatrix} A(w) + BK(w) & BH(w) & BK(w)Q(w) \\ G(w) & F(w) & G(w)Q(w) \\ 0 & 0 & P(w) \end{bmatrix} \begin{bmatrix} x(k) \\ z_c(k) \\ r(k) \end{bmatrix}.$$

The block triangular structure implies closed-loop stability if the blocks on the diagonal are stable. The latter stability condition can be always guaranteed according to the results of sections 3 and 4. Indeed, the first block $(= A_{aug}(w))$ satisfies (28), while the second $(= P(w))$ given by (51) can be taken stable because of the assumed linear detectability. □

6. Example. As an example we consider the discrete-time system described by the vertex matrices

$$A^{(1)} = \begin{bmatrix} 1 & .25 & 0 \\ .25 & 1 & -.2 \\ 0 & 0 & -.16 \end{bmatrix}, \qquad A^{(2)} = \begin{bmatrix} 1 & .25 & 0 \\ .25 & 1 & -.05 \\ 0 & 0 & .16 \end{bmatrix},$$

$$A^{(3)} = \begin{bmatrix} 1 & .32 & 0 \\ .32 & 1 & -.05 \\ 0 & 0 & -.16 \end{bmatrix}, \qquad A^{(4)} = \begin{bmatrix} 1 & .32 & 0 \\ .32 & 1 & -.2 \\ 0 & 0 & .16 \end{bmatrix}$$

and

$$B = \begin{bmatrix} 0 \\ 0 \\ .3 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}.$$

As far as the gain scheduling stabilization problem is concerned, we applied Procedure 3.1 with $\lambda_c = .96$, which led to the matrices

$$X = \begin{bmatrix} 1 & 1 & .0017 \\ -1 & -.3666 & 1 \\ 0 & 4.17 & 9.75 \end{bmatrix},$$

$$P^{(1)} = \begin{bmatrix} .75 & .908 & .253 \\ 0 & 0 & 0 \\ 0 & -.042 & -.698 \end{bmatrix}, \qquad P^{(2)} = \begin{bmatrix} .75 & 0 & 0 \\ 0 & .908 & .251 \\ 0 & .008 & .605 \end{bmatrix},$$

$$P^{(3)} = \begin{bmatrix} .68 & 0 & 0 \\ 0 & .883 & .321 \\ 0 & .068 & .63 \end{bmatrix}, \qquad P^{(4)} = \begin{bmatrix} .68 & .880 & .323 \\ 0 & .003 & 0 \\ 0 & 0 & -.628 \end{bmatrix},$$

and

$$U^{(1)} = \begin{bmatrix} 0 & .8587 & -17.4958 \end{bmatrix}, \qquad U^{(2)} = \begin{bmatrix} 0 & 10.6599 & 17.9611 \end{bmatrix},$$
$$U^{(3)} = \begin{bmatrix} 0 & 16.7125 & 30.1538 \end{bmatrix}, \qquad U^{(4)} = \begin{bmatrix} 0 & -2.1828 & -25.6258 \end{bmatrix}.$$

Since X is square, the compensator (22)–(23) reduces to a static compensator with $K^{(h)} = U^{(h)} X^{-1}$:

$$\begin{bmatrix} K^{(1)} \\ K^{(2)} \\ K^{(3)} \\ K^{(4)} \end{bmatrix} = \begin{bmatrix} 40.6617 & 40.6617 & -5.9683 \\ 14.5357 & 14.5357 & 0.3486 \\ 18.6315 & 18.6315 & 1.1778 \\ 42.7795 & 42.7795 & -7.0191 \end{bmatrix}.$$

As far as the state estimation is concerned, we determined the matrices of the estimator by determining a gain scheduling controller for the dual system with convergence speed $\lambda_o = .3$. The matrices derived for the observer were

$$F = \begin{bmatrix} 0 & 0 & 1.0000 \\ -3.2400 & .9334 & -.0666 \\ -3.2800 & .9248 & .0752 \end{bmatrix},$$

$$H^{(1)} = \begin{bmatrix} -.1600 & 0 & 0 \\ -.1672 & .1322 & 0 \\ -.1895 & .1123 & 0 \end{bmatrix}, \qquad H^{(2)} = \begin{bmatrix} .1600 & 0 & 0 \\ -.0485 & .1322 & 0 \\ -.0267 & .1123 & 0 \end{bmatrix},$$

$$H^{(3)} = \begin{bmatrix} -.1600 & 0 & 0 \\ -.0276 & 0 & -.1118 \\ -.0481 & 0 & -.1349 \end{bmatrix}, \qquad H^{(4)} = \begin{bmatrix} .1600 & 0 & 0 \\ -.1889 & 0 & -.1118 \\ -.1628 & 0 & -.1349 \end{bmatrix},$$

and

$$\begin{bmatrix} Y^{(1)} & Y^{(2)} & Y^{(3)} & Y^{(4)} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 2.5783 & 2.5783 & 3.3080 & 3.3080 \\ 2.6850 & 2.6850 & 3.4267 & 3.4267 \end{bmatrix}.$$

The resulting gain scheduling output feedback controller is hence given by

$$\begin{aligned} z_o(k+1) &= \left(\sum_{i=1}^{4} H^{(i)} w_i(k)\right) z_o(k) - \left(\sum_{i=1}^{4} Y^{(i)} w_i(k)\right) y(k) + FBu(k), \\ \hat{x}(k) &= F^{\dagger} z_o(k), \\ u(k) &= \left(\sum_{i=1}^{4} K^{(i)} w_i(k)\right) \hat{x}(k). \end{aligned}$$

The overall behavior was simulated with a random $w(\cdot)$ starting from the initial condition $x(1) = \begin{bmatrix} 1 & -2 & 1 \end{bmatrix}^T$. Figures 1 and 2 depict the state and error time evolution during the first 12 time-steps.

We let the reader note that, though at first surprising, the peak of x_3 is essentially due to the fact that x_3 is the "virtual" input for the uncertain subsystem associated with the first two components and the magnitude of the elements $(3, 2)$ in $A(w)$ can be very small.

7. Concluding discussions. In this paper we focused our attention on the stabilization of linear parameter varying (LPV) discrete-time systems. We showed by very simple examples that the robust stabilization and the gain scheduling stabilization problems, different from the continuous-time case, are rather different problems. We also showed the existence of a duality relation between the gain scheduling control and the gain scheduling linear observation problem for LPV. Finally, a separation principle to derive linear output feedback stabilizing controllers was derived.

Among the limits of the paper we remind the reader that the matrices C and B are assumed constant. If we assume that $B = \sum_{i=1}^{m} w_h \, B^{(h)}$, we have trouble with (18) since its natural extension $A^{(h)}X + B^{(h)}U^{(h)} = XP^{(h)}$, with $\|P^{(h)}\|_1 < 1$, does not even ensure stabilizability of the system. Therefore future research directions in this area are the investigation of discrete-time gain scheduling observers when the matrices B and C are also time-dependent, discrete-time robust observers, and the properties of continuous-time gain scheduling observers. In particular, as far as it concerns the continuous time case, the construction procedures proposed in this paper can be immediately used to produce continuous-time compensators by means of the Euler approximating system (see [8])

$$x(k+1) = [I + \tau A(w)]x(t) + \tau Bu$$

obtained from the continuous-time model $\dot{x} = A(w)x + Bu$. Thus the design procedure proposed in [10] for GSSF design can be complemented by the state observer design procedure proposed here.

However, some theoretical questions remain open. For instance, in [10] it is shown that, in the continuous-time case, there is no advantage for the compensator in the

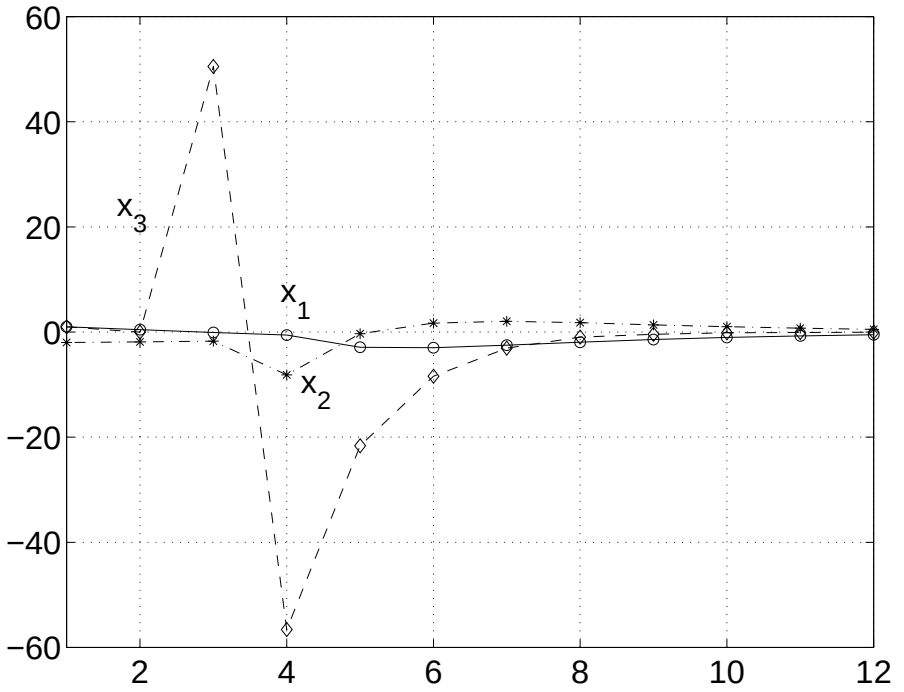


FIG. 1. *System state evolution.*

knowledge of $w(t)$ if $x(t)$ is available to the controller. It is not clear if or under which conditions this property holds in the output feedback case. Also we have shown, by means of a simple counterexample, that in the discrete-time case there can exist gain scheduling nonlinear observers for systems which do not admit linear gain scheduling observers. It is not obvious how to generate these examples in the continuous-time case (if it is possible at all). We think that these questions may be interesting subjects of investigation.

Appendix A. Proof of (i) $\Rightarrow$ (iv) in Theorem 3.2. Before starting with the proof, we present its main idea. If there exists a continuous stabilizing control, it is obviously stabilizing if we consider only the extremal sequences $w(k)$ with values in the vertices of $\mathcal{W}$, namely, if we assume $w(k) \in vert\{\mathcal{W}\}$ for every $k \geq 0$. If we take as initial conditions the vertices of a polytope including the origin as an interior point, and we consider the convex hull of all possible trajectories of the state (associated with extremal sequences) emanated by these vertices, we achieve a new polytope in the augmented state space. The key point is that this polytope as well as its projection on the plant state space are controlled-invariant, namely, they can be rendered positively invariant by a suitable controller (which is not necessarily the one whose existence is assumed). From such a region we can obtain a symmetric one by considering the convex hull of the union of it and its opposite. By introducing a perturbation parameter, we prove the existence of a solution for the considered equation.

Proof. Assume that there exists a continuous gain scheduling state feedback compensator of the form (4) and let $\mathcal{P}_0$ be an arbitrary polytope in the $x - z$ space

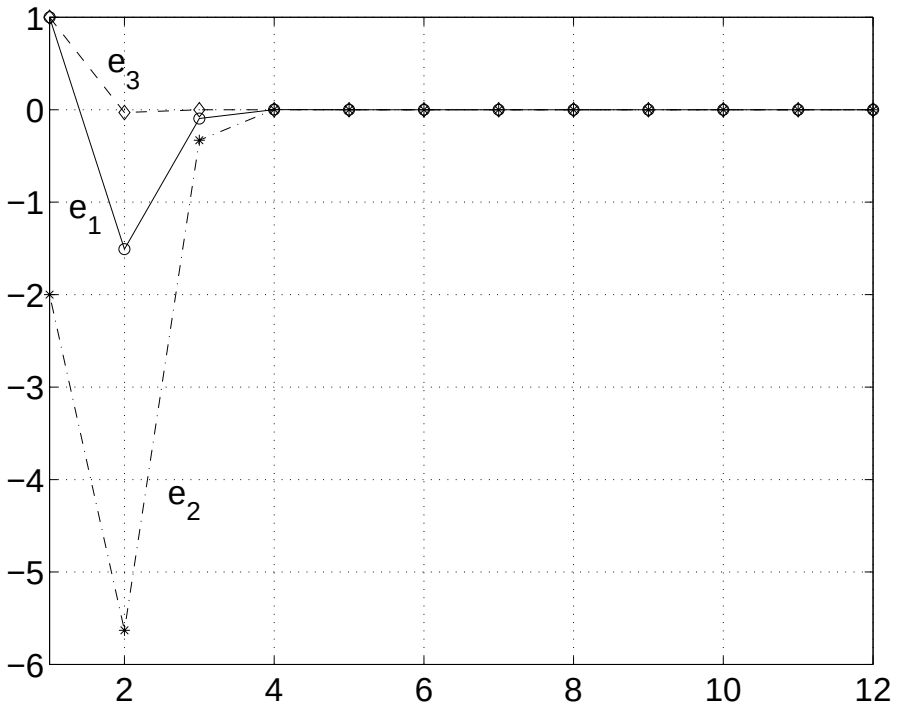


FIG. 2. *System error evolution.*

including the origin in its interior. The closed-loop system is

$$\begin{array}{rcl} x(k+1) & = & \lambda^{-1}[A(w(k))x(k) + Bu(k)], \\ z(k+1) & = & \lambda^{-1}[F(z(k), x(k), w(k))], \\ u(k) & = & G(z(k), x(k), w(k)), \end{array} \tag{61}$$

where λ, currently set to $\lambda = 1$, is a fictitious parameter whose significance will be soon explained. Let

$$Reach(\mathcal{P}_0, \lambda, k)$$

be the *discrete* set formed by all the possible states of the system (61) which can be reached in k steps, for all possible initial conditions taken on the vertices of $\mathcal{P}_0$, namely, $[x(0)^T \;\; z(0)^T]^T \in vert\{\mathcal{P}_0\}$, and when we have restricted

$$w(k) \in vert\{\mathcal{W}\}.$$

In view of the assumed stability there exists $\bar{k}$ such that

$$Reach(\mathcal{P}_0, 1, \bar{k}) \subset \text{int}\{\mathcal{P}_0\}.$$

Now, by continuity, there exists $0 \leq \lambda < 1$ such that

$$Reach(\mathcal{P}_0, \lambda, \bar{k}) \subset \text{int}\{\mathcal{P}_0\}.$$

Denote by $\mathcal{P}$ the polytope which is the convex hull of all the extremal trajectories

$$\mathcal{P} \doteq conv\left\{\bigcup_{k=0}^{\bar{k}} Reach(\mathcal{P}_0, \lambda, k)\right\}.$$

By construction for every vertex $\left[\begin{array}{cc} x^T & z^T \end{array}\right]^T$ of $\mathcal{P}$ and every vertex w of $\mathcal{W}$, there exists $u = G(z, x, w)$ such that

$$\left[\begin{array}{c} \lambda^{-1}[A(w)x + Bu] \\ \lambda^{-1}F(z, x, w) \end{array}\right] \in \mathcal{P} \Longrightarrow \left[\begin{array}{c} x' \\ z' \end{array}\right] = \left[\begin{array}{c} A(w)x + Bu \\ F(z, x, w) \end{array}\right] \in \lambda\mathcal{P}. \tag{62}$$

If we denote by $\mathcal{X}$ the projection of $\mathcal{P}$ on the plant state space (thus $\lambda\mathcal{X}$ is projection of $\lambda\mathcal{P}$), then, since the vertices of $\mathcal{X}$ are the projections of some of the vertices of $\mathcal{P}$, for each vertex x of $\mathcal{X}$ and $w \in vert\{\mathcal{W}\}$ there exist u such that

$$A(w)x + Bu \in \lambda\mathcal{X}, \tag{63}$$

being one such u of the control $u = G(z, x, w)$ associated with the vertex $[x^T\ z^T]^T$, whose projection is x. If we consider the opposite set, $-\mathcal{X}$, we obtain the opposite relation

$$A(w)(-x) + B(-u) \in \lambda(-\mathcal{X}).$$

Therefore, if we consider the 0-*symmetric* polytope $\tilde{\mathcal{X}} \doteq conv\{-\mathcal{X} \bigcup \mathcal{X}\}$ (whose vertices $vert\{\tilde{X}\} = \{x_1, \ldots, x_l, -x_1, \ldots, -x_l\}$ form a subset of $vert\{\mathcal{X}\} \bigcup vert\{-\mathcal{X}\}$ and can be stacked columnwise into the matrix $[X, -X] = [x_1, \ldots, x_l, -x_1, \ldots, -x_l]$), we have that, for each $w^{(h)} \in vert\{\mathcal{W}\}$,

$$A^{(h)}x_j + Bu_j^{(h)} \in \lambda\tilde{\mathcal{X}}, \qquad j = 1, \ldots, l.$$

Denoting by $\Psi(\cdot)$ the Minkowski functional of $\tilde{\mathcal{X}}$, this in turn means that $\Psi(A^{(h)}x_j + Bu_j^{(h)}) \leq \lambda$. In view of (16), this can be written as

$$A^{(h)}x_j + Bu_j^{(h)} = Xp_j^{(h)},$$

where $\|p_j^{(h)}\|_1 \leq \lambda$. Grouping together these equations, with $U^{(h)} = [u_1^{(h)} \ldots u_l^{(h)}]$ and $P^{(h)} = [p_1^{(h)} \ldots p_l^{(h)}]$, we get

$$A^{(h)}X + BU^{(h)} = XP^{(h)},$$

which concludes the proof. □

Appendix B. Proof of Lemma 4.2. To prove the lemma we recall the notion of invariant subspace for a family of matrices.

DEFINITION B.1. *Given a family of matrices $P(w)$, $w \in \mathcal{W}$, we say that a subspace $\mathcal{S}$ of $\mathbb{R}^n$ is $P(w)$-invariant if $x \in \mathcal{S}$ implies $P(w)x \in \mathcal{S}$ for all $w \in \mathcal{W}$. Furthermore we call the kernel of $P(w)$ the set of all vectors r such that $P(w)r = 0$ for all $w \in \mathcal{W}$.*

We first show that if an observer of the form (43) exists, there is an "equivalent" one satisfying the following assumption.

Assumption B.1. There are no $P(w)$-invariant subspaces included in the kernel of $Q(w)$.

Note that for known P and Q this is just an observability assumption. In our case this assumption means that

$$z(k+1) = P(w(k))z(k), \tag{64}$$

$$p(k) = Q(w(k))z(k) \tag{65}$$

cannot be reduced by the transformation $[S\ \tilde{S}]$, where S is a basis of $\mathcal{S}$, the invariant subspace, and $\tilde{S}$ is a complement matrix to the form

$$\hat{P}(w(k)) = \begin{bmatrix} \hat{P}_{11}(w(k)) & \hat{P}_{12}(w(k)) \\ 0 & \hat{P}_{22}(w(k)) \end{bmatrix}, \qquad \hat{Q}(w(k)) = \begin{bmatrix} 0 & \hat{Q}_2(w(k)) \end{bmatrix}.$$

Now, if the decomposition above can be achieved, then we can eliminate the "nonobservable" part of the system and consider an observer of the form (43) with $P(w(k)) = \hat{P}_{22}(w(k))$ and $Q(w(k)) = \hat{Q}_2(w(k))$ that fulfills the assumption. With this we are ready to show that, under such an assumption, T_1 must be constant.

Proof. Consider the variable $r(k) = z(k) - T_1(w(k))x(k)$. With simple computations, if we replace $r(k)$ in (43) in view of (44) which, as we have seen, must hold for each $w \in \mathcal{W}$, we get the following equation:

$$\begin{array}{rcl} r(k+1) & = & P(w(k))r(k) + [T_1(w(k)) - T_1(w(k+1))]x(k+1), \\ \hat{x}(k) - x(k) & = & Q(w(k))r(k). \end{array} \tag{66}$$

We show that $[T_1(w(k+1)) - T_1(w(k))] = 0$ for all $w(k), w(k+1) \in \mathcal{W}$. By contradiction assume that there exists w_0 and w_1 such that $T_1(w_0) \neq T_1(w_1)$. Let $x(0) = 0$, $z(0) = 0$ (thus $r(0) = 0$). From (ii) in Definition 4.1 we should have

$$\hat{x}(k) - x(k) = 0 \qquad \text{for all} \quad k > 0. \tag{67}$$

Assume that $w(0) = w_0$ and $w(1) = w_1$ and take x such that $(T_1(w_0) - T_1(w_1))x = r_1 \neq 0$. Then if we take $v(0) = x$, we get $x(1) = x$ and $r(1) = r_1$. Note that $v(k)$ can always be such that $x(k) = 0$, $k \geq 2$, and we assume that this is the case. Thus we consider system (66) with zero input. Consider the set of all vectors that can be reached from r_1, namely, the set

$$\mathit{Reach}(r_1) = \{x(k) = A(w(k))A(w(k-1)) \dots A(w(1))r_1 \quad \text{for some } k \geq 1\},$$

and let $\mathcal{S}$ be the smallest subspace that includes $\mathit{Reach}(r_1)$. $\mathcal{S}$ is necessarily $A(w)$-invariant because it admits as a basis a finite set of vectors $r_i \in \mathit{Reach}(r_1)$. From (67) for all vectors $r \in \mathit{Reach}(r_1)$ we must have $Q(w)r = 0$ for all $w \in \mathcal{W}$; then $Q(w)r = 0$ for all $r \in \mathcal{S}$ in contradiction with the assumption.

The final step is to show that if there exists an observer, with constant T_1, then there exists one with T_1 full column rank. This is trivial. It is sufficient to fictitiously augment (46) as

$$\begin{array}{rcll} z(k+1) & = & P(w(k))z(k) - L(w(k))y(k) & + T_1 v(k), \\ \tilde{z}(k+1) & = & & + \tilde{T}_1 v(k), \\ \hat{x}(k) & = & Q(w(k))z(k) + R(w(k))y(k), & \end{array}$$

with $\left[\begin{smallmatrix} T_1 \\ \tilde{T}_1 \end{smallmatrix}\right]$ full column rank. ∎

Acknowledgment. The authors are grateful to the reviewers for their careful reading and their numerous comments that strongly improved the quality of the paper.

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Detectability and stabilizability of time-varying discrete-time linear systems*, SIAM J. Control Optim., 19 (1981), pp. 20–32.
[2] N. E. Barabanov, *Lyapunov indicator of discrete inclusion.* I., Automat. Remote Control, 49 (1988), pp. 152–157.
[3] N. E. Barabanov, *Lyapunov indicator of discrete inclusion.* II., Automat. Remote Control, 49 (1988), pp. 283–287.
[4] N. E. Barabanov, *Lyapunov indicator of discrete inclusion.* III., Automat. Remote Control, 49 (1988), pp. 558–565.
[5] G. Becker and A. Packard, *Robust performance of linear parametrically varying systems using parametrically-dependent linear feedback*, Systems Control Lett., 23 (1994), pp. 205–215.
[6] G. Bitsoris, *On the positive invariance of polyhedral sets for discrete-time systems*, Systems Control Lett., 11 (1988), pp. 243–248.
[7] F. Blanchini, *Ultimate boundedness control for discrete-time uncertain system via set-induced Lyapunov functions*, IEEE Trans. Automat. Control, 39 (1994), pp. 428–433.
[8] F. Blanchini, *Nonquadratic Lyapunov function for robust control*, Automatica, 31 (1995), pp. 451–461.
[9] F. Blanchini, *Set invariance in control – a survey*, Automatica, 35 (1999), pp. 1747–1767.
[10] F. Blanchini, *The gain scheduling and the robust state feedback stabilization problems*, IEEE Trans. Automat. Control, 45 (2000), pp. 2061–2070.
[11] F. Blanchini and A. Megretski, *Robust state feedback control of LTV systems: Nonlinear is better than linear*, IEEE Trans. Automat. Control, 44 (1999), pp. 802–807.
[12] F. Blanchini and S. Miani, *Piecewise-linear functions in robust control*, Robust Control via Variable Structure and Lyapunov Methods, Lecture Notes in Control and Inform. Sci. 217, F. Garofalo and L. Glielmo, eds., Springer–Verlag, London, 1996, pp. 213–240.
[13] F. Blanchini and S. Miani, *Discussion on: "(A,B)-invariance conditions of polyhedral domains for continuous–time systems" by C. E. T. Dórea and J. C. Hennet*, Eur. J. Control, 5 (1999), pp. 82–86.
[14] R. K. Brayton and C. H. Tong, *Constructive stability and asymptotic stability of dynamical systems*, IEEE Trans. Circ. Syst., 27 (1980), pp. 1121–1130.
[15] C. E. T. Dorea and J. C. Hennet, *(A, B)–invariance conditions of polyhedral domains for continuous–time systems*, Eur. J. Contr., 5 (1999), pp. 70–81.
[16] A. Helmerson, *μ synthesis and LFT gain scheduling with real uncertainties*, Int. J. Rob. Nonlin. Contr., 8 (1998), pp. 631–642.
[17] D. A. Lawrence and W. J. Rugh, *Gain scheduling dynamic controllers for a nonlinear plant*, Automatica, 31 (1995), pp. 381–390.
[18] S. R. Lay, *Convex Sets and Their Applications*, John Wiley, New York, 1982.
[19] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York-London-Sydney, 1969.
[20] D. G. Luenberger, *An introduction to observers*, IEEE Trans. Automat. Control, 16 (1971), pp. 596–602.
[21] A. Megretski, *How conservative is the circle criterion?*, in Open Problems in Mathematical System and Control Theory, Comm. Control Engrg. Ser., V. D. Blondel, E. D. Sontag, M. Vidyasagar, and J. C. Willems, eds., Springer, London, 1999, pp. 149–151.
[22] A. M. Meilakhs, *Design of stable systems subject to parametric perturbations*, Automat. Remote Control, 39 (1979), pp. 1409–1418.
[23] A. N. Michel, B. H. Nam, and V. Vittal, *Computer generated Lyapunov functions for interconnected systems: Improved results with applications to power systems*, IEEE Trans. Circ. Syst., 31 (1984), pp. 189–198.
[24] Y. Ohta, H. Imanishi, L. Gong, and H. Haneda, *Computer generated Lyapunov functions for a class of nonlinear systems*, IEEE Trans. Circ. Syst., 40 (1993), pp. 343–354.
[25] J. O'Reilly, *Observers for Linear Systems*, Academic Press, London, New York, 1983.
[26] W. J. Rugh and J. S. Shamma, *Research on gain scheduling*, Automatica, 36 (2000), pp. 1401–1425.
[27] G. Scorletti and L. El Ghaoui, *Improved LMI conditions for gain scheduling and related control problems*, Int. J. Rob. Nonlin. Contr., 8 (1992), pp. 845–877.

[28] S. M. Shahruz and S. Behtash, *Design controllers for linear parameter–varying systems by the gain scheduling technique*, J. Math. Anal. Appl., 168 (1992), pp. 195–217.

[29] J. S. Shamma and M. Athans, *Analysis of gain scheduled control for nonlinear plants*, IEEE Trans. Automat. Control, 35 (1990), pp. 898–907.

[30] J. S. Shamma and M. Athans, *Guaranteed properties of gain scheduled control for linear parameter–varying plants*, Automatica, 27 (1991), pp. 559–564.

[31] J. S. Shamma and D. Xiong, *Control of rate constrained linear parameter varying systems*, in Proceedings of the 34th IEEE Conference on Decision and Control, IEEE, Piscataway, NJ, 1995, pp. 2515–2520.

[32] C. C. Tsui, *A new design approach to unknown input observers*, IEEE Trans. Automat. Control, 41 (1996), pp. 464–468.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 98–122

OPTIMIZATION-BASED STABILIZATION OF SAMPLED-DATA NONLINEAR SYSTEMS VIA THEIR APPROXIMATE DISCRETE-TIME MODELS*

LARS GRÜNE† AND DRAGAN NEŠIĆ‡

Abstract. We present results on numerical regulator design for sampled-data nonlinear plants via their approximate discrete-time plant models. The regulator design is based on an approximate discrete-time plant model and is carried out either via an infinite horizon optimization problem or via a finite horizon with terminal cost optimization problem. In both cases, we discuss situations when the sampling period T and the integration period h used in obtaining the approximate discrete-time plant model are the same or they are independent of each other. We show that, using this approach, practical and/or semiglobal stability of the exact discrete-time model is achieved under appropriate conditions.

Key words. controller design, asymptotic controllability, stabilization, numerical methods, optimal control

AMS subject classifications. 93D15, 49N35, 65P40

PII. S036301290240258X

1. Introduction.

Stabilization of controlled systems is one of the central topics in control theory that has lead to a wealth of different stabilization techniques. An important set of stabilization methods is based on optimization techniques, such as receding horizon control (RHC) or model predictive control (MPC) (see [15, 8] and references therein). In optimization-based stabilization methods, one can compute control signals either on-line, like in MPC algorithms, or off-line, like in [9, 10, 14]. In either case, it is common to implement the controller using a computer with A/D and D/A converters (sampler and zero-order hold), which leads to the investigation of sampled-data nonlinear systems.

One of the main issues in sampled-data nonlinear control is the fact that the control designer usually cannot compute the exact discrete-time model of the plant and has to use an approximate discrete-time model when designing a stabilizing controller. The approximate model is obtained by numerically integrating the continuous-time plant dynamics over one sampling interval while keeping the control constant (if a zero-order hold is used). However, it is typically assumed in the optimization-based stabilization literature that the exact discrete-time plant model is available for controller design (see, for instance, [6, 15, 14, 13, 12, 1]). Hence there are gaps in the literature between the developed theory that is based on exact discrete-time models and the actual implementation of algorithms that invariably make use of approximate discrete-time models to compute control actions (see Example 1 in [3], section V in [6], and section IV in [14]). It is the purpose of this paper to present a careful investigation of the effects that numerical errors in approximating the model may have on the stabilization of the exact discrete-time model.

*Received by the editors February 12, 2002; accepted for publication (in revised form) September 3, 2002; published electronically March 19, 2003.
http://www.siam.org/journals/sicon/42-1/40258.html

†Mathematisches Institut, Fakultät für Mathematik und Physik, Universität Bayreuth, 95440 Bayreuth, Germany (lars.gruene@uni-bayreuth.de).

‡Department of Electrical and Electronic Engineering, The University of Melbourne, Victoria 3010, Australia (d.nesic@ee.mu.oz.au).

While it may seem that any controller that stabilizes a sufficiently "good" approximate model would always stabilize the exact model for sufficiently small values of integration and/or sampling period, this issue is much more subtle than it may appear at a first glance. Indeed, a number of counterexamples illustrating different mechanisms that cause instability of exact models when controlled by controllers that stabilize approximate models have been presented in [16, 19]. Moreover, results in [16, 19] present a set of general sufficient conditions on the continuous-time plant model, the approximate discrete-time plant model, and the designed controller that guarantee that controllers that stabilize the approximate model would also stabilize the exact model for a sufficiently small sampling and/or integration period. Moreover, backstepping results in [18] show that controller design within the framework proposed in [16] may lead to considerable performance improvement as opposed to controller design based on the continuous-time plant model that is followed by discretization of the controller (emulation design).

Results in [16, 19] present a framework for controller design via approximate discrete-time models, but they do not explain how the actual controller design can be carried out within this framework. It is the purpose of this paper to investigate several situations in which the optimization-based stabilization is done within the framework of [16, 19]. In particular, we consider the following problem:

> Suppose we are given a family of approximate discrete-time plant models
>
> $$x(k+1) = F^a_{T,h}(x(k), u(k))$$
>
> that are parameterized with the sampling period T and a modeling parameter h, which is typically the integration period of the underlying integration scheme. Given a family of cost functions $J_{T,h}$, suppose that a family of controllers
>
> $$u(k) = u^{a,*}_{T,h}(x(k))$$
>
> minimizes the given family of costs and is stabilizing for the family of approximate models. When would the same family of controllers stabilize the family of exact models
>
> $$x(k+1) = F^e_{T,h}(x(k), u(k))$$
>
> for sufficiently small values of the modeling parameter h?

We present conditions that guarantee that the family of controllers $u^{a,*}_{T,h}$ stabilizes in an appropriate sense the family of exact models for sufficiently small values of the modeling parameter. Two important situations are considered:

(i) $J_{T,h}$ is an infinite horizon cost function;
(ii) $J_{T,h}$ is a finite horizon cost function with a terminal cost.

In either case, we discuss two important subcases:

(i) T and h are independent of each other. This case is important when the sampling period T is fixed and the family of approximate models is generated via a numerical integration method with adjustable integration step h. This case usually produces better results, but the numerical computations required are more intensive (see, for instance, [14, 6]).
(ii) $T = h$ and T can be arbitrarily adjusted. This case is often used in the literature. The main motivation for using this approach is a reduced computational

burden in obtaining the approximate model, but it will be shown below that this method requires much stronger conditions than the first method (see [3]).

While our results do not cover all possible costs $J_{T,h}$ of interest, the presented proofs can be adapted to cover many other important situations. Moreover, the results we present are important in cases when the approximation of the plant model comes from a completely different mechanism than numerical integration of the plant dynamics. For example, the modeling parameter h may capture the size of the cells used in the space discretization that is usually needed in numerical calculation of the controller via optimization techniques, such as dynamic programming (see [14]). The modeling parameter h can be, in general, a vector capturing several different approximation mechanisms in obtaining the plant model, and our results can be extended to cover this important case.

The paper is organized as follows. In section 2, we present several motivating examples. Preliminaries are presented in section 3. Several results from [16, 19] that we use to prove our main results are presented in section 4. Infinite horizon and finite horizon optimization-based stabilization problems are considered, respectively, in sections 5 and 6. Conclusions are presented in the last section, and some auxiliary lemmas are stated and proved in the appendix.

2. Motivation. In this section, we present two examples for which a family of optimal control laws is designed to stabilize the family of approximate models, but the exact discrete-time model is destabilized for sufficiently fast sampling by the same family of controllers. These examples strongly motivate the results of our paper.

Example 2.1. We consider the sampled-data control of the triple integrator (this example was taken from [19])

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = x_3, \quad \dot{x}_3 = u.$$

While the exact discrete-time model of this system can be computed, we base our control algorithm on the family of Euler approximate discrete-time models in order to illustrate possible pitfalls in optimal control design based on approximate discrete-time models. The family of Euler approximate discrete-time models is

$$\begin{aligned} x_1(k+1) &= x_1(k) + Tx_2(k), \\ x_2(k+1) &= x_2(k) + Tx_3(k), \\ x_3(k+1) &= x_3(k) + Tu(k). \end{aligned} \tag{2.1}$$

Denote $x_i := x_i(0), i = 1, 2, 3$, $x := (x_1\ x_2\ x_3)^T$, and $\bar{u} := \{u(0), u(1), u(2), \ldots\}$. A minimum time dead beat controller for the Euler discrete-time model is designed to minimize the cost

$$\begin{aligned} J_T(x, \bar{u}) = x^T(3)x(3) &= (x_1 + 3Tx_2 + 3T^2x_3 + T^3u(0))^2 \\ &\quad + (x_2 + 3Tx_3 + 2T^2u(0) + T^2u(1))^2 + (x_3 + Tu(0) + Tu(1) + Tu(2))^2, \end{aligned}$$

and we obtain the optimal controller

$$u_T^*(x) = \left(-\frac{x_1}{T^3} \ -\frac{3x_2}{T^2} \ -\frac{3x_3}{T} \right). \tag{2.2}$$

The closed loop system (2.1)–(2.2) has all poles equal to zero for all $T > 0$, and hence this discrete-time Euler-based closed loop system is asymptotically stable for

all $T > 0$. On the other hand, the closed loop system consisting of the exact discrete-time model of the triple integrator and the optimal controller (2.2) has a pole at ≈ -2.644 for all $T > 0$. Hence, the optimal controller for the approximate model destabilizes the exact model for any sampling period.

Example 2.2. Consider the scalar linear system

$$\dot{x} = x + u, \tag{2.3}$$

whose exact discrete-time model is

$$x(k+1) = e^T x(k) + (e^T - 1)u(k). \tag{2.4}$$

We use the Euler model

$$x(k+1) = (1+T)x(k) + Tu(k) = F_T x(k) + G_T u(k) \tag{2.5}$$

for controller design. Consider the cost

$$J_T(x, \bar{u}) = \sum_{k=0}^{\infty} (Q_T x^2(k) + R_T u^2(k)), \tag{2.6}$$

where $Q_T = T$ and $R_T = T^3(1-T)^3$. Obviously, the instantaneous cost $Q_T x^2 + R_T u^2$ is a positive definite function of x, u for all $T \in (0, 1)$. Using [2, pp. 53–54], we can obtain the family of optimal controllers for (2.5) as

$$u_T^*(x) = \frac{G_T F_T S_T}{G_T^2 S_T + R_T}\, x, \tag{2.7}$$

where S_T is the solution of the following Ricatti equation:

$$S_T = F_T^2 \left(S_T - \frac{S_T^2 G_T^2}{G_T^2 S_T + R_T} \right) + Q_T. \tag{2.8}$$

Using the computer algebra system MAPLE, we computed the family of optimal control laws to be

$$u_T^*(x) = \left(-1 - \frac{5}{2} T^2 + O(T^3) \right) x,$$

which, for sufficiently small T, yields the stable approximate closed loop

$$x(k+1) = \left(1 - \frac{5}{2} T^3 + O(T^4) \right) x(k).$$

However, the same family of controllers yields the unstable exact closed loop

$$x(k+1) = \left(1 + \frac{1}{2} T^2 + O(T^3) \right) x(k)$$

for all sufficiently small T. Again, the family of optimal controllers for the family of approximate models is destabilizing for the family of exact models for all sufficiently small sampling periods T.

Remark 2.3. Note that the optimal controller gain in the first example is not uniformly bounded in T, and, in particular, as $T \to 0$, we have for any $x \neq 0$

that $|u_T(x)| \to \infty$. It may appear that this is the only reason why instability of the exact model occurs. However, in the second example, we have that the optimal controller gain is bounded uniformly in T, and yet instability occurs. Additional similar examples that do not use optimal control laws can be found in [19].

In both examples above, we can say that the used cost $J_T(x, \bar{u})$ is ill parameterized with T, and this causes instability of the exact closed loop. In what follows, we present conditions for well-parameterized costs that avoid problems presented in the examples.

Remark 2.4. The interpretation of the above results is as follows. One cannot first find a sufficiently "good" approximate plant model with a sufficiently small sampling and/or integration period and then assume that the optimal controller for the approximate model with respect to any given cost would stabilize the exact model. Indeed, because of the fact that we are considering parameterized systems and costs, the examples illustrate that, given an arbitrarily small sampling period (and hence an arbitrarily "good" plant model), there exists a cost function for which the controller that is optimal for the approximate model would destabilize the exact model. Hence a careful investigation of stability is needed to avoid situations presented in the examples.

3. Preliminaries. $\mathbb{R}$ and $\mathbb{N}$ denote, respectively, the sets of real and natural numbers. We also denote $\mathbb{N}_0 := \{0\} \cup \mathbb{N}$. In the Euclidean space $\mathbb{R}^n$, $\|\cdot\|$ denotes the usual Euclidean norm, and B_r and $\overline{B}_r$ denote, respectively, the open and closed ball with radius r around the origin. A continuous function $\gamma : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ is of class $\mathcal{K}$ if it is strictly increasing with $\gamma(0) = 0$; it is of class $\mathcal{K}_\infty$ if it is of class $\mathcal{K}$ and unbounded. A continuous function $\beta : \mathbb{R}_0^+ \times \mathbb{R}_0^+ \to \mathbb{R}_0^+$ is of class $\mathcal{KL}$ if it is of class $\mathcal{K}$ in the first argument and strictly decreasing to 0 in the second.

Consider a continuous-time plant given by

$$\dot{x} = f(x, u), \tag{3.1}$$

where $x \in \mathbb{R}^n$ and $u \in U \subseteq \mathbb{R}^m$ with $0 \in U$. The plant is to be controlled via a computer that is connected to the plant via a sampler and a zero-order hold. We assume that f is locally Lipschitz, which guarantees that the solutions of (3.1) exist locally in time. Let $\phi(t, x_0, u)$ denote the solution trajectory for time t, initial value x_0, and constant control function $u \in U$. Suppose that, for a given T, x, u, the solution $\phi(t, x, u)$, exists for all $t \in [0, T]$. Then we can introduce the exact discrete-time model of the system

$$x_{k+1} = F_T^e(x(k), u(k)), \tag{3.2}$$

where $F_T^e(x, u) := \phi(T, x, u)$. Note that the trajectories of (3.1) may have finite escape time, in which case $F_T^e(x, u)$ might not be defined for all $x \in \mathbb{R}^n$, $u \in U$. However, since f is assumed to be locally Lipschitz, we have that for each $\Delta > 0$ there exists $T_\Delta^* > 0$ such that $F_T^e(x, u)$ exists for all $x \in \overline{B}_\Delta$, $u \in \overline{B}_\Delta$, and $T \in (0, T_\Delta^*]$. The set of all control sequences is denoted by $\mathcal{U}$; members of $\mathcal{U}$ will be denoted by $\overline{u} = (u(k))_{k \in \mathbb{N}_0}$.

We note that, since f is typically nonlinear, F_T^e in (3.2) is not known in most cases. Hence, if we want to carry out controller design for the sampled-data plant (3.1) via its discrete-time model, we need to use instead an approximate discrete-time model

$$x_{k+1} = F_{T,h}^a(x(k), u(k)), \tag{3.3}$$

where $T \in (0, T^*]$ is the sampling rate with some upper bound $T^* > 0$ and $h \in (0, T]$

is a parameter for the accuracy of the approximate model, e.g., the integration step for some underlying numerical one-step approximation.

Remark 3.1. The map $F^a_{T,h}$ defining the approximate model is typically interpreted as a numerical approximation of F^e_T using some suitable numerical scheme. For instance, $F^a_{T,h}$ might be constructed using multiple steps of a one-step Runge–Kutta scheme Φ_{h_i} with integration step sizes h_i, $i = 1, \dots, m$, satisfying $h_i \le h$ and $\sum_{i=1}^m h_i = T$; i.e.,

$$x_\circ = x, \;\; x_{i+1} = \Phi_{h_i}(x_i, u), \;\; F^a_{T,h}(x, u) = x_m.$$

Note that, for constant control functions u, system (3.1) is an autonomous ODE, and hence all numerical schemes for autonomous ODEs are applicable; see, e.g., [11], [23], or [24] for a description of suitable numerical methods. In the simplest case, Φ_{h_i} could be chosen as the Euler method $\Phi_{h_i}(x, u) = x + hf(x, u)$. Note that, for any T, h, the numerical scheme $F^a_{T,h}(x, u)$ will normally exist for all x, u because the computation of $F^a_{T,h}$ is usually based on finitely many evaluations of f only.

Given a family of cost functions $J_{T,h}(x, \overline{u})$, we will design a family of optimal control laws for the approximate model

$$u(k) = u^{a,*}_{T,h}(x(k)) \tag{3.4}$$

and investigate when they stabilize the family of exact models (3.2) for all small h.

In general, it is useful to consider exact models that are also parameterized by a modeling parameter (for motivation, see [19])

$$x_{k+1} = F^e_{T,h}(x(k), u(k)). \tag{3.5}$$

In this case, however, h is not interpreted as a numerical integration step. We write $F_{T,h}$ if we refer to a general discrete-time parameterized system

$$x_{k+1} = F_{T,h}(x(k), u(k)); \tag{3.6}$$

in particular, $F_{T,h}$ may stand for both $F^e_{T,h}$ and $F^a_{T,h}$. The special case $T = h$ has received much attention in the literature, and, in this case, we will write F_T instead of $F_{T,T}$. Given $\overline{u}$ and $x_\circ$, the trajectories of the systems (3.5) and (3.3) are denoted, respectively, by $\phi^e_{T,h}(k, x_\circ, \overline{u})$ and $\phi^a_{T,h}(k, x_\circ, \overline{u})$. Again, if we refer to a generic system (3.6), we use the notation $\phi_{T,h}(k, x_\circ, \overline{u})$, and, if $T = h$, we write ϕ_T instead of $\phi_{T,T}$.

Assumption 3.2. We assume that both $F^e_{T,h}$ and $F^a_{T,h}$ are continuous in u and satisfy a local Lipschitz condition of the following type: for each $\Delta > 0$, there exist $T > 0$, $L > 0$, and $h^* > 0$, such that

$$\|F_{T,h}(x, u) \; - \; F_{T,h}(y, u)\| \le e^{LT}\|x - y\| \tag{3.7}$$

holds for all $u \in \overline{B}_\Delta$, all $h \in (0, h^*]$, and all x, $y \in \overline{B}_\Delta$.

For the exact model, this property is easily verified using Gronwall's lemma (if $F^e_{T,h}$ is well defined), while, for the approximate model, it depends on the properties of the numerical scheme in use. For Runge–Kutta schemes, e.g., it is verified by induction using the property $\|\Phi_{h_i}(x, u) - \Phi_{h_i}(y, u)\| \le (1 + Lh_i)\|x - y\|$ (cf. [24]) and the inequality $1 + Lh_i \le e^{Lh_i}$.

4. Definitions and background results. In [16, 19], sufficient conditions based on the Lyapunov second method were presented that guarantee that the family of controllers that stabilizes (3.3) would also stabilize (3.5) for sufficiently small h. Here the control laws under consideration do not need to come from optimal control problems; however, they will still be parameterized by the parameters T and h. The results in this section will be used in the rest of this paper. In order to state these results, we need several definitions.

DEFINITION 4.1. *Let strictly positive real numbers (T, Δ_1, Δ_2) be given. If there exists $h^* > 0$ such that*

$$\sup_{\{x \in B_{\Delta_1} \ , \ h \in (0,h^*]\}} |u_{T,h}(x)| \leq \Delta_2, \tag{4.1}$$

then we say that the family of controllers (3.4) *is (T, Δ_1, Δ_2)-uniformly bounded. Moreover, if $T = h$ and if, for any strictly positive Δ_1, there exist strictly positive (Δ_2, h^*) so that* (4.1) *holds, then we say that the family of controllers* (3.4) *is semiglobally uniformly bounded.*

The following "consistency" property is central in our developments, and it is an appropriate adaptation and generalization of a consistency property used in the numerical analysis literature (see [24]).

DEFINITION 4.2. *Let a triple of strictly positive numbers (T, Δ_1, Δ_2) be given, and suppose that there exist $\gamma \in \mathcal{K}$ and $h^* > 0$ such that*

$$(x,u) \in B_{\Delta_1} \times B_{\Delta_2}, h \in (0,h^*] \implies \|F^a_{T,h}(x,u) - F^e_{T,h}(x,u)\| \leq T\gamma(h). \tag{4.2}$$

Then we say that the family $F^a_{T,h}$ is (T, Δ_1, Δ_2)-consistent with $F^e_{T,h}$. Moreover, if $T = h$ and if, for any pair of strictly positive numbers (Δ_1, Δ_2), there exist $\gamma \in \mathcal{K}$ and $h^* > 0$ such that (4.2) *holds, then we say that $F^a_{T,h}$ is semiglobally consistent with $F^e_{T,h}$.*

Sufficient checkable conditions for consistency properties can be found in [16, 19].

DEFINITION 4.3. *Let a pair of strictly positive real numbers (T, D), a family of functions $V_{T,h} : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, functions $\sigma_1, \sigma_2 \in \mathcal{K}_\infty$, and a positive definite function $\sigma_3 : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be given. Suppose, for any pair of strictly positive real numbers (δ_1, δ_2) with $\delta_2 < D$, that there exist $h^* > 0$ and $c > 0$ such that, for all $x \in B_D$, $h \in (0, h^*]$, we have*

$$\sigma_1(\|x\|) \leq V_{T,h}(x) \leq \sigma_2(\|x\|), \tag{4.3}$$

$$V_{T,h}(F^a_{T,h}(x, u_{T,h}(x))) - V_{T,h}(x) \leq -T\sigma_3(\|x\|) + T\delta_1, \tag{4.4}$$

and, for all $x_1, x_2 \in B_D - B_{\delta_2}$, with $\|x_1 - x_2\| \leq c$, we have

$$|V_{T,h}(x_1) - V_{T,h}(x_2)| \leq \delta_1. \tag{4.5}$$

Then we say that the family (3.6), (3.4) *is (T, D)-practically stable with a continuous Lyapunov function. Moreover, if $T = h$ and, for any triple of strictly positive real numbers (D, δ_1, δ_2) with $\delta_2 < D$, there exist $h^* > 0$ and $L > 0$ such that, for all $x, x_1, x_2 \in B_D$, $h \in (0, h^*]$, we have that* (4.3), (4.4), *and*

$$|V_T(x_1) - V_T(x_2)| \leq L\|x_1 - x_2\| \tag{4.6}$$

hold, then we say that the family (3.6), (3.4) *is semiglobally stable with a Lipschitz Lyapunov function.*

The following two theorems from [16, 19] play a central role in our developments.

THEOREM 4.4. *Suppose that there exist a triple of strictly positive numbers* (T, D, M) *such that the following hold:*

(i) *The family of closed loop systems* $(F^a_{T,h}, u^a_{T,h})$ *is* (T, D)*-practically stable with a continuous Lyapunov function.*
(ii) *The family of controllers* $u^a_{T,h}$ *is* (T, D, M)*-uniformly bounded.*
(iii) *The family* $F^a_{T,h}$ *is* (T, D, M)*-consistent with* $F^e_{T,h}$.

Then there exist $\beta \in \mathcal{KL}$, $D_1 \in (0, D)$, *and, for any* $\delta > 0$, *there exists* $h^* > 0$ *such that, for all* $x_\circ \in B_{D_1}$ *and* $h \in (0, h^*]$, *the solutions of the family* $(F^e_{T,h}, u^a_{T,h})$ *satisfy*

$$\|\phi^e_{T,h}(k, x_\circ)\| \le \beta(\|x_\circ\|, kT) + \delta \qquad \forall k \in \mathbb{N}_0. \tag{4.7}$$

THEOREM 4.5. *Suppose that* $T = h$ *and the following conditions hold:*

(i) *The family of closed loop systems* (F^a_T, u^a_T) *is semiglobally stable with a Lipschitz Lyapunov function.*
(ii) *The family of controllers* u^a_T *is semiglobally uniformly bounded.*
(iii) *The family* F^a_T *is semiglobally consistent with* F^e_T.

Then there exists $\beta \in \mathcal{KL}$ *such that, for any* $D_1 > 0$ *and* $\delta > 0$, *there exists* $T^* > 0$ *such that, for all* $x_\circ \in B_{D_1}$ *and* $T \in (0, T^*]$ *the solutions of the family* (F^e_T, u^a_T) *satisfy*

$$\|\phi^e_T(k, x_\circ)\| \le \beta(\|x_\circ\|, kT) + \delta \qquad \forall k \in \mathbb{N}_0. \tag{4.8}$$

Remark 4.6. Theorems 4.4 and 4.5 provide general conditions on the controller, approximate model, and continuous-time plant that guarantee that the controllers that are designed via the approximate model would also stabilize the exact model for sufficient values of the small modeling parameter. In what follows, we investigate the conditions under which control laws that are optimal in some sense for the approximate model satisfy all conditions of Theorems 4.4 and 4.5.

All conditions of Theorems 4.4 and 4.5 are checkable in principle since they do not require the knowledge of the exact discrete-time model. Indeed, the consistency can be checked using the properties of the approximate model (3.3) and the continuous-time plant model (3.1), and it holds for most reasonable numerical integration schemes such as consistent Runge–Kutta methods. Classical numerical texts [11, 23, 24] present a range of consistent numerical schemes that can be used within our framework. Boundedness of the control law is relatively easy to check. Finally, the Lyapunov conditions are hard to check in general, but there is a range of important situations where this is possible to do, as illustrated in [17, 18, 7] and references therein. The present paper presents another class of problems where the Lyapunov conditions can be verified.

5. Infinite horizon problems. In the first part of this section, we assume that $T \ne h$ and h can be assigned arbitrarily and independently of T, which is arbitrary but fixed. In the second part, we consider the case when $T = h$ and T can be assigned arbitrarily.

5.1. Stabilization with a fixed sampling rate T. We consider the optimal control problem

$$\min_{\overline{u} \in \mathcal{U}} \sum_{k=0}^{\infty} T l_h(\phi_{T,h}(k, x, \overline{u}), u(k)), \tag{5.1}$$

where the running cost l_h satisfies the following assumption.

Assumption 5.1. The following hold:

(i) l_h is continuous with respect to x and u, uniformly in small h.

(ii) There exist $h^* > 0$ and two class $\mathcal{K}_\infty$ functions ρ_1 and ρ_2 such that the inequality

$$\rho_1(\|x\| + \|u\|) \le l_h(x,u) \le \rho_2(\|x\| + \|u\|) \tag{5.2}$$

holds for all x, u, and $h \in (0, h^*]$.

(iii) For each $\Delta > 0$, there exist $N > 0$ and $h^* > 0$ such that

$$|l_h(x,u) - l_h(y,u)| \le N\|x - y\|$$

for all $h \in (0, h^*]$, $x, y \in \mathbb{R}^n$, and all $u \in U$ with $\|x\|, \|y\|, \|u\| \le \Delta$.

Note that the sum in (5.1) may diverge; hence it may take the value ∞. We make the convention that this sum takes the value ∞ if the trajectory $\phi_{T,h}(\cdot, x, \overline{u})$ does not exist for some $k \in \mathbb{N}_0$.

We denote the optimal cost functions related to the exact and the approximate system by

$$W^e_{T,h}(x) := \min_{\overline{u}\in\mathcal{U}} \sum_{k=0}^{\infty} Tl_h(\phi^e_{T,h}(k, x, \overline{u}), u(k)),$$

$$W^a_{T,h}(x) := \min_{\overline{u}\in\mathcal{U}} \sum_{k=0}^{\infty} Tl_h(\phi^a_{T,h}(k, x, \overline{u}), u(k)),$$

again using $W_{T,h}(x)$ if we want to refer to a general system and W_T if $T = h$. Note that $W_{T,h}(x) = \infty$ is possible, so we will have to formulate conditions such that $W_{T,h}$ is finite at least for compact subsets of the state space.

It is clear that not every plant would allow for a meaningful solution of the optimal control problem (5.1). However, if the plant model satisfies the following asymptotic controllability assumption, we will prove in Theorem 5.4 that a solution to (5.1) exists under certain assumptions.

DEFINITION 5.2. *Let $T > 0$, $\beta \in \mathcal{KL}$, and $\Delta > 0$ be given. The family of systems* (3.6) *is called (T, Δ, β)-asymptotically controllable to the origin with vanishing controls if there exists $h^* > 0$ such that, for all $h \in (0, h^*]$ and each $x \in \overline{B}_\Delta$, there exists $\overline{u} \in \mathcal{U}$ such that*

$$\|\phi_{T,h}(k, x, \overline{u})\| + \|u(k)\| \le \beta(\|x\|, Tk), \qquad k \in \mathbb{N}_0.$$

Asymptotic controllability has been introduced in [21], and we have adapted the definition from [14] to be applicable to families of discrete-time systems. Note that this definition, in particular, requires $\|u(k)\| \le \beta(\|x\|, Tk)$. This assumption is mainly needed in order to simplify some of the following arguments and could be relaxed in various ways, e.g., to $\|u(k)\| \le \delta + \beta(\|x\|, Tk)$ for some $\delta > 0$, provided that Assumption 5.1 (ii) is suitably adjusted also. The following result is used in what follows.

PROPOSITION 5.3 (see [22]). *Given an arbitrary $\beta \in \mathcal{KL}$, there exist two functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ such that the following holds:*

$$\beta(s,t) \le \alpha_2\left(\alpha_1(s)e^{-t}\right) \qquad \forall s, t \ge 0. \tag{5.3}$$

Note that, using Proposition 5.3, there is no loss of generality if we assume that $\beta(s,t)$ in Definition 5.2 is replaced by $\alpha_2(\alpha_1(s)e^{-t})$. The following theorem shows conditions under which the optimal feedback law for the approximate model exists and can be used to stabilize the exact closed loop system.

THEOREM 5.4. *Let strictly positive real numbers (Δ, T) and functions $\beta \in \mathcal{KL}$ and $l_h(\cdot,\cdot)$ satisfying Assumption 5.1 be given. Let β generate $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ using Proposition 5.3, and let l_h generate $\rho_1, \rho_2 \in \mathcal{K}_\infty$ via (5.2). Suppose the following:*

(i) *The family of approximate models $F^a_{T,h}$ satisfies Assumption 3.2.*

(ii) *The family of approximate models $F^a_{T,h}$ is (T, Δ, β)-asymptotically controllable to the origin with vanishing controls.*

(iii) *There exists $C > 0$ such that*

$$\int_0^1 \frac{\rho_2 \circ \alpha_2(s)}{s} ds \le C. \tag{5.4}$$

Then, for the family of systems $F^a_{T,h}$, there exists a solution to the family of optimal control problems

$$\min_{\overline{u} \in \mathcal{U}} \sum_{k=0}^{\infty} T l_h(\phi^a_{T,h}(k, x, \overline{u}), u(k))$$

of the form

$$u(k) = u^{a,*}_{T,h}(x(k)) \tag{5.5}$$

and numbers $D \in (0, \Delta)$, $M > 0$ such that the following hold:

(i') *The family of controllers $u^{a,*}_{T,h}$ is (T, D, M)-uniformly bounded.*

(ii') *The family $(F^a_{T,h}, u^{a,*}_{T,h})$ is (T, D)-practically stable with continuous Lyapunov function.*

Suppose, moreover, that the following additional condition holds:

(iii') *The family of approximate models $F^a_{T,h}$ is (T, D, M)-consistent with $F^e_{T,h}$.*

Then, there exists $D_1 \in (0, D)$ and $\beta_1 \in \mathcal{KL}$, and, for any $\delta > 0$, there exists $h^* > 0$ such that, for all $x_\circ \in B_{D_1}$ and all $h \in (0, h^*]$, the solutions of the family $(F^e_{T,h}, u^{a,*}_{T,h})$ satisfy

$$\|\phi^e_{T,h}(k, x_\circ)\| \le \beta_1(\|x_\circ\|, kT) + \delta \qquad \forall k \in \mathbb{N}_0.$$

Proof. Let all the conditions of Theorem 5.4 be satisfied. We will prove that conditions (i), (ii), and (iii) imply conditions (i') and (ii'). Then the last statement follows immediately from (i'), (ii'), and (iii') via Theorem 4.4.

(i)+(ii)+(iii) $\Rightarrow$ (ii'). We use the optimal value of the cost $W^a_{T,h}(x)$ as the Lyapunov function for the approximate closed loop system, which is standard in optimization literature. We now show that $W^a_{T,h}$ satisfies (4.3), (4.4), and (4.5) of Definition 4.3.

It is immediate from (5.2) that, for any x and $h \in (0, h^*]$, we have

$$\sigma_1(\|x\|) := T\rho_1(\|x\|) \le W^a_{T,h}(x). \tag{5.6}$$

Let $x \in \overline{B}_\Delta$ and $h \in (0, h^*]$. Using the definition of the cost, the bound (5.2), and condition (ii), we obtain for $\overline{u}$ from Definition 5.2:

$$\begin{aligned}
W^a_{T,h}(x) &\le \sum_{k=0}^{\infty} T l_h(\phi^a_{T,h}(k,x,\overline{u}),u(k)) \\
&\le \sum_{k=0}^{\infty} T\rho_2(\|\phi^a_{T,h}(k,x,\overline{u})\| + \|u(k)\|) \\
&\le \sum_{k=0}^{\infty} T\rho_2(\beta(\|x\|,kT)) \\
&\le \sum_{k=0}^{\infty} T\rho_2(\alpha_2(e^{-kT}\alpha_1(\|x\|))) \\
&= T\rho_2\circ\alpha_2\circ\alpha_1(\|x\|) + \sum_{k=1}^{\infty} T\rho_2(\alpha_2(e^{-kT}\alpha_1(\|x\|))) \\
&\le T\rho_2(\alpha_2(\alpha_1(\|x\|))) + \int_0^{\infty} \rho_2(\alpha_2(e^{-t}\alpha_1(\|x\|)))dt.
\end{aligned}$$

It was shown in [4, Proof of Theorem 1] that, under condition (iii), the integral term in the last inequality can be bounded by $\tilde{\sigma}(\|x\|)$ for some $\tilde{\sigma} \in \mathcal{K}_\infty$. Hence, if we define $\sigma_2(r) := T\rho_2(\alpha_2(\alpha_1(r))) + \tilde{\sigma}(r)$, we can write, for all $x \in \overline{B}_\Delta$ and $h \in (0, h^*]$, that

$$W^a_{T,h}(x) \le \sigma_2(\|x\|). \tag{5.7}$$

Hence (5.6) and (5.7) show that (4.3) holds.

Let an arbitrary $\delta_1 > 0$ be given. We show now that, for the given (Δ, δ_1), there exist $D \in (0, \Delta]$, $c > 0$, and $h^* > 0$ such that the implication

$$x \in \overline{B}_D, \|x-y\| \le c, h \in (0,h^*] \Rightarrow \ |W^a_{T,h}(x) - W^a_{T,h}(y)| \le \delta_1 \tag{5.8}$$

holds, which proves that (4.5) is satisfied.[1]

For the rest of the proof, we use lemmas that are presented and proved in the appendix. Let $\rho_1, \rho_2 \in \mathcal{K}_\infty$, and $h^*_1 > 0$ come from Assumption 5.1. Define the following numbers:

$$\begin{aligned}
S &:= \sigma_1(\Delta) + \delta_1/4, \\
\widetilde{\Delta} &:= \rho_1^{-1}(S/T), \\
\alpha &= \sigma_2^{-1}\left(\frac{\delta_1}{8}\right).
\end{aligned}$$

Let $(S, \alpha/2)$ generate via Lemma A.2 the number $\tau > 0$. Let $\widetilde{\Delta}$ generate via (3.7) the numbers $N > 0$ and $h^*_2 > 0$. Let $(\tilde{\Delta}, \tau, T)$ and $\delta := \min\{\alpha/2, \frac{\delta_1}{2N\tau}\}$ generate via Lemma A.4 the numbers $c > 0$ and $h^*_3 > 0$. Let $h^* := \min\{h^*_1, h^*_2, h^*_3\}$. Let $D := \sigma_2^{-1} \circ \sigma_1(\Delta)$.

In all calculations below, we consider arbitrary $x \in \overline{B}_D$, $h \in (0, h^*]$, and $\|x-y\| \le c$. Let $\overline{u}$ be a control sequence such that

$$\sum_{k=0}^{\infty} l_h(\phi^a_{T,h}(k,x,\overline{u}),u(k)) \le W^a_{T,h}(x) + \delta_1/4,$$

[1] Note that this is a stronger condition than what is needed in Definition 4.3 since we have $\delta_2 = 0$.

which implies from $\|x\| \le D$ and the definition of S that $\sum_{k=0}^{\infty} l_h(\phi_{T,h}(k,x,\overline{u}),u(k)) \le S$. From Lemma A.1 and the definition of $\widetilde{\Delta}$, we have

$$\|\phi^a_{T,h}(k,x,\overline{u})\| + \|u(k)\| \le \widetilde{\Delta} \qquad \forall k \in \mathbb{N}_0.$$

From the definition of α and (5.7), we have $W^a_{T,h}(x) \le \delta_1/8$ for all $x \in \overline{B}_\alpha$. From our choice of τ, it follows from Lemma A.2 that, for some $j \in \mathbb{N}_0$ with $Tj \le \tau$, we have $\|\phi^a_{T,h}(j,x,u)\| \le \alpha/2$. Moreover, from Lemma A.4 and our choice of δ, it follows that $\|\phi^a_{T,h}(j,x,\overline{u}) - \phi^a_{T,h}(j,y,\overline{u})\| \le \delta \le \alpha/2$ and, consequently, $\|\phi^a_{T,h}(j,y,\overline{u})\| \le \alpha$, which implies from the choice of α that

$$W^a_{T,h}(\phi^a_{T,h}(j,y,\overline{u})) \le \delta_1/8\ .$$

Abbreviating $\tilde{y} = \phi^a_{T,h}(j,y,\overline{u})$, we can choose a control sequence $\overline{u}^*$ satisfying

$$\sum_{k=0}^{\infty} l_h(\phi^a_{T,h}(k,\tilde{y},\overline{u}^*),u^*(k)) \le W^a_{T,h}(\tilde{y}) + \delta_1/8 \le \delta_1/4.$$

Replacing $u(k)$, $k = j, j+1, \ldots$, by u^*_{k-j}, we thus obtain

$$\begin{aligned} W^a_{T,h}(y) &\le \sum_{k=0}^{j-1} T l_h(\phi^a_{T,h}(k,y,\overline{u}),u(k)) + \sum_{k=j}^{\infty} T l_h(\phi^a_{T,h}(k,y,\overline{u}),u(k)) \\ &= \sum_{k=0}^{j-1} T l_h(\phi^a_{T,h}(k,y,\overline{u}),u(k)) + \sum_{k=0}^{\infty} T l_h(\phi^a_{T,h}(k,\tilde{y},\overline{u}^*),u^*(k)) \\ &\le \sum_{k=0}^{j-1} T l_h(\phi^a_{T,h}(k,y,\overline{u}),u(k)) + \delta_1/4. \end{aligned} \tag{5.9}$$

Again, using Lemma A.4, the Lipschitz property of l_h, the fact that $jT \le \tau$, and our choice of δ, we can conclude that

$$\sum_{k=0}^{j-1} T(l_h(\phi^a_{T,h}(k,y,\overline{u}),u(k)) - l_h(\phi^a_{T,h}(k,x,\overline{u}),u(k))) \le N\tau\delta \le \delta_1/2. \tag{5.10}$$

The definition of $W^a_{T,h}$, the choice of $\overline{u}$, and the positive definiteness of l_h imply

$$W^a_{T,h}(x) \ge \sum_{k=0}^{j-1} T l_h(\phi^a_{T,h}(k,x,\overline{u})) - \delta_1/4. \tag{5.11}$$

We now combine (5.9), (5.10), and (5.11):

$$\begin{aligned} W^a_{T,h}(y) - W^a_{T,h}(x) &\le \sum_{k=0}^{j-1} T l_h(\phi^a_{T,h}(k,y,\overline{u}),u(k)) \\ &\qquad - \sum_{k=0}^{j-1} T l_h(\phi^a_{T,h}(k,x,\overline{u}),u(k)) + \delta_1/4 + \delta_1/4 \\ &\le \delta_1/2 + \delta_1/4 + \delta_1/4 = \delta_1. \end{aligned}$$

Since the corresponding estimate for $W_T^a(x) - W_T^a(y)$ follows by symmetry, this completes the proof of (4.5).

Finally, with the given (Δ, δ_1), we show that (4.4) is satisfied. For any fixed T and h, standard optimal control arguments show that $W_{T,h}^a$ satisfies the dynamic programming equation

$$W_{T,h}^a(x) = \inf_{u \in U} \{Tl_h(x,u) + W_{T,h}^a(F_{T,h}^a(x,u))\}.$$

Since $F_{T,h}^a$ and l_h are continuous in u, $W_{T,h}^a$ is continuous in x, and l_h is positive definite, the "inf" is actually a "min," and we can define the desired $u_{T,h}^{a,*}(x)$ by choosing it such that

$$Tl_h(x, u_{T,h}^{a,*}(x)) + W_{T,h}^a(F_{T,h}^a(x, u_{T,h}^{a,*}(x))) = \min_{u \in U}\{Tl_h(x,u) + W_{T,h}^a(F_{T,h}^a(x,u))\}.$$

Combining the given bounds above and using (5.2), we obtain

$$\begin{aligned} W_{T,h}^a(F_{T,h}^a(x, u_{T,h}^{a,*}(x))) - W_{T,h}^a(x) &= -Tl_h(x, u_{T,h}^{a,*}(x)) \\ &\le -T\rho_1(\|x\|) \\ &\le -T\rho_1(\|x\|) + T\delta_1, \end{aligned}$$

which proves (4.4) and completes the proof of (ii').

(i)+(ii)+(iii) $\Rightarrow$ (i'). Since, for all $x \in \overline{B}_D$, we have $Tl_h(x, u_{T,h}^{a,*}(x)) \le W_{T,h}^a(x)$ and since (5.2) holds, we can write that

$$\begin{aligned} \|u_{T,h}^{a,*}(x)\| &\le \rho_1^{-1}(l_h(x, u_{T,h}^{a,*}(x))) \\ &\le \rho_1^{-1}\left(\frac{1}{T} W_{T,h}^a(x)\right) \\ &\le \rho_1^{-1}\left(\frac{1}{T}\sigma_2(\|x\|)\right) \\ &\le \rho_1^{-1}\left(\frac{\sigma_2(D)}{T}\right) =: M, \end{aligned} \tag{5.12}$$

which proves that (i') holds. □

Remark 5.5. The conditions in Theorem 5.4 are relatively hard to verify. We have already commented on condition (i) after Assumption 3.2. The asymptotic controllability (condition (ii)) is a necessary condition for any stabilization problem, and most well-designed control systems will satisfy this condition. Finally, the condition (iii) suggests that, if the system is asymptotically controllable with a certain α_2, then we can always adjust the cost $l_h(x,u)$ so that $\rho_2(s) \le \alpha_2^{-1}(Cs)$ for some $C > 0$. Hence, by reducing ρ_2 sufficiently, we can always make condition (iii) hold for any asymptotically controllable system. Since we are interested only in stabilization, the procedure for computing a stabilizing controller for asymptotically controllable systems may be as follows: start with a certain pair of $l_h(x,u), \rho_2$, and attempt the numerical optimization. If the optimal solution has been obtained, use this optimal controller. If the optimal solution was not obtained, introduce new $\tilde{\rho}$ such that $\tilde{\rho}(s) < \rho(s)$ for all $s > 0$ and a new $\tilde{l}_h(x,u)$, and repeat the procedure.

Remark 5.6. It is often the case that, instead of the optimal controller, one can only compute a certain suboptimal controller with a corresponding value function. Our results still apply in this case if the suboptimal controller and the resulting value

function satisfy the first two conditions of Theorems 4.4 or 4.5 and the approximate model satisfies the third condition in these theorems. Hence, under these conditions, the suboptimal controller for the approximate model will still be stabilizing the exact model in an appropriate sense. Due to space reasons, we did not pursue this issue in more detail.

Remark 5.7. Note that if T can be adjusted arbitrarily and independent of h, and, moreover, if for any arbitrary $\Delta > 0$ there exists T so that the system is (T, Δ, β)-asymptotically controllable with vanishing controls, and if all other conditions of Theorem 5.4 hold, then all conclusions of Theorem 5.4 hold. Hence, for T varying and independent of h, we can modify the statement of Theorem 5.4 to obtain a result on semiglobal practical stabilization. However, if $T = h$, we need much stronger conditions to achieve semiglobal practical stabilization, which is discussed in more detail in the next subsection.

Remark 5.8. Neither of the examples of section 2 satisfies Assumption 5.1, and that is the reason why the controllers $u_T^{a,*}$ do not stabilize the family of exact models F_T^e.

Remark 5.9. It is possible under mild conditions to obtain $\mathcal{KL}$ stability bounds for the solutions of the sampled-data system from the $\mathcal{KL}$ stability bounds for the exact discrete-time model and bounds on the intersample behavior, as illustrated in [20].

5.2. Stabilization with varying sampling rate $T = h$. The case in which $T = h$ is sometimes considered in the literature (see Example 1 in [3]), and we discuss it next. For instance, some authors use the Euler approximate model

$$x(k+1) = F_T^a(x(k), u(k)) = x(k) + Tf(x(k), u(k))$$

in MPC of a continuous-time plant $\dot{x} = f(x, u)$. While this approach is very attractive because of the reduced computational effort in obtaining the approximate discrete-time model F_T^a, we show below that it may have serious limitations.

Note that for $T = h$ we need to use Theorem 4.5, which requires (among other things) the following:

C1. a lower bound on the optimal value function that is uniform in small T; cf. Theorem 5.12 (iv);

C2. boundedness of the optimal controller $u_T^{a,*}$ on compact sets uniform in small T; cf. Theorem 5.12 (iii');

C3. W_T^a locally Lipschitz and uniformly in small T; cf. Theorem 5.12 (v).

It is well known from optimal control theory that, even for fixed $T > 0$, one cannot expect W_T^a to be locally Lipschitz in general, and hence condition C3 usually does not hold. Moreover, note that the inequalities (5.6) and (5.12) seem to suggest that, in general, for any fixed x, we may have that $W_T^a(x) \to 0$ and $\|u_T^{a,*}(x)\| \to \infty$ as $T \to 0$, which violates conditions C1 and C2. The next example shows that this can indeed happen when $T = h$.

Example 5.10. Consider the scalar system

$$\dot{x} = u^3$$

with $u \in U = \mathbb{R}$ and the running cost $l(x, u) = \|x\|^2 + \|u\|^2$. The corresponding exact discrete-time model is given by

$$x(k+1) = x(k) + Tu^3(k) =: F_T(x(k), u(k)),$$

so the control sequence $\overline{u}$ induced by the state feedback law

$$u_T^*(x) = -(x/T)^{1/3}$$

yields

$$\sum_{k=0}^{\infty} Tl(\phi_T(k, x(0), \overline{u}), u(k)) = T(\|x(0)\|^2 + (x(0)/T)^{2/3}) = T\|x(0)\|^2 + T^{1/3}\|x(0)\|^{2/3}.$$

Consequently, we obtain

$$W_T(x(0)) \le T\|x(0)\|^2 + T^{1/3}\|x(0)\|^{2/3}.$$

Setting $W_T(x(0)) = T\|x(0)\|^2 + T^{1/3}\|x(0)\|^{2/3}$, one sees that the equality

$$l(x, u_T^*(x)) + W_T(F_T(x, u_T^*(x))) = \inf_{u\in U}\{l(x,u) + W_T(F_T(x,u))\}$$

holds. (One verifies that, for all x, T, the term on the right-hand side has only two local minima located at $u = 0$ and $u = u_T^*(x)$, and the latter yields a smaller value.) Hence the feedback law $u_T^*(x)$ is optimal for this problem.

Note that, for any fixed $x \neq 0$, we have $T \to 0 \implies W_T(x) \to 0$ and $\|u_T^*(x)\| \to \infty$.

While, in the example discussed above, $u_T^*(x)$ still asymptotically stabilizes the exact model (due to the fact that, for this simple system, the exact discrete-time model and its Euler approximation coincide), in general, this phenomenon poses a serious problem, and $u_T^{a,*}$ may, in general, destabilize the family F_T^e. Several examples illustrating this phenomenon can be found in [19].

As a result of the above discussion, it is obvious that one can either search for conditions on f, F_T^a, and l_T to guarantee that C1, C2, and C3 hold or simply assume that they hold. While it is apparent that the first approach poses interesting and relevant questions, we did not pursue it in this paper. Using the second approach, we can state Theorem 5.4. Before we state the theorem, we need to slightly modify the definition of asymptotic controllability as follows.

DEFINITION 5.11. *Let $\beta \in \mathcal{KL}$ be given. The family of systems $x(k+1) = F_T(x(k), u(k))$ is called semiglobally asymptotically controllable to the origin with vanishing controls if, for each $\Delta > 0$, there exists $T^* > 0$ such that, for all $T \in (0, T^*]$ and each $x \in \overline{B}_\Delta$, there exists $\overline{u} \in \mathcal{U}$ such that*

$$\|\phi_T(k, x, \overline{u})\| + \|u(k)\| \le \beta(\|x\|, Tk).$$

THEOREM 5.12. *Let $T = h$. Let $\beta \in \mathcal{KL}$ and $l_T(\cdot,\cdot)$ satisfying Assumption 5.1 be given. Let β generate $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ using Proposition 5.3, and let l_T generate $\rho_1, \rho_2 \in \mathcal{K}_\infty$ via (5.2). Suppose the following:*

(i) *The family of approximate models F_T^a satisfies the following: for any $\Delta > 0$, there exist $N > 0$ and $T^* > 0$ such that, for all $T \in (0, T^*]$ and $x \in \overline{B}_\Delta$, we have*

$$\|F_T^a(x,u) - F_T^a(y,u)\| \le e^{NT}\|x - y\|.$$

(ii) *The family of approximate models F_T^a is semiglobally asymptotically controllable to the origin with vanishing controls.*

(iii) *There exists $C > 0$ such that condition* (5.4) *holds.*

(iv) *Condition* C1 *holds; i.e., there exist $\sigma_1 \in \mathcal{K}_\infty$ and $T^* > 0$ such that, for all x and $T \in (0, T^*]$, we have*

$$\sigma_1(\|x\|) \le W_T^a(x).$$

(v) *Condition* C3 *holds; that is, for any $\Delta > 0$, there exist $T^* > 0$ and $L > 0$ such that*

$$|W_T^a(x) - W_T^a(y)| \le L\|x - y\|$$

for all $x, y \in \overline{B}_\Delta$, $T \in (0, T^*]$.

Then, for the family of systems F_T^a, there exists a solution to the family of optimal control problems

$$\min_{\overline{u} \in \mathcal{U}} \sum_{k=0}^{\infty} Tl_T(\phi_T^a(k, x, \overline{u}), u(k))$$

of the form

$$u(k) = u_T^{a,*}(x(k)) \tag{5.13}$$

such that the following holds:

(i') *The family $(F_T^a, u_T^{a,*})$ is semiglobally practically stable with a Lipschitz Lyapunov function.*

Suppose, moreover, that the following additional conditions hold:

(ii') *The family of approximate models F_T^a is semiglobally consistent with F_T^e.*

(iii') *Condition* C2 *holds; i.e., for any $\Delta > 0$, there exist $T^* > 0$ and $M > 0$ such that, for all $\|x\| \le \Delta$, $T \in (0, T^*)$,*

$$\|u_T^{a,*}(x)\| \le M.$$

Then there exists $\beta_1 \in \mathcal{KL}$ such that, for any strictly positive (D_1, δ), there exists $T^* > 0$ such that, for all $x_\circ \in B_{D_1}$ and all $T \in (0, T^*]$, the solutions of the family $(F_T^e, u_T^{a,*})$ satisfy

$$\|\phi_T^e(k, x_\circ)\| \le \beta_1(\|x_\circ\|, kT) + \delta \qquad \forall k \in \mathbb{N}_0.$$

Proof. We provide only a sketch of the proof since it is very similar to the proof of Theorem 5.4. The only thing to prove is that (i)–(v) imply (i') since the rest of the proof follows immediately from Theorem 4.5.

Note that the condition (v) implies (4.6) and the condition (iv) implies the lower bound in (4.3). The upper bound in (4.3) is established in the same way as that in the proof of Theorem 5.4. The inequality (4.4) is established in the same way as in the proof of Theorem 5.4, which completes the proof. ∎

6. Finite horizon with terminal cost problems. In practice, the optimal control problem under consideration will often not be solved over an infinite time horizon but by using a suitable terminal cost. There are various ways to introduce a terminal cost (see, e.g., [5, sections III.3 and IV.3]), and we believe that our approach can be adjusted in order to cope with most of them. In order to illustrate this procedure, we consider the special type of terminal cost introduced by Kreisselmeier and Birkhölzer in [14].

We consider a family of continuous and positive definite functions $\overline{W}_{T,h} : \mathbb{R}^n \to \mathbb{R}_0^+$ for $T \in (0, T^*]$ and $h \in (0, T]$ and define the following family of finite horizon optimal control problems with terminal costs:

$$W^a_{T,h}(x) := \inf_{\overline{u}\in\mathcal{U},\, k'\in\mathbb{N}_0} \left\{ \sum_{k=0}^{k'-1} Tl_h(\phi^a_{T,h}(k, x, \overline{u}), u(k)) + \overline{W}_{T,h}(\phi^a_{T,h}(k', x, \overline{u})) \right\}. \tag{6.1}$$

Using our continuity assumptions on $F^a_{T,h}$ and l_h in u, it is easily seen that there always exists a feedback law $u^{a,*}_{T,h} : \mathbb{R}^n \to U$ satisfying

$$Tl_h(x, u^{a,*}_{T,h}(x)) + W^a_{T,h}(F^a_{T,h}(x, u^{a,*}_{T,h}(x))) = \min_{u\in U}\{Tl_h(x, u) + W^a_{T,h}(F^a_{T,h}(x, u))\}. \tag{6.2}$$

Moreover, observe that, using (6.2), the dynamic programming equation for $W^a_{T,h}(x)$ reads

$$W^a_{T,h}(x) = \min\{Tl_h(x, u^{a,*}_{T,h}(x)) + W^a_{T,h}(F^a_{T,h}(x, u^{a,*}_{T,h}(x))),\ \overline{W}_{T,h}(x)\}. \tag{6.3}$$

6.1. Stabilization with fixed sampling rate T. In this section, we consider T as an arbitrary but fixed positive sampling rate. In order to derive a stabilization result, we need the following assumption on $\overline{W}_{T,h}$.

Assumption 6.1. The following hold:

(i) $\overline{W}_{T,h}$ is continuous, uniformly in small h.

(ii) There exist $h^* > 0$ and two class $\mathcal{K}_\infty$ functions γ_1 and γ_2 such that the inequality

$$\gamma_1(\|x\|) \le \overline{W}_{T,h}(x) \le \gamma_2(\|x\|) \tag{6.4}$$

holds for all x and $h \in (0, h^*]$.

THEOREM 6.2. *Let strictly positive real numbers (Δ, T) and the family of functions $\overline{W}_{T,h}(\cdot)$ satisfying Assumption* 6.1 *and the family of functions $l_h(\cdot,\cdot)$ satisfying Assumption* 5.1 *be given.*

Suppose the following:

(i) *The family of approximate models $F^a_{T,h}$ satisfies Assumption* 3.2.

(ii) *For any $d > 0$, there exists $h^* > 0$ such that, for all $h \in (0, h^*]$, there exists a solution to the optimization problem* (6.1) *that satisfies*

$$W^a_{T,h}(x) < \overline{W}_{T,h}(x) \qquad \forall x \in B_\Delta - B_d,\ h \in (0, h^*]. \tag{6.5}$$

Then there exists $M > 0$ such that $u^{a,*}_{T,h}(\cdot)$ from (6.2) *satisfies the following properties for $D = \Delta$:*

(i') *The family of controllers $u^{a,*}_{T,h}$ is (T, D, M)-uniformly bounded.*

(ii') *The family $(F^a_{T,h}, u^{a,*}_{T,h})$ is (T, D)-practically stable with a continuous Lyapunov function.*

Suppose, moreover, that the following additional condition holds:

(iii') *The family of approximate models $F^a_{T,h}$ is (T, D, M)-consistent with $F^e_{T,h}$.*

Then there exists $D_1 \in (0, D)$ and $\beta_1 \in \mathcal{KL}$, and, for any $\delta > 0$, there exists $h^* > 0$ such that, for all $x_\circ \in \overline{B}_{D_1}$ and all $h \in (0, h^*]$, the solutions of the family $(F^e_{T,h}, u^{a,*}_{T,h})$ satisfy

$$\|\phi^e_{T,h}(k, x_\circ)\| \le \beta_1(\|x_\circ\|, kT) + \delta \qquad \forall k \in \mathbb{N}_0.$$

Proof. Similar to the proof of Theorem 5.4, the main task is to prove that conditions (i), (ii), and (iii) imply conditions (i') and (ii'). Then, again, the last statement follows immediately from (i'), (ii'), and (iii') via Theorem 4.4.

(i)+(ii)+(iii) $\Rightarrow$ (ii'). We use the optimal value function $W^a_{T,h}(x)$ as a Lyapunov function candidate and verify the conditions (4.3), (4.4), and (4.5) of Definition 4.3 for the family $(F^a_{T,h}, u^{a,*}_{T,h})$.

Let Assumption 6.1 generate $h_1^* > 0$ and $\gamma_1, \gamma_2 \in \mathcal{K}_\infty$. Let Assumption 5.1 generate $h_2^* > 0$ and $\rho_1, \rho_2 \in \mathcal{K}_\infty$. Let (T, Δ) come from conditions of the theorem. Define $D := \Delta$, and let δ_1 be an arbitrarily strictly positive real number.[2] Let d be such that

$$T\rho_2(d) + \gamma_2(e^{LT}d) \le T\delta_1.$$

Let (D, d) generate $h_3^* > 0$ using condition (ii) of the theorem. Let $h^* := \min\{h_1^*, h_2^*, h_3^*\}$. In the rest of the proof, we consider arbitrary $x \in \overline{B}_D$ and $h \in (0, h^*]$.

First, we prove that (4.3) holds. Using the definition of $W^a_{T,h}$, we obtain the inequality

$$W^a_{T,h}(x) \le \overline{W}_{T,h}(x) \ \le \gamma_2(\|x\|) =: \sigma_2(\|x\|).$$

For the lower bound, observe from (6.3) that we have either

$$W^a_{T,h}(x) = \overline{W}_{T,h}(x) \ge \gamma_1(\|x\|)$$

or

$$W^a_{T,h}(x) \ge T l_h(x, u^{a,*}_{T,h}(x)) \ge T\rho_1(\|x\|),$$

and hence

$$W^a_{T,h}(x) \ge \min\{\gamma_1(\|x\|), T\rho_1(\|x\|)\} =: \sigma_1(\|x\|),$$

which completes the proof of (4.3).

Next we show (4.4) for the family $(F^a_{T,h}, u^{a,*}_{T,h})$. From our choice of x and h, for any $x \in B_D - B_d$, we obtain that the "min" in (6.3) is attained in the first term; hence

$$W^a_{T,h}(F^a_{T,h}(x, u^{a,*}_{T,h}(x))) - W^a_{T,h}(x) = -T l_h(x, u^{a,*}_{T,h}(x)). \tag{6.6}$$

For $x \in B_d$, observe that inequality (3.7) and $F^a_{T,h}(0,0) = 0$ imply $\|F^a_{T,h}(x,0)\| \le e^{LT}\|x\|$. Hence from (6.2) we obtain

$$\begin{aligned} T l_h(x, u^{a,*}_{T,h}(x)) + W^a_{T,h}(F^a_{T,h}(x, u^{a,*}_{T,h}(x))) &\le T l_h(x,0) + W^a_{T,h}(F^a_{T,h}(x,0)) \\ &\le T\rho_2(\|x\|) + \gamma_2(e^{LT}\|x\|) \\ &\le T\rho_2(d) + \gamma_2(e^{LT}d) \\ &\le T\delta_1 \ . \end{aligned} \tag{6.7}$$

Since $W^a_{T,h}(x) \ge 0$, this implies

$$W^a_{T,h}(F^a_{T,h}(x, u^{a,*}_{T,h}(x))) - W^a_{T,h}(x) \le -T l_h(x, u^{a,*}_{T,h}(x)) + T\delta_1. \tag{6.8}$$

[2] Like in the previous section, we prove that all conditions of Definition 4.3 hold with $\delta_2 = 0$.

Then, for $x \in B_D$, either (6.6) or (6.8) holds, which implies

$$\begin{aligned} W^a_{T,h}(F^a_{T,h}(x, u^{a,*}_{T,h}(x))) - W^a_{T,h}(x) &\le -Tl_h(x, u^{a,*}_{T,h}(x)) + T\delta_1 \\ &\le -T\rho_1(\|x\|) + T\delta_1 \\ &=: -T\sigma_3(\|x\|) + T\delta_1; \end{aligned}$$

i.e., the desired estimate (4.4) holds.

In order to show the continuity property (4.5), first observe that, by the continuity condition on $\overline{W}_{T,h}$ from Assumption 6.1, for any given $\tilde{\delta} > 0$, we find $\tilde{c} > 0$ such that, for all $x, y \in B_D$ with $\|x - y\| \le \tilde{c}$, we obtain

$$|\overline{W}^a_{T,h}(x) - \overline{W}^a_{T,h}(y)| \le \tilde{\delta}. \tag{6.9}$$

Consider the (arbitrary) $\delta_1 > 0$, which has been chosen above. Then, for any $x \in B_D$, we find a control sequence $\overline{u}$ and a value $\overline{k} \in \mathbb{N}_0$ such that

$$W^a_{T,h}(x) + \delta_1/4 \ge \sum_{k=0}^{\overline{k}-1} Tl_h(\phi^a_{T,h}(k, x, \overline{u}), u(k)) + \overline{W}_{T,h}(\phi^a_{T,h}(\overline{k}, x, \overline{u})).$$

Proceeding similarly as in the proof of Theorem 5.4, we find a constant $c > 0$ and, from Lemma A.2, a time $\tau > 0$ such that either $\overline{k} - 1 \le \tau/T$ or, for any $y \in B_D$ with $\|x - y\| \le c$, the inequality

$$W^a_{T,h}(\phi^a_{T,h}(j, y, \overline{u})) \le \delta_1/8$$

holds for some $j \in \mathbb{N}_0$ with $j \le \tau/T$. If $\overline{k} - 1 > \tau/T$ holds, we can exactly follow the proof of Theorem 5.4 to obtain that the implication

$$\|x - y\| \le c \;\Rightarrow\; W^a_{T,h}(y) - W^a_{T,h}(x) \le \delta_1$$

holds. Otherwise, i.e., when $\overline{k} - 1 \le \tau/T$, using the values defined in the proof of Theorem 5.4, for any $y \in B_D$ with $\|x - y\| \le c$, we obtain

$$\begin{aligned} W^a_{T,h}(y) - W^a_{T,h}(x) \le &\sum_{k=0}^{\overline{k}-1} Tl_h(\phi^a_{T,h}(k, y, \overline{u}), u(k)) + \overline{W}_{T,h}(\phi^a_{T,h}(\overline{k}, y, \overline{u})) \\ &- \sum_{k=0}^{\overline{k}-1} Tl_h(\phi^a_{T,h}(k, x, \overline{u}), u(k)) - \overline{W}_{T,h}(\phi^a_{T,h}(\overline{k}, x, \overline{u})) + \delta_1/4 \\ &\le \delta_1/2 + \delta_1/4 + \overline{W}_{T,h}(\phi^a_{T,h}(\overline{k}, y, \overline{u})) - \overline{W}_{T,h}(\phi^a_{T,h}(\overline{k}, x, \overline{u})). \end{aligned}$$

Setting $\tilde{\delta}$ from (6.9) equal to $\delta_1/4$, we find $\tilde{c}$ such that

$$\overline{W}_{T,h}(\phi^a_{T,h}(\overline{k}, y, \overline{u})) - \overline{W}_{T,h}(\phi^a_{T,h}(\overline{k}, x, \overline{u})) \le \delta_1/4$$

if $\|\phi^a_{T,h}(\overline{k}, x, \overline{u}) - \phi^a_{T,h}(\overline{k}, y, \overline{u})\| \le \tilde{c}$. Using that $\overline{k} - 1 \le \tau/T$ and that the control sequence $u^{a,*}_{T,h}$ is bounded by Lemma A.1, by reducing $c > 0$, if necessary, Lemma A.4 guarantees that $\|x - y\| \le c$ implies $\|\phi^a_{T,h}(\overline{k}, x, \overline{u}) - \phi^a_{T,h}(\overline{k}, y, \overline{u})\| \le \tilde{c}$. Thus, also in the case in which $\overline{k} - 1 \le \tau/T$, we obtain the implication

$$\|x - y\| \le c \;\Rightarrow\; W^a_{T,h}(y) - W^a_{T,h}(x) \le \delta_1.$$

Since the same estimate for $W^a_{T,h}(x) - W^a_{T,h}(y)$ follows by symmetry, this proves (4.5) and thus the proof of (ii').

(i)+(ii)+(iii) $\Rightarrow$ (i'). For $x \in B_D - B_d$, this follows exactly as in the proof of Theorem 5.4, and we have that (5.12) holds. For $x \in B_d$, from inequality (6.7), we obtain

$$T\rho_1(\|u^{a,*}_{T,h}(x)\|) \le Tl_h(x, u^{a,*}_{T,h}(x)) \le T\rho_2(d) + \gamma_2(e^{LT}d),$$

implying that, for all $x \in B_D$ and $h \in (0, h^*]$, we have

$$\|u^{a,*}_{T,h}(x)\| \le \max\left\{\rho_1^{-1}\left(\frac{\sigma_2(D)}{T}\right), \rho_1^{-1}\left(\rho_2(d) + \frac{\gamma_2(e^{LT}d)}{T}\right)\right\} =: M,$$

which completes the proof. □

Remark 6.3. The results of Theorem 6.2 provide a rigorous framework for optimization-based stabilization via a finite horizon with terminal cost problem. The same optimization problem was considered in [14] under the assumption that the exact discrete-time plant model is known. In the same reference in section IV, an inverted pendulum example was successfully stabilized using this method, where the discrete-time model of the plant was obtained by integrating the continuous-time dynamics over one sampling interval. Strictly, the theoretical results of [14] do not directly apply to the example considered since an approximate (instead of exact) discrete-time model was used for the example. Hence our Theorem 6.2 provides a rigorous proof that justifies the use of the approximate discrete-time model for the inverted pendulum example in [14].

6.2. Stabilization with varying sampling rate $T = h$. We now state conditions that guarantee that optimization-based stabilization of sampled-data systems via their approximate discrete-time models with $T = h$ can be successfully carried out. Consider the following version of Assumption 6.1 for $T = h$.

Assumption 6.4. There exists $T^* > 0$ such that the following hold:

(i) $\overline{W}_T$ is continuous, uniformly in $T \in (0, T^*]$.

(ii) There exist two class $\mathcal{K}_\infty$ functions γ_1 and γ_2 such that the inequality

$$\gamma_1(\|x\|) \le \overline{W}_T(x) \le \gamma_2(\|x\|) \tag{6.10}$$

holds for all x and $T \in (0, T^*]$.

THEOREM 6.5. *Let functions $\overline{W}_T(\cdot)$ satisfying Assumption* 6.4 *and $l_T(\cdot,\cdot)$ satisfying Assumption* 5.1 *be given.*

Suppose the following:

(i) *The family of approximate models F^a_T satisfies Assumption* 3.2 *and satisfies that, for each $\Delta > 0$, there exist $\gamma \in \mathcal{K}$ and $T^* > 0$ such that the inequality*

$$\|F_T(x, 0) - x\| \le T\gamma(\|x\|)$$

holds for all $\|x\| \le \Delta$ and all $T \in (0, T^*]$.

(ii) *For any pair of positive real values (D, d), there exists $T^* > 0$ such that, for all $T \in (0, T^*]$, there exists a solution to the optimization problem* (6.1) *that satisfies*

$$W^a_T(x) < \overline{W}_T(x) \qquad \forall x \in B_D - B_d,\ T \in (0, T^*]. \tag{6.11}$$

(iii) *For any $\Delta > 0$, there exist $L > 0$ and $T^* > 0$ such that*

$$|W_T^a(x) - W_T^a(y)| \le L\|x - y\|$$

for all $x, y \in B_\Delta$, $T \in (0, T^*]$.

(iv) *There exist $\sigma_1 \in \mathcal{K}_\infty$ and $T^* > 0$ such that, for all x and $T \in (0, T^*]$, we have*

$$\sigma_1(\|x\|) \le W_T^a(x).$$

Then we have the following:

(i') *The family $(F_T^a, u_T^{a,*})$ is semiglobally stable with a Lipschitz Lyapunov function.*

Suppose, moreover, that the following additional conditions hold:

(ii') *The family of approximate models F_T^a is semiglobally consistent with F_T^e.*

(iii') *The family of controllers $u_T^{a,*}$ is semiglobally uniformly bounded.*

Then there exists $\beta_1 \in \mathcal{KL}$ such that, for any strictly positive (D_1, δ), there exists $T^* > 0$ such that, for all $x_\circ \in \overline{B}_{D_1}$ and all $T \in (0, T^*]$, the solutions of the family $(F_T^e, u_T^{a,*})$ satisfy

$$\|\phi_T^e(k, x_\circ)\| \le \beta_1(\|x_\circ\|, kT) + \delta \qquad \forall k \in \mathbb{N}_0.$$

Proof. It suffices to prove that (i)–(iv) imply (i') because then the statement follows from (i'), (ii'), and (iii') by applying Theorem 4.5.

Since the condition (iii) implies (4.6), for proving (i'), we have only to show (4.3) and (4.4).

Let (D, δ_1) be given.[3] Let $\overline{W}_T$ generate via Assumption 6.4 the functions $\gamma_1, \gamma_2 \in \mathcal{K}_\infty$ and $T_1^* > 0$. Let l_T generate via Assumption 5.1 the functions $\rho_1, \rho_2 \in \mathcal{K}_\infty$ and $T_2^* > 0$. Let the condition (iv) generate σ_1 and $T_3^* > 0$. Let $\Delta = D$ generate via the condition (i) the function $\gamma \in \mathcal{K}$ and $T_4^* > 0$. Let $\Delta_1 = D + \gamma(D)$ generate $L > 0$ and $T_5^* > 0$ via the condition (iii). Let $d > 0$ be such that

$$\rho_2(d) + L\gamma(d) \le \delta_1.$$

Let (D, d) generate $T_6^* > 0$ via the condition (ii). Let $T^* = \min\{1, T_1^*, T_2^*, T_3^*, T_4^*, T_5^*, T_6^*\}$. Consider arbitrary $T \in (0, T^*]$ and $x \in \overline{B}_D$.

In order to prove (4.3), observe that the lower bound follows from the condition (iv), while the upper bound follows immediately from the inequality $W_T^a(x) \le \overline{W}_T(x)$ and Assumption 6.4. Recall from (6.3) that the dynamic programming equation for $W_T^a(x)$ reads

$$W_T^a(x) = \min\{Tl_T(x, u_T^{a,*}(x)) + W_T^a(F_T^a(x, u_T^{a,*}(x))),\ \overline{W}_T(x)\}.$$

For all $T \in (0, T^*]$ and all $x \in \overline{B}_D - \overline{B}_d$, we obtain that the "min" is attained in the first term; hence

$$W_T^a(F_T^a(x, u_T^{a,*}(x))) - W_T^a(x) = -Tl_T(x, u_T^{a,*}(x)). \tag{6.12}$$

[3] We will again prove (4.4) with $\delta_2 = 0$.

For $x \in \overline{B}_d$, recall Assumption (i), which yields $\|F^a_{T,h}(x,0) - x\| \le T\gamma(\|x\|)$. Hence from (6.2) we obtain

$$\begin{aligned} Tl_T(x, u_T^{a,*}(x)) + W_T^a(F_T^a(x, u_T^{a,*}(x))) - W_T^a(x) &\le Tl_T(x,0) + W_T^a(F_T^a(x,0)) - W_T^a(x) \\ &\le T\rho_2(\|x\|) + LT\gamma(\|x\|) \\ &\le T(\rho_2(d) + L\gamma(d)) \ \le \ T\delta_1. \end{aligned} \tag{6.13}$$

Hence, for $x \in \overline{B}_D$, either (6.12) or (6.13) holds, which implies

$$W_T^a(F_T^a(x, u_T^{a,*}(x))) - W_T^a(x) \le -Tl_T(x, u_T^{a,*}(x)) + T\delta_1 \le -T\rho_1(\|x\|) + T\delta_1,$$

i.e., the desired estimate (4.4) with $\alpha_3(r) := \rho_1(r)$. □

7. Conclusion and outlook. Results on optimization-based stabilization of sampled-data systems via approximate discrete-time plant models are presented. Infinite horizon and finite horizon with terminal cost optimization problems were considered. In both cases, it was shown under reasonable assumptions that, when integration period h is independent of the sampling period T, then one can use an approximate discrete-time plant model in the controller design to achieve stability of the exact discrete-time plant model. On the other hand, if $T = h$, then optimization-based stabilization of sampled-data systems via approximate discrete-time models requires much stronger assumptions to produce a stabilizing controller for the exact discrete-time plant model. Several examples are presented to illustrate the most common problems with this approach.

Apart from the optimal control problem we have considered in this paper, one of the most important optimal control techniques is RHC, which is often used in MPC schemes; cf. the references in the introduction. Due to the special structure of RHC and MPC techniques, our results in this paper are not directly applicable. We do, however, think that analysis techniques similar to those we have used here can be applied also to these kinds of controllers. Future research will include the derivation of rigorous results in this direction.

Appendix.

LEMMA A.1. *Let l_h satisfy* (5.2) *with some $\rho_1, \rho_2 \in \mathcal{K}_\infty$, and $h^* > 0$. Then, for any strictly positive (S,T), $h \in (0, h^*]$, and $x \in \mathbb{R}^n, \overline{u} \in \mathcal{U}, \overline{k} \in \mathbb{N}_0$ satisfying*

$$\sum_{k=0}^{\overline{k}} Tl_h(\phi^a_{T,h}(k,x,\overline{u}), u(k)) \le S, \tag{A.1}$$

we have

$$\|\phi^a_{T,h}(k,x,\overline{u})\| + \|u(k)\| \le \rho_1^{-1}(S/T) \qquad \forall k \in \mathbb{N}_0,\ k \le \overline{k}.$$

Proof. Let (A.1) hold, and assume the existence of $k \in \mathbb{N}_0$ with

$$\|\phi^a_{T,h}(k,x,\overline{u})\| + \|u(k)\| > \rho_1^{-1}(S/T).$$

This implies, using (5.2), that

$$Tl_h(\phi^a_{T,h}(k,x,\overline{u}), u(k)) \ge T\rho_1(\|\phi^a_{T,h}(k,x,\overline{u})\| + \|u(k)\|) > S,$$

which contradicts (A.1). □

LEMMA A.2. *Let l_h satisfy (5.2) with some $\rho_1, \rho_2 \in \mathcal{K}_\infty$, and $h^* > 0$. Then, for any pair of strictly positive numbers (C, ε), there exists $\tau = \tau(C, \varepsilon) > 0$ such that, for any $x \in \mathbb{R}^n, \overline{u} \in \mathcal{U}$, any $T > 0$ and $h \in (0, h^*]$, and any $\overline{k} \in \mathbb{N}$ with $\overline{k}T > \tau$ satisfying*

$$\sum_{k=0}^{\overline{k}} T l_h(\phi_{T,h}(k, x, \overline{u}), u(k)) < C, \tag{A.2}$$

there exists $j \in \mathbb{N}_0$ with $j \le \tau/T$ such that $\|\phi^a_{T,h}(j, x, \overline{u})\| + \|u_j\| \le \varepsilon$.

Proof. We define $\tau := C/\rho_1(\varepsilon)$. Now assume $\|\phi^a_{T,h}(j, x, \overline{u})\| + \|u_j\| \ge \varepsilon$ for all $j \in \mathbb{N}_0$ with $j \le \tau/T$. Denoting by $[\tau/T]$ the integer part of τ/T and using (5.2), we can conclude that

$$\begin{aligned}\sum_{k=0}^{\overline{k}} T l_h(\phi^a_{T,h}(k, x, \overline{u}), u(k)) &\ge \sum_{k=0}^{[\tau/T]} T l_h(\phi^a_{T,h}(k, x, \overline{u}), u(k)) \ge [\tau/T + 1] T \rho_1(\varepsilon) \\ &\ge \tau \rho_1(\varepsilon) \ge C,\end{aligned}$$

which contradicts (A.2). ∎

The following lemma is a consequence of the consistency property. Similar results can be found in numerical analysis literature (see, for example [24, Theorems 6.2.1 and 6.2.2]), and the proof is provided below for completeness.

LEMMA A.3. *Let a 4-tuple of strictly positive numbers $(\Delta_1, \Delta_2, T, \tau)$ be given, and let $F^a_{T,h}$ be (T, Δ_1, Δ_2)-consistent with $F^e_{T,h}$. Let Assumption 3.2 hold. Consider any $x \in \mathbb{R}^n, \overline{u} \in \mathcal{U}$ satisfying*

$$kT \in [0, \tau] \Longrightarrow \|\phi^a_{T,h}(k, x, \overline{u})\| \le \Delta_1 \text{ and } \|u(k)\| \le \Delta_2. \tag{A.3}$$

Then, for any $\delta > 0$, there exist strictly positive numbers $h^* = h^*(\Delta_1, \Delta_2, \delta, T) > 0$ and $L = L(\Delta_1, \Delta_2, \delta)$ and $\gamma \in \mathcal{K}$ such that, if $h \in (0, h^*]$, then $kT \in [0, \tau]$ implies that

$$\|\phi^e_{T,h}(k, x, \overline{u})\| \le \Delta_1 + \delta \tag{A.4}$$

and

$$\|\phi^a_{T,h}(k, x, \overline{u}) - \phi^e_{T,h}(k, x, \overline{u})\| \le T\gamma(h) \frac{e^{L(\tau+T)} - 1}{LT}. \tag{A.5}$$

Moreover, an analogous estimate holds if we exchange the roles of $\phi^a_{T,h}$ and $\phi^e_{T,h}$.

Proof. Let all conditions of the lemma hold. Let $L > 0$ be the Lipschitz constant from (3.7) on the set $\overline{B}_{\max\{\Delta_1+\delta, \Delta_2\}}$, and let $\gamma \in \mathcal{K}$ come from the consistency property for the same set. Define

$$h^* = \gamma^{-1}\left(\frac{\delta L}{e^{L(\tau+T)} - 1}\right).$$

Consider now arbitrary $x, \overline{u}$ satisfying (A.3), and let $h \in (0, h^*]$. We abbreviate $x(k)^a = \phi^a_{T,h}(k, x, \overline{u})$, $x(k)^e = \phi^e_T(k, x, \overline{u})$ and show the assertion by induction. For $k = 0$, there is nothing to show. Pick $k \ge 1$ such that $kT \in [0, \tau]$, and assume for the purpose of induction that for the step $k - 1$ the following holds:

$$\|\phi^a_{T,h}(k-1, x, \overline{u}) - \phi^e_T(k-1, x, \overline{u})\| \le \gamma(h) T \sum_{i=0}^{k-1} e^{LTi} \;\le\; \delta. \tag{A.6}$$

We can conclude using Assumption 3.2 and consistency that

$$\begin{aligned}
&\|x(k)^a - x(k)^e\| \\
&= \|F^a_{T,h}(x^a_{k-1}, u_{k-1}) - F^e_{T,h}(x^e_{k-1}, u_{k-1})\| \\
&\leq \|F^a_{T,h}(x^a_{k-1}, u_{k-1}) - F^e_{T,h}(x^a_{k-1}, u_{k-1})\| + \|F^e_{T,h}(x^a_{k-1}, u_{k-1}) - F^e_{T,h}(x^e_{k-1}, u_{k-1})\| \\
&\leq \gamma(h)T + e^{LT}\gamma(h)T\sum_{i=0}^{k-1} e^{LTi} \\
&\leq \gamma(h)T + \gamma(h)T\sum_{i=1}^{k} e^{LTi} \\
&\leq \gamma(h)T\sum_{i=0}^{k} e^{LTi} \ ,
\end{aligned}$$

which shows that (A.6) holds for the step k. Finally, since $T\sum_{i=0}^{k} e^{LTi} \leq \frac{e^{LT(k+1)}-1}{L}$ and because of our choice of h^*, it follows that, for all $h \in (0, h^*]$ and all $kT \in [0, \tau]$, we have

$$\begin{aligned}
\|x(k)^a - x(k)^e\| &\leq T\gamma(h)\frac{e^{L(\tau+T)} - 1}{LT} \\
&\leq \delta,
\end{aligned}$$

which proves that (A.5) holds. Finally, since $x^a_k \in \overline{B}_{\Delta_1}$ for all $kT \in [0, \tau]$ and (A.7) holds, this implies that (A.4) is satisfied, which completes the proof. □

LEMMA A.4. *Let an arbitrary triple of strictly positive numbers (Δ, τ, T) be given. Let $k_0 \in \mathbb{N}$ be such that $k_0 T \in [0, \tau]$. Let Assumption 3.2 hold for the family $F_{T,h}$. Then, for each $\delta > 0$, there exist $c > 0$, $L > 0$, and $h^* > 0$ such that, for each $h \in (0, h^*]$, each two points $x, y \in \mathbb{R}^n$ with $\|x - y\| \leq c$, and each $\overline{u} \in \mathcal{U}$ satisfying*

$$\|\phi_{T,h}(k, x, \overline{u})\| + \|u(k)\| \leq \Delta \quad \forall k = 0, 1, \ldots, k_0,$$

the inequalities

$$\|\phi_{T,h}(k, y, \overline{u})\| \leq \Delta + \delta$$

and

$$\|\phi_{T,h}(k, x, \overline{u}) - \phi_{T,h}(k, y, \overline{u})\| \leq \|x - y\| e^{LTk} \leq \delta$$

hold for all $k = 0, 1, \ldots, k_0$.

Proof. The proof follows with arguments similar to those in the proof of Lemma A.3. □

REFERENCES

[1] M. ALAMIR AND G. BORNARD, *On the stability of receding horizon control of nonlinear discrete-time systems*, Systems Control Lett., 23 (1994), pp. 291–296.

[2] B. D. O. ANDERSON AND J. MOORE, *Optimal Control: Linear Quadratic Methods*, Prentice–Hall, Englewood Cliffs, NJ, 1990.

[3] D. Angeli and E. Mosca, *Command governors for constrained nonlinear systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 816–818.

[4] M. Arcak, D. Angeli, and E. D. Sontag, *Stabilization of cascades using integral input-to-state stability*, in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, FL, 2001, pp. 3814–3819.

[5] M. Bardi and I. Capuzzo Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Boston, 1997.

[6] G. de Nicolao, L. Magni, and R. Scattolini, *Stabilizing receding-horizon control of nonlinear time-varying systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 1030–1036.

[7] D. S. Laila and D. Nešić, *Changing supply rates for input-output to state stable discrete-time nonlinear systems with applications*, Automatica J. IFAC, to appear.

[8] F. A. C. C. Fontes, *A general framework to design stabilizing nonlinear model predictive controllers*, Systems Control Lett., 42 (2001), pp. 127–143.

[9] L. Grüne, *Discrete feedback stabilization of semilinear control systems*, ESAIM Control Optim. Calc. Var., 1 (1996), pp. 207–224.

[10] L. Grüne, *Homogeneous state feedback stabilization of homogeneous systems*, SIAM J. Control Optim., 38 (2000), pp. 1288–1308.

[11] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations*, Cambridge Texts Appl. Math., Cambridge University Press, Cambridge, UK, 1996.

[12] S. S. Keerthy and E. G. Gilbert, *An existence theorem for discrete-time infinite horizon optimal control problems*, IEEE Trans. Automat. Control, 30 (1985), pp. 907–909.

[13] S. S. Keerthy and E. G. Gilbert, *Optimal infinite horizon feedback laws for a general class of constrained discrete-time systems: Stability and moving horizon approximations*, J. Optim. Theory Appl., 57 (1988), pp. 265–293.

[14] G. Kreisselmeier and T. Birkhölzer, *Numerical nonlinear regulator design*, IEEE Trans. Automat. Control, 39 (1994), pp. 33–46.

[15] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, *Constrained model predictive control: Stability and optimality*, Automatica J. IFAC, 36 (2000), pp. 789–814.

[16] D. Nešić, A. R. Teel, and P. V. Kokotović, *Sufficient conditions for stabilization of sampled-data nonlinear systems via discrete-time approximations*, Systems Control Lett., 38 (1999), pp. 259–270.

[17] D. Nešić and A. R. Teel, *Set stabilization of nonlinear sampled-data differential inclusions via their approximate discrete-time models*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia, 2000, pp. 2112–2117.

[18] D. Nešić and A. R. Teel, *Backstepping on the Euler approximate model for stabilization of sampled-data nonlinear systems*, in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, FL, 2001, pp. 1737–1742.

[19] D. Nešić and A. R. Teel, *A framework for stabilization of nonlinear sampled-data systems based on their approximate discrete-time models*, IEEE Trans. Automat. Control, to appear.

[20] D. Nešić, A. R. Teel, and E. D. Sontag, *Formulas relating KL stability estimates of discrete-time and sampled-data nonlinear systems*, Systems Control Lett., 38 (1999), pp. 49–60.

[21] E. D. Sontag, *A Lyapunov-like characterization of asymptotic controllability*, SIAM J. Control Optim., 21 (1983), pp. 462–471.

[22] E. D. Sontag, *Comments on integral variants of ISS*, Systems Control Lett., 34 (1998), pp. 93–100.

[23] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

[24] A. M. Stuart and A. R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, UK, 1996.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 123–137

INDEFINITE STOCHASTIC RICCATI EQUATIONS*

YING HU† AND XUN YU ZHOU‡

Abstract. This paper is concerned with stochastic Riccati equations (SREs), which are a class of matrix-valued, nonlinear backward stochastic differential equations (BSDEs). The SREs under consideration are, in general, indefinite, in the sense that certain parameter matrices are indefinite. This kind of equations arises from the stochastic linear-quadratic (LQ) optimal control problem with random coefficients and indefinite state and control weighting costs, the latter having profound implications in both theory and applications. While the solvability of the SREs is the key to solving the indefinite stochastic LQ control, it remains, in general, an extremely difficult, open problem. This paper attempts to solve the problem of existence and uniqueness of solutions to the indefinite SREs for a number of special, yet important, cases.

Key words. stochastic Riccati equation, backward stochastic differential equation, stochastic linear-quadratic optimal control

AMS subject classifications. 65C30, 93E20

PII. S0363012901391330

1. Introduction.

In this paper, we study the following matrix-valued equation, called a stochastic Riccati equation (SRE), on a finite time horizon $[0,T]$ (the time variable t is suppressed):

(1.1)
$$\begin{cases} dP = -\Big\{P\,A + A'\,P + \sum_{j=1}^{k}(\Lambda_j\,C_j + C_j'\,\Lambda_j + C_j'\,P\,C_j) + Q \\ \qquad\quad - \Big[P\,B + \sum_{j=1}^{k}(C_j'P + \Lambda_j)\,D_j\Big]\,K^{-1}\,\Big[B'P + \sum_{j=1}^{k} D_j'\,(PC_j + \Lambda_j)\Big]\Big\}\,dt \\ \qquad\quad + \sum_{j=1}^{k}\Lambda_j\,dW^j, \qquad t \in [0,\,T], \\ P(T) = H, \\ K = R + \sum_{j=1}^{k} D_j'\,P\,D_j \;>\; 0. \end{cases}$$

Note that the last (matrix) positive definiteness constraint is *part* of the equation that must be satisfied by any solution. Thus, strictly speaking, this is an equation with mixed equality/inequality constraints. The first two constraints constitute what is known as a nonlinear backward stochastic differential equation (BSDE). The unknown of the SRE is the matrix-valued stochastic process $(P(t), \Lambda_1(t), \ldots, \Lambda_k(t))$ adapted to the filtration $\mathcal{F}_t$ that is generated by the underlying Brownian motion $W(t) = (W^1(t), \ldots, W^k(t))$. The coefficients of this equation, $A(t), B(t), C_j(t), D_j(t), Q(t)$, and $R(t)$, are matrix-valued $\mathcal{F}_t$-adapted stochastic processes, and H is an $\mathcal{F}_T$-measurable random matrix.

The SRE (1.1) relates intimately to the following optimal linear-quadratic (LQ) control problem with *random* coefficients:

*Received by the editors June 23, 2001; accepted for publication (in revised form) September 9, 2002; published electronically March 19, 2003. This research was supported by RGC Earmarked Grant CUHK 4435/99E.

http://www.siam.org/journals/sicon/42-1/39133.html

†Institut de Recherche Mathématique de Rennes, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex, France (hu@maths.univ-rennes1.fr). Part of this work was completed when this author was visiting the Chinese University of Hong Kong, whose hospitality is greatly appreciated.

‡Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, NT, Hong Kong (xyzhou@se.cuhk.edu.hk).

Minimize the quadratic cost functional

$$J(x_0, u(\cdot)) = E\Big\{ \int_0^T \Big(x(t)'Q(t)x(t) + u(t)'R(t)u(t) \Big)\, dt + x(T)'Hx(T) \Big\}, \tag{1.2}$$

subject to the stochastic linear system dynamics

$$\begin{cases} dx(t) = [A(t)\, x(t) + B(t)\, u(t)\,]\, dt \\ \qquad\qquad + \sum_{j=1}^{k} [C_j(t)\, x(t) + D_j(t)\, u(t)]\, dW^j(t), \qquad t \in [0,\ T], \\ x(0) = x_0. \end{cases} \tag{1.3}$$

In general, if the SRE (1.1) admits a solution, then the LQ problem (1.2)–(1.3) is solvable, and an optimal control can be represented explicitly in terms of the solution to (1.1) (see section 3 for details). Bismut [5] was the first to study the stochastic LQ problem with random coefficients and the associated SRE. Using a fixed-point argument, he proved the existence and uniqueness of solutions to the SRE under the assumptions (among others) that the random coefficients are independent of the Brownian motion and that R is uniformly positive definite and Q and H are positive semidefinite. The definiteness assumption is particularly crucial in proving the result, though the assumption per se seems rather natural because it had been taken for granted in the stochastic LQ literature [3, 4], which was presumably inherited from its deterministic counterpart [2]. Note, however, the SRE investigated by Bismut is not really a BSDE in the sense of Pardoux and Peng [16] since the coefficients are independent of, rather than adapted to, the driving Brownian motion. Later, Peng [17] studied the SRE (1.1) as a nonlinear BSDE, but again under the definiteness condition, i.e., $R > 0$, $Q \geq 0$, $H \geq 0$, in addition to a strong assumption that $D = 0$.

Recently, it was found by Chen, Li, and Zhou [6] that a stochastic LQ problem where R is possibly *indefinite* may still be solvable as long as the corresponding SRE (1.1) is solvable. This finding has triggered an extensive research on the so-called *indefinite* LQ control problem [9, 1, 7, 8, 18]. The problem not only stands out on its own as an interesting theoretical problem, but also has promising applications in practical areas, especially in finance. In the mean-variance portfolio selection [20, 14], or hedging [13], problems, for example, the matrix R is inherently zero, which is a special indefinite case. Moreover, a pollution control model where R is *negative* definite is formulated in [6].

The indefinite stochastic LQ problem leads to an *indefinite* SRE, adding new twists and greater difficulty to the study of the backward SRE, an already hard problem in terms of the existence of its solutions. To make it more precise, in the definite case, i.e., $R > 0$, $Q \geq 0$, $H \geq 0$, the third inequality constraint normally becomes redundant which can therefore be eliminated from consideration. In the indefinite case, though, this constraint becomes a hard one that must be taken care of. In [6], a special case when all the coefficients are deterministic (so the equation reduces to an ordinary differential equation) and $C = 0$ is solved. However, even that case is nontrivial due to the indefiniteness of R. Most of the subsequent research has centered around solving the indefinite SRE, analytically or computationally, with deterministic coefficients [9, 1, 13, 18]. For the case with random coefficients a very special indefinite SRE was solved in [14] where P is scalar-valued, $C_j = Q = 0$, R is a zero matrix, and $H = 1$. This special SRE arises from a mean-variance portfolio selection problem for a market with random parameters. The existence of a solution to this special SRE is proved in [14] using an ad hoc approach. On the other hand,

local solvability (i.e., existence of solution in a small neighborhood of the terminal time) of the general SRE (1.1) is established in [7, 8]. Note, however, that the local solvability is not very useful in view of the LQ control application because the time horizon in a control problem is typically given a priori.

The *global* existence of the general, indefinite SRE (1.1) on the entire time interval $[0, T]$ remains an extremely challenging, if not at all insurmountable, open problem due to the following reasons. First of all, it is a highly nonlinear BSDE, especially in view of the matrix inverse term $(R+D'PD)^{-1}$, for which the normal Lipschitz/linear growth conditions for solvability [16, 15, 19] are not valid. Second, the indefiniteness of R makes possible the singularity of the term $R + D'PD$ when one tries to use the typical approximation scheme to construct a solution. Third, the final constraint in (1.1) must be satisfied by any solution. This is a feature not typically dealt with in the BSDE literature. Finally, (1.1) is a *matrix* equation. Hence certain terms do not commute which adds substantial difficulty to the analysis.

This paper represents the *first* systematic attempt to tackle the indefinite SRE (1.1), where Q, R, and H are all allowed to be indefinite. While the results of this paper are still far away from being a complete solution to the solvability of (1.1) in its greatest generality, we are able to identify some special cases where a global solution is available.

It should be mentioned that, after finishing this work, we have received the preprints [11, 12] by Kohlmann and Tang. In these, Peng's result for definite SRE [17] is improved to the case where R is possibly singular, i.e., $R \geq 0$, $Q \geq 0$, $H \geq 0$, albeit with various other assumptions. The main approach of [11, 12] is an approximation scheme, which does not apply to the indefinite case due to the possible singularity mentioned above.

The remainder of this paper is organized as follows: In section 2, we give preliminaries including two known results needed in the subsequent study. In section 3 we recall the origin of the SRE, namely, the stochastic LQ optimal control problem, via which we also address the uniqueness of solutions to SRE. Sections 4 and 5 are devoted to the study of one-dimensional SREs and higher-dimensional SREs, respectively. Finally, section 6 concludes the paper.

2. Preliminaries. We assume throughout that $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, P)$ is a given complete, filtered probability space and that $W(\cdot)$ is a k-dimensional standard Brownian motion on this space with $W(0) = 0$. In addition, we assume that $\mathcal{F}_t$ is the augmentation of $\sigma\{W(s) \,|\, 0 \leq s \leq t\}$ by all the P-null sets of $\mathcal{F}$.

Throughout this paper, we denote by $\mathbb{R}^{m\times n}$ the set of $m \times n$ real matrices, and by $\mathbb{S}^n$ the set of symmetric $n \times n$ real matrices. If $M = (m_{ij}) \in \mathbb{R}^{m\times n}$, we denote its norm by $||M|| = \sqrt{\sum_{i,j} m_{ij}^2}$. If $M \in \mathbb{S}^n$ is positive (positive semi)definite, we write $M > (\geq)\ 0$. Suppose $\eta : \Omega \to \mathbb{R}^n$ is an $\mathcal{F}_T$-random variable. We write $\eta \in L^2_{\mathcal{F}_T}(\Omega; \mathbb{R}^n)$ if η is square integrable (i.e., $E|\eta|^2 < \infty$), and $\eta \in L^\infty_{\mathcal{F}_T}(\Omega; \mathbb{R}^n)$ if η is uniformly bounded. Consider now the case when $f : [0, T] \times \Omega \to \mathbb{R}^n$ is an $\{\mathcal{F}_t\}_{t\geq 0}$ adapted process. If $f(\cdot)$ is square integrable (i.e., $E\int_0^T |f(t)|^2\, dt < \infty$) we shall write $f(\cdot) \in L^2_{\mathcal{F}}(0, T; \mathbb{R}^n)$; if $f(\cdot)$ is uniformly bounded (i.e., ess $\sup_{(t,\omega)\in[0,T]\times\Omega} |f(t)| < \infty$), then $f(\cdot) \in L^\infty_{\mathcal{F}}(0, T; \mathbb{R}^n)$. If $f(\cdot)$ has (P-a.s.) continuous sample paths and $E \sup_{t\in[0,T]} |f(t)|^2 < \infty$, we write $f(\cdot) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R}^n))$. These definitions generalize in the obvious way to the case when $f(\cdot)$ is $\mathbb{R}^{n\times m}$- or $\mathbb{S}^n$-valued. Finally, we say that $N \in L^2_{\mathcal{F}}(0, T; \mathbb{S}^n)$ is positive (positive semi)definite, which is sometimes denoted simply by $N > (\geq)\ 0$, if $N(t, \omega) > (\geq)\ 0$ for a.e. $t \in [0, T]$ and P-a.s. ω, and

we say that N is uniformly positive definite if $N \geq \delta I$ for a.e. $t \in [0, T]$ and P-a.s. ω with some given $\delta > 0$.

DEFINITION 2.1. *A stochastic process* $(P, \Lambda_1, \ldots, \Lambda_k) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{S}^n)) \times \big(L^2_{\mathcal{F}}(0, T; \mathbb{S}^n)\big)^k$ *is called a solution to the SRE* (1.1) *if it satisfies the first equation of* (1.1) *in the Itô sense as well as the second (the terminal condition) and third (the positive definiteness) constraints of* (1.1). *A solution* $(P, \Lambda_1, \ldots, \Lambda_k)$ *of* (1.1) *is called bounded if* $P \in L^\infty_{\mathcal{F}}(0, T; \mathbb{S}^n)$, *and is called positive (positive semi)definite if* $P > (\geq)\, 0$.

Throughout this paper, we impose the following assumptions on the parameters of the SRE (1.1).

Assumption.

(A1)

$$\begin{cases} A,\, C_j \in L^\infty_{\mathcal{F}}(0, T; \mathbb{R}^{n\times n}), & j = 1, 2, \ldots, k, \\ B,\, D_j \in L^\infty_{\mathcal{F}}(0, T; \mathbb{R}^{n\times m}), & j = 1, 2, \ldots, k, \\ Q \in L^\infty_{\mathcal{F}}(0, T; \mathbb{S}^n), \\ R \in L^\infty_{\mathcal{F}}(0, T; \mathbb{S}^m), \\ H \in L^\infty_{\mathcal{F}_T}(\Omega; \mathbb{S}^n). \end{cases}$$

In particular, we do not assume that $Q \geq 0$, $R > 0$, or $H \geq 0$. In addition, A, B, C_j, D_j, Q, R, H are random. Later, we shall impose other specific assumptions for various special cases considered in this paper.

Next, we collect two known results needed in our subsequent study of indefinite SREs. The first result, due to Kobylanski [10], concerns the existence of solution and comparison theorem for one-dimensional BSDEs with quadratic growth.

Let $\alpha_0, \beta_0, b \in \mathbb{R}$, and $c : [0, +\infty) \to [0, +\infty)$ be a continuous increasing function. We say that the coefficient F satisfies condition (H1) with α_0, β_0, b, c, if F is continuous, and for all $(t, y, z, \omega) \in [0, +\infty) \times \mathbb{R} \times \mathbb{R}^k \times \Omega$,

$$F(t, y, z, \omega) = a_0(t, y, z, \omega)y + F_0(t, y, z, \omega),$$

where $a_0(\cdot, y, z, \cdot)$ and $F_0(\cdot, y, z, \cdot)$ are $\{\mathcal{F}_t\}_{t\geq 0}$ adapted processes for fixed $(y, z) \in \mathbb{R} \times \mathbb{R}^k$, and

$$\beta_0 \leq a_0(t, y, z, \omega) \leq \alpha_0, \qquad P\text{-a.s. } \omega,$$

$$|F_0(t, y, z, \omega)| \leq b + c(|y|)||z||^2, \qquad P\text{-a.s. } \omega.$$

LEMMA 2.1. *Let* (F, ξ) *be a set of parameters of the following BSDE:*

$$Y(t) = \xi + \int_t^T F(s, Y(s), Z(s))ds - \int_t^T Z(s)dW(s), \qquad 0 \leq t \leq T. \tag{2.1}$$

Suppose that the coefficient F *satisfies* (H1) *with* α_0, β_0, b, c, *and* $\xi \in L^\infty_{\mathcal{F}_T}(\Omega)$. *Then,*

(i) *the BSDE* (2.1) *has at least one solution* $(Y, Z) \in [L^\infty_{\mathcal{F}}(0, T; \mathbb{R}) \cap L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R}))] \times L^2_{\mathcal{F}}(0, T; \mathbb{R}^k)$;

(ii) *there exists a maximal solution* (Y^*, Z^*) *of* (2.1) *in the following sense: For any set of parameters* (G, η), *where* G *satisfies* (H1) *with* α_0, β_0, b, c, *if*

$$F \geq G \text{ and } \xi \geq \eta,$$

then for any solution (Y_G, Z_G) *of the BSDE* (2.1) *with the parameters* (G, η), *one must have*

$$Y^* \geq Y_G.$$

The second result, due to Peng [17, Theorem 5.1], is about the unique solvability of the SRE (1.1) in the definite case (i.e., $Q \geq 0$, $H \geq 0$, $R > 0$), where $D = 0$ is additionally assumed.

LEMMA 2.2. *The following SRE with $Q \geq 0$, $H \geq 0$, $R > 0$,*

$$\begin{cases} dP = -\Big\{P\,A + A'\,P + \sum_{j=1}^{k}(\Lambda_j\,C_j + C_j'\,\Lambda_j + C_j'\,P\,C_j) + Q - P\,BR^{-1}\,B'P\Big\}\,dt \\ \qquad\quad + \sum_{j=1}^{k}\Lambda_j\,dW^j, \qquad t \in [0,\,T], \\ P(T) = H, \end{cases}$$

admits a unique bounded positive semidefinite solution (P, Λ).

3. Stochastic LQ control and uniqueness of solutions to SRE. In this section we recall the connection between the SRE (1.1) and the stochastic LQ control problem. A general result of the uniqueness of solutions to SRE will also be addressed via LQ control.

We first recall the formulation of the stochastic LQ control [6]. Let $(\Omega,\,\mathcal{F},\,\{\mathcal{F}_t\}_{t\geq 0},\,P)$ be a complete filtered probability space on which a k-dimensional standard Brownian motion $W(\cdot)$ is defined such that $\{\mathcal{F}_t\}_{t\geq 0}$ is the natural filtration of $W(t)$ augmented by all the P-null sets of $\mathcal{F}$. For any given $(s,\xi) \in [0,T)\times L^2_{\mathcal{F}_s}(\Omega;\mathbb{R}^n)$, consider the following linear SDE on $[s,T]$:

$$\begin{cases} dx(t) = [A(t)\,x(t) + B(t)\,u(t)\,]\,dt \\ \qquad\quad + \sum_{j=1}^{k}[C_j(t)\,x(t) + D_j(t)\,u(t)]\,dW^j(t), \qquad t \in [s,\,T], \\ x(s) = \xi, \end{cases} \tag{3.1}$$

where A, B, C_j, and D_j are the same parameters as appearing in the SRE (1.1). The class of *admissible controls* is the set $\mathcal{U} = L^2_{\mathcal{F}}(0,\,T;\,\mathbb{R}^m)$. If $u(\cdot) \in \mathcal{U}$ and $x(\cdot)$ is the associated solution of (3.1), then we refer to $(x(\cdot),\,u(\cdot))$ as an *admissible pair*.

Suppose that the cost functional is given by

$$J(s,\xi;u(\cdot)) = E\left\{\int_0^T \Big(x(t)'Q(t)x(t) + u(t)'R(t)u(t)\Big)\,dt + x(T)'Hx(T)\Big|\mathcal{F}_s\right\}. \tag{3.2}$$

Again, Q, R, and H are the same as given in (1.1). To summarize, the *stochastic LQ control problem* associated with (3.1)–(3.2) is as follows:

$$\begin{cases} \min J(s,\xi;u(\cdot)) \\ \text{subject to } (x(\cdot),\,u(\cdot)) \text{ admissible for (3.1).} \end{cases} \tag{3.3}$$

The problem (3.3) is said to be *solvable* if for any $(s,\xi) \in [0,T)\times L^2_{\mathcal{F}_s}(\Omega;\mathbb{R}^n)$ there exists a control $u^*(\cdot) \in \mathcal{U}$ such that

$$-\infty < J(s,\xi;u^*(\cdot)) \leq J(s,\xi;u(\cdot)), \qquad P\text{-a.s. } \omega \quad \forall\, u(\cdot) \in \mathcal{U}.$$

In this case, the control $u^*(\cdot)$ is referred to as an *optimal control*.

Suppose $(P, \Lambda_1, \dots, \Lambda_k)$ is a solution to (1.1). We introduce the following (forward) SDE (the argument t is suppressed):

(3.4)
$$\begin{cases} dx = \Big[A - BK^{-1}(B'P + \sum_{j=1}^k D_j'(P\,C_j + \Lambda_j))\Big]x\,dt \\ \qquad\quad + \sum_{j=1}^k \Big[C_j - D_jK^{-1}(B'P + \sum_{j=1}^k D_j'(P\,C_j + \Lambda_j))\Big]x\,dW^j, \quad t \in [s,\,T], \\ x(s) = \xi. \end{cases}$$

LEMMA 3.1. *Assume that* (A1) *holds. Suppose that the BSDE* (1.1) *has a solution* $(P,\,\Lambda) \in L^\infty_{\mathcal{F}}(0,\,T;\mathbb{R}^{n\times n}) \times L^2_{\mathcal{F}}(0,\,T;\mathbb{R}^{n\times n})^k$, *where* $\Lambda = (\Lambda_1, \dots, \Lambda_k)$, *and the SDE* (3.4) *has a solution* $x(\cdot) \in L^2_{\mathcal{F}}(\Omega;\,C(0,\,T;\,\mathbb{R}^n))$. *Then the stochastic LQ problem* (3.3) *is solvable. A (unique) optimal feedback control is*

$$u^*(t) = -K(t)^{-1}\left[\left(B(t)'P(t) + \sum_{j=1}^k D_j(t)'(P(t)\,C_j(t) + \Lambda_j(t))\right)\,x(t)\right], \tag{3.5}$$

where $K(t) = R(t) + \sum_{j=1}^k D_j(t)'\,P(t)\,D_j(t)$, *and the associated optimal cost is*

$$\inf_{u(\cdot)\in\mathcal{U}} J(s,\xi;u(\cdot)) = \xi'\,P(s)\,\xi, \quad P\text{-a.s. } \omega. \tag{3.6}$$

Proof. This result has been proved in [6, Theorem 3.1] along with [6, Remark 3.1]. The only minor difference is that in [6] the initial state ξ is deterministic, but the argument there works for the case when ξ is $\mathcal{F}_s$-measurable, since ξ is almost surely deterministic under the the probability measure $P(\cdot|\mathcal{F}_s)$. □

THEOREM 3.2. *Assume that* (A1) *holds. If* (1.1) *has two solutions* (P^i, Λ^i), $i = 1, 2$, *such that the corresponding SDE* (3.4) *has solutions* $x_i(\cdot) \in L^2_{\mathcal{F}}(\Omega;\,C(0,\,T;\,\mathbb{R}^n))$, $i = 1, 2$, *then* $(P^1(s), \Lambda^1(s)) = (P^2(s), \Lambda^2(s))$, *P-a.s.* ω, *a.e.* $s \in [0, T]$.

Proof. By Lemma 3.1, $\xi' P^1(s)\xi = \xi' P^2(s)\xi$, for all $(s, \xi) \in [0, T) \times L^2_{\mathcal{F}_s}(\Omega;\mathbb{R}^n)$. Therefore $P^1(s) = P^2(s)$, P-a.s. ω, for all $s \in [0, T]$. Now applying Itô's formula to $|P^1(s) - P^2(s)|^2$ and using their respective equations, we get

$$\begin{aligned} 0 \equiv d|P^1(s) - P^2(s)|^2 &= 2(P^1(s) - P^2(s))d(P^1(s) - P^2(s)) + |\Lambda^1(s) - \Lambda^2(s)|^2 ds \\ &= |\Lambda^1(s) - \Lambda^2(s)|^2 ds. \end{aligned}$$

Hence $\Lambda^1(s) = \Lambda^2(s)$, P-a.s. ω, a.e. $s \in [0, T]$. □

Remark 3.1. Theorem 3.2 essentially gives the uniqueness to the SRE (1.1) among such solutions that (3.4) admits solutions. As stipulated in [6, Remark 3.1] the solvability of (3.4) depends on some moment estimates of its coefficients and may be available in some special cases.

4. Existence: The case when $n = 1$. In this section, we consider the case when $n = 1$ (therefore the unknown P of the SRE (1.1) is a scalar). However, it is still allowed that $m > 1$, $k > 1$. Note that this situation is typically encountered in many financial problems, where the state variable is the wealth, which is scalar-valued; see, e.g., [20, 13, 14].

When $n = 1$, (1.1) reduces to

$$\begin{cases} dP = -\Big\{(2A + \sum_{j=1}^k C_j^2)P + 2\sum_{j=1}^k C_j\Lambda_j + Q \\ \qquad\qquad - \Big[(B + \sum_{j=1}^k C_jD_j)P + \sum_{j=1}^k D_j\Lambda_j\Big]K^{-1} \\ \qquad\qquad \times \Big[(B' + \sum_{j=1}^k C_jD_j')P + \sum_{j=1}^k D_j'\Lambda_j\Big]\Big\}dt \\ \qquad + \sum_{j=1}^k \Lambda_j dW^j, \qquad t \in [0,\, T], \\ P(T) = H, \\ K = R + P\sum_{j=1}^k D_j'D_j > 0. \end{cases} \tag{4.1}$$

Let us set

$$\alpha = 2A + \sum_{j=1}^k C_j^2, \qquad \beta \equiv (\beta_1, \beta_2, \dots, \beta_k)' = (2C_1, 2C_2, \dots, 2C_k)',$$
$$\Gamma = B + \sum_{j=1}^k C_jD_j, \quad D = (D_1', D_2', \dots, D_k')', \quad \Lambda = (\Lambda_1, \Lambda_2, \dots, \Lambda_k).$$

Then, the above equation can be rewritten as

$$\begin{cases} dP = -\Big\{\alpha P + \Lambda\beta + Q - (\Gamma P + \Lambda D)(R + PD'D)^{-1}(\Gamma P + \Lambda D)'\Big\}dt \\ \qquad + \Lambda dW, \qquad t \in [0, T], \\ P(T) = H, \\ R + PD'D > 0. \end{cases} \tag{4.2}$$

4.1. Standard case. The standard case (i.e., $H \geq 0$, $R > 0$, $Q \geq 0$) can be treated as an application of Lemma 2.1. In fact, we are going to study this case for a generalized version of (4.2), which in turn will be useful for the case when R is possibly indefinite (see section 4.3 below). To be specific, we add a parameter Δ, which is an m-dimensional-row-vector-valued, essentially bounded $\mathcal{F}_t$-adapted process, in (4.2):

$$\begin{cases} dP = -\Big\{\alpha P + \Lambda\beta + Q \\ \qquad - (\Gamma P + \Lambda D + \Delta)(R + PD'D)^{-1}(\Gamma P + \Lambda D + \Delta)'\Big\}dt \\ \qquad + \Lambda dW, \qquad t \in [0, T], \\ P(T) = H, \\ R + PD'D > 0. \end{cases} \tag{4.3}$$

THEOREM 4.1. *Assume that $H \geq 0$, $R > 0$, $D'D > 0$, $Q - \Delta R^{-1}\Delta' \geq 0$, $R^{-1} \in L^\infty_{\mathcal{F}}(0,\, T; \mathbb{R}^{m\times m})$, and $(D'D)^{-1} \in L^\infty_{\mathcal{F}}(0,\, T; \mathbb{R}^{m\times m})$. Then* (4.3) *admits a bounded nonnegative solution. In particular, the Riccati equation* (4.2) *admits a bounded nonnegative solution if $H \geq 0$, $R > 0$, $D'D > 0$, $Q \geq 0$, $R^{-1} \in L^\infty_{\mathcal{F}}(0,\, T; \mathbb{R}^{m\times m})$, and $(D'D)^{-1} \in L^\infty_{\mathcal{F}}(0,\, T; \mathbb{R}^{m\times m})$.*

Proof. Let us consider the following equation:

$$\begin{cases} dP = -F(t, P, \Lambda)dt + \Lambda dW, \qquad t \in [0, T], \\ P(T) = H, \end{cases} \tag{4.4}$$

where

$$\begin{aligned} &F(t, P, \Lambda) \\ &\quad = \alpha(t)P + \Lambda\beta(t) + Q(t) \\ &\qquad - (\Gamma(t)P^+ + \Lambda D(t) + \Delta(t))(R + P^+D'D)^{-1}(\Gamma(t)P^+ + \Lambda D(t) + \Delta(t))'. \end{aligned}$$

In the above, $P^+ = \max(P, 0)$. Noting the following inequalities (by virtue of the assumptions that $R > 0$ and $D'D > 0$),

$$||(R + P^+D'D)^{-1}|| \le ||R^{-1}||, \qquad ||(R + P^+D'D)^{-1}|| \le \frac{||(D'D)^{-1}||}{|P^+|} \ \text{(for } P \neq 0),$$

we can easily check that (4.4) satisfies the assumption of Lemma 2.1. Hence it admits a bounded maximal solution (P, Λ). Moreover, we have (t is suppressed)

$$\begin{aligned} F(t, P, \Lambda) = {} & \alpha P + \Lambda\beta + Q - 2\Delta(R + P^+D'D)^{-1}(\Gamma P^+ + \Lambda D)' \\ & + \Delta[R^{-1} - (R + P^+D'D)^{-1}]\Delta' \\ & - (\Gamma P^+ + \Lambda D)(R + P^+D'D)^{-1}(\Gamma P^+ + \Lambda D)' \\ & + Q - \Delta R^{-1}\Delta'. \end{aligned}$$

On the other hand, the following BSDE

$$\begin{cases} dP = -\Big\{\alpha P + \Lambda\beta - 2\Delta(R + P^+D'D)^{-1}(\Gamma P^+ + \Lambda D)' \\ \qquad + \Delta[R^{-1} - (R + P^+D'D)^{-1}]\Delta' \\ \qquad - (\Gamma P^+ + \Lambda D)(R + P^+D'D\ )^{-1}(\Gamma P^+ + \Lambda D)'\Big\}dt \\ \qquad + \Lambda dW, t \in [0, T], \\ P(T) = 0 \end{cases}$$

satisfies the assumption of Lemma 2.1 and admits a bounded solution $(0, 0)$. Applying Lemma 2.1(ii) to (4.4), we deduce that $P \ge 0$. Hence, P is also a bounded nonnegative solution of (4.3). The assertion regarding the original equation (4.2) is straightforward. □

4.2. Case when $R = 0$. The case when $R = 0$ is studied in [14] under the additional assumptions that $C = 0$ (i.e., $\beta = 0$) and $Q = 0$, which arises naturally in a mean-variance portfolio selection problem. Now we discuss this case without those additional assumptions.

When $R = 0$, (4.2) specializes to

$$\begin{cases} dP = -\Big\{\alpha P + \Lambda\beta + Q - \frac{1}{P}(\Gamma P + \Lambda D)(D'D)^{-1}(\Gamma P + \Lambda D)'\Big\}dt \\ \qquad + \Lambda dW, \qquad t \in [0, T], \\ P(T) = H, \\ P > 0. \end{cases} \tag{4.5}$$

The key idea is that if (4.5) has a bounded solution, then, assuming that $H > 0$, $H^{-1} \in L^\infty(\Omega, \mathcal{F}_T, P; \mathbb{R})$, and $D'D > 0$, $(D'D)^{-1} \in L^\infty_{\mathcal{F}}(0, T; \mathbb{R}^{m\times m})$, the process $(Y, Z) = (\frac{1}{P}, -\frac{\Lambda}{P^2})$ should satisfy the following equation (by virtue of the Itô formula):

$$\begin{cases} dY = -\Big\{[\Gamma(D'D)^{-1}\Gamma' - \alpha]Y + Z[\beta - 2D(D'D)^{-1}\Gamma] \\ \qquad\qquad - QY^2 - \frac{Z[D(D'D)^{-1}D' - I]Z'}{Y}\Big\}dt \\ \qquad + ZdW, \qquad t \in [0, T], \\ Y(T) = \frac{1}{H}. \end{cases} \tag{4.6}$$

THEOREM 4.2. *Suppose that $H > 0$, $H^{-1} \in L^\infty(\Omega, \mathcal{F}_T, P; \mathbb{R})$, $Q \ge 0$, $m = k$, and that $D'D > 0$, $(D'D)^{-1} \in L^\infty_{\mathcal{F}}(0, T; \mathbb{R}^{m\times m})$. Then (4.5) admits a unique bounded, uniformly positive solution.*

Proof. Since $m = k$ and D is invertible, the last term of the drift coefficient of (4.6) vanishes. Thus (4.6) is a type of SRE (1.1) for which Lemma 2.2 applies. So it has a unique bounded nonnegative solution. Rewrite this equation as

$$\begin{cases} dY = -\Big\{[\Gamma(D'D)^{-1}\Gamma' - \alpha - QY]Y + Z[\beta - 2D(D'D)^{-1}\Gamma]\Big\}dt \\ \qquad\quad + ZdW, \qquad t \in [0,T], \\ Y(T) = \frac{1}{H}. \end{cases} \tag{4.7}$$

Put

$$\tilde{\alpha} = \Gamma(D'D)^{-1}\Gamma' - \alpha - QY, \quad \tilde{\beta} = \beta - 2D(D'D)^{-1}, \quad \tilde{W}_t = W_t - \int_0^t \tilde{\beta}(s)ds.$$

Then from the well-known Girsanov theorem, $\tilde{W}$ is a Brownian motion under a certain probability measure $\tilde{P}$. The above equation now becomes

$$\begin{cases} dY = -\tilde{\alpha}Y dt + Z d\tilde{W}, \qquad t \in [0,T], \\ Y(T) = \frac{1}{H}. \end{cases}$$

Hence,

$$Y(t) = \tilde{E}\left(\frac{1}{H} e^{\int_t^T \tilde{\alpha}_s ds} | \mathcal{F}_t\right) \geq \frac{1}{||H||_\infty} e^{-||\tilde{\alpha}||_\infty (T-t)} \geq \frac{1}{||H||_\infty} e^{-||\tilde{\alpha}||_\infty T} > 0.$$

Thus we can set

$$(P, \Lambda) = \left(\frac{1}{Y}, -\frac{Z}{Y^2}\right).$$

Applying Itô's formula, we deduce easily that (P, Λ) is a bounded solution of (4.5) with P uniformly positive. Finally, the uniqueness follows from that of (4.7). □

Remark 4.1. Under the assumption that $D'D$ is invertible, it is necessary that $m \leq k$ ($m = \text{rank}(D'D) \leq \text{rank } D \leq k$). When $m < k$, it means that, in the context of the stochastic control that leads to the SRE (1.1), the number of the independent controllable directions is less than that of the independent random sources. In the special case of a stochastic market, $m < k$ implies that the number of the stocks available for selection is less than that of the independent random sources that constitute the market or, in other words, the market is incomplete. Hence, assuming $m = k$ (and $D'D$ is invertible) really stipulates that we are in the realm of the complete market in the context of the finance problem. On the other hand, when $m < k$, then (4.6) suggests that one should handle the last term of its drift coefficient (in particular, the matrix-valued process $I - D(D'D)^{-1}D'$, which is only positive semidefinite with possible zero eigenvalues).

Remark 4.2. In [11], results similar to Theorems 4.1 and 4.2 are derived using approximation techniques which are quite involved. Here we provide completely different yet much simpler proofs. It should be emphasized, however, that our ultimate objective is to prove the existence for some indefinite SREs, for which Theorems 4.1 and 4.2 will be utilized. See the next subsection for details.

4.3. Case when R is indefinite. The case with an indefinite R is more complicated. Unfortunately we are able to only treat the case when $m = k = 1$. In this

case, (4.2) further simplifies to

$$\begin{cases} dP = -\Big\{\alpha P + \Lambda\beta + Q - \frac{(\Gamma P+\Lambda D)^2}{R+D^2P}\Big\}dt + \Lambda dW, \qquad t \in [0,T], \\ P(T) = H, \\ R + D^2P > 0, \end{cases} \tag{4.8}$$

where

$$\alpha = 2A + C^2, \qquad \beta = 2C, \qquad \Gamma = B + CD.$$

In this subsection we give two sets of sufficient conditions that guarantee the solvability of the Riccati equation (4.8).

THEOREM 4.3. *Assume that* $\Gamma = 0$, $D \neq 0$, $Q - \frac{\alpha R}{D^2} \geq 0$, $\frac{R}{D^2} + H > 0$, *and* $(\frac{R}{D^2} + H)^{-1} \in L^\infty(\Omega, \mathcal{F}_T, P; \mathbb{R})$. *Moreover, assume that* $\frac{R}{D^2}$ *is constant. Then the Riccati equation* (4.4) *admits a unique bounded solution.*

Proof. Let us consider the following BSDE:

$$\begin{cases} dY = -\Big\{ -\alpha Y + \beta Z - (Q - \frac{\alpha R}{D^2})Y^2\Big\}dt + ZdW, \qquad t \in [0,T], \\ Y(T) = \frac{D^2}{R+HD^2}. \end{cases}$$

As in the proof of Theorem 4.2, this equation has a unique bounded solution by virtue of the assumptions. Moreover, we can prove in the same manner that there exists a constant $\delta > 0$ such that $Y \geq \delta$. Now, we set

$$(P, \Lambda) = \left(\frac{1}{Y} - \frac{R}{D^2}, -\frac{Z}{Y^2}\right).$$

Applying Itô's formula and noting that $\frac{R}{D^2}$ is constant, we deduce easily that (P, Λ) is a solution of (4.4), which is bounded. The uniqueness comes from the inverse procedure. □

Remark 4.3. From the above result we can see that the SRE may admit a solution even when *both* Q and R are negative. (In the context of the stochastic LQ control, this amounts to saying that an LQ control may be solvable even when both the running state and control costs are negative.) To see this, take an example where all the assumptions of Theorem 4.3 are satisfied. Moreover, let $\alpha = 2A + C^2 > 0$. Then we see that the critical condition $Q - \frac{\alpha R}{D^2} \geq 0$ may be satisfied even when both Q and R are negative (but then H must be positive and large enough).

In the above result it is assumed that $\Gamma \equiv B + CD = 0$, which is rather strict. The next theorem replaces this condition by others.

THEOREM 4.4. *Assume that there exists a constant* $\epsilon > 0$ *such that*

$$|D| > 0, \;\; H + \frac{R}{D^2} \geq \epsilon, \quad Q + \alpha\left(\epsilon - \frac{R}{D^2}\right) - \frac{\Gamma^2(\epsilon - \frac{R}{D^2})^2}{D^2\epsilon} \geq 0. \tag{4.9}$$

Moreover, assume that $\frac{R}{D^2}$ *is constant, and* $D^{-1} \in L^\infty_{\mathcal{F}}(0, T; \mathbb{R})$. *Then there exists a bounded solution for the Riccati equation* (4.8).

Proof. Consider the following Riccati equation:

$$\begin{cases} dY = -\Big\{\alpha Y + \beta Z + Q + \alpha(\epsilon - \frac{R}{D^2}) - \frac{[\Gamma Y + \Lambda D + \Gamma(\epsilon - \frac{R}{D^2})]^2}{D^2\epsilon + D^2 Y}\Big\}dt \\ \qquad\quad + ZdW, \qquad t \in [0,T], \\ Y(T) = H + \frac{R}{D^2} - \epsilon. \end{cases}$$

This equation is of the same type as (4.3) with $\Delta = \Gamma(\epsilon - \frac{R}{D^2})$. The assumption of Theorem 4.1 is satisfied due to (4.9). Thus by Theorem 4.1 the above equation admits a bounded nonnegative solution $(Y, Z) \in L^\infty_{\mathcal{F}}(0, T; \mathbb{R}) \times L^2_{\mathcal{F}}(0, T; \mathbb{R})$.

We set

$$(P, \Lambda) = \left(Y + \epsilon - \frac{R}{D^2}, Z \right).$$

It is straightforward to verify that (P, Λ) is a bounded solution of the Riccati equation (4.4). Moreover, since $\frac{R}{D^2}$ is a constant, the solution $P = Y + \epsilon - \frac{R}{D^2}$ is bounded. □

Remark 4.4. Condition (4.9) gives the overall requirement for the coefficients of the SRE in order for it to admit a solution. The positiveness/nonnegativeness of *individual* coefficients is no longer required.

Example 4.1. Let us take an example to illustrate Theorem 4.4. Take $H \equiv 1$, $D \equiv 1$, and R being negative with $1 + R \geq \epsilon$. In this case, the conditions (4.9) specialize to

$$Q + (\epsilon - R) \left(\alpha - \Gamma^2 + \frac{R\Gamma^2}{\epsilon} \right) \geq 0. \tag{4.10}$$

Now if $\alpha \equiv 2A + C^2 \leq \Gamma^2 \equiv (B + C)^2$, then Q has to be positive in order for the above inequality holds. However, if $\alpha > \Gamma^2$, then Q can also be negative while still satisfying (4.10). Indeed, as long as $\alpha > \frac{\Gamma^2}{\epsilon}$ one can show that $(\epsilon - R)(\alpha - \Gamma^2 + \frac{R\Gamma^2}{\epsilon}) > 0$. Hence there is some room for Q to be negative.

5. Existence: The case when $n > 1$. In this section, we will investigate the case when the unknown P is a matrix. For technical reasons, we assume that $k = 1$ and $m = n$. Note that even in the standard case (i.e., $Q \geq 0$, $H \geq 0$, $R > 0$), the solvability of the Riccati equation remains generally unsolved. In this section we shall study two cases where $R = 0$ and R is indefinite, respectively.

5.1. Case when $R = 0$. In this subsection, we treat the case when $R = 0$. We need to assume in addition that $C = 0$.

First we have the following technical lemma.

LEMMA 5.1. *We have the following assertions.*

(i) *If X is a square matrix, then $X + X' + 2||X||I \geq 0$.*

(ii) *If X is a positive definite matrix, then $X^{-1} - ||X||^{-1} I \geq 0$.*

Proof. (i) For any column vector x of appropriate size, we have by the Cauchy–Schwarz inequality

$$-(x'Xx + x'X'x) \leq |x'Xx| + |x'X'x| \leq 2||x||^2||X||,$$

or equivalently

$$x'(X + X' + 2||X||I)x \geq 0 \qquad \forall x.$$

This proves the claim.

(ii) For any column vector x of appropriate size, again by the Cauchy–Schwarz inequality

$$(X^{-\frac{1}{2}}x)'X(X^{-\frac{1}{2}}x) \leq ||X||||X^{-\frac{1}{2}}x||^2 = ||X||(x'X^{-1}x),$$

or equivalently

$$x'(X^{-1} - ||X||^{-1}I)x \geq 0 \qquad \forall x.$$

This proves the claim. □

THEOREM 5.2. *Suppose that $Q \geq 0$, $H > 0$, $H^{-1} \in L^{\infty}_{\mathcal{F}_T}(\Omega; \mathbb{S}^n)$, D is nonsingular, and $D^{-1} \in L^{\infty}_{\mathcal{F}}(0,T;\mathbb{R}^{n\times n})$. Then the Riccati equation* (5.1) *admits a unique bounded, uniformly positive definite solution.*

Proof. Let us first assume that $D \equiv I$. Then the Riccati equation becomes

(5.1)
$$\begin{cases} dP = -\Big\{PA + A'P + Q - (PB + \Lambda)P^{-1}(B'P + \Lambda)\Big\}dt + \Lambda dW, \quad t \in [0,T], \\ P(T) = H, \\ P > 0. \end{cases}$$

Consider the following BSDE:

$$\begin{cases} dY = -\Big\{ - AY - YA' - BZ - ZB' + BYB' - YQY\Big\}dt + ZdW, \quad t \in [0,T], \\ Y(T) = H^{-1}. \end{cases}$$

This equation admits a unique bounded positive semidefinite solution according to Lemma 2.2. Rewrite the equation as

$$\begin{cases} dY = -\Big\{Y\tilde{A} + \tilde{A}'Y + \tilde{C}'Y\tilde{C} + Z\tilde{C} + \tilde{C}'Z\Big\}dt + ZdW, \qquad t \in [0,T], \\ Y(T) = H^{-1}, \end{cases} \tag{5.2}$$

where $\tilde{A} = -A' - \frac{1}{2}QY$ and $\tilde{C} = -B'$.

We need to prove that $Y > 0$. To this end, introduce

$$\bar{Y} = ||H||_{\infty}^{-1} e^{-2(T-t)||\tilde{A}||_{\infty}} I, \qquad \bar{Z} = 0,$$

where

$$||\tilde{A}||_{\infty} = \operatorname{esssup}_{(t,\omega)} ||\tilde{A}(t,\omega)||, \qquad ||H||_{\infty} = \operatorname{esssup}_{\omega} ||H(\omega)||.$$

Then $(\bar{Y}, \bar{Z})$ is a bounded solution of the following BSDE:

$$\begin{cases} d\bar{Y} = -\Big\{ - 2||\tilde{A}||_{\infty}\bar{Y} + \bar{Z}\tilde{C} + \tilde{C}'\bar{Z}\Big\}dt + \bar{Z}dW, \qquad t \in [0,T], \\ \bar{Y}(T) = ||H||_{\infty}^{-1} I. \end{cases}$$

Now we put

$$\hat{Y} = Y - \bar{Y}, \quad \hat{Z} = Z - \bar{Z}.$$

Then $(\hat{Y}, \hat{Z})$ satisfies the following BSDE:

$$\begin{cases} d\hat{Y} = -\Big\{\hat{Y}\tilde{A} + \tilde{A}'\hat{Y} + \hat{Z}\tilde{C} + \tilde{C}'\hat{Z} + \tilde{C}'\bar{Y}\tilde{C} \\ \qquad\qquad + \bar{Y}(\tilde{A} + \tilde{A}' + 2||\tilde{A}||_{\infty} I)\Big\}dt + \hat{Z}dW, \qquad t \in [0,T], \\ \hat{Y}(T) = H^{-1} - ||H||_{\infty}^{-1} I. \end{cases}$$

By Lemma 5.1, we have

$$\tilde{A}(t,\omega)+\tilde{A}'(t,\omega)+2||\tilde{A}||_\infty I\geq 0 \qquad \forall(t,\omega)$$

and

$$H(\omega)^{-1}-||H||_\infty^{-1}I\geq 0 \qquad \forall\omega.$$

It follows then from Lemma 2.2 that

$$\hat{Y}\geq 0,$$

namely,

$$Y\geq \bar{Y}>0.$$

Now set

$$(P,\Lambda)=(Y^{-1},-Y^{-1}ZY^{-1}).$$

It is easy to check, via the Itô formula, that (P,Λ) is a bounded, uniformly positive definite solution of (5.1). Again, the uniqueness follows from the inverse procedure.

Now, for a general nonsingular D, the Riccati equation can be rewritten as

$$\begin{cases} dP=-\Big\{PA+A'P+Q-(PBD^{-1}+\Lambda)P^{-1}((D^{-1})'B'P+\Lambda)\Big\}dt \\ \qquad\quad +\Lambda dW, \qquad t\in[0,T], \\ P(T)=H, \\ P>0. \end{cases}$$

This is in the same form of (5.1). The proof is completed. □

5.2. Case when $B=0$, $C=0$. In this subsection we study another special case when $B=0$, $C=0$. We also need to assume that R is a constant matrix. However, R is allowed to be *indefinite*.

THEOREM 5.3. *Assume that $Q-RA-A'R\geq 0$, $R+H>0$, $(R+H)^{-1}\in L^\infty(\Omega,\mathcal{F}_T,P;\mathbb{S}^n)$, D is nonsingular, and $D^{-1}\in L^\infty_{\mathcal{F}}(0,T;\mathbb{R}^{n\times n})$. Then the Riccati equation* (5.3) *admits a unique bounded solution.*

Proof. First let us take $D=I$. Then the Riccati equation under consideration is

$$\begin{cases} dP=-\Big\{PA+A'P+Q-\Lambda(R+P)^{-1}\Lambda\Big\}dt+\Lambda dW, \qquad t\in[0,T], \\ P(T)=H, \\ R+P>0. \end{cases} \tag{5.3}$$

Consider the following BSDE:

$$\begin{cases} dY=-\Big\{-AY-YA'-Y(Q-RA-A'R)Y\Big\}dt+ZdW, \qquad t\in[0,T], \\ Y(T)=(R+H)^{-1}. \end{cases} \tag{5.4}$$

By Lemma 2.2, this equation admits a unique bounded solution $(Y,Z)\in L^\infty_{\mathcal{F}}(0,T;\mathbb{S}^n)\times L^2_{\mathcal{F}}(0,T;\mathbb{S}^n)$. As in the proof of the preceding theorem, there exists a

constant $c > 0$, such that $Y \geq cI$. Hence we may set $(P, \Lambda) = (Y^{-1} - R, -Y^{-1}ZY^{-1})$. It is easy to verify via the Itô formula that (P, Λ) is a bounded solution of (5.3). On the other hand, the uniqueness comes from the inverse procedure.

Now for a general nonsingular D, the Riccati equation can be rewritten as

$$\begin{cases} dP = -\Big\{PA + A'P + Q - \Lambda((D^{-1})'RD^{-1} + P)^{-1}\Lambda\Big\}dt + \Lambda dW, \qquad t \in [0, T], \\ P(T) = H, \\ (D^{-1})'RD^{-1} + P > 0. \end{cases}$$

This is in the same form of (5.3), which completes the proof. □

Remark 5.1. In [12] the unique solvability of a *definite* SRE is proved under the assumptions (among others) that $B = C = 0$ and R is uniformly positive definite.

6. Concluding remarks. In this paper we have investigated the stochastic Riccati equation (1.1) with random coefficients where the matrix-valued stochastic processes Q, R, and H are possibly indefinite. The existence of its solution is a prerequisite for solving the corresponding indefinite stochastic linear-quadratic control problem, which in turn has important applications in many applied areas especially in finance. Here we have identified several special cases where the existence is proved. The general global existence remains an extremely challenging open problem.

Acknowledgment. We thank the two anonymous referees for their careful reading of an earlier version of the paper and for their helpful comments.

REFERENCES

[1] M. Ait Rami and X. Zhou, *Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic controls,* IEEE Trans. Automat. Control, 45 (2000), pp. 1131–1143.

[2] B.D.O. Anderson and J.B. Moore, *Optimal Control—Linear Quadratic Methods,* Prentice–Hall, Englewood Cliffs, NJ, 1989.

[3] M. Athens, *Special issues on linear–quadratic–Gaussian problem*, IEEE Trans. Automat. Control, 16 (1971), pp. 527–869.

[4] A. Bensoussan, *Lecture on stochastic control, part* I, in Nonlinear Filtering and Stochastic Control, Lecture Notes in Math. 972, S.K. Mitter and A. Moro, eds., Springer, New York, 1983, pp. 1–39.

[5] J.-M. Bismut, *Linear quadratic optimal stochastic control with random coefficients,* SIAM J. Control Optim., 14 (1976), pp. 419–444.

[6] S. Chen, X. Li, and X.Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs,* SIAM J. Control Optim., 36 (1998), pp. 1685–1702.

[7] S. Chen and J. Yong, Stochastic linear quadratic optimal control problems with random coefficients, Chinese Ann. Math. Ser. B, 21 (2000), pp. 323–338.

[8] S. Chen and J. Yong, *Stochastic linear quadratic optimal control problems,* Appl. Math. Optim., 43 (2001), pp. 21–45.

[9] S. Chen and X.Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs.* II, SIAM J. Control Optim., 39 (2000), pp. 1065–1081.

[10] M. Kobylanski, *Backward stochastic differential equations and partial differential equations with quadratic growth,* Ann. Probab., 28 (2000), pp. 558–602.

[11] M. Kohlmann and S. Tang, *Global adapted solution of one-dimensional backward stochastic Riccati equations, with application to the mean–variance hedging,* Stochastic Process. Appl., 97 (2002), pp. 255–288.

[12] M. Kohlmann and S. Tang, *Multi-Dimensional Backward Stochastic Riccati Equations, with Applications,* Working paper, 2001.

[13] M. Kohlmann and X.Y. Zhou, *Relationship between backward stochastic differential equations and stochastic controls: A linear-quadratic approach,* SIAM J. Control Optim., 38 (2000), pp. 1392–1407.

[14] A.E.B. Lim and X.Y. Zhou, *Mean-variance portfolio selection with random parameters in a complete market*, Math. Oper. Res., 27 (2002), pp. 101–120.
[15] J. Ma and J. Yong, *Forward-Backward Stochastic Differential Equations and Their Applications*, Lecture Notes in Math. 1702, Springer, Berlin, 1999.
[16] E. Pardoux and S. Peng, *Adapted solution of a backward stochastic differential equation,* Systems Control Lett., 14 (1990), pp. 55–61.
[17] S. Peng, *Stochastic Hamilton–Jacobi–Bellman equations,* SIAM J. Control Optim., 30 (1992), pp. 284–304.
[18] D. Yao, S. Zhang, and X.Y. Zhou, *Stochastic linear-quadratic control via semidefinite programming,* SIAM J. Control Optim., 40 (2001), pp. 801–823.
[19] J. Yong and X.Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer, New York, 1999.
[20] X.Y. Zhou and D. Li, *Continuous-time mean-variance portfolio selection: A stochastic LQ framework,* Appl. Math. Optim., 42 (2000), pp. 19–33.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 138–154

SUFFICIENT SECOND-ORDER OPTIMALITY CONDITIONS FOR A PARABOLIC OPTIMAL CONTROL PROBLEM WITH POINTWISE CONTROL-STATE CONSTRAINTS*

A. RÖSCH† AND F. TRÖLTZSCH†

Abstract. An optimal control problem for a semilinear parabolic equation is investigated, where pointwise constraints are given on the control and the state. The state constraints are of mixed (bottleneck) type, where associated Lagrange multipliers can be assumed to be bounded and measurable functions. Based on this property, a second-order sufficient optimality condition is established that considers strongly active constraints.

Key words. optimal control, heat equation, parabolic differential equation, sufficient second-order optimality condition, pointwise mixed control-state constraints, bottleneck constraints

AMS subject classifications. 49K20, 90C48

PII. S0363012902403262

1. Introduction.

In this paper we consider the optimal control problem to minimize

(1.1)
$$F(y,u) = \int_\Omega \omega(x, y(T,x))\, dx + \int_\Sigma \sigma(t,x,y(t,x),u(t,x))\, d\Gamma dt + \int_Q q(t,x,y(t,x))\, dxdt$$

subject to the state equations

$$\begin{aligned} y_t + Ay &= 0 && \text{in } Q = (0,T)\times\Omega, \\ \partial_n y &= b(t,x,y,u) && \text{in } \Sigma = (0,T)\times\Gamma, \\ y(0,x) &= y_o(x) && \text{in } \Omega \end{aligned} \tag{1.2}$$

and subject to the mixed control-state constraints

$$0 \le u(t,x) \le c(t,x) + \gamma(t,x)y(t,x) \quad \text{for } (t,x)\in\Sigma. \tag{1.3}$$

The main task of our paper is to establish second-order sufficient optimality conditions that are close to the associated necessary ones. For control-constrained problems, this issue was discussed quite completely in literature for semilinear elliptic and parabolic equations. We mention Bonnans [3], Casas, Tröltzsch, and Unger [5], Goldberg and Tröltzsch [8], and Heinkenschloss and Tröltzsch [9].

The main difficulty in our problem is the presence of the pointwise control-state constraint $u(t,x) \le c(t,x) + \gamma(t,x)y(t,x)$ in (1.3). If pointwise state constraints are given, then the theory of sufficient second-order conditions is faced with specific difficulties that are still far from being solved. In particular, these problems arise for pointwise state constraints of the type $y(t,x) \le c$. Here, the difficulties are caused by the low regularity of Lagrange multipliers associated with the pointwise state

*Received by the editors February 27, 2002; accepted for publication (in revised form) September 11, 2002; published electronically March 19, 2003.
http://www.siam.org/journals/sicon/42-1/40326.html
†Technische Universität Berlin, Fakultät II – Mathematik und Naturwissenschaften, Str. des 17. Juni 136, D-10623 Berlin, Germany (roesch@math.tu-berlin.de, troeltzsch@math.tu-berlin.de).

constraints. The multipliers are Borel measures. We refer to Casas, Tröltzsch, and Unger [6] and Raymond and Tröltzsch [12].

In our problem (1.1)–(1.3), the situation is slightly simpler, since the constraint (1.3) is a *mixed* control-state constraint of bottleneck type. In this case, the Lagrange multipliers are more regular, and they can be assumed to be functions of $L^\infty(\Sigma)$; see Bergounioux and Tröltzsch [2] or Arada and Raymond [1].

Higher regularity of the multipliers is the main advantage enabling us to establish second-order conditions. Moreover, the second-order conditions should require minimum assumptions, i.e., they should be as close as possible to associated necessary conditions. Often, this task is accomplished by considering strongly active sets (see [7] for control-constrained optimal control of ordinary differential equations). Here, we extend this technique to our case of mixed constraints. Our analysis will show that this is by far not an easy problem. It indicates again that pointwise state constraints of more general type will give rise to even more difficult techniques. Our paper extends the results of [15], where second-order conditions were derived for a weakly singular integral state equation. This problem covered the one-dimensional parabolic case.

The paper is organized as follows: In section 2 we formulate first- and second-order optimality conditions and state the main result. Section 3 contains several auxiliary results. The proof that our second-order conditions are sufficient for local optimality is presented in section 4.

In the paper we use the following notations: By $b'(t,x,y,u)$ and $b''(t,x,y,u)$ we denote the gradient and the Hessian matrix of b with respect to (y,u):

$$b'(t,x,y,u) = \begin{pmatrix} b_y(t,x,y,u) \\ b_u(t,x,y,u) \end{pmatrix}, \quad b''(t,x,y,u) = \begin{pmatrix} b_{yy}(t,x,y,u) & b_{yu}(t,x,y,u) \\ b_{yy}(t,x,y,u) & b_{uu}(t,x,y,u) \end{pmatrix}.$$

Here, the notations $b_y(t,x,y,u) = D_y b(t,x,y,u)$ and $b_{yy}(t,x,y,u) = D_{yy} b(t,x,y,u)$ are used. The norms $|b'|$, $|b''|$ are defined by adding the absolute values of all entries of b' and b'', respectively. By ∂_n we denote the outward normal derivative at Γ.

The following assumptions are required:

(A1) The function $b : \Sigma \times \mathbb{R}^2 \to \mathbb{R}$, $b = b(t,x,y,u)$, satisfies the following Carathéodory condition: b is of class C^2 with respect to (y,u). Moreover, for all $(y,u) \in \mathbb{R}^2$, it is measurable with respect to (t,x).

For all $M > 0$, there are a constant $C_M > 0$ and a continuous, monotone increasing function $\eta \in C(\mathbb{R}^+ \cup \{0\})$ with $\eta(0) = 0$ such that:

$$|b(t,x,0,0)| + |b'(t,x,0,0)| + |b''(t,x,0,0)| \le C_M,$$
$$|b''(t,x,y_1,u_1) - b''(t,x,y_2,u_2)| \le \eta(|y_1 - y_2| + |u_1 - u_2|)$$

for almost all (t,x) and all $|y|, |u|, |y_1|, |y_2|, |u_1|, |u_2| \le M$. The same conditions are imposed for $\sigma = \sigma(t,x,y,u)$.

In addition, we suppose on $\Sigma \times \mathbb{R}^2$ that

$$b_u(t,x,y,u) \ge 0, \quad b_y(t,x,y,u) \le 0, \quad |b(t,x,y_1,u) - b(t,x,y_2,u)| \le L|y_1 - y_2|$$

holds for all $|y|, |u| \le M$ and $y_1, y_2 \in \mathbb{R}$. Notice that b is supposed to be globally Lipschitz with respect to y. The constant L does not depend on M.

(A2) The function $\omega : \Omega \times \mathbb{R} \to \mathbb{R}$, $\omega = \omega(x,y)$, is of class C^2 with respect to y. Furthermore, ω is measurable with respect to x for all $y \in \mathbb{R}$. We assume $\omega(\cdot,0) \in L^\infty(\Omega)$, $\omega_y(\cdot,0) \in L^\infty(\Omega)$, $\omega_{yy}(\cdot,0) \in L^\infty(\Omega)$, and

$$|\omega_{yy}(x,y_1) - \omega_{yy}(x,y_2)| \le \eta(|y_1 - y_2|)$$

for almost all $x \in \Omega$ and all $|y_1| \le M$, $|y_2| \le M$.

The function $q = q(t,x,y)$ is assumed to satisfy the assumptions on ω, where Q is substituted for Ω and (t,x) is substituted for x.

(A3) We assume that $c, \gamma \in C(\bar\Sigma)$ and $c(t,x) > 0$, $\gamma(t,x) \geq 0$ for all $(t,x) \in \bar\Sigma$. In addition, we require $y_0 \in C(\bar\Omega)$, $y_0(x) \geq 0$ for all $x \in \Omega$.

(A4) The domain $\Omega \subset \mathbb{R}^m$ has a boundary Γ of class C^2. The elliptic operator A is defined by

$$Ay(x) = -\sum_{i,j=1}^{m} D_i(a_{ij}(x)D_j y(x)),$$

where $a_{ij} \in C^{1,\nu}$ satisfy, for some positive m_0, the condition of ellipticity

$$\sum_{i,j=1}^{m} a_{ij}(x)\xi_i\xi_j \geq m_0|\xi|^2.$$

Other estimates of b, ω, σ, q and their first derivatives can be derived from (A1), (A2) by the mean value theorem. For convenience, in (A1) we assume a global Lipschitz continuity for b with respect to y. This is not really a strong assumption. The maximum principle of the parabolic equation ensures a priori bounds on the solution of the parabolic equation. Therefore, the Lipschitz continuity with respect to y is only needed on a bounded set that is predetermined by the given data.

2. First- and second-order optimality conditions. First, we introduce the spaces $V = H^1(\Omega)$ and $W(0,T) = \{v \in L^2(0,T;V) : v_t \in L^2(0,T;V^*)\}$. Since $W(0,T)$ is not embedded in the space $C(\bar Q)$, which is needed to differentiate the superposition operators associated with the nonlinear functions ω, σ, q, and b, we fix $\alpha > \frac{m}{2} + 1$, $s > m + 1$ and introduce the state space

$$Y = \{y \in W(0,T) | y_t + Ay \in L^\alpha(Q), \partial_n y \in L^s(\Sigma), y(0) \in C(\bar\Omega)\}$$

endowed with the norm

$$\|y\|_Y = \|y\|_{W(0,T)} + \|y_t + Ay\|_{L^\alpha(Q)} + \|\partial_n y\|_{L^s(\Sigma)} + \|y(0)\|_{C(\bar\Omega)}.$$

Due to the choice of α and s, the embedding of Y into $C(\bar Q)$ is continuous [4], [13].

We denote by $a = a[y,v] : V \times V \to \mathbb{R}$ the bilinear form associated with A:

$$a[y,v] = \int_\Omega \sum_{i,j=1}^{m} a_{ij}(x)D_j y(x)D_i v(x)\, dx.$$

A function $y \in Y$ is said to be a (weak) solution of (1.2) if y satisfies the initial value problem

$$\begin{aligned} \frac{d}{dt}(y(t),v)_{L^2(\Omega)} + a[y(t),v] &= (b(t,\cdot,y(t),u(t)),v)_{L^2(\Gamma)}, \\ y(0,\cdot) &= y_0 \end{aligned} \tag{2.1}$$

for almost all t and all $v \in V$.

LEMMA 2.1. *For each $u \in L^\infty(\Sigma)$, (1.2) admits a unique solution $y \in Y$.*

For the proof we refer to [11] and [13]. In these papers, the authors use a weak solution approach. It is also possible to get a similar result by semigroup techniques; see, for instance, [14].

By Lemma 2.1, a solution mapping $G : L^\infty(\Sigma) \to C(\bar{Q})$ is defined that assigns to $u \in L^\infty(\Sigma)$ the solution y of (1.2). The boundary values of y are of particular importance for us. Thus we define the mapping $S : L^\infty(\Sigma) \to C(\bar{\Sigma})$ with $S = \tau G$ that assigns to u the boundary values of y. It is known from literature that G and S are twice continuously Fréchet-differentiable. Nevertheless, for our further results it is useful to briefly sketch the proof. Let $(\bar{y}, \bar{u}) \in Y \times L^\infty(\Sigma)$ be a fixed reference pair. Later, this couple will stand for a local minimum of (1.1)–(1.3). Below we use the abbreviations $\bar{b}_u = b_u(t, x, \bar{u}(t, x), \bar{y}(t, x))$, $\bar{b}_y = b_y(t, x, \bar{u}(t, x), \bar{y}(t, x))$ with $\bar{y} = S(\bar{u})$. In the same way $\bar{b}_{yy}$, $\bar{b}_{uu}$, $\bar{\sigma}_y$, etc. are defined.

LEMMA 2.2. *The nonlinear mapping $S : L^\infty(\Sigma) \to C(\bar{Q})$ is of class C^1. Its Fréchet derivative $S'(\bar{u})$ at $\bar{u}$ in direction u is given by $S'(\bar{u})u = w|_\Sigma$, where w is the solution of the initial-boundary value problem*

$$\begin{aligned} w_t + Aw &= 0 && \text{in } Q, \\ \partial_n w - \bar{b}_y\, w &= \bar{b}_u\, u && \text{in } \Sigma, \\ w(0, x) &= 0 && \text{in } \Omega. \end{aligned} \tag{2.2}$$

Proof. Let w be the solution of (2.2) with $u = \tilde{u} - \bar{u}$ and set

$$z := \tilde{y} - \bar{y} - w = S(\tilde{u}) - S(\bar{u}) - w. \tag{2.3}$$

Next, we perform a Taylor expansion for $b(t, x, \tilde{u}(t, x), \tilde{y}(t, x))$:

$$\begin{aligned} b(t, x, \tilde{u}(t, x), \tilde{y}(t, x)) &= b(t, x, \bar{u}(t, x), \bar{y}(t, x)) + \bar{b}_u(\tilde{u}(t, x) - \bar{u}(t, x)) \\ &\quad + \bar{b}_y(\tilde{y}(t, x) - \bar{y}(t, x)) + r(t, x). \end{aligned} \tag{2.4}$$

The remainder term $r = r(t, x)$ depends on the point $\bar{u}$ and on the direction h. It is known that

$$\frac{\|r(\bar{u}, h)\|_{L^\infty(\Sigma)}}{\|h\|_{L^\infty(\Sigma)}} \to 0 \quad \text{as } \|h\|_{L^\infty(\Sigma)} \to 0. \tag{2.5}$$

One can easily verify that z solves the initial-boundary value problem

$$\begin{aligned} z_t + Az &= 0 && \text{in } Q, \\ \partial_n z - \bar{b}_y\, z &= r && \text{in } \Sigma, \\ z(0, x) &= 0 && \text{in } \Omega. \end{aligned} \tag{2.6}$$

The estimate (2.5) of the remainder r implies a similar property for z,

$$\frac{\|z(\bar{u}, h)\|_{C(\bar{Q})}}{\|h\|_{L^\infty(\Sigma)}} \to 0 \quad \text{as } \|h\|_{L^\infty(\Sigma)} \to 0, \tag{2.7}$$

and the differentiability of S is readily seen from $S(\tilde{u}) = S(\bar{u}) + w + z(\bar{u}, h)$. ∎

It is possible to extend the operator $S'(\bar{u})$ to a linear continuous operator in $\mathcal{L}(L^2(\Sigma))$. From now on, we consider $S'(\bar{u})$ in this way. The known property

$$\frac{\|r(\bar{u}, h)\|_{L^2(\Sigma)}}{\|h\|_{L^2(\Sigma)}} \to 0 \quad \text{as } \|h\|_{L^\infty(\Sigma)} \to 0$$

(see Maurer [10]) implies a similar property for $z|_\Sigma$:

$$\frac{\|z|_\Sigma(\bar{u}, h)\|_{L^2(\Sigma)}}{\|h\|_{L^2(\Sigma)}} \to 0 \quad \text{as } \|h\|_{L^\infty(\Sigma)} \to 0.$$

Next, we introduce the L^2-adjoint operator $S'(\bar{u})^* \in \mathcal{L}(L^2(\Sigma))$. This operator is given by $S'(\bar{u})^*\mu = \varphi|_\Sigma$, where φ is the solution of the well-posed parabolic backward problem

$$\begin{aligned} -\varphi_t + A^*\varphi &= 0 && \text{in } Q, \\ \partial_n\varphi - \bar{b}_y\ \varphi &= \bar{b}_u\ \mu && \text{in } \Sigma, \\ \varphi(T,x) &= 0 && \text{in } \Omega, \end{aligned} \tag{2.8}$$

where A^* is the formal adjoint operator to A. In all that follows let $(\bar{y},\bar{u})$ be a locally optimal reference solution of (1.1)–(1.3). Let us set up the associated first-order necessary optimality conditions in form of a Karush–Kuhn–Tucker-type theorem.

To this aim, we introduce the Lagrange functional $L : Y \times L^\infty(\Sigma) \times Y \times L^\infty(\Sigma)^2 \to \mathbb{R}$,

$$\begin{aligned} L(y,u,p,\mu_1,\mu_2) = F(y,u) &- \int_Q (y_t + Ay)p\ dxdt - \int_\Sigma (\partial_n y - b)p\ d\Gamma dt \\ &- \int_\Sigma \mu_1 u\ d\Gamma dt + \int_\Sigma (u - c - \gamma y)\mu_2\ d\Gamma dt, \end{aligned}$$

where $d\Gamma$ denotes the Lebesgue surface measure induced on Γ with respect to x.

Let us now comment on this choice for L. The heat equation (1.2) is considered in Y, while the inequality constraints (1.3) are posed in $L^\infty(\Sigma)$. Knowing the general Karush–Kuhn–Tucker theory in Banach spaces, one expects associated Lagrange multipliers $p \in Y^*$ and $\mu_i \in (L^\infty(\Sigma))^*$, together with a related quite complicated Lagrange functional. In contrast to this, special techniques for optimal control problems of bottleneck type have shown that, under natural assumptions, the Lagrange multipliers can be expressed by regular functions, i.e., $p \in W(0,T)\cap C(\bar{Q})$ and $\mu_i \in L^\infty(\Sigma)$; see Bergounioux and Tröltzsch [2] and Arada and Raymond [1]. This well-known advantage of bottleneck-type problems is our key idea to establish special second-order sufficient optimality conditions, which are hardly to be expected for $\mu_i \in (L^\infty(\Sigma))^*$. The existence of such regular multipliers can be shown under a Slater-type condition and the assumption $\gamma(t,x) \geq 0$. Here, the nonnegativity of γ plays a crucial role.

Therefore we are justified to *assume* that an adjoint state $\bar{p} \in W(0,T)\cap C(\bar{Q})$ and Lagrange multipliers $\bar{\mu}_i \in L^\infty(\Sigma)$ exist such that $(\bar{y},\bar{u},\bar{p},\bar{\mu}_1,\bar{\mu}_2)$ satisfies the following first-order necessary optimality system (FON):

$$\text{(FON)} \quad \begin{cases} D_y L(\bar{y},\bar{u},\bar{p},\bar{\mu}_1,\bar{\mu}_2) &= 0, \\ D_u L(\bar{y},\bar{u},\bar{p},\bar{\mu}_1,\bar{\mu}_2) &= 0, \\ \text{and for almost all } (t,x) \in \Sigma & \\ \bar{\mu}_1(t,x) &\geq 0, \\ \bar{\mu}_2(t,x) &\geq 0, \\ \bar{u}(t,x)\bar{\mu}_1(t,x) &= 0, \\ (\bar{u}(t,x) - c(t,x) - \gamma(t,x)\bar{y}(t,x))\bar{\mu}_2(t,x) &= 0. \end{cases}$$

The last two conditions of (FON) are the well-known *complementary slackness conditions*. They imply $\bar{\mu}_1(t,x) > 0 \ \Rightarrow \bar{u}(t,x) = 0$ and $\bar{\mu}_2(t,x) > 0 \ \Rightarrow \bar{u}(t,x) = c(t,x) + \gamma(t,x)\bar{y}(t,x)$. Let us express these optimality conditions also in terms of

partial differential equations. As it is well known, the first equation of (FON) is represented by the adjoint equation

$$\begin{aligned} -\bar{p}_t + A^*\bar{p} &= q_y(t,x,\bar{y}) && \text{in } Q,\\ \partial_n\bar{p} - b_y(t,x,\bar{y},\bar{u})\bar{p} &= \sigma_y(t,x,\bar{y},\bar{u}) - \gamma\bar{\mu}_2 && \text{in } \Sigma,\\ \bar{p}(T,x) &= \omega_y(x,\bar{y}(T,x)) && \text{in } \Omega. \end{aligned} \tag{2.9}$$

The second equation of (FON) is equivalent to

$$\sigma_u(t,x,\bar{y},\bar{u}) + b_u(t,x,\bar{y},\bar{u})\bar{p} - \bar{\mu}_1 + \bar{\mu}_2 = 0. \tag{2.10}$$

Next, we discuss a sufficient second-order optimality condition (SSC). For this purpose, we define *strongly active sets* and the associated *critical subspace*. Assume that $(\bar{y},\bar{u},\bar{p},\bar{\mu}_1,\bar{\mu}_2)$ fulfills (FON).

DEFINITION 2.3. *Let $\delta_1,\delta_2 > 0$ be real numbers and $\bar{\mu}_1,\bar{\mu}_2 \in L^\infty(\Sigma)$ be the Lagrange multipliers introduced in* (FON). *The sets*

$$A_1(\delta_1) := \{(t,x) \in \Sigma : \bar{\mu}_1(t,x) \geq \delta_1\}, \tag{2.11}$$

$$A_2(\delta_2) := \{(t,x) \in \Sigma : \bar{\mu}_2(t,x) - (S'(\bar{u})^*\gamma\bar{\mu}_2)(t,x) \geq \delta_2\} \tag{2.12}$$

are called strongly active sets.

All further arguments hold true for an arbitrary choice of δ_1 and δ_2. Later, these numbers will be chosen such that a second-order sufficient optimality condition is satisfied.

On $A_1(\delta_1)$ we have $\bar{\mu}_1 > 0$; hence the complementary slackness conditions yield $\bar{u}(t,x) = 0$. On $A_2(\delta_2)$, it holds $\bar{\mu}_2 > \delta_2 + S'(\bar{u})^*\gamma\bar{\mu}_2 \geq \delta_2 > 0$, hence the complementary slackness conditions give $\bar{u}(t,x) = c(t,x) + \gamma(t,x)\bar{y}(t,x)$. Notice that $A_1(\delta_1) \cap A_2(\delta_2) = \emptyset$, because $\bar{y} \geq 0$ and

$$c(t,x) + \gamma(t,x)\bar{y}(t,x) \geq c(t,x) > 0$$

holds true. Therefore we cannot have $\bar{u} = 0$ and $\bar{u} = c + \gamma\bar{y}$ at the same time.

DEFINITION 2.4. *We say that $(y,u) \in C(\bar{Q}) \times L^\infty(\Sigma)$ belongs to the critical subspace if*

$$u = 0 \qquad \text{on } A_1, \tag{2.13}$$

$$u = \gamma y|_\Sigma \qquad \text{on } A_2 \tag{2.14}$$

and

$$\begin{aligned} y_t + Ay &= 0 && \text{in } Q,\\ \partial_n y - \bar{b}_y\, y &= \bar{b}_u\, u && \text{in } \Sigma,\\ y(0,x) &= 0 && \text{in } \Omega. \end{aligned} \tag{2.15}$$

Notice that (2.15) implies $y|_\Sigma = S'(\bar{u})u$. In (2.12), the expression $S'(\bar{u})^*\gamma\bar{\mu}_2$ can be evaluated by solving the backward problem

$$\begin{aligned} -\kappa_t + A^*\kappa &= 0 && \text{in } Q,\\ \partial_n\kappa - \bar{b}_y\, \kappa &= \bar{b}_u\, \gamma\,\bar{\mu}_2 && \text{in } \Sigma,\\ \kappa(T,x) &= 0 && \text{in } \Omega. \end{aligned} \tag{2.16}$$

The boundary values of κ deliver the desired expression, $\kappa|_\Sigma = S'(\bar u)^*\gamma\bar\mu_2$. Knowing κ, it is easy to determine the strongly active set A_2.

Before we formulate the second-order sufficient optimality condition, we mention for convenience the explicit expression of $L''_{(u,y)}(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2)[h_y,h_u]^2$:

$$\begin{aligned} L''_{(u,y)}(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2)[h_y,h_u]^2 &= \int_\Omega \bar\omega_{yy}h_y^2\,dx + \int_Q \bar q_{yy}h_y^2\,dxdt \\ &\quad + \int_\Sigma (\bar\sigma_{yy}h_y^2 + 2\bar\sigma_{yu}h_yh_u + \bar\sigma_{uu}h_u^2)\,d\Gamma dt \\ &\quad + \int_\Sigma (\bar b_{yy}h_y^2 + 2\bar b_{yu}h_yh_u + \bar b_{uu}h_u^2)\bar p\,d\Gamma dt. \end{aligned} \tag{2.17}$$

Here, h_y, h_u denote arbitrary increments of y and u, respectively. Now we state the main result of our paper, the second-order sufficient condition.

(SSC). There exist positive numbers δ, δ_1, δ_2 such that the coercivity condition

$$L''_{(u,y)}(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2)[h_y,h_u]^2 \ge \delta\|h_u\|^2_{L^2(\Sigma)} \tag{2.18}$$

holds true for all (h_y,h_u) belonging to the critical subspace defined upon δ_1, δ_2.

THEOREM 2.5 (second-order sufficiency). *Assume that $(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2)$ fulfills the first-order optimality system* (FON). *If the second-order condition* (SSC) *is satisfied, then there exist $\delta_s>0$ and $\varepsilon>0$ such that the quadratic growth condition*

$$F(y,u) - F(\bar y,\bar u) \ge \delta_s\|u-\bar u\|^2_{L^2(\Sigma)} \tag{2.19}$$

holds for all admissible pairs (y,u) with $\|u-\bar u\|_{L^\infty(\Sigma)}<\varepsilon$. Therefore, $\bar u$ is a locally optimal control in the norm of $L^\infty(\Sigma)$.

The proof is contained in section 4.

3. Auxiliary results.

LEMMA 3.1. *Let $\beta\in L^\infty(\Sigma)$ and $f\in L^2(\Sigma)$ be given and let v be the solution of the initial boundary value problem*

$$\begin{aligned} v_t + Av &= 0 && \text{in } Q,\\ \partial_n v + \beta v &= f && \text{in } \Sigma,\\ v(0,x) &= 0 && \text{in } \Omega. \end{aligned}$$

If $f\ge 0$ holds a.e. on Σ, then also $v\ge 0$ holds true a.e. on Σ.

For a proof of this *comparison principle* we refer to Raymond and Zidani [13].

DEFINITION 3.2. *A continuous linear operator A in $\mathcal{L}(L^2(\Sigma))$ is said to be nonnegative if $u\ge 0$ a.e. on Σ implies $Au\ge 0$ a.e. on Σ. In this case, we write $A\ge 0$.*

LEMMA 3.3 (comparison principle). *Under the assumptions* (A1)–(A4), *the nonnegativity properties*

$$S'(\bar u) \ge 0, \tag{3.1}$$

$$(I-\gamma S'(\bar u))^{-1} \ge 0 \tag{3.2}$$

hold true.

Proof. The operator $S'(\bar u): u\mapsto w_\Sigma$, is defined upon (2.2). In (A1) we have assumed $b_u\ge 0$. Hence Lemma 3.1, applied with $\beta=-\bar b_y$, $f=\bar b_u u$, yields that $u\ge 0$ implies $w|_\Sigma\ge 0$ and (3.1) is proved.

To prove (3.2), we apply Lemma 3.1 to the system

$$\begin{aligned} v_t + Av &= 0 && \text{in } Q, \\ \partial_n v - (\bar{b}_y + \gamma \bar{b}_u)v &= \bar{b}_u u && \text{in } \Sigma, \\ v(0,x) &= 0 && \text{in } \Omega. \end{aligned} \tag{3.3}$$

Invoking Lemma 3.1 again, the implication $u \geq 0 \Rightarrow v|_\Sigma \geq 0$ holds. A comparison of (3.3) with (2.2) shows that

$$S'(\bar{u})(\gamma v + u) = v$$

holds. Setting $z = \gamma v + u$, we get $z = \gamma S'(\bar{u})(\gamma v + u) + u = \gamma S'(\bar{u})z + u$, hence $z = (I - \gamma S'(\bar{u}))^{-1}u$. Thanks to the implication $u \geq 0 \Rightarrow v|_\Sigma \geq 0$, (A3), and $z = \gamma v + u$, we obtain $u \geq 0 \Rightarrow z \geq 0$. This proves (3.2). □

COROLLARY 3.4. *The property* (3.1) *extends to the adjoint operator* $S'(\bar{u})^*$,

$$S'(\bar{u})^* \geq 0. \tag{3.4}$$

In what follows, we repeatedly need controls u defined as follows: Let M_1, M_2 be disjoint measurable subsets of Σ such that $M_1 \cup M_2 = \Sigma$, and let $f \in L^\infty(\Sigma)$ be given. Then we would like to define u by

$$u(t,x) = \begin{cases} f(t,x) & \text{on } M_1, \\ f(t,x) + \gamma(t,x)(S'(\bar{u})u)(t,x) & \text{on } M_2. \end{cases} \tag{3.5}$$

The next lemma shows that this setting is correct.

LEMMA 3.5. *For all disjoint measurable subsets* M_1, M_2 *of* Σ *with* $M_1 \cup M_2 = \Sigma$ *and all* $f \in L^\infty(\Sigma)$, *there is exactly one function* $u \in L^\infty(\Sigma)$ *that satisfies condition* (3.5). *In addition, the implication*

$$f \geq 0 \quad \Rightarrow \quad u \geq 0 \tag{3.6}$$

holds true. Moreover, the estimates

$$\|u\|_{L^\infty(\Sigma)} \leq c_1 \|f\|_{L^\infty(\Sigma)}, \tag{3.7}$$

$$\|u\|_{L^2(\Sigma)} \leq c_2 \|f\|_{L^2(\Sigma)} \tag{3.8}$$

hold with certain constants c_1, c_2 *that do not depend on* M_1, M_2, *and* f.

Proof. Suppose that $u \in L^\infty(\Sigma)$ satisfies (3.5). Put $v := S'(\bar{u})u$. Then v satisfies the heat equation with homogeneous initial data and the boundary condition

$$\partial_n v - \bar{b}_y v = \begin{cases} \bar{b}_u f & \text{on } M_1, \\ \bar{b}_u(f + \gamma v) & \text{on } M_2; \end{cases} \tag{3.9}$$

that is,

$$\partial_n v - (\bar{b}_y v + \chi_{M_2} \bar{b}_u \gamma)v = \bar{b}_u f \quad \text{on } \Sigma. \tag{3.10}$$

This solution v is unique. Therefore, if u satisfies (3.5), then $v = S'(\bar{u})u$ is unique; hence u is unique, because of

$$u = \begin{cases} f & \text{on } M_1, \\ f + \gamma v & \text{on } M_2. \end{cases} \tag{3.11}$$

On the other hand, starting from M_1, M_2, and f, the solution v of the heat equation with homogeneous initial data and boundary condition (3.10) is defined, and the function u given by (3.11) satisfies (3.5), since, by definition of v, $u = S'(\bar u)v$.

Finally, by Lemma 3.1 applied to (3.10) with $f := \bar b_u f$, the relation $f \ge 0$ implies $v \ge 0$, hence also $u \ge 0$. The estimates (3.7) and (3.8) follow immediately from those for v and (3.11). □

To prove the main result, we later have to compare the reference pair $(\bar y, \bar u)$ with another admissible pair (y,u), where $y = S(u)$. Then we have to estimate the difference

$$y - \bar y = S(u) - S(\bar u) = S'(\bar u)(u - \bar u) + r_1(\bar u, u - \bar u),$$

where r_1 stands for the associated first-order remainder term. In the following, we denote for short the remainder $r_1(\bar u, u - \bar u)$ and the derivative $S'(\bar u)$ by r_1 and S', respectively, if there is no risk of notational confusion.

Before continuing our analysis of second-order sufficiency, we briefly discuss the main difficulties and our main ideas to resolve them. We start with the case of pure control constraints, which is covered by our setting for $\gamma(t,x) \equiv 0$. Then the constraints are simple box constraints,

$$0 \le u(t,x) \le c(t,x).$$

On A_1, we have $\bar u(t,x) \equiv 0$, hence $u - \bar u \ge 0$ on A_1, while $\bar u(t,x) = c(t,x)$ holds on A_2; thus $u - \bar u \le 0$ on A_2. The associated terms in the Lagrange functional can be estimated as

$$\begin{aligned}\int_{A_1} \bar\mu_1(u - \bar u)\, d\Gamma dt - \int_{A_2} \bar\mu_2(u - \bar u)\, d\Gamma dt &\ge \int_{A_1} \delta_1(u - \bar u)\, d\Gamma dt + \int_{A_2} \delta_2(u - \bar u)\, d\Gamma dt \\ &= \delta_1 \|u - \bar u\|_{L^1(A_1)} + \delta_2 \|u - \bar u\|_{L^1(A_2)}.\end{aligned}$$

In the proof of the sufficiency theorem the L^1-norms on the right-hand side will compensate for the lack of coercivity, since (2.18) does not help on $A_1 \cup A_2$.

Now we return to the given mixed control-state constraints

$$0 \le u(t,x) \le c(t,x) + \gamma(t,x)\, y(t,x).$$

To simplify our explanation, assume for a while that the control-state mapping is linear. This holds for $y_0 = 0$ functions b being linear with respect to y and u. Then $S' = S$; hence

$$0 \le u \le c + \gamma S' u \tag{3.12}$$

holds for any admissible control u. On A_1, we have again $0 = \bar u \le u$, hence $u - \bar u \ge 0$ on A_1. However, in contrast to the case of pure control constraints, the relation $u \le \bar u$ cannot be expected on A_2 now. If $u > \bar u$ holds somewhere on $\Sigma \setminus A_2$, then $S'u > S'\bar u$ can hold on A_2. Then the right-hand side of (3.12) is greater than $c + \gamma S'\bar u$ and $u > \bar u$ can happen.

To overcome this difficulty, we represent u in the form $u = u_1 + u_2$, such that $u_1 \le \bar u$ can be shown on A_2 and u_2 stands for the additional margin of freedom that is caused by $u > \bar u$ outside of A_2. Hence we split u in two parts, $u = u_1 + u_2$ on Σ, where

$$\begin{aligned} &u_1 = \bar u, && u_2 = u - \bar u && \text{on } \Sigma \setminus A_2, \\ &u_2 = \gamma(S'u_2 + r_1), && u_1 = u - u_2 && \text{on } A_2. \end{aligned} \tag{3.13}$$

The functions u_1 and u_2 are well defined. To see this, we apply Lemma 3.5, where $M_1 = \Sigma \setminus A_2$ and $M_2 = A_2$. On M_1, u_2 is given by $u - \bar{u}$. On M_2, $u_2 = \gamma r_1 + \gamma S' u_2$; hence

$$u_2 = \begin{cases} f & \text{on } M_1, \\ f + \gamma S' u_2 & \text{on } M_2, \end{cases}$$

where $f = u - \bar{u}$ on M_1, $f = \gamma r_1$ on M_2. Then u_2 is well defined by Lemma 3.5. Note that $S'u_2 = S'(\bar{u})(\chi_{M_1}(u - \bar{u}) + \chi_{M_2} u_2)$. From (3.10) and the properties of the remainder r_1 we easily get

$$\|u_2\|_{L^\infty(\Sigma)} \le c_3 \|u - \bar{u}\|_{L^\infty(\Sigma)}.$$

Therefore, we find

$$\begin{aligned} \|u_1 - \bar{u}\|_{L^\infty(A_2)} &\le \|u - \bar{u}\|_{L^\infty(A_2)} + \|u_2\|_{L^\infty(A_2)} \\ &\le c_4 \|u - \bar{u}\|_{L^\infty(\Sigma)}. \end{aligned} \tag{3.14}$$

LEMMA 3.6. *Assume that $(\bar{y}, \bar{u}, \bar{p}, \bar{\mu}_1, \bar{\mu}_2)$ fulfills the first-order optimality system* (FON) *and take any u satisfying the mixed control-state constraints. Let u be split by $u = u_1 + u_2$ according to formula* (3.13). *Then it holds that*

$$\bar{u} - u_1 \ge 0 \quad \text{a.e. on } \Sigma. \tag{3.15}$$

Proof. Let A_1 and A_2 be the strongly active sets of $\bar{u}$ defined by δ_1 and δ_2. On A_2, the inequality $\bar{\mu}_2 \ge \bar{\mu}_2 - \gamma S'(\bar{u})^* \bar{\mu}_2 \ge \delta_2 > 0$ holds. Therefore, (FON) implies $\bar{u} = c + \gamma \bar{y}$ there. In addition, we know on A_2 that $u - \gamma y \le c = \bar{u} - \gamma \bar{y}$; hence

$$u - \gamma S(u) \le \bar{u} - \gamma S(\bar{u})$$

holds there. In view of this, we find on A_2

$$\begin{aligned} u - \gamma(S(u) - S(\bar{u})) &\le \bar{u}, \\ u - \gamma(S'(\bar{u})(u - \bar{u}) + r_1) &\le \bar{u}, \\ u_1 - \gamma S' u_1 + (u_2 - \gamma(S' u_2 + r_1)) &\le \bar{u} - \gamma S' \bar{u}, \\ u_1 - \gamma S' u_1 &\le \bar{u} - \gamma S' \bar{u}, \\ (I - \gamma S')(u_1 - \bar{u}) &\le 0, \end{aligned} \tag{3.16}$$

where we have inserted the definition of u_2. Outside of A_2, it holds by definition $u_1 = \bar{u}$. We are now again in the situation that was described in Lemma 3.5. Indeed, taking $M_1 := \Sigma \setminus A_2$, $M_2 = A_2$, $f = 0$ on M_1, and $f = (I - \gamma S')(\bar{u} - u_1)$ on M_2, we have $f \ge 0$. Applying (3.6), we obtain

$$\bar{u} - u_1 \ge 0 \quad \text{a.e. on } \Sigma,$$

which is just inequality (3.15). ∎

LEMMA 3.7. *Assume that $(\bar{y}, \bar{u}, \bar{p}, \bar{\mu}_1, \bar{\mu}_2)$ fulfill the first-order optimality system* (FON). *Then the estimates*

$$\int_\Sigma (u - \bar{u}) \bar{\mu}_1 \, d\Gamma dt \ge \frac{\delta_1}{\varepsilon} \|u - \bar{u}\|^2_{L^2(A_1)}, \tag{3.17}$$

$$-\int_\Sigma (u - \bar{u} - \gamma(y - \bar{y})) \bar{\mu}_2 \, d\Gamma dt \ge \frac{\delta_2}{c_4 \varepsilon} \|u_1 - \bar{u}\|^2_{L^2(A_2)} \tag{3.18}$$

are valid for all $\varepsilon > 0$ and all admissible pairs (u, y) satisfying $\|u - \bar{u}\|_{L^\infty(\Sigma)} < \varepsilon$.

Proof. (i) Because of (FON), $\bar\mu_1 > 0$ can only hold if $\bar u = 0$. If $\bar u > 0$, then $\bar\mu_1 = 0$. Moreover, u is admissible, hence $u \ge 0$ and we have almost everywhere

$$(u - \bar u)\bar\mu_1 \ge 0.$$

Therefore we get by (2.11)

$$\int_\Sigma (u - \bar u)\bar\mu_1 \, d\Gamma dt \ge \int_{A_1} (u - \bar u)\bar\mu_1 \, d\Gamma dt \ge \delta_1 \|u - \bar u\|_{L^1(A_1)}.$$

By our assumption, we have $\|u - \bar u\|_{L^\infty(\Sigma)} < \varepsilon$. In particular, this inequality includes $\|u - \bar u\|_{L^\infty(A_1)} < \varepsilon$. Consequently,

$$\int_\Sigma (u - \bar u)\bar\mu_1 \, d\Gamma dt \ge \delta_1 \|u - \bar u\|_{L^1(A_1)} \frac{\|u - \bar u\|_{L^\infty(A_1)}}{\varepsilon} \ge \frac{\delta_1}{\varepsilon} \|u - \bar u\|^2_{L^2(A_1)},$$

and (3.17) is proved.

(ii) Next, we discuss the integral in (3.18). Because of (FON), $\bar\mu_2 > 0$ can only hold for $\bar u - c - \gamma\bar y = 0$. In addition, (y, u) is admissible, hence in particular $u \le c + \gamma y$. Therefore, we obtain almost everywhere

$$-(u - \bar u - \gamma(y - \bar y))\bar\mu_2 \ge 0$$

and

$$-\int_\Sigma (u - \bar u - \gamma(y - \bar y))\bar\mu_2 \, d\Gamma dt \ge -\int_{A_2} (u - \bar u - \gamma(y - \bar y))\bar\mu_2 \, d\Gamma dt.$$

Let us discuss this integral more detailed. Expressing $y - \bar y$ in terms of the controls,

$$\int_{A_2} (u - \bar u - \gamma(y - \bar y))\bar\mu_2 \, d\Gamma dt = \int_{A_2} (u - \bar u - \gamma(S'(\bar u)(u - \bar u) + r_1))\bar\mu_2 \, d\Gamma dt \tag{3.19}$$

is found. The definition of u_1 and u_2 yields on A_2

$$u - \gamma(S'u + r_1) = u_1 + u_2 - \gamma S'u_1 - \gamma S'u_2 - \gamma r_1 = u_1 - \gamma S'u_1.$$

Inserting the last equation in (3.19), we continue by

$$\begin{aligned}
\int_{A_2} (u - \bar u - \gamma(y - \bar y))\bar\mu_2 \, d\Gamma dt &= \int_{A_2} (u_1 - \bar u - \gamma(S'(\bar u)(u_1 - \bar u)))\bar\mu_2 \, d\Gamma dt \\
&= \int_\Sigma (u_1 - \bar u - \gamma(S'(\bar u)(u_1 - \bar u)))\chi_{A_2}\bar\mu_2 \, d\Gamma dt \\
&= \int_\Sigma (u_1 - \bar u)(I - (\gamma S')^*)(\chi_{A_2}\bar\mu_2) \, d\Gamma dt \\
&= \int_{A_2} (u_1 - \bar u)(I - (\gamma S')^*)(\chi_{A_2}\bar\mu_2) \, d\Gamma dt. \qquad (3.20)
\end{aligned}$$

To deduce the last equation, we used $\bar u - u_1 = 0$ outside of A_2. Now we discuss the expression $(I - (\gamma S')^*)(\chi_{A_2}\bar\mu_2)$ in (3.20). On A_2 we have

$$(I - (\gamma S')^*)(\chi_{A_2}\bar\mu_2) = \chi_{A_2}\bar\mu_2 - (\gamma S')^*(\chi_{A_2}\bar\mu_2) = \bar\mu_2 - (\gamma S')^*(\chi_{A_2}\bar\mu_2).$$

Using the nonnegativity of S'^* following from (3.4), together with $\chi_{A_2}\bar\mu_2 \le \bar\mu_2$, we obtain

$$(\gamma S')^*(\chi_{A_2}\bar\mu_2) = (S')^*(\gamma\chi_{A_2}\bar\mu_2) \le (S')^*(\gamma\bar\mu_2) = (\gamma S')^*\bar\mu_2.$$

Combining these results, we continue by

$$(I-(\gamma S')^*)(\chi_{A_2}\bar\mu_2) = \bar\mu_2 - (\gamma S')^*(\chi_{A_2}\bar\mu_2) \ge (I-(\gamma S')^*)\bar\mu_2 \ge \delta_2, \tag{3.21}$$

where the last inequality follows from the definition (2.12) of A_2. Inserting (3.15) and (3.21) in (3.20), we infer

$$\begin{aligned} -\int_{A_2}(u-\bar u-\gamma(y-\bar y))\bar\mu_2\,d\Gamma dt &= -\int_{A_2}(u_1-\bar u)(I-(\gamma S')^*)(\chi_{A_2}\bar\mu_2)\,d\Gamma dt \\ &\ge \delta_2\|\bar u-u_1\|_{L^1(A_2)}. \end{aligned}$$

Invoking again $\|u-\bar u\|_{L^\infty(\Sigma)} < \varepsilon$ and (3.14), we obtain

$$\begin{aligned} -\int_{A_2}(u-\bar u-\gamma(y-\bar y))\bar\mu_2\,d\Gamma dt &\ge \delta_2\|u_1-\bar u\|_{L^1(A_2)}\cdot\frac{\|u-\bar u\|_{L^\infty(A_2)}}{\varepsilon} \\ &\ge \frac{\delta_2}{c_4\varepsilon}\|u_1-\bar u\|^2_{L^2(A_2)}, \end{aligned}$$

implying inequality (3.18). □

If $A_1\cup A_2=\Sigma$, then the critical subspace contains only the zero-function. Then the assumptions of Theorem 2.5 are trivially fulfilled. In this case, (3.17) and (3.18) express the so-called *first-order sufficient optimality conditions*.

4. Second-order sufficient optimality condition. This section includes the proof of the sufficiency Theorem 2.5. We start from an admissible control u in a sufficiently small L^∞-neighborhood of $\bar u$ and have to show that $F(y,u)\ge F(\bar y,\bar u)$. Let us introduce the increments $\delta u := u-\bar u$ and $\delta y := S'(\bar u)\delta u$. We split $\delta u = u_0+u_+$, where

$$\begin{array}{lll} u_0=0, & u_+=\delta u & \text{on } A_1,\\ u_0=\delta u, & u_+=0 & \text{on } \Sigma\setminus(A_1\cup A_2),\\ u_0=\gamma S'(\bar u)u_0, & u_+=\delta u-u_0 & \text{on } A_2. \end{array}$$

Thanks to Lemma 3.5, the definition of u_0 and hence u_+ are correct. We take $M_1=\Sigma\setminus A_2$, $M_2=A_2$, $f:=0$ on $A_1\cup A_2$, $f:=\delta u$ on $\Sigma\setminus(A_1\cup A_2)$. The part u_0 belongs to the critical subspace, while u_+ is the part of δu that accounts for the effects of first-order sufficiency. Furthermore, we define $y_0 := S'u_0$ and $y_+ := S'u_+$. By the linearity of S', we have $\delta y = y_0+y_+$.

Below, we estimate the difference $L(y,u,\bar p,\bar\mu_1,\bar\mu_2)-L(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2)$. Let us write for short $L(y,u)-L(\bar y,\bar u)$, since $(\bar p,\bar\mu_1,\bar\mu_2)$ remains fixed in all the following considerations. We also do not explicitly indicate the point $(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2)$ where all derivatives are taken, i.e., we write $L_u u$ instead of $(D_uL)(\bar y,\bar u,\bar p,\bar\mu_1,\bar\mu_2)u$.

LEMMA 4.1. *Under the assumptions of Theorem* 2.5,

$$L(y,u)-L(\bar y,\bar u)\ge\frac{\delta}{4}\|u_0\|^2_{L^2(\Sigma)}-\frac{c_s}{2}\|u_+\|^2_{L^2(\Sigma)}+r_2+\tilde r_2 \tag{4.1}$$

holds, where r_2, $\tilde r_2$ are second-order remainder terms with

$$\frac{|r_i|}{\|u-\bar u\|^2_{L^2(\Sigma)}}\to 0 \quad \text{if } \|u-\bar u\|_{L^\infty(\Sigma)}\to 0.$$

Proof. Using a Taylor expansion, in view of (FON) we get

$$\begin{aligned} L(y,u) - L(\bar y,\bar u) &= L_u[u-\bar u] + L_y[y-\bar y] + \frac{1}{2}(L_{uu}[u-\bar u]^2 \\ &\quad + 2L_{uy}[u-\bar u, y-\bar y] + L_{yy}[y-\bar y]^2) + r_2 \\ &= \frac{1}{2}(L_{uu}[u-\bar u]^2 + 2L_{uy}[u-\bar u, y-\bar y] + L_{yy}[y-\bar y]^2) + r_2. \end{aligned} \tag{4.2}$$

The following property of the remainder is known:

$$\frac{|r_2(\bar u,h)|}{\|h\|^2_{L^2(\Sigma)}} \to 0 \quad \text{as } \|h\|_{L^\infty(\Sigma)} \to 0.$$

For the proof we refer to [16]. According to the notation of Lemma 3.5, we get $y-\bar y = \delta y + r_1$. Replacing $y-\bar y$ by δy in (4.2), we cause a small error of second order

$$\begin{aligned} \tilde r_2 := \frac{1}{2}(L_{uu}[u-\bar u]^2 + 2L_{uy}[u-\bar u, y-\bar y] + L_{yy}[y-\bar y]^2) \\ - \frac{1}{2}(L_{uu}[\delta u]^2 + 2L_{uy}[\delta u, \delta y] + L_{yy}[\delta y]^2). \end{aligned}$$

It is easy to show that

$$\frac{|\tilde r_2|}{\|u-\bar u\|^2_{L^2(\Sigma)}} \to 0 \quad \text{as } \|u-\bar u\|_{L^\infty(\Sigma)} \to 0.$$

With these notations, we can express (4.2) in the form

$$L(y,u) - L(\bar y,\bar u) = \frac{1}{2}(L_{uu}[\delta u]^2 + 2L_{uy}[\delta u, \delta y] + L_{yy}[\delta y]^2) + r_2 + \tilde r_2. \tag{4.3}$$

We continue by splitting the Lagrange functional,

$$\begin{aligned} L_{uu}[\delta u]^2 + 2L_{uy}[\delta u, \delta y] + L_{yy}[\delta y]^2 &= L_{uu}[u_0]^2 + 2L_{uy}[u_0, y_0] + L_{yy}[y_0]^2 \\ &\quad + L_{uu}[u_+]^2 + 2L_{uy}[u_+, y_+] + L_{yy}[y_+]^2 \\ &\quad + 2L_{uu}[u_0, u_+] + 2L_{uy}[u_0, y_+] \\ &\quad + 2L_{uy}[u_+, y_0] + 2L_{yy}[y_0, y_+]. \end{aligned}$$

As u_0 belongs to the critical subspace, the second-order condition (SSC) yields

$$L''[u_0,y_0]^2 = L_{uu}[u_0]^2 + 2L_{uy}[u_0,y_0] + L_{yy}[y_0]^2 \ge \delta \|u_0\|^2_{L^2(\Sigma)}.$$

The other terms are easily estimated by $\|y_0\|^2_{L^2(\Sigma)} \le \|S'\|^2 \|u_0\|^2_{L^2(\Sigma)}$, $\|y_+\|^2_{L^2(\Sigma)} \le \|S'\|^2 \|u_+\|^2_{L^2(\Sigma)}$, and by means of Young's inequality,

$$\begin{aligned} |L_{uu}[u_+]^2 + 2L_{uy}[u_+, y_+] + L_{yy}[y_+]^2 \\ + 2L_{uu}[u_0, u_+] + 2L_{uy}[u_0, y_+] \\ + 2L_{uy}[u_+, y_0] + 2L_{yy}[y_0, y_+]| \le \frac{\delta}{2}\|u_0\|^2_{L^2(\Sigma)} + c_s \|u_+\|^2_{L^2(\Sigma)}. \end{aligned}$$

In this setting, c_s is a certain (large) constant. Combining the last two results, we arrive at

$$L_{uu}[\delta u]^2 + 2L_{uy}[\delta u, \delta y] + L_{yy}[\delta y]^2 \ge \frac{\delta}{2}\|u_0\|^2_{L^2(\Sigma)} - c_s \|u_+\|^2_{L^2(\Sigma)}.$$

Returning to (4.3), we end up with

$$L(y,u) - L(\bar{y},\bar{u}) \geq \frac{\delta}{4}\|u_0\|^2_{L^2(\Sigma)} - \frac{c_s}{2}\|u_+\|^2_{L^2(\Sigma)} + r_2 + \tilde{r}_2,$$

which is exactly the assertion. □

In the next lemma, we estimate the term $\|u_+\|^2_{L^2(\Sigma)}$ in (4.1).

LEMMA 4.2. *Under the assumptions of Theorem* 2.5,

$$\left(\frac{c_s}{2} + \frac{\delta}{4}\right)\|u_+\|^2_{L^2(\Sigma)} \leq c_5\|u_1 - \bar{u}\|^2_{L^2(A_2)} + c_6\|r_1\|^2_{L^2(\Sigma)} + c_7\|u - \bar{u}\|^2_{L^2(A_1)} \tag{4.4}$$

holds with certain positive constants c_5, c_6, *and* c_7.

Proof. First, we get on A_1

$$\|u_+\|_{L^2(A_1)} = \|\delta u\|_{L^2(A_1)} = \|u - \bar{u}\|_{L^2(A_1)}. \tag{4.5}$$

On the whole set Σ we have

$$u_+ + u_0 = \delta u = u - \bar{u}.$$

We apply the operator $I - \gamma S'$ to this equation and consider the image only on the set A_2. Using $u_0 = \gamma S' u_0$ on A_2, we find

$$u_+ - \gamma S' u_+ = u - \gamma S' u - (\bar{u} - \gamma S'\bar{u}) \text{ on } A_2.$$

Now, u is again replaced by $u_1 + u_2$ (see (3.13)) to obtain on A_2

$$u_+ - \gamma S' u_+ = u_1 - \gamma S' u_1 + u_2 - \gamma S' u_2 - (\bar{u} - \gamma S'\bar{u}).$$

On A_2, by definition, the equation $u_2 - \gamma S' u_2 = r_1$ is satisfied. Therefore, here we are able to continue by

$$u_+ - \gamma S' u_+ = u_1 - \bar{u} - (\gamma S'(\bar{u})(u_1 - \bar{u})) + r_1 \text{ on } A_2.$$

Due to our definitions, $u_+ = \delta u = u - \bar{u}$ holds on A_1. In addition, u_+ vanishes on $\Sigma \setminus (A_1 \cup A_2)$. Therefore, we find

$$u_+ = \begin{cases} u_1 - \bar{u} + \gamma S'(\bar{u})(u_+ - u_1 + \bar{u}) + r_1 & \text{on } A_2, \\ u - \bar{u} & \text{on } A_1, \\ 0 & \text{on } \Sigma \setminus (A_1 \cup A_2). \end{cases}$$

Again we have a construction that was investigated in Lemma 3.5. Setting $M_2 = A_2$, $M_1 = \Sigma \setminus A_2$, and applying (3.11), we get the inequality

$$\|u_+\|_{L^2(\Sigma)} \leq c_2\|f\|_{L^2(\Sigma)},$$

where f is defined by

$$f = \begin{cases} r_1 + (u_1 - \bar{u}) - \gamma S'(\bar{u})(u_1 - \bar{u}) & \text{on } A_2, \\ u - \bar{u} & \text{on } A_1, \\ 0 & \text{on } \Sigma \setminus (A_1 \cup A_2). \end{cases}$$

Therefore, we obtain

$$\|u_+\|_{L^2(\Sigma)} \leq c_2(\|u - \bar{u}\|_{L^2(A_1)} + c_8\|u_1 - \bar{u}\|_{L^2(\Sigma)} + \|r_1\|_{L^2(A_2)}),$$

where the positive constant c_8 is related to $\|S'\|$. Using $\|u_1-\bar u\|_{L^2(\Sigma)} = \|u_1-\bar u\|_{L^2(A_2)}$,

$$\|u_+\|_{L^2(\Sigma)} \le c_9\|u_1-\bar u\|_{L^2(A_2)} + c_2\|r_1\|_{L^2(A_2)} + c_2\|u-\bar u\|_{L^2(A_1)}$$

is found. Young's inequality yields

$$\|u_+\|^2_{L^2(\Sigma)} \le 3c_9\|u_1-\bar u\|^2_{L^2(A_2)} + 3c_2\|r_1\|^2_{L^2(\Sigma)} + 3c_2\|u-\bar u\|^2_{L^2(A_1)}.$$

A multiplication by $(\frac{c_s}{2}+\frac{\delta}{4})$,

$$\left(\frac{c_s}{2}+\frac{\delta}{4}\right)\|u_+\|^2_{L^2(\Sigma)} \le c_5\|u_1-\bar u\|^2_{L^2(A_2)} + c_6\|r_1\|^2_{L^2(\Sigma)} + c_7\|u-\bar u\|^2_{L^2(A_1)},$$

concludes the proof of the lemma. □

Now we are able to prove our main result, Theorem 2.5.

Proof of Theorem 2.5. Inserting (4.4) in (4.1),

$$\begin{aligned} L(y,u)-L(\bar y,\bar u) \ge{}& \frac{\delta}{4}(\|u_0\|^2_{L^2(\Sigma)} + \|u_+\|^2_{L^2(\Sigma)}) + r_2 + \tilde r_2 \\ &- c_7\|u-\bar u\|^2_{L^2(A_1)} - c_5\|u_1-\bar u\|^2_{L^2(A_2)} - c_6\|r_1\|^2_{L^2(\Sigma)} \end{aligned}$$

is obtained. Next, we return to the objective F,

$$\begin{aligned} L(y,u)-L(\bar y,\bar u) &= F(y,u)-F(\bar y,\bar u) \\ &\quad - \int_Q (y_t+Ay)\bar p\,dxdt - \int_\Sigma (\partial_n y - b)\bar p\,d\Gamma dt \\ &\quad + \int_Q (\bar y_t+A\bar y)\bar p\,dxdt + \int_\Sigma (\partial_n \bar y - \bar b)\bar p\,d\Gamma dt \\ &\quad - \int_\Sigma \bar\mu_1 u\,d\Gamma dt + \int_0^T \bar\mu_1\bar u\,d\Gamma dt \\ &\quad + \int_\Sigma (u-c-\gamma y)\bar\mu_2\,d\Gamma dt \\ &\quad - \int_\Sigma (\bar u - c - \gamma\bar y)\bar\mu_2\,d\Gamma dt \\ &= F(y,u)-F(\bar y,\bar u) \\ &\quad - \int_\Sigma \bar\mu_1(u-\bar u)\,d\Gamma dt \\ &\quad + \int_\Sigma (u-\bar u-\gamma(y-\bar y)\bar\mu_2\,d\Gamma dt. \end{aligned}$$

Using Lemma 3.7, we find

$$\begin{aligned} F(y,u)-F(\bar y,\bar u) \ge{}& \frac{\delta}{4}(\|u_0\|^2_{L^2(\Sigma)} + \|u_+\|^2_{L^2(\Sigma)}) + r_2 + \tilde r_2 \\ &+ \left(\frac{\delta_1}{\varepsilon} - c_7\right)\|u-\bar u\|^2_{L^2(A_1)} + \left(\frac{\delta_2}{c_4\varepsilon} - c_5\right)\|u_1-\bar u\|^2_{L^2(A_2)} \\ &- c_6\|r_1\|^2_{L^2(\Sigma)}. \end{aligned} \tag{4.6}$$

Next, $\|\delta u\|_{L^2(\Sigma)} = \|u_0 + u_+\|^2_{L^2(\Sigma)} \le 2\|u_0\|^2_{L^2(\Sigma)} + 2\|u_+\|^2_{L^2(\Sigma)}$ is applied to continue by

$$\begin{aligned} F(y,u) - F(\bar y,\bar u) \ge{}& \frac{\delta}{8}\|\delta u\|^2_{L^2(\Sigma)} + r_2 + \tilde r_2 \\ &+ \left(\frac{\delta_1}{\varepsilon} - c_7\right)\|u-\bar u\|^2_{L^2(A_1)} + \left(\frac{\delta_2}{c_4\varepsilon} - c_5\right)\|u_1-\bar u\|^2_{L^2(A_2)} \\ &- c_6\|r_1\|^2_{L^2(\Sigma)}. \end{aligned} \tag{4.7}$$

Now take ε sufficiently small such that

$$\frac{\delta_1}{\varepsilon} - c_7 \ge 0 \quad \text{and} \quad \frac{\delta_2}{c_4\varepsilon} - c_5 \ge 0.$$

Then we can omit the associated terms in (4.6),

$$F(y,u) - F(\bar y,\bar u) \ge \frac{\delta}{8}\|\delta u\|^2_{L^2(\Sigma)} + r_2 + \tilde r_2 - c_6\|r_1\|^2_{L^2(\Sigma)}. \tag{4.8}$$

Due to the discussions during the proof, all terms of the right-hand side (except the first one) are small with respect to $\|u-\bar u\|^2_{L^2(\Sigma)}$. Therefore

$$F(y,u) - F(\bar y,\bar u) \ge \frac{\delta}{16}\|u-\bar u\|^2_{L^2(\Sigma)} \tag{4.9}$$

holds if $\|u-\bar u\|_{L^\infty(\Sigma)} < \varepsilon$ and ε is sufficiently small. The quadratic growth condition is proved. We can choose $\delta_s = \delta/16$. □

REFERENCES

[1] N. Arada and J.-P. Raymond, *Optimal control problems with mixed control-state constraints*, SIAM J. Control Optim., 39 (2000), pp. 1391–1407.

[2] M. Bergounioux and F. Tröltzsch, *Optimal control of semilinear parabolic equations with state-constraints of bottleneck type*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 595–608.

[3] F. Bonnans, *Second order analysis for control constrained optimal control problems of semilinear elliptic systems*, Appl. Math. Optim., 38 (1998), pp. 303–325.

[4] E. Casas, *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations*, SIAM J. Control Optim., 35 (1997), pp. 1297–1327.

[5] E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for a nonlinear elliptic control problem*, Z. Anal. Anwendungen., 15 (1996), pp. 687–707.

[6] E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for some state-constrained control problems of semilinear elliptic equations*, SIAM J. Control Optim., 38 (2000), pp. 1369–1391.

[7] A. Dontchev, W. Hager, A. Poore, and B. Yang, *Optimality, stability, and convergence in nonlinear control*, Appl. Math. Optim., 31 (1995), pp. 297–326.

[8] H. Goldberg and F. Tröltzsch, *Second-order sufficient optimality conditions for a class of nonlinear parabolic boundary control problems*, SIAM J. Control Optim., 31 (1993), pp. 1007–1025.

[9] M. Heinkenschloss and F. Tröltzsch, *Analysis of an SQP Method for the Control of the Phase Field Equation*, ICAM Report, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1995.

[10] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Study, 14 (1981), pp. 163–177.

[11] J.-P. Raymond, *Nonlinear boundary control of semilinear parabolic problems with pointwise state constraints*, Discrete Contin. Dynam. Systems, 3 (1997), pp. 341–370.

[12] J.-P. Raymond and F. Tröltzsch, *Second order sufficient optimality conditions for nonlinear parabolic control problems with state constraints*, Discrete Contin. Dynam. Systems, 6 (2000), pp. 431–450.

[13] J.-P. Raymond and H. Zidani, *Hamiltonian Pontryagin's principles for control problems governed by semilinear parabolic equations*, Appl. Math. Optim., 39 (1999), pp. 143–177.

[14] A. Rösch and F. Tröltzsch, *An optimal control problem arising from the identification of nonlinear heat transfer laws*, Arch. Control Sci., 1 (1992), pp. 183–195.

[15] A. Rösch and F. Tröltzsch, *Sufficient second order optimality condition for a state-constrained optimal control problem of a weakly singular integral equation*, Numer. Funct. Anal. Optim., 23 (2002), pp. 173–193.

[16] F. Tröltzsch, *Approximation of non–linear parabolic boundary control problem by the Fourier method – convergence of optimal controls*, Optimization, 22 (1991), pp. 83–98.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 155–177

LOCALLY CONVERGENT NONLINEAR OBSERVERS*

ARTHUR J. KRENER† AND WEI KANG‡

Abstract. We introduce a new method for the design of observers for nonlinear systems using backstepping. The method is applicable to a class of nonlinear systems slighter larger than those treated by Gauthier, Hammouri, and Othman [*IEEE Trans. Automat. Control,* 27 (1992), pp. 875–880]. They presented an observer design method that is globally convergent using high gain. In contrast to theirs, our observer is not high gain, but it is only locally convergent. If the initial estimation error is not too large, then the estimation error goes to zero exponentially. A design algorithm is presented.

Key words. nonlinear estimation, nonlinear observers

AMS subject classifications. 93C10, 93B50, 93E11

PII. S0363012900368612

1. Introduction. The problem of estimating the state of a dynamical system from partial and possibly noisy measurements has a long history. In its nonlinear state space form, one assumes that the dynamics satisfies a known nonlinear differential equation with unknown initial condition and the measurement is a known nonlinear function of the state

$$\begin{aligned} \dot{x} &= f(x), \\ x(0) &= x^0, \\ y &= h(x). \end{aligned} \tag{1.1}$$

The linear form of the problem is

$$\begin{aligned} \dot{x} &= Ax, \\ x(0) &= x^0, \\ y &= Cx. \end{aligned} \tag{1.2}$$

One is given an estimate $\hat{x}^0$ of x^0 and the observations $y(s),\ 0 \le s \le t$, up to time t. The problem is to generate an estimate $\hat{x}(t)$ of $x(t)$ in real time, as the process evolves. The estimate should converge to the true state as $t \to \infty$. Ideally the estimation process should be robust to noise both in the dynamics and in the observations, to the initial state error, and also to modeling errors in the functions $f,\ h$. Furthermore, the error should converge to zero quickly.

One way of approaching this problem is to assume that the dynamics, the initial condition, and the observations are corrupted by noises with known distributions and then to find the conditional density of the state given the past observations. If the dynamics and observations are linear functions of the state and if the noises and

*Received by the editors March 2, 2000; accepted for publication (in revised form) September 6, 2002; published electronically March 19, 2003.
http://www.siam.org/journals/sicon/42-1/36861.html

†Department of Mathematics, University of California, Davis, CA 95616-8633 (ajkrener@ucdavis.edu). This author's research was supported in part by AFOSR-49620-95-1-0409 and NSF-DMS-9709452.

‡Department of Mathematics, Naval Postgraduate School, Monterey, CA (wkang@math.nps.navy.mil).

the initial condition are independent and Gaussian, then the conditional density is Gaussian and explicitly computable. Wiener [28] solved this problem for stationary Gaussian processes using the method of spectral factorization. Kalman [12], [13] extended this to nonstationary Gaussian processes and reduced the problem to solving an off-line Riccati equation and an on-line linear differential equation driven by the observations.

When the dynamics and/or observations are nonlinear the unnormalized conditional density satisfies the Zakai equation, a parabolic PDE driven by the observations [4]. It is a very difficult task to accurately compute its solution in real time for all but the smallest state dimensions.

The extended Kalman filter [10] is a widely used alternative method for estimating the state of a nonlinear system. It is obtained by linearizing the nonlinear dynamics and the observation along the trajectory of the estimate. It requires that the on-line solution of a Riccati differential equation and a linear differential equation be driven by the observations. The extended Kalman filter is globally defined but it is only a local method. Under certain conditions, the estimate converges to the true state if the initial estimation error is not too large [1], [23].

There are several nonstochastic approaches to state estimation. For linear systems (1.2), Luenberger [19] developed the concept of an observer. This is another linear dynamical system that is driven by the observations in such a way that the error dynamics is asymptotically stable.

Several nonstochastic methods have been proposed for nonlinear estimation. Some of these are surveyed by Misawa and Hedrick [21]. Other methods include linearization [16], [2], [17], H_∞ methods [15], bilinear systems [29], and high gain observers [3], [5], [6], [7], [8], [9], [24], [25], [26], [27].

This paper describes a simple and efficient method for the design of observers for a broad class of nonlinear systems based on backstepping. The backstepping technique has been used extensively to design stabilizing state feedback control laws [18], [20]. The assumptions on the system are that it be smooth and observable in an appropriate sense. The method is applicable to systems whose error dynamics are not necessarily linearizable by a change of coordinates and input/output injection [16], [2], [17]. It is applicable to a slightly larger class of systems than the high gain observer of Gauthier, Hammouri, and Othman [8]. The latter result can be applied to systems that can be globally described in observable form while the backstepping approach requires only a local observable form. Moreover, backstepping is not a high gain design procedure and hence only local convergence is guaranteed. An explicit formula for the observer gain is derived. The gains are functions of the state of the observer. The gains can be derived off-line through an algorithm presented below. The observer is defined on an arbitrarily large compact subset of the state space but is only locally convergent. We shall prove that the estimate converges exponentially to the true state if the state starts in a compact positively invariant set and the initial estimation error is not too large.

The paper is organized as follows. In section 2 the backstepping approach to observer design is illustrated for a scalar output system without inputs in observable form. In section 3 this is generalized to systems in observable form with vector output and no inputs. In section 4, this technique is generalized to systems that locally can be described in observable form. Systems with inputs are discussed in section 5. In section 6, the relative perfomance of the high gain observer and the backstepping observer are discussed. We close with examples in section 7.

2. The backstepping observer. Consider a smooth nonlinear system in observable form:

$$\begin{aligned} y &= x_1, \\ \dot{x}_1 &= x_2, \\ \dot{x}_2 &= x_3, \\ &\vdots \\ \dot{x}_{n-1} &= x_n, \\ \dot{x}_n &= f_n(x). \end{aligned} \tag{2.1}$$

The state x is n dimensional and the output y is one dimensional. There is no input. Later we shall relax these assumptions. By smooth we mean C^r for r sufficiently large.

The backstepping observer will be in the following form:

$$\begin{aligned} \dot{\hat{x}}_1 &= \hat{x}_2 + \psi_1(\hat{x})(x_1 - \hat{x}_1), \\ \dot{\hat{x}}_2 &= \hat{x}_3 + \psi_2(\hat{x})(x_1 - \hat{x}_1), \\ &\vdots \\ \dot{\hat{x}}_{n-1} &= \hat{x}_n + \psi_{n-1}(\hat{x})(x_1 - \hat{x}_1), \\ \dot{\hat{x}}_n &= f_n(\hat{x}) + \psi_n(\hat{x})(x_1 - \hat{x}_1). \end{aligned} \tag{2.2}$$

The error $e = x - \hat{x}$ dynamics is given by

$$\begin{aligned} \dot{e}_1 &= e_2 - \psi_1(\hat{x})e_1, \\ \dot{e}_2 &= e_3 - \psi_2(\hat{x})e_1, \\ &\vdots \\ \dot{e}_{n-1} &= e_n - \psi_{n-1}(\hat{x})e_1, \\ \dot{e}_n &= f_n(x) - f_n(\hat{x}) - \psi_n(\hat{x})e_1. \end{aligned} \tag{2.3}$$

The problem of observer design is to find gains $\psi_i(\hat{x})$, $1 \le i \le n$, so that $e(t) \to 0$ as $t \to \infty$.

Notice that the error dynamics (2.3) is dependent on both e and $\hat{x}$. The combined system, consisting of the system and its observer, can be described in $x, \hat{x}$ coordinates (2.1), (2.2), in $e, \hat{x}$ coordinates (2.2), (2.3), or in x, e coordinates (2.1), (2.3).

Suppose that K is a compact subset of x space, which is positively invariant under (2.1), i.e., if a trajectory starts in K, then it remains in K for all future times. The set $K \times \{e = 0\}$ is a positively invariant set of the combined system (2.1), (2.3). Using a backstepping approach [18], we will construct a local Lyapunov function for the combined system to prove local exponential convergence to this positively invariant set. The observer gains ψ_i will be chosen in the course of this construction.

We employ the following notation: an error term $O(e)^k$ is a function of $\hat{x}, e$ such that on any compact subset L of $\hat{x}$ space there exists a constant $N > 0, \delta > 0$ such that

$$|O(e)^k| \le N|e|^k \tag{2.4}$$

for all $\hat{x} \in L$ and all $|e| < \delta$. We abbreviate $O(e)^1$ as $O(e)$.

We now proceed with the construction of the backstepping observer on a compact, positively invariant set K and show its local convergence.

Define $z_1 = e_1$ and $V_1 = \frac{1}{2}z_1^2$; then

$$\dot{V}_1 = z_1\dot{z}_1 = -c_1 z_1^2 + z_1 z_2 + O(e)^3,$$

where $c_1 > 0$ and z_2 is the linear function of e that satisfies $z_2 = c_1 z_1 + \dot{z}_1 + O(e)^2$.

If $n = 1$, we choose

$$\psi_1(\hat{x}) = c_1 + \frac{df_1}{dx_1}(\hat{x}_1)$$

so that the auxiliary variable $z_2 = 0$ and

$$\dot{V}_1 = z_1\dot{z}_1 = -c_1 z_1^2 + O(e)^3. \tag{2.5}$$

If $n \geq 2$, then z_1, z_2 and e_1, e_2 are linearly related by

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ b_{2,1} - \psi_1 & 1 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}, \tag{2.6}$$

where

$$b_{2,1} = c_1. \tag{2.7}$$

Define $V_2 = V_1 + \frac{1}{2}z_2^2$; then

$$\dot{V}_2 = -c_1 z_1^2 - c_2 z_2^2 + z_2 z_3 + O(e)^3,$$

where $c_2 > 0$ and z_3 is the linear function of e that satisfies $z_3 = z_1 + c_2 z_2 + \dot{z}_2 + O(e)^2$. Notice that z_2, z_3 depend on the as yet unspecified observer gains $\psi_1(\hat{x}), \psi_2(\hat{x})$.

If $n = 2$, then we would like to choose the gains so that the auxiliary variable z_3 is 0, for then

$$\dot{V}_2 = -c_1 z_1^2 - c_2 z_2^2 + O(e)^3. \tag{2.8}$$

Now

$$\begin{aligned} z_3 &= z_1 + c_2 z_2 + \dot{z}_2 + O(e)^2 \\ &= (b_{3,1} + f_{2;1} - \psi_2)e_1 + (b_{3,2} + f_{2,2} - \psi_1)e_2, \end{aligned}$$

where

$$f_{n;i} = \frac{\partial f_n}{\partial x_i}(\hat{x}),$$

$$\begin{aligned} b_{3,1} &= 1 + c_2(b_{2,1} - \psi_1) + (b_{2,1} - \psi_1)' - (b_{2,1} - \psi_1)\psi_1, \\ b_{3,2} &= c_1 + c_2. \end{aligned} \tag{2.9}$$

We denote differentiation along the observer dynamics when $e_1 = 0$ by $'$. For an n dimensional observer (2.2), the operation $'$ is defined on functions $\phi(\hat{x})$ by

$$\phi'(\hat{x}) = \sum_{j=1}^{n-1} \frac{\partial \phi}{\partial \hat{x}_j}(\hat{x})\hat{x}_{j+1} + \frac{\partial \phi}{\partial \hat{x}_n}(\hat{x}) f_n(\hat{x}).$$

Notice that $'$ does not involve the gains ψ_i and

$$\phi' = \dot{\phi} + O(e).$$

If $n = 2$, we successively define

$$\psi_1(\hat{x}) = b_{3,2}(\hat{x}) + f_{2;2}(\hat{x}),$$
$$\psi_2(\hat{x}) = b_{3,1}(\hat{x}) + f_{2;1}(\hat{x});$$

then (2.8) holds. Notice that the gains are functions of $\hat{x}$ alone.

If $n \geq 3$, we define z_1, z_2 as before (2.6) and z_3 as a linear function of e so that

$$z_3 = z_1 + c_2 z_2 + \dot{z}_2 + O(e)^2.$$

By a calculation similar to the above, we see that

$$z_3 = (b_{3,1} - \psi_2)e_1 + (b_{3,2} - \psi_1)e_2 + e_3, \tag{2.10}$$

where $b_{3,1},\ b_{3,2}$ are given by (2.9). Define $V_3 = V_2 + \frac{1}{2}z_3^2$; then

$$\dot{V}_3 = -c_1 z_1^2 - c_2 z_2^2 - c_3 z_3^2 + z_3 z_4 + O(e)^3,$$

where $c_3 > 0$ and z_4 is the linear function of e that satisfies $z_4 = z_2 + c_3 z_3 + \dot{z}_3 + O(e)^2$.

If $n = 3$, we would like to choose the observer gains so that the auxiliary variable z_4 is 0. After a straightforward calculation, one finds that

$$\begin{aligned} z_4 = {} & (b_{4,1} + f_{3;1} - \psi_3)e_1 + (b_{4,2} + f_{3;2} - \psi_2)e_2 \\ & + (b_{4,3} + f_{3;3} - \psi_1)e_3, \end{aligned} \tag{2.11}$$

where $b_{4,j} = b_{4,j}(\hat{x})$ are functions only of $\hat{x}$ and $b_{4,j}$ depends only on c, ψ_k for $1 \leq k < 3 - j$ and $b_{r,s}$ for $1 < r < 4,\ 1 \leq r - s \leq 4 - j$:

$$\begin{aligned} b_{4,1} &= b_{2,1} - \psi_1 + c_3(b_{3,1} - \psi_2) + (b_{3,1} - \psi_2)' - (b_{3,1} - \psi_2)\psi_1 - (b_{3,2} - \psi_1)\psi_2, \\ b_{4,2} &= 1 + c_3(b_{3,2} - \psi_1) + (b_{3,2} - \psi_1)' + b_{3,1}, \\ b_{4,3} &= c_3 + b_{3,2}. \end{aligned} \tag{2.12}$$

Hence we can successively solve (2.12) for the desired observer gains,

$$\psi_1 = b_{4,3} + f_{3;3},$$
$$\psi_2 = b_{4,2} + f_{3;2},$$
$$\psi_3 = b_{4,1} + f_{3;1}.$$

Turning to the n dimensional system in observable form, the variables $z_1, \ldots, z_{n+1}$ are defined as follows:

$$\begin{aligned} z_1 &= e_1, \\ z_2 &= c_1 z_1 + \dot{z}_1 + O(e)^2, \\ z_3 &= z_1 + c_2 z_2 + \dot{z}_2 + O(e)^2, \\ &\vdots \\ z_i &= z_{i-2} + c_{i-1} z_{i-1} + \dot{z}_{i-1} + O(e)^2, \\ &\vdots \\ z_{n+1} &= z_{n-1} + c_n z_n + \dot{z}_n + O(e)^2, \end{aligned} \tag{2.13}$$

where $c_i > 0$ and the error terms are chosen so that z is a linear function of e. A straightforward calculation yields

(2.14)
$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \\ z_{n+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ b_{2,1} - \psi_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ b_{n,1} - \psi_{n-1} & b_{n,2} - \psi_{n-2} & \cdots & 1 \\ b_{n+1,1} + f_{n;1} - \psi_n & b_{n+1,2} + f_{n;2} - \psi_{n-1} & \cdots & b_{n+1,n} + f_{n;n} - \psi_1 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix},$$

where $b_{i,j} = b_{i,j}(\hat{x})$ are functions of $\hat{x}$ given by (2.7), (2.9) and for $4 \le i \le n+1$ and $2 \le j \le i-3$,

$$\begin{aligned} b_{i,1} &= b_{i-2,1} - \psi_{i-3} + c_{i-1}(b_{i-1,1} - \psi_{i-2}) + (b_{i-1,1} - \psi_{i-2})' \\ &\quad - \sum_{j=1}^{i-2} (b_{i-1,j} - \psi_{i-j-1})\psi_j, \end{aligned}$$

$$\begin{aligned} b_{i,j} &= b_{i-2,j} - \psi_{i-j-2} + c_{i-1}(b_{i-1,j} - \psi_{i-j-1}) + (b_{i-1,j} - \psi_{i-j-1})' \\ &\quad + b_{i-1,j-1}, \end{aligned} \tag{2.15}$$

$$b_{i,i-2} = 1 + c_{i-1}(b_{i-1,i-2} - \psi_1) + (b_{i-1,i-2} - \psi_1)' + b_{i-1,i-3},$$

$$b_{i,i-1} = c_{i-1} + b_{i-1,i-2}.$$

In the backstepping observer we choose the observer gains to zero the last row of the matrix in (2.14),

$$\begin{aligned} \psi_1 &= b_{n+1,n} + f_{n;n}, \\ &\vdots \\ \psi_{n-1} &= b_{n+1,2} + f_{n;2}, \\ \psi_n &= b_{n+1,1} + f_{n;1}, \end{aligned} \tag{2.16}$$

so that $z_{n+1} = 0$.

By induction one sees that $b_{i,j}$ depends only on the quantities

$$\begin{aligned} &c_1, \dots, c_{i-1}, \\ &\psi_1, \dots, \psi_{i-j-1}, \\ &b_{r,s}, \quad 1 < r < i, \quad 1 \le r - s < i - j, \end{aligned} \tag{2.17}$$

and so $b_{i,j}$ can be computed down the diagonals of (2.14). We start with the diagonal just below the main one and successively compute $b_{2,1}, b_{3,2}, \dots, b_{n+1,n}$, which yields

$$b_{i,i-1} = c_1 + \cdots + c_{i-1}.$$

Then we define ψ_1 by (2.16). Going down the diagonal two below the main we compute $b_{3,1}, b_{4,2}, \dots, b_{n+1,n-1}$ and then ψ_2, etc.

THEOREM 2.1. *Suppose that K is a compact, positively invariant set for the system* (2.1). *Consider the observer* (2.2) *with backstepping gains* (2.16) *and error dynamics* (2.3). *There exist constants $M > 0$, $\epsilon > 0$, $\gamma > 0$ such that if $x(0) \in K$ and $|e(0)| < \epsilon$, then*

$$|e(t)| < M|e(0)| \exp(-\gamma t).$$

Proof. Define

$$V = \frac{1}{2}\sum_{i=1}^{n} z_i^2; \tag{2.18}$$

then from (2.13)

$$\dot{V} = -\sum_{i=1}^{n} c_i z_i^2 + z_n z_{n+1} + O(e)^3$$

and

$$\dot{V} = -\sum_{i=1}^{n} c_i z_i^2 + O(e)^3. \tag{2.19}$$

Now let U_r be the $r > 0$ neighborhood of K; then its closure $\bar{U}_r$ is a compact subset. Hence there exist constants $N > 0, \epsilon > 0$ such that the error term in (2.19) satisfies

$$|O(e)^3| \leq N|e|^3 \tag{2.20}$$

for all $\hat{x} \in \bar{U}_r, |e| < \epsilon$. Redefine ϵ to be the smaller of r and ϵ.

From (2.14) we know that there exist constants $M_1 > 0, M_2 > 0$ such that for all $\hat{x} \in \bar{U}_r$ and all e, z,

$$M_1|e| \leq |z| \leq M_2|e|. \tag{2.21}$$

Since $c_i > 0$ there exists a constant $\gamma > 0$ such that

$$4\gamma|z|^2 \leq \sum_{i=1}^{n} c_i z_i^2. \tag{2.22}$$

Hence there is an $\epsilon > 0$ sufficiently small so that the error term in (2.19) satisfies

$$|O(e)^3| \leq \frac{1}{2}\sum_{i=1}^{n} c_i z_i^2 \tag{2.23}$$

for all $\hat{x} \in \bar{U}_r$ and all $|e| < \epsilon$. For these $\hat{x}, e$

$$\dot{V} \leq -\frac{1}{2}\sum_{i=1}^{n} c_i z_i^2 \tag{2.24}$$

$$\leq -2\gamma V. \tag{2.25}$$

Consider the set $D = \{(x, e) : x \in K, V(z(e)) < M_1\epsilon/2\}$; this is a neighborhood of $K \times \{0\}$ in x, e space. From (2.21) we see that on D, we have $|e| < \epsilon$, so $\dot{V} < -2\gamma V$, so D is positively invariant, and by Gronwall's inequality

$$V(t) \leq \exp(-2\gamma t)V(0). \tag{2.26}$$

From (2.21) we obtain

$$|e(t)| \leq \frac{M_2}{M_1}\exp(-\gamma t)|e(0)| \qquad \square$$

Remark 1. There are other possible choices of the Lyapunov function (2.18)—this one was chosen to simplify the calculations. The constants c_i appearing in (2.14) can be chosen as functions of $y, \hat{x}$ as long as they are positive and bounded away from zero.

3. Vector output systems in observable form. The above result generalizes immediately to vector output systems with all observability indices the same. These are systems of the form

$$\begin{aligned} y_l &= x_{1,l}, \\ \dot{x}_{1,l} &= x_{2,l}, \\ \dot{x}_{2,l} &= x_{3,l}, \\ &\vdots \\ \dot{x}_{k,l} &= f_{k,l}(x_1, \dots, x_k), \end{aligned} \tag{3.1}$$

where $l = 1, \dots, p$. The output $y = (y_1, \dots, y_p)$ is p dimensional and so is each $x_i = (x_{i,1}, \dots, x_{i,p})$. The state dimension is then $n = pk$ dimensional. The construction of the observer proceeds exactly as before except that the previously scalar quantities $\hat{x}_i, e_i, z_i, \psi_i$ are now p dimensional, z_i^2 is replaced by $|z_i|^2$, $z_i z_j$ is replaced by $z_i \cdot z_j$, and $\psi_i, b_{i,j}, f_{n,j}$ are $p \times p$ dimensional.

More generally we consider systems of the form

$$\begin{aligned} y_l &= x_{1,l}, \\ \dot{x}_{1,l} &= x_{2,l}, \\ \dot{x}_{2,l} &= x_{3,l}, \\ &\vdots \\ \dot{x}_{k_l,l} &= f_{k_l,l}(x_1, \dots, x_{k_l}), \end{aligned} \tag{3.2}$$

where y is p dimensional, x is $n = \sum k_l$ dimensional, and without loss of generality $k_1 \geq k_2 \geq \cdots \geq k_p$. The indices $k_1, \dots, k_p$ are the observability indices of the system [17]. The dual indices are $m_1, \dots, m_{k_1}$, where m_i is the number of k_l's that are greater than or equal to i. The subvectors x_i are defined as $x_i = (x_{i,1}, \dots, x_{i,m_i})$.

The observer is of the form

$$\begin{aligned} \dot{\hat{x}}_{1,l} &= \hat{x}_{2,l} + \psi_{1,l}(\hat{x}_1, \dots, \hat{x}_{k_l})(x_1 - \hat{x}_1), \\ \dot{\hat{x}}_{2,l} &= \hat{x}_{3,l} + \psi_{2,l}(\hat{x}_1, \dots, \hat{x}_{k_l})(x_1 - \hat{x}_1), \\ &\vdots \\ \dot{\hat{x}}_{k_l,l} &= f_{k_l,l}(\hat{x}_1, \dots, \hat{x}_{k_l}) + \psi_{k_l,l}(\hat{x}_1, \dots, \hat{x}_{k_l})(x_1 - \hat{x}_1), \end{aligned} \tag{3.3}$$

where $\psi_{r,l}(\hat{x}_1, \dots, \hat{x}_{k_l})$ is $1 \times p$ dimensional.

The error dynamics is given by

$$\begin{aligned} \dot{e}_{1,l} &= e_{2,l} - \psi_{1,l}(\hat{x}_1, \dots, \hat{x}_{k_l})e_1, \\ \dot{e}_{2,l} &= e_{3,l} - \psi_{2,l}(\hat{x}_1, \dots, \hat{x}_{k_l})e_1, \\ &\vdots \\ \dot{e}_{k_l,l} &= f_{k_l,l}(x_1, \dots, x_{k_l}) - f_{k_l,l}(\hat{x}_1, \dots, \hat{x}_{k_l}) - \psi_{k_l,l}(\hat{x}_1, \dots, \hat{x}_{k_l})e_1. \end{aligned} \tag{3.4}$$

The method is a modification of the previous approach but the notation is cumbersome. The subvector x_j is m_j dimensional and so are the subvectors $\hat{x}_j$ and e_j.

The subvector z_j is m_{j-1} dimensional and is defined by a modification of (2.13),

$$
\begin{aligned}
z_{1,l} &= e_{1,l}, && 1 \le l \le p,\\
z_{2,l} &= c_{1,l}z_{1,l} + \dot{z}_{1,l} + O(e)^2, && 1 \le l \le m_1,\\
z_{3,l} &= z_{1,l} + c_{2,l}z_{2,l} + \dot{z}_{2,l} + O(e)^2, && 1 \le l \le m_2,\\
&\vdots &&\\
z_{r+1,l} &= z_{r-1,l} + c_{r,l}z_{r,l} + \dot{z}_{r,l} + O(e)^2, && 1 \le l \le m_r,\\
&\vdots &&\\
z_{k_1+1,l} &= z_{k_1-1,l} + c_{k_1,l}z_{k_1,l} + \dot{z}_{k_1,l} + O(e)^2, && 1 \le l \le m_{k_1}.
\end{aligned} \tag{3.5}
$$

The auxiliary variables are the extra components of z, namely $z_{k_1+1,1}, \dots, z_{k_p+1,p}$, and the observer gains are determined by setting them to zero. If

$$V = \frac{1}{2}\sum_{l=1}^{p}\sum_{r=1}^{k_l} z_{r,l}^2,$$

then

$$\dot{V} = -\sum_{l=1}^{p}\sum_{r=1}^{k_l} c_{r,l}z_{r,l}^2 + O(e)^3$$

and the argument proceeds as before.

We illustrate with an example. Consider a three dimensional system

$$
\begin{aligned}
y_1 &= x_{1,1},\\
y_2 &= x_{1,2},\\
\dot{x}_{1,1} &= x_{2,1},\\
\dot{x}_{1,2} &= f_{1,2}(x_{1,1}, x_{1,2}),\\
\dot{x}_{2,1} &= f_{2,1}(x_{1,1}, x_{1,2}, x_{2,1}).
\end{aligned} \tag{3.6}
$$

The indices are $k_1 = 2, k_2 = 1$ and the dual indices are $m_1 = 2, m_2 = 1$. The observer is of the form

$$
\begin{aligned}
\dot{\hat{x}}_{1,1} &= \hat{x}_{2,1} + \psi_{1,1,1}(\hat{x}_{1,1}, \hat{x}_{1,2}, \hat{x}_{2,1})(x_{1,1} - \hat{x}_{1,1}) + \psi_{1,1,2}(\hat{x}_{1,1}, \hat{x}_{1,2}, \hat{x}_{2,1})(x_{1,2} - \hat{x}_{1,2}),\\
\dot{\hat{x}}_{1,2} &= f_{1,2}(\hat{x}_{1,1}, \hat{x}_{1,2}) + \psi_{1,2,1}(\hat{x}_{1,1}, \hat{x}_{1,2})(x_{1,1} - \hat{x}_{1,1}) + \psi_{1,2,2}(\hat{x}_{1,1}, \hat{x}_{1,2})(x_{1,2} - \hat{x}_{1,2}),\\
\dot{\hat{x}}_{2,1} &= f_{2,1}(\hat{x}_{1,1}, \hat{x}_{1,2}, \hat{x}_{2,1}) + \psi_{2,1,1}(\hat{x}_{1,1}, \hat{x}_{1,2}, \hat{x}_{2,1})(x_{1,1} - \hat{x}_{1,1})\\
&\quad + \psi_{2,1,2}(\hat{x}_{1,1}, \hat{x}_{1,2}, \hat{x}_{2,1})(x_{1,2} - \hat{x}_{1,2}),
\end{aligned} \tag{3.7}
$$

and the error dynamics is

$$
\begin{aligned}
\dot{e}_{1,1} &= e_{2,1} - \psi_{1,1,1}(\hat{x}_{1,1}, \hat{x}_{1,2}, \hat{x}_{2,1})e_{1,1} - \psi_{1,1,2}(\hat{x}_{1,1}, \hat{x}_{1,2}, \hat{x}_{2,1})e_{1,2},\\
\dot{e}_{1,2} &= f_{1,2}(x_{1,1}, x_{1,2}) - f_{1,2}(\hat{x}_{1,1}, \hat{x}_{1,2}) - \psi_{1,2,1}(\hat{x}_{1,1}, \hat{x}_{1,2})e_{1,1} - \psi_{1,2,2}(\hat{x}_{1,1}, \hat{x}_{1,2})e_{1,2},\\
\dot{e}_{2,1} &= f_{2,1}(x_{1,1}, x_{1,2}, x_{2,1}) - f_{2,1}(\hat{x}_{1,1}, \hat{x}_{1,2}, \hat{x}_{2,1})\\
&\quad - \psi_{2,1,1}(\hat{x}_{1,1}, \hat{x}_{1,2}, \hat{x}_{2,1})e_{1,1} - \psi_{2,1,2}(\hat{x}_{1,1}, \hat{x}_{1,2}, \hat{x}_{2,1})e_{1,2}.
\end{aligned} \tag{3.8}
$$

From (3.5) we obtain

$$
\begin{aligned}
z_{1,1} &= e_{1,1},\\
z_{1,2} &= e_{1,2},\\
z_{2,1} &= (c_{1,1} - \psi_{1,1,1})e_{1,1} - \psi_{1,1,2}e_{1,2} + e_{2,1},\\
z_{2,2} &= \left(\frac{\partial f_{1,2}}{\partial \hat{x}_{1,1}} - \psi_{1,2,1}\right) e_{1,1} + \left(c_{1,2} + \frac{\partial f_{1,2}}{\partial \hat{x}_{1,2}} - \psi_{1,2,2}\right) e_{1,2},\\
z_{3,1} &= \Bigg(1 + c_{1,1}c_{2,1} - (c_{1,1} + c_{2,1})\psi_{1,1,1} - \psi'_{1,1,1} + \psi^2_{1,1,1}\\
&\quad - \psi_{1,1,2}\left(\frac{\partial f_{1,2}}{\partial \hat{x}_{1,1}} - \psi_{1,2,1}\right) + \frac{\partial f_{2,1}}{\partial \hat{x}_{1,1}} - \psi_{2,1,1}\Bigg)e_{1,1}\\
&\quad + \Bigg(- (c_{1,1+}c_{2,1})\psi_{1,1,2} - \psi'_{1,1,2} + \psi_{1,1,1}\psi_{1,1,2}\\
&\quad - \psi_{1,1,2}\left(\frac{\partial f_{1,2}}{\partial \hat{x}_{1,2}} - \psi_{1,2,2}\right) + \frac{\partial f_{2,1}}{\partial \hat{x}_{1,2}} - \psi_{2,1,2}\Bigg) e_{1,2}\\
&\quad + \left(c_{1,1} + c_{2,1} + \frac{\partial f_{2,1}}{\partial \hat{x}_{2,1}} - \psi_{1,1,1}\right) e_{2,1}.
\end{aligned}
\tag{3.9}
$$

Setting the auxiliary variables $z_{2,2} = z_{3,1} = 0$, we obtain a solution

$$
\begin{aligned}
\psi_{1,2,1} &= \frac{\partial f_{1,2}}{\partial \hat{x}_{1,1}},\\
\psi_{1,2,2} &= c_{1,2} + \frac{\partial f_{1,2}}{\partial \hat{x}_{1,2}},\\
\psi_{2,1,1} &= 1 + c_{1,1}c_{2,1} - (c_{1,1} + c_{2,1})\psi_{1,1,1} - \psi'_{1,1,1} + \psi^2_{1,1,1}\\
&\qquad -\psi_{1,1,2}\left(\frac{\partial f_{1,2}}{\partial \hat{x}_{1,1}} - \psi_{1,2,1}\right) + \frac{\partial f_{2,1}}{\partial \hat{x}_{1,1}},\\
\psi_{2,1,2} &= -(c_{1,1+}c_{2,1})\psi_{1,1,2} - \psi'_{1,1,2} + \psi_{1,1,1}\psi_{1,1,2} + \frac{\partial f_{2,1}}{\partial \hat{x}_{1,2}},\\
\psi_{1,1,1} &= c_{1,1} + c_{2,1} + \frac{\partial f_{2,1}}{\partial \hat{x}_{2,1}},\\
\psi_{1,1,2} &= 0.
\end{aligned}
\tag{3.10}
$$

There are other solutions with $\psi_{1,1,2} \neq 0$.

A computationally simpler approach [22] is to add extra states so as to make all the observability indices the same. We illustrate with the three dimensional system (3.6) above. We imbed it in the four dimensional system

$$
\begin{aligned}
y_1 &= x_{1,1},\\
y_2 &= x_{1,2},\\
\dot{x}_{1,1} &= x_{2,1},\\
\dot{x}_{1,2} &= x_{2,2} + f_{1,2}(x_{1,1}, x_{1,2}),\\
\dot{x}_{2,1} &= f_{2,1}(x_{1,1}, x_{1,2}, x_{2,1}),\\
\dot{x}_{2,2} &= 0.
\end{aligned}
\tag{3.11}
$$

In the new coordinates

$$\begin{aligned}\bar{x}_{1,1} &= x_{1,1},\\ \bar{x}_{1,2} &= x_{1,2},\\ \bar{x}_{2,1} &= x_{2,1},\\ \bar{x}_{2,2} &= x_{2,2} + f_{1,2}(x_{1,1}, x_{1,2}),\end{aligned} \tag{3.12}$$

the system is in observable form

$$\begin{aligned} y_1 &= \bar{x}_{1,1},\\ y_2 &= \bar{x}_{1,2},\\ \dot{\bar{x}}_{1,1} &= \bar{x}_{2,1},\\ \dot{\bar{x}}_{1,2} &= \bar{x}_{2,2},\\ \dot{\bar{x}}_{2,1} &= f_{2,1}(\bar{x}_{1,1}, \bar{x}_{1,2}, \bar{x}_{2,1}),\\ \dot{\bar{x}}_{2,2} &= \frac{\partial f_{1,2}}{\partial \bar{x}_{1,1}}(\bar{x}_{1,1}, \bar{x}_{1,2})\bar{x}_{2,1} + \frac{\partial f_{1,2}}{\partial \bar{x}_{1,2}}(\bar{x}_{1,1}, \bar{x}_{1,2}) f_{1,2}(\bar{x}_{1,1}, \bar{x}_{1,2}).\end{aligned} \tag{3.13}$$

This process can be repeated to make all the observability indices identical. An observer can be constructed for the higher dimensional system and since it is convergent it will yield convergent estimates for the original system.

4. Systems locally in observable form. In this section we construct an observer for a nonlinear system with scalar output of the form

$$\begin{aligned}\dot{\xi} &= f(\xi),\\ y &= h(\xi),\end{aligned} \tag{4.1}$$

where $\xi \in \mathbb{R}^n, y \in \mathbb{R}$.

Following [6] and [8], we say a system is *uniformly observable* if the mapping

$$\xi \mapsto \begin{bmatrix} h(\xi) \\ L_f h(\xi) \\ \vdots \\ L_f^{n-1} h(\xi) \end{bmatrix} \tag{4.2}$$

is a global diffeomorphism, where $L_f^j h(\xi)$ is the j-fold Lie derivative of h by f,

$$\begin{aligned} L_f h(\xi) &= \frac{\partial h}{\partial \xi}(\xi) f(\xi),\\ L_f^j h(\xi) &= \frac{\partial L_f^{j-1} h}{\partial \xi}(\xi) f(\xi).\end{aligned} \tag{4.3}$$

A system can be transformed globally into observable form iff it is uniformly observable. The high gain observer of Gauthier, Hammouri, and Othman [8] requires that the system be uniformly observable while our observer requires only that the system be locally uniformly observable.

A system is *locally uniformly observable* at ξ^0 if the mapping (4.2) is a local diffeomorphism on a neighborhood of ξ^0. A system is *locally uniformly observable* on a set K if it is *locally uniformly observable* at every $\xi^0 \in K$.

If a system is locally uniformly observable at ξ^0, then we can define new local coordinates around ξ^0:

$$x(\xi) = \begin{bmatrix} h(\xi) \\ L_f h(\xi) \\ \vdots \\ L_f^{n-1} h(\xi) \end{bmatrix}. \tag{4.4}$$

In these coordinates the system is in observable form (2.1) with

$$f_n(x) = L_f^n h(\xi(x)). \tag{4.5}$$

LEMMA 4.1. *Suppose that the system* (4.1) *is locally uniformly observable on a compact subset K of ξ space; then there exist an $\epsilon > 0$ and constants $M_1 > 0, M_2 > 0$ such that for all $\xi, \zeta \in K$, $|\xi - \zeta| < \epsilon$*

$$M_1|\xi - \zeta| \leq |x(\xi) - x(\zeta)| \leq M_2|\xi - \zeta|. \tag{4.6}$$

Proof. The map (4.2) is a local diffeomorphism so at any $\zeta \in K$ there exist $\delta(\zeta) > 0, M_1(\zeta) > 0, M_2(\zeta) > 0$ such that

$$M_1(\zeta)|\xi_1 - \xi_2| \leq |x(\xi_1) - x(\xi_2)| \leq M_2(\zeta)|\xi_1 - \xi_2| \tag{4.7}$$

for all $|\xi_i - \zeta| < \delta(\zeta)$. Let $B(\zeta)$ denote the open ball around ζ of radius $\delta(\zeta)/2$. These balls form an open cover of the compact set K so there exists a finite subcover $B(\zeta_1), \ldots, B(\zeta_k)$. Define

$$\begin{aligned} \epsilon &= \frac{1}{2} \min\{\delta(\zeta_1), \ldots, \delta(\zeta_k)\}, \\ M_1 &= \min\{M_1(\zeta_1), \ldots, M_1(\zeta_k)\}, \\ M_2 &= \max\{M_2(\zeta_1), \ldots, M_2(\zeta_k)\}. \end{aligned}$$

If $|\xi_1 - \xi_2| < \epsilon$, then there exists a j such that $|\xi_i - \zeta_j| < \delta(\zeta_j)$ for $i = 1, 2$, so the conclusion follows from (4.7). □

The observer for (4.1) will be of the form

$$\begin{aligned} \dot{\hat{\xi}} &= f(\hat{\xi}) + \phi(\hat{\xi})\,(y - \hat{y})\,, \\ \hat{y} &= h(\hat{\xi}). \end{aligned} \tag{4.8}$$

THEOREM 4.2. *Suppose the system* (4.1) *is locally uniformly observable on a compact positively invariant set K. There exists an observer* (4.8) *and constants $M > 0$, $\epsilon > 0$, $\gamma > 0$ such that if $\xi(0) \in K$ and $|\xi(0) - \hat{\xi}(0)| < \epsilon$, then*

$$|\xi(t) - \hat{\xi}(t)| < M|\xi(0) - \hat{\xi}(0)| \exp(-\gamma t).$$

Proof. Notice that the mapping (4.4) is globally defined on the compact positively invariant set K. It may fail to define global coordinates on K but it is locally one to one and so defines valid local coordinates. In these local coordinates the system is in observable form (2.1) and so we can proceed as in section 2. In the local x coordinates the observer (4.8) takes the form (2.2) and the local error dynamics is given by (2.3). It is important to note that the x variables are globally defined as are $\hat{x}, e$, although

they may be valid coordinates only locally. This allows us to construct the observer in the x variables exactly as before and its local convergence is guaranteed by the above lemma. The observer in the ξ coordinates is given by (4.8), where

$$\phi(\hat{\xi}) = \left[\frac{\partial x}{\partial \xi}(\hat{\xi})\right]^{-1} \psi(x(\hat{\xi})). \tag{4.9}$$

Note that this does not require inverting $x = x(\hat{\xi})$, but it does require inverting the Jacobian matrix $\frac{\partial x}{\partial \xi}(\hat{\xi})$ at each $\hat{\xi}$. □

5. Systems with inputs. In this section we consider systems with inputs

$$\begin{aligned} \dot{\xi} &= f(\xi, u), \\ y &= h(\xi). \end{aligned} \tag{5.1}$$

The state trajectory of a system in observable form (2.1) is completely determined by the output trajectory. The generalization of observable form to a system with inputs is one of the form

$$\begin{aligned} y &= x_1, \\ \dot{x}_1 &= x_2 + g_1(x_1, u), \\ \dot{x}_2 &= x_3 + g_2(x_1, x_2, u), \\ &\vdots \\ \dot{x}_{n-1} &= x_n + g_{n-1}(x_1, \ldots, x_{n-1}, u), \\ \dot{x}_n &= f_n(x) + g_n(x, u). \end{aligned} \tag{5.2}$$

Such a system is said to be *uniformly observable for any input* [9]. Regardless of what input $u = u(t)$ is chosen, the system is observable in the sense that the output and input trajectories uniquely determine the state trajectory,

$$\begin{aligned} x_1 &= y, \\ x_2 &= \dot{x}_1 - g_1(x_1, u), \\ x_3 &= \dot{x}_2 - g_2(x_1, x_2, u), \\ &\vdots \\ x_n &= \dot{x}_{n-1} - g_{n-1}(x_1, \ldots, x_{n-1}, u). \end{aligned}$$

We assume that the state estimate from the observer will be used in a feedback law $u = \kappa(\hat{x})$ to control the system. For a system that is uniformly observable for any input (5.2), the observer will be in the following form:

$$\begin{aligned} \dot{\hat{x}}_1 &= \hat{x}_2 + g_1(\hat{x}_1, \kappa(\hat{x})) + \psi_1(\hat{x})(x_1 - \hat{x}_1), \\ \dot{\hat{x}}_2 &= \hat{x}_3 + g_2(\hat{x}_1, \hat{x}_2, \kappa(\hat{x})) + \psi_2(\hat{x})(x_1 - \hat{x}_1), \\ &\vdots \\ \dot{\hat{x}}_{n-1} &= \hat{x}_n + g_{n-1}(\hat{x}_1, \ldots, \hat{x}_{n-1}, \kappa(\hat{x})) + \psi_{n-1}(\hat{x})(x_1 - \hat{x}_1), \\ \dot{\hat{x}}_n &= f_n(\hat{x}) + g_n(\hat{x}, \kappa(\hat{x})) + \psi_n(\hat{x})(x_1 - \hat{x}_1). \end{aligned} \tag{5.3}$$

The error $e = x - \hat{x}$ dynamics is given by

$$\begin{aligned}
\dot{e}_1 &= e_2 + g_1(x_1, \kappa(\hat{x})) - g_1(\hat{x}_1, \kappa(\hat{x})) - \psi_1(\hat{x})e_1,\\
\dot{e}_2 &= e_3 + g_2(x_1, x_2, \kappa(\hat{x})) - g_2(\hat{x}_1, \hat{x}_2, \kappa(\hat{x})) - \psi_2(\hat{x})e_1,\\
&\vdots\\
\dot{e}_{n-1} &= e_n + g_{n-1}(x_1, \dots, x_{n-1}, \kappa(\hat{x})) - g_{n-1}(\hat{x}_1, \dots, \hat{x}_{n-1}, \kappa(\hat{x})) - \psi_{n-1}(\hat{x})e_1,\\
\dot{e}_n &= f_n(x) - f_n(\hat{x}) + g_n(x, \kappa(\hat{x})) - g_n(\hat{x}, \kappa(\hat{x})) - \psi_n(\hat{x})e_1.
\end{aligned} \tag{5.4}$$

The observer is constructed as before; define variables $z_1, \dots, z_{n+1}$ by (2.13), where the error terms are chosen so that z is a linear function of e, (2.14). The coefficients $b_{i,j}(\hat{x})$ are given by the generalization of (2.7), (2.9), and (2.15),

$$\begin{aligned}
b_{2,1} &= c_1 + g_{1,1},\\
b_{3,1} &= 1 + c_2(b_{2,1} - \psi_1) + (b_{2,1} - \psi_1)' + (b_{2,1} - \psi_1)(g_{1,1} - \psi_1) + g_{2,1},\\
b_{3,2} &= c_2 + b_{2,1} + g_{2,2},
\end{aligned} \tag{5.5}$$

and for $4 \le i \le n+1$ and $2 \le j \le i-3$

$$\begin{aligned}
b_{i,1} &= b_{i-2,1} - \psi_{i-3} + c_{i-1}(b_{i-1,1} - \psi_{i-2}) + (b_{i-1,1} - \psi_{i-2})'\\
&\quad - \sum_{j=1}^{i-2}(b_{i-1,j} - \psi_{i-j-1})(\psi_j - g_{j;1}) + g_{i-1;1},
\end{aligned}$$

$$\begin{aligned}
b_{i,j} &= b_{i-2,j} - \psi_{i-j-2} + c_{i-1}(b_{i-1,j} - \psi_{i-j-1}) + (b_{i-1,j} - \psi_{i-j-1})'\\
&\quad + b_{i-1,j-1} + \sum_{k=j}^{i-2}(b_{i-1,k} - \psi_{i-k-1})g_{k;j} + g_{i-1;j},
\end{aligned}$$

$$\begin{aligned}
b_{i,i-2} &= 1 + c_{i-1}(b_{i-1,i-2} - \psi_1) + (b_{i-1,i-2} - \psi_1)' + b_{i-1,i-3}\\
&\quad + (b_{i-1,i-2} - \psi_1)g_{i-2;i-2} + g_{i-1;i-2},
\end{aligned}$$

$$b_{i,i-1} = c_{i-1} + b_{i-1,i-2} + g_{i-1;i-1}, \tag{5.6}$$

where

$$g_{i;j}(\hat{x}) = \frac{\partial g_i}{\partial x_j}(\hat{x}, \kappa(\hat{x})) \tag{5.7}$$

and the operation $'$ is defined on functions $\phi(\hat{x})$ by

$$\phi'(\hat{x}) = \sum_{j=1}^{n-1} \frac{\partial \phi}{\partial \hat{x}_j}(\hat{x})\left(\hat{x}_{j+1} + g_j(\hat{x}, \kappa(\hat{x}))\right) + \frac{\partial \phi}{\partial \hat{x}_n}\left(f_n(\hat{x}) + g_n(\hat{x}, \kappa(\hat{x}))\right). \tag{5.8}$$

Notice that as before that $'$ does not involve the gains ψ_i and

$$\phi' = \dot{\phi} + O(e).$$

Define

$$V = \frac{1}{2}\sum_{i=1}^{n} z_i^2; \tag{5.9}$$

then from (2.13)

$$\dot{V} = -\sum_{i=1}^{n} c_i z_i^2 + z_n z_{n+1} + O(e)^3. \tag{5.10}$$

We choose the observer gains

$$\begin{aligned} \psi_1 &= b_{n+1,n} + f_{n;n}, \\ &\vdots \\ \psi_{n-1} &= b_{n+1,2} + f_{n;2}, \\ \psi_n &= b_{n+1,1} + f_{n;1} \end{aligned} \tag{5.11}$$

so that $z_{n+1} = 0$ and

$$\dot{V} = -\sum_{i=1}^{n} c_i z_i^2 + O(e)^3. \tag{5.12}$$

A system (5.1) is said to *locally uniformly observable for any input* if around every ξ^0 the transformation (4.2) locally carries it to the form (5.2). For such systems the above algorithm will yield an observer on any compact positively invariant set.

If the system is not locally uniformly observable for any input, then one can still attempt to design the observer by the above algorithm by defining variables $z_1, \dots, z_{n+1}$ by (2.13), where the error terms are chosen so that z is a linear function of e, (2.14). The triangular structure of (2.14) will be lost so there is no guarantee that the transformation from e to z is invertible. If it is, then the algorithm yields a locally convergent observer.

Suppose $U(x)$ is a Lyapunov function for the system (5.1) under full state feedback $u = \kappa(x)$,

$$\dot{U}(x) = \frac{\partial U}{\partial x}(x) f(x, \kappa(x)) \le 0,$$

and suppose that the Lipschitz conditions

$$\begin{aligned} \left| \frac{\partial U}{\partial x}(x) \left(f(x,u) - f(x,\kappa(x)) \right) \right| &\le M \left| u - \kappa(x) \right|, \\ \left| \kappa(x) - \kappa(\hat{x}) \right| &\le M \left| x - \hat{x} \right| \end{aligned}$$

hold for some constant M. If we are able to design an observer using the backstepping technique, then we can choose $c_1 = \cdots = c_n = N$ so that $U(x) + V^{\frac{1}{2}}(e, \hat{x})$ is a Lyapunov function for the combined system. For $|e|$ sufficiently small by (2.21) and (2.23),

$$\begin{aligned} \frac{d}{dt}\left(U + V^{\frac{1}{2}} \right) &= \frac{\partial U}{\partial x}(x) f(x, \kappa(\hat{x})) + \frac{1}{2} V^{-\frac{1}{2}} \dot{V} \\ &= \frac{\partial U}{\partial x}(x) f(x, \kappa(x)) + (M^2 - 2^{-\frac{1}{2}} N M_1) |e| \le 0 \end{aligned}$$

if N is sufficiently large, where M_1 satisfies (2.21). Hence for small initial estimation errors, the output feedback certainty equivalence control inherits the stability of the state feedback control.

6. The high gain observer. In this section, we compare the high gain observer of [8] with high gain, the same observer using low gain, and the backstepping observer. Consider a smooth, scalar output nonlinear system in observable form:

$$\begin{aligned} y &= x_1, \\ \dot{x}_1 &= x_2, \\ \dot{x}_2 &= x_3, \\ &\vdots \\ \dot{x}_{n-1} &= x_n, \\ \dot{x}_n &= f_n(x). \end{aligned} \tag{6.1}$$

The observer proposed in [8] is of the form

$$\begin{aligned} \dot{\hat{x}}_1 &= \hat{x}_2 + L_1(y - \hat{x}_1), \\ \dot{\hat{x}}_2 &= \hat{x}_3 + L_2(y - \hat{x}_1), \\ &\vdots \\ \dot{\hat{x}}_{n-1} &= \hat{x}_n + L_{n-1}(y - \hat{x}_1), \\ \dot{\hat{x}}_n &= f_n(\hat{x}) + L_n(y - \hat{x}_1) \end{aligned} \tag{6.2}$$

with

$$L_i = \begin{pmatrix} n \\ i \end{pmatrix} \theta^i,$$

where θ is a parameter that must be chosen "sufficiently large" to insure global convergence—just how large is never explicitly stated. A mathematical proof of the convergence of the high gain observer with sufficiently high gain has been given in [8] for systems without noise. But because of the high gain even relatively small noise can degrade the performance of the high gain observer.

Suppose the variables and functions are of order one and hence one might choose θ to be an order of magnitude bigger, say $\theta = 10$, so as to be "sufficiently large." If the system is three dimensional, then the largest gain is 1000! If the other variables are of order one, then the right side of the observer dynamics (6.2) is completely dominated by its gain times innovation term. The innovation is $y - \hat{x}_1$. If there is any observation noise, then this is magnified by the gains in the error dynamics. For example, suppose there is observation noise of order $\epsilon = 0.01$ so the signal to noise ratio is 100—not a bad situation. However, the noise in the gain times innovation term of the last state is of order 10 while the state is of order 1. When $\hat{x}_1 \approx x_1$, the signal to noise ratio in the observer dynamics is 0.1—hardly conducive to accurate estimation. Even if there is no observation noise, driving noises can have similar but less dramatic effects.

Many successful applications of the high gain observer have been reported in the literature [8], [6], [7], [9]. In most of these applications, a high gain is not actually used. The method of Gauthier, Hammouri, and Othman [8] is used to design an observer but the gain parameter, θ, is chosen to be relatively small. No attempt is made to determine how large θ must be to guarantee global convergence. It appears that the high gain observer with low gain is actually an excellent local observer and this is why it has been successful in applications. It would nice to have a theoretical explanation for why this is so.

TABLE 6.1
Mean square errors of different observers.

State error	High gain $\theta = 8$	Low gain $\theta = 2$	Backstepping $c_i = 1$
e_1	1.32e-05	3.31e-06	6.27e-07
e_2	6.49e-04	1.22e-05	8.93e-06
e_3	4.18e-03	3.16e-05	1.15e-04

We give a simple example exhibiting this problem. Consider a three dimensional system

$$y = Cx + v, \tag{6.3}$$

$$\dot{x} = Ax + g(x), \tag{6.4}$$

where

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -11 & -6 \end{bmatrix},$$

$$C = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix},$$

$$g(x) = \begin{bmatrix} 0 \\ 0 \\ 7\sin(x_1^2 + x_2^2 + x_3^2) \end{bmatrix}.$$

There is one equilibrium at $x = 0$ which is asymptotically stable as the eigenvalues of A are $\{-1, -2, -3\}$. The nonlinear term g is bounded. The observation noise v is assumed to be small, band limited, Gaussian noise.

The observer with noise is

$$\dot{\hat{x}} = A\hat{x} + g(\hat{x}) + L(x_1 - \hat{x}_1 + v),$$

where the observer gain is

$$L = \begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix} = \begin{bmatrix} 3\theta \\ 3\theta^2 \\ \theta^3 \end{bmatrix}.$$

To estimate how large the gain parameter θ should be, we started the system at $x(0) = (0, 0, 0)$ and the observer at $\hat{x}(0) = (0, 0, 1)$. The noiseless observer did not converge to the true value when $\theta = 7$ but did converge when $\theta = 8$, so we chose the latter value. We simulated three observers with small observation noise starting from the true state $\hat{x}(0) = x(0) = (0, 0, 0)$. We used white Gaussian noise of covariance 1e$-$04, sampled and held for 0.1 second. The first was the observer [8] with a high gain $\theta = 8$, the second was the observer [8] with a relatively low gain $\theta = 2$ and the third was the backstepping observer that is presented above with all the design parameters $c_i = 1$. As can be seen from Table 6.1, the high gain observer performs poorly as compared with the low gain and backstepping observers, which are comparable in performance. Table 6.2 contains the errors of high and low gain observers relative to the backstepping observer.

It should be noted that in the absence of noise there is no assurance that $\theta = 8$ is high enough so that the high gain observer converges globally or even locally. On

TABLE 6.2
Relative errors of different observers.

State error	High gain $\theta = 8$	Low gain $\theta = 2$	Backstepping $c_i = 1$
e_1	21.0	5.3	1
e_2	72.7	1.4	1
e_3	36.2	0.3	1

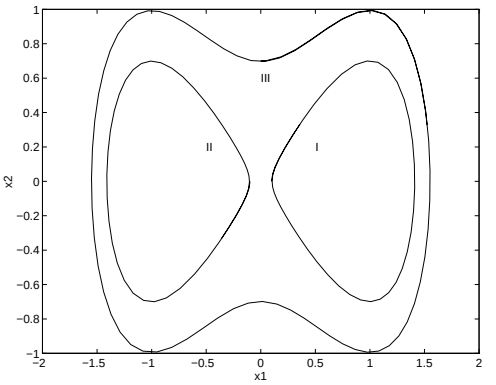


FIG. 1. *The trajectories of Duffing's equation.*

the other hand we have proven that the backstepping observer converges locally for any $c_i > 0$.

One might ask how the example of [8] avoids this high gain difficulty in its the noisy simulations. The answer is very simple: by taking small noise and $\theta = 1,\ n = 2$, it is not really high gain.

7. Examples. The design algorithm described in section 2 has been implemented in MAPLE. This is applied to some examples.

7.1. Duffing's equation. Duffing's equation [11], [14] is

$$\begin{aligned} y &= \xi_1 + \frac{1}{2}\xi_2, \\ \dot{\xi}_1 &= \xi_2, \\ \dot{\xi}_2 &= \xi_1 - \xi_1^3. \end{aligned} \tag{7.1}$$

It is a conservative system with the energy function

$$E(\xi) = \frac{1}{2}\xi_2^2 - \frac{1}{2}\xi_1^2 + \frac{1}{4}\xi_1^4.$$

It has three equilibrium points, $\xi^0 = (0,0)$, $\xi^I = (1,0)$, and $\xi^{II} = (-1,0)$. There are three typical trajectories (see Figure 1): one is around ξ^I (type I), one is around ξ^{II} (type II), and the third one encloses all three equilibria (type III). We define the compact positively invariant region K to be the area enclosed by a trajectory of type III. The system is locally uniformly observable on $\mathbb{R}^2$. The observation function, $h(\xi) = \xi_1 + \frac{1}{2}\xi_2$, was chosen so that the system is not in observable form. The system can be transformed globally into observable form but cannot be transformed into the

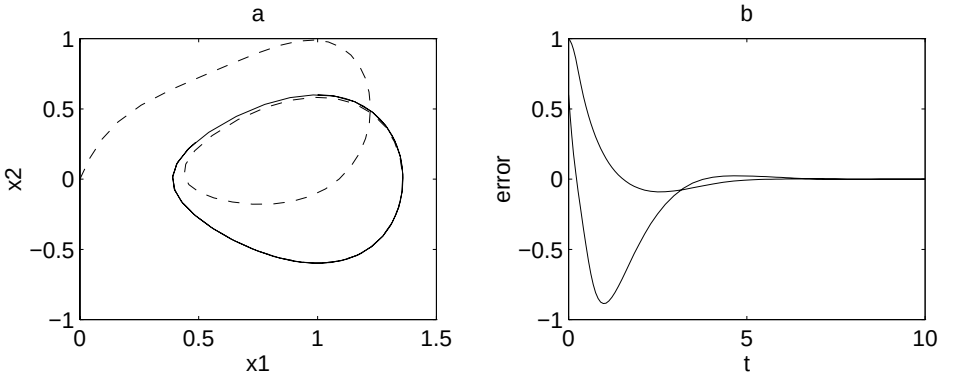


FIG. 2. *Estimation of trajectory of type* I *for Duffing's equation.*

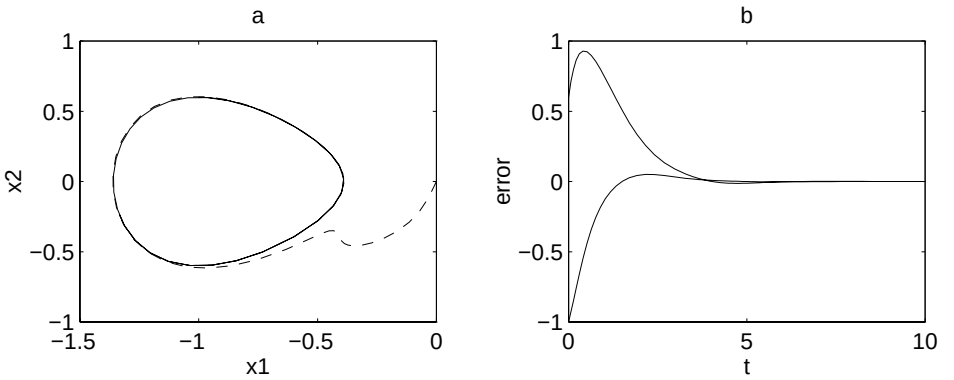


FIG. 3. *Estimation of trajectory of type* II *for Duffing's equation.*

observer form by output injection and change of coordinates [17]. When $c_1 = c_2 = 1$, the observer gain is

$$\phi_1(\hat{\xi}) = -\frac{-2 - 14\,\hat{\xi}_1^2 - 14\,\hat{\xi}_1^4 - 8\,\hat{\xi}_1\hat{\xi}_2 - 8\,\hat{\xi}_1^3\hat{\xi}_2 - 4\,\hat{\xi}_2^2 + 12\,\hat{\xi}_2^2\hat{\xi}_1^2 - 2\,\hat{\xi}_1^6}{3\left(1 + \hat{\xi}_1^2\right)^3},$$

$$\phi_2(\hat{\xi}) = \frac{8 + 8\,\hat{\xi}_1^2 + 8\,\hat{\xi}_1^4 - 4\,\hat{\xi}_1\hat{\xi}_2 + 8\,\hat{\xi}_1^3\hat{\xi}_2 + 8\,\hat{\xi}_1^6 + 12\,\hat{\xi}_1^5\hat{\xi}_2 - 8\,\hat{\xi}_2^2 + 24\,\hat{\xi}_2^2\hat{\xi}_1^2}{3\left(1 + \hat{\xi}_1^2\right)^3}.$$

Three simulations of the system and the backstepping observer are shown in Figures 2, 3, and 4 for trajectory types I, II, and III. Notice that all the simulations use the same observer with the same gain and the same initial estimate but different initial states. The state trajectories are of different types around different equilibrium points.

The solid and dotted curves in Figures 2a, 3a, and 4a are the graphs of the trajectories of the system and the observer. The curves in Figures 2b, 3b, and 4b show the errors $e_1 = \xi_1 - \hat{\xi}_1$ and $e_2 = \xi_2 - \hat{\xi}_2$.

7.2. Homoclinic bifurcation. The backstepping approach can be used to design observers for systems with parameters. Such systems can undergo bifurcations.

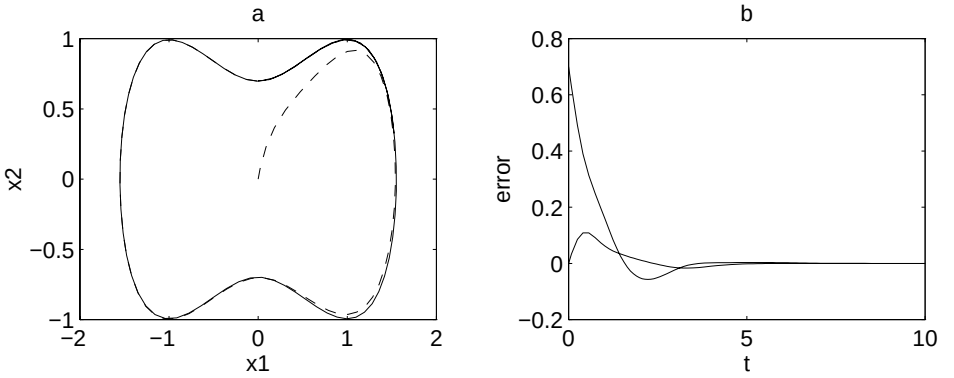


FIG. 4. *Estimation of trajectory of type* III *for Duffing's equation.*

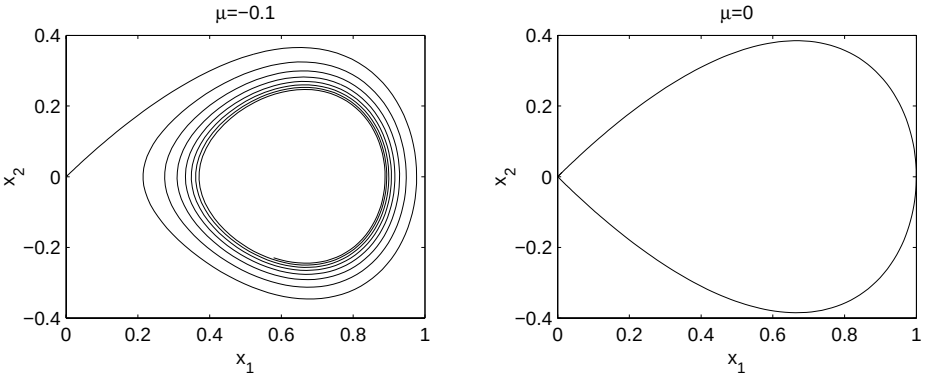


FIG. 5. *The homoclinic bifurcation.*

Consider the following system from [11]:

$$\begin{aligned} y &= \xi_1, \\ \dot{\xi}_1 &= 2\xi_2, \\ \dot{\xi}_2 &= 2\xi_1 - 3\xi_1^2 - \xi_2(\xi_1^3 - \xi_1^2 + \xi_2^2 - \mu). \end{aligned} \tag{7.2}$$

The system depends on a parameter c. For all values of the parameter, there is a saddle at $(0,0)$ with one stable and one unstable direction and an unstable source at $(2/3,0)$. For $-4/27 < \mu < 0$, there is an asymptotically stable periodic orbit around the unstable source. At $\mu = 0$, the periodic orbit becomes a homoclinic orbit consisting of branches of the stable and unstable manifolds of the saddle. For $\mu > 0$, there are no periodic orbits nearby (see Figure 5). For $\mu < 0$ we can find a compact positively invariant set K containing the attracting limit cycle, and for $\mu = 0$ we can take as K the compact set consisting of the homoclinic orbit and it interior. Because of the parameter μ, (7.2) represents a family of systems. However, the computational algorithm for the observer gain is implemented symbolically and μ can be treated as a parameter in the observer. Notice the construction of the observer does not depend on K.

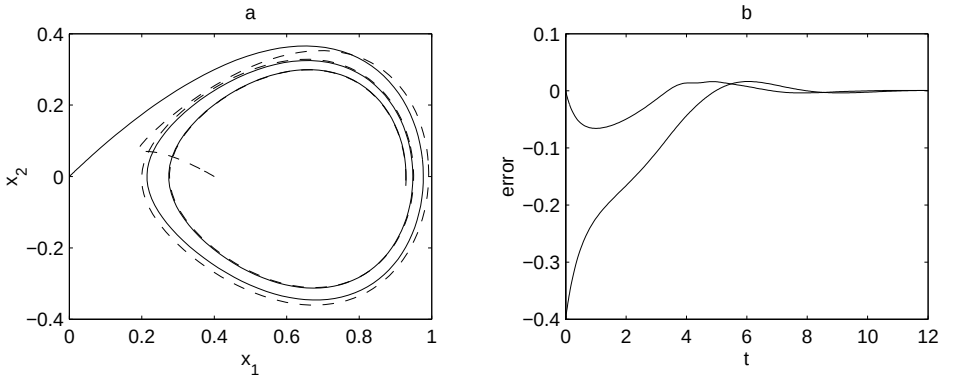


FIG. 6. *Estimation around the periodic solution with* $\mu = -0.1$.

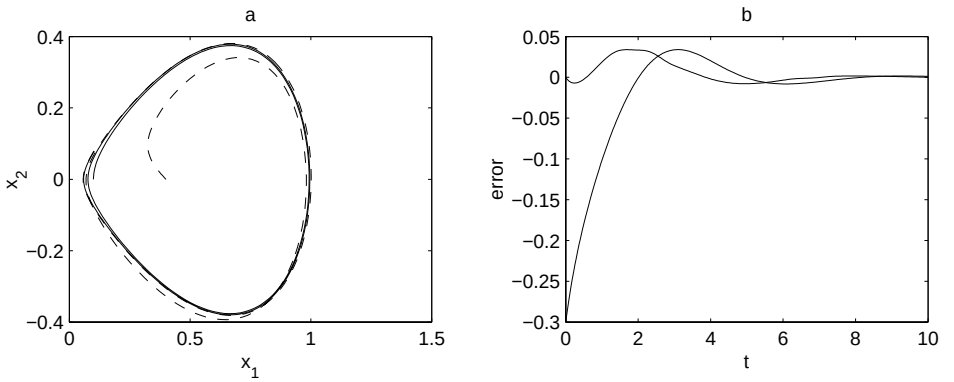


FIG. 7. *Initial state and estimate inside the homoclinic loop with* $\mu = 0$.

If we set $c_1 = c_2 = 1$, the observer gains are

$$\phi_1(\hat{\xi}) = 2 - \hat{\xi}_1^3 + \hat{\xi}_1^2 - 3\,\hat{\xi}_2^2 + \mu,$$
$$\phi_2(\hat{\xi}) = 6\,\hat{\xi}_2\hat{\xi}_1 - 6\,\hat{\xi}_1 + \hat{\xi}_1^2 - \hat{\xi}_1^3 - 3\,\hat{\xi}_2^2 + \mu + \frac{3\,\hat{\xi}_2^4}{2} + \frac{\hat{\xi}_1^6}{2} - \hat{\xi}_1^5 + \frac{\hat{\xi}_1^4}{2} - \hat{\xi}_1^3\mu + \hat{\xi}_1^2\mu + \frac{\mu^2}{2} - 9\,\hat{\xi}_2\hat{\xi}_1^2 + 3.$$

The performance of the observer for $\mu = -0.1$ and $\mu = 0$ are shown in Figures 6, 7, and 8. In Figures 6a, 7a, and 8a the trajectories of (7.2) (solid curves) and the trajectories of the observer (dotted curves) are shown. The estimation error is plotted in Figures 6b, 7b, and 8b. Notice that, in Figure 8, the state starts inside the homoclinic orbit, the estimate starts outside where the system is unstable, and the observer still converges.

8. Conclusion. We have presented a method for designing observers for nonlinear systems based on the backstepping. The method is broadly applicable and the observer error exponentially converges to zero provided the initial error is not too large. It is applicable to a slightly broader class of systems than the high gain observer of Gauthier, Hammouri, and Othman [8] but differs in that the gain is not high and the convergence is only local. The method is easily implemented in a symbolic

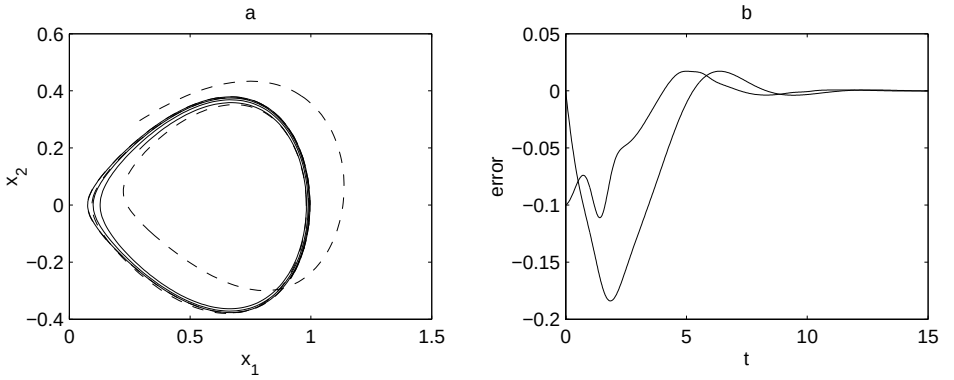


FIG. 8. *Initial state inside the homoclinic loop and initial estimate outside the homoclinic loop with* $\mu = 0$.

computational package such as MAPLE.

REFERENCES

[1] J. S. BARAS, A. BENSOUSSAN, AND M. R. JAMES, *Dynamic observers as asymptotic limits of recursive filters: Special cases*, SIAM J. Appl. Math., 48 (1988), pp. 1147–1158.

[2] D. BESTLE AND M. ZEITZ, *Canonical form observer design for non-linear time-variable systems*, Internat. J. Control, 38 (1983), pp. 419–431.

[3] G. BORNARD AND H. HAMMOURI, *A high gain observer for a class of uniformly observable systems*, in Proceedings of the 30th IEEE Conference on Decision and Control, 1991, pp. 1494–1496.

[4] M. H. A. DAVIS AND S. I. MARCUS, *An introduction to nonlinear filtering*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., D. Reidel Publishing, Dordrecht, The Netherlands, 1981, pp. 53–76.

[5] F. DEZA AND J. P. GAUTHIER, *A simple and robust nonlinear estimator*, in Proceedings of the 30th IEEE Conference on Decision and Control, 1991, pp. 453–454.

[6] J. P. GAUTHIER AND G. BORNARD, *Observability for any* $u(t)$ *of a class of nonlinear systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 922–926.

[7] J. P. GAUTHIER, H. HAMMOURI, AND I. KUPKA, *Observers for nonlinear systems*, in Proceedings of the 30th IEEE Conference on Decision and Control, 1991, pp. 1483–1489.

[8] J. P. GAUTHIER, H. HAMMOURI, AND S. OTHMAN, *A simple observer for nonlinear systems with applications to bioreactors*, IEEE Trans. Automat. Control, 37 (1992), pp. 875–880.

[9] J. P. GAUTHIER AND I. A. K. KUPKA, *Observability and observers for nonlinear systems*, SIAM J. Control Optim., 32 (1994), pp. 975–994.

[10] A. GELB, *Applied Optimal Estimation*, MIT Press, Cambridge, MA, 1974.

[11] J. HALE AND H. KOÇAK, *Dynamics and Bifurcations*, Springer–Verlag, New York, 1991.

[12] R. E. KALMAN, *A new approach to linear filtering and prediction problems*, Trans. ASME Ser. D. J. Basic Engrg., 82 (1960), pp. 35–45.

[13] R. E. KALMAN AND R. S. BUCY, *New results in linear filtering and prediction theory*, Trans. ASME Ser. D. J. Basic Engrg., 83 (1961), pp. 95–108.

[14] A. J. KRENER, *Nonlinear stabilizability and detectability*, in Systems and Networks: Mathematical Theory and Applications, U. Helmke, R. Mennicken, and J. Saurer, eds., Akademie Verlag, Berlin, 1994, pp. 231–250.

[15] A. J. KRENER, *Necessary and sufficient conditions for nonlinear worst case (H-infinity) control and estimation,* J. Math. Systems Estimation Control, 7 (1997), pp. 81–106; summary appeared in J. Math. Systems Estimation Control, 4 (1994), pp. 485–488.

[16] A. J. KRENER AND A. ISIDORI, *Linearization by output injection and nonlinear observers*, Systems Control Lett., 3 (1983), pp. 47–52.

[17] A. J. KRENER AND W. RESPONDEK, *Nonlinear observers with linearizable error dynamics*, SIAM J. Control Optim., 23 (1985), pp. 197–216.

[18] M. Krstić, I. Kanellakopoulos, and P. Kokotović, *Nonlinear and Adaptive Control Design,* John Wiley, New York, 1995.
[19] D. G. Luenberger, *Observing the state of a linear system*, IEEE Trans. Military Electronics, 8 (1964), pp. 74–80.
[20] R. Marino and P. Tomei, *Nonlinear Control Design: Geometric, Adaptive, and Robust*, Prentice–Hall, Englewood Cliffs, NJ, 1995.
[21] E. A. Misawa and J. K. Hedrick, *Nonlinear observers, a state of the art survey*, Trans. ASME J. Dynamic Systems Measurement Control, 111 (1989), pp. 344–352.
[22] A. R. Phelps, *On constructing nonlinear observers*, SIAM J. Control Optim., 29 (1991), pp. 516–534.
[23] Y. Song and J. W. Grizzle, *The extended Kalman filter as a local asymptotic observer for nonlinear discrete-time systems*, J. Math. Systems Estimation Control, 5 (1995), pp. 59–78.
[24] A. Teel and L. Praly, *Global stabilizability and observability imply semi-global stabilizability by output feedback*, Systems Control Lett., 22 (1994), pp. 313–326.
[25] F. Thau, *Observing the state of a non-linear dynamic system*, Internat. J. Control, 17 (1973), pp. 471–479.
[26] A. Tornambé, *Use of asymptotic observers having high-gain in the state and parameter estimation*, in Proceedings of the 28th IEEE Conference on Decision and Control, 1989, pp. 13–15.
[27] J. Tsinias, *Further results on the observer design problem*, Systems Control Lett., 14 (1990), pp. 411–418.
[28] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series, with Engineering Applications*, Technology Press of MIT, Cambridge, MA, 1949.
[29] D. Williamson, *Observation of bilinear systems with application to biological control*, Automatica, 13 (1977), pp. 243–254.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 178–217

SEQUENCING AND ROUTING IN MULTICLASS QUEUEING NETWORKS PART II: WORKLOAD RELAXATIONS*

SEAN P. MEYN†

Abstract. Part II continues the development of policy synthesis techniques for multiclass queueing networks based upon a linear fluid model. The following are shown:

(i) A relaxation of the fluid model based on workload leads to an optimization problem of lower dimension. An analogous workload-relaxation is introduced for the stochastic model. These relaxed control problems admit pointwise optimal solutions in many instances.

(ii) A translation to the original fluid model is *almost optimal*, with vanishing relative error as the network load ρ approaches one. It is pointwise optimal after a short transient period, provided a pointwise optimal solution exists for the relaxed control problem.

(iii) A translation of the optimal policy for the fluid model provides a policy for the stochastic network model that is almost optimal in heavy traffic, over all solutions to the relaxed stochastic model, again with vanishing relative error. The regret is of order $|\log(1-\rho)|$.

Key words. queueing networks, routing, scheduling, optimal control

AMS subject classifications. Primary 90B35, 68M20, 90B15; Secondary 93E20, 60J20

PII. S036301290138376X

1. Introduction.

Variability in a queueing system may have significant impact on performance. Kingman's bound implies that in heavy traffic, when the load ρ is close to unity, even small variations in service times or interarrival times can lead to long delays [3]. This is one reason for the much-publicized difficulties in the air-traffic, highway, and power industries, where real-life networks in heavy traffic are experienced each day by pilots, passengers, and home-owners (see, e.g., [45, 5]).

Variability often plays a smaller role in *relative performance* in network models, when comparing two candidate policies for network regulation (although this depends upon the specific network topology and the performance metric under consideration). For this reason, variability often plays a minor role in many aspects of network design and analysis. Stability of a network under a particular policy is determined by first order statistics (mean arrival and service rates and routing probabilities), except in pathological examples [8, 10, 9, 34, 36]. Part I primarily concerns policy synthesis in queueing networks. It is shown that robustly stabilizing policies can be constructed by appropriately translating a policy for the associated linear fluid model, which is defined only by first order statistics (see [36] for a bibliography).

This paper continues the development of part I. We focus primarily on policy synthesis and network optimization because of the intrinsic importance of these issues, and because it is likely that a deeper understanding will lead to improved methods for addressing many other issues in design, such as performance approximation and network planning. Among the issues not addressed in [36, 35] are the following:

(i) *The role of variability in design.* It is known that an understanding of variability is important in the determination of safety stocks to prevent unwanted idleness. Is this the only use of high order statistical information in policy synthesis?

*Received by the editors January 11, 2001; accepted for publication (in revised form) September 7, 2002; published electronically March 26, 2003. This work was supported in part by NSF grants DMI 00 85165 and ECS 99 72957.

http://www.siam.org/journals/sicon/42-1/38376.html

†Coordinated Science Laboratory and the University of Illinois, 1308 W. Main Street, Urbana, IL 61801 (s-meyn@uiuc.edu, http://black1.csl.uiuc.edu:80/∼meyn).

(ii) *Complexity management.* For example, is it possible to construct optimal policies for the fluid model when the network is large?

(iii) *Performance validation.* Will a translation lead to an optimal allocation for the physical network?

Some answers to these questions are provided here.

A series of recent papers in the stochastic network literature show that a combination of "resource pooling" and "state space collapse" occur in heavy traffic, where $\rho \sim 1$ [38, 39, 43, 29, 25, 4]. See also the recent monographs [6, 30]. State space collapse can transform a network with hundreds of buffers into a far simpler model that retains most of the essential information required for the design of efficient policies. All of these prior results are based on a reflected Brownian motion (RBM) model to approximate the network of interest. This approach is not pursued here for several reasons: Technicalities arising in a proof of weak convergence to an RBM model are avoided, and as pointed out in part I, it is not necessary to assume that the network is *balanced* (i.e., loads at all stations are comparable). This allows significantly greater flexibility in modelling. In this paper we also find that the "Brownian motion scaling" may wash away too many details. By avoiding any scaling, relative bounds on performance are obtained that are far stronger than reported previously in any examples.

As in part I, the primary model considered here is the linear fluid model (2.5). One of the main contributions of the present paper is to introduce a *workload-relaxation* of the fluid model that may be viewed as a generalization of state space collapse, as formulated in the aforementioned references. The significant model reduction obtained in a workload-relaxation provides a framework for addressing many aspects of (i)–(iii).

We show in particular that very strong solidarity exists between respective optimal control solutions. Let c denote a norm on the state space of buffer-levels $\mathsf{X} := \mathbb{R}_+^\ell$—in the results below we eventually specialize to piecewise linear functions on X. Suppose that $\boldsymbol{Q}$ is any queue length process evolving on X defined by some admissible policy. Kingman's bound will then give a steady-state bound of the form

$$\mathsf{E}[c(Q(t))] \geq O\Big(\frac{1}{1-\rho}\Big).$$

Suppose that $\boldsymbol{Q}^\circ$ is the process on X obtained through tracking the optimal fluid model trajectories, as described below. Under general geometric conditions (including uniqueness of solutions to the fluid-model optimal-control problem), we show in Theorem 4.3 that $\boldsymbol{Q}^\circ$ *is approximately optimal, with logarithmic regret*: as $\rho \uparrow 1$,

$$\frac{1}{T}\int_0^T c(Q^\circ(t;x))\,dt \leq \frac{1}{T}\int_0^T c(Q(t;x))\,dt + O(\log\big((1-\rho)^{-1}\big)), \quad 0 \leq T \leq \frac{1}{(1-\rho)^3},$$

where $\boldsymbol{Q}$ is any other solution. We also find that no formulation of sample-path optimality is feasible in heavy traffic under complementary geometric conditions. Consequently, extensions of the results reported here require comparison of a mean-performance metric, rather than sample path bounds (see [7] for recent results in this direction).

The remainder of the paper is organized as follows. Section 2 provides a description of a stochastic network model and the linear fluid model. A reduced order model based on "workload-relaxation" is developed in section 3, and optimal policies for the relaxation are constructed.

Section 4 concerns models in heavy traffic, where $\rho \sim 1$. A policy is constructed based on a translation of the optimal solution to the relaxed fluid-model optimal-control problem. It is shown that this translation is almost optimal for the original fluid model, with bounded error as the system load approaches unity. When a reflected Brownian motion limit exists in heavy traffic, then this "state space collapse" coincides with that observed in the aforementioned references. Similar results hold for a general stochastic model: it is shown that this policy is approximately optimal for a stochastic model, with logarithmic regret, over all solutions to a relaxation of the associated stochastic optimal-control problem.

Section 5 contains conclusions and poses various possible extensions.

2. Models and control. As in [36], this paper is based on a stochastic, bursty model, and a linear fluid model that may be interpreted as a scaled version of its bursty counterpart.

2.1. The stochastic model. The network model described here is a version of the stochastic processing network developed in [23, 24]. We denote by $\boldsymbol{Q}$ the stochastic process evolving on $\mathsf{X} = \mathbb{R}_+^\ell$ whose components indicate buffer levels for the stochastic network model. For example, the network shown in Figure 1 is a simple manufacturing model in which $\ell = 16$, and four of these buffers are *virtual*, corresponding to backlog or excess inventory.

For a given initial condition $Q(0;x) = x \in \mathsf{X}$ the dynamics of $\boldsymbol{Q}$ are expressed

$$Q(t;x) = x - S(Z(t;x)) + R(Z(t;x)) + A(t), \qquad t \geq 0. \tag{2.1}$$

The vector-valued stochastic process $\boldsymbol{Z}$ is the *allocation* (or *control*) evolving on $\mathbb{R}_+^{\ell_u}$ for some integer ℓ_u. The ith component $Z_i(t;x)$ gives the cumulative time that the activity i has run up to time t, $1 \leq i \leq \ell_u$. Activities may include a combination of *sequencing* of various jobs at a particular station and *routing* those jobs to other stations once service is completed. Several examples are given in [36].

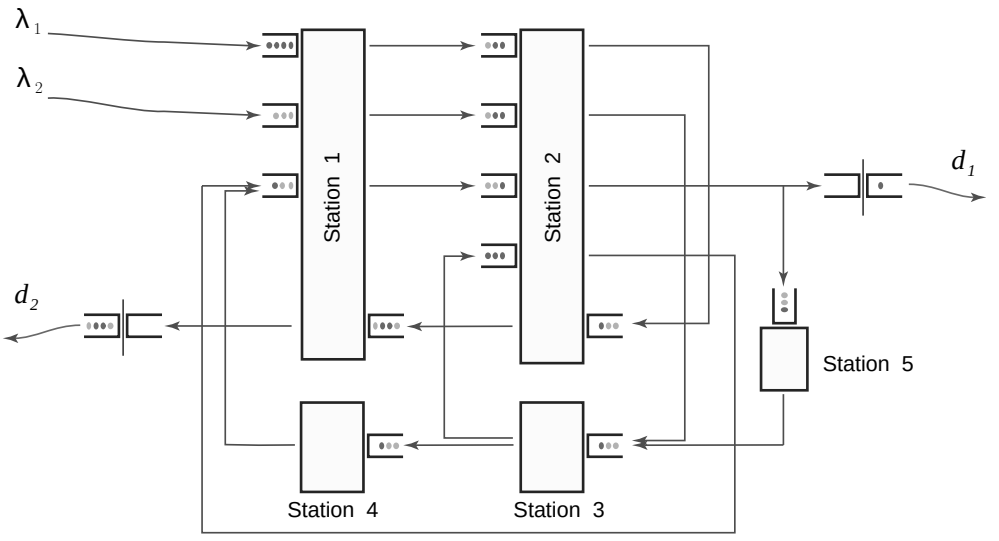


FIG. 1. *A network with many buffers, controlled routing, uncontrolled routing, multiple demands, and virtual buffers.*

The allocation rates are subject to linear constraints

$$Z(t;x) - Z(s;x) \ge \boldsymbol{\theta}\,, \quad C[Z(t;x) - Z(s;x)] \le [t-s]\mathbf{1}\,, \qquad t \ge s \ge 0, \tag{2.2}$$

where the *constituency matrix* C is an $\ell_m \times \ell_u$ matrix with binary entries; $\boldsymbol{\theta}$ is a vector of zeros; and $\mathbf{1}$ is a vector of ones. The *ith resource* $\mathcal{R}_i$ is defined to be the set of activities j such that $C_{ij} = 1$. The constraint (2.2) expresses the condition that resources are shared, and they are limited.

The process $\boldsymbol{A}$ may denote a combination of exogenous arrivals to the network and exogenous demands for materials *from* the network. The function $S(\,\cdot\,)$ represents possibly random service times, and the function $R(\,\cdot\,)$ represents the effects of a combination of possibly uncontrolled, possibly random routing and random service times.

Specific statistical assumptions on $\{\boldsymbol{A}, \boldsymbol{R}, \boldsymbol{S}\}$ are given in section 4.2 where the stochastic model is considered in detail. Many of the variables $\{A_i(\,\cdot\,), R_i(\,\cdot\,), S_i(\,\cdot\,)\}$ will be null in general, and they are typically highly correlated.

The average-cost optimization problem is concerned with minimizing the long-run average cost,

$$\Gamma(x) = \limsup_{T\to\infty} \mathsf{E}\Big[\frac{1}{T}\int_0^T c(Q(t;x))\,dt\Big]\,, \tag{2.3}$$

subject to the constraints given above, where $c\colon \mathbb{R}^\ell \to \mathbb{R}_+$ is a convex function that vanishes only at the origin. In section 4.2 we consider generalizations in which $c(\,\cdot\,)$ is also a function of $\boldsymbol{Z}$. In this case the cost function may be chosen to reflect the desire to maximize utilization of some resources, while minimizing utilization of others.

It is clear that an exact optimal solution to (2.3) will not be found except in very special cases.

2.2. The linear fluid model. Assumption S, to be imposed in section 4.2, implies that the law of large numbers holds: For some $\ell \times \ell_u$ matrix B, a vector $\alpha \in \mathbb{R}^\ell_+$, and any $z \in \mathbb{R}^{\ell_u}_+$,

$$Bz = \lim_{T\to\infty} \frac{1}{T}\int_0^T [-S(zt) + R(zt)]\,dt, \qquad \alpha = \lim_{T\to\infty}\frac{1}{T}\int_0^T A(t)\,dt\,. \tag{2.4}$$

This provided motivation in [36] for the fluid analogue of (2.1) given by

$$q(t;x) = x + Bz(t;x) + \alpha t, \qquad t \ge 0.$$

The vector $\zeta(t;x) = \frac{d}{dt} z\,(t;x)$ denotes allocation rates, and $q(t;x)$ is a vector of buffer levels. This is also expressed as the controlled, linear ordinary differential equation

$$\frac{d}{dt} q(t;x) = B\zeta(t;x) + \alpha\,, \quad t \ge 0\,, \qquad q(0;x) = x\,, \tag{2.5}$$

where throughout the paper the symbol "$\frac{d}{dt}$" denotes a *right derivative*.

It is convenient to envision (2.5) as a differential inclusion:

(i) The state $\boldsymbol{q}$ is constrained to evolve in the polyhedron $\mathsf{X} = \mathbb{R}^\ell_+$.

(ii) The allocation rates $\boldsymbol{\zeta}$ evolve in a polyhedron $\mathsf{U} \subseteq \mathbb{R}^{\ell_u}_+$, defined by

$$\mathsf{U} := \{\zeta \in \mathbb{R}^{\ell_u} : \zeta \ge \boldsymbol{\theta}, C\zeta \le \mathbf{1}\}.$$

(iii) The velocity $\frac{d}{dt}q$ is constrained to lie in the polyhedron

$$\mathsf{V} := \{v = B\zeta + \alpha : \zeta \in \mathsf{U}\}.$$

The assumptions below imply that the network can be controlled so that, starting empty, it will remain empty. This means that V contains the origin, or equivalently, there exists at least one solution $\zeta^{\text{ss}} \in \mathsf{U}$ to the equilibrium equation

$$B\zeta^{\text{ss}} = -\alpha.$$

Section 2.3 is concerned with the existence of equilibria and simple formulations of *optimality* for ζ^{ss}.

Two dynamic optimization problems are singled out because of their mathematical and economic interest:

Time-optimal control. For any initial condition $q(0) = x$, find an allocation $\boldsymbol{z}$ that minimizes

$$T(x) = \min\{t : q(t; x) = \boldsymbol{\theta}\}.$$

The minimal draining time is denoted $T^*(x)$, with the convention that the minimum over an empty set is interpreted as infinity.

Total-cost optimal control. For any initial condition $q(0) = x$, find an allocation $\boldsymbol{z}$ that minimizes

$$J(x) = \int_0^T c(q(t; x))\, dt\,. \tag{2.6}$$

We consider primarily the infinite-horizon case in which $T = \infty$, and in this case we let J^* denote the "optimal cost" (i.e., the infimum over all policies).

The fluid model is called *stabilizable* if $T^*(x) < \infty$ for all $x \in \mathsf{X}$. If the model is stabilizable, then there exists a time-optimal allocation that is *linear*. For any $x \in \mathsf{X}$, if $\boldsymbol{z}$ is any time-optimal allocation, then we write

$$\zeta^\circ = \frac{z(T^*(x); x)}{T^*(x)} \in \mathsf{U}. \tag{2.7}$$

The allocation $z^\circ(t; x) = t\zeta^\circ$, $0 \le t \le T^*(x)$, is evidently feasible and time-optimal. This linear policy and stochastic translations are considered in [11], and generalizations are treated in [17, 14].

The infinite-horizon cost criterion is more closely aligned with the average-cost optimization problem. Computing J^* and an optimal allocation z^* can be formulated as an infinite-dimensional linear program when the cost c is piecewise linear [37]. Algorithms are available that solve this control problem for models of moderate complexity [32, 42].

In the remainder of the paper we take c to be piecewise linear, of the form

$$c(x) = \max_{1 \le i \le \ell_c} \langle c^i, x \rangle, \qquad x \in \mathbb{R}^\ell, \tag{2.8}$$

where $\langle \cdot, \cdot \rangle$ denotes the usual inner product on $\mathbb{R}^\ell$. We can approximate any norm on $\mathbb{R}^\ell$ through an appropriate choice of $\{c^i\} \subset \mathbb{R}^\ell$. A lower bound on performance,

at a specific time t, given a specific initial condition $x \in \mathsf{X}$, is found by solving the linear program

$$
\begin{aligned}
&\min \quad \gamma \\
&\text{subject to} \quad \begin{aligned} \gamma &\geq \langle c^i, y \rangle, \quad 1 \leq i \leq \ell_c, \\ y &= x + Bz + \alpha t, \\ Cz &\leq t\mathbf{1}, \\ y, z &\geq \boldsymbol{\theta}\,. \end{aligned}
\end{aligned} \tag{2.9}
$$

We denote the value of this linear program by $\underline{c}^*(t;x)$. A feasible state trajectory $\boldsymbol{q}^*$ starting from x is called *pointwise optimal* if $c(q^*(t;x)) = \underline{c}^*(t;x)$ for every t. A pointwise optimal trajectory is always time-optimal, and it is also *greedy*: The derivative $\frac{d}{dt}c(q(t;x))$ is minimized over all allocation rates at each time t.

It is rare to find a model for which a pointwise optimal solution exists from each initial condition. However, in section 3 general conditions are formulated which ensure that $c(q^*(t;x)) = \underline{c}^*(t;x)$ for all t following a short transient period.

A first step towards optimization is stabilizability: When are T^* and J^* finite-valued? What is the network load?

2.3. Capacity and time-optimal control. If the fluid model is stabilizable, then the origin is an equilibrium for the model, which means that $\boldsymbol{\theta} \in \mathsf{V}$. We let U_{ss} denote the set of all allocation rates that achieve this: $\mathsf{U}_{\text{ss}} = \{\zeta : B\zeta + \alpha = \boldsymbol{\theta}\}$. In the classical *scheduling* problem there is a unique activity associated with each buffer. This implies that the matrix B is square, and stabilizability ensures that B is full-rank. It then follows that U_{ss} contains a unique vector of steady-state allocation rates given by $\zeta^{\text{ss}} := -B^{-1}\alpha$. We then define the *vector load* by

$$
\boldsymbol{\rho} = (\rho_1, \ldots, \rho_\ell)^{\scriptscriptstyle T} = -CB^{-1}\alpha = C\zeta^{\text{ss}}, \tag{2.10}
$$

and the *system load* is the maximum, $\rho = \max_i \rho_i$.

In other models the matrix B may not be square. The set of equilibrium rates U_{ss} may be large, and some may impose a greater "load" on the system than others. The following is taken from [23], following [29, 22]. The network load ρ is defined as the solution to the linear program

$$
\begin{aligned}
&\min \quad \rho \\
&\text{subject to} \quad \begin{aligned} B\zeta + \alpha &= \boldsymbol{\theta}, \\ C\zeta &\leq \rho\mathbf{1}, \\ \zeta &\geq \boldsymbol{\theta}. \end{aligned}
\end{aligned} \tag{2.11}
$$

The idea is that we consider all allocation rates ζ^{ss} that provide an equilibrium and choose among these the one that has minimal overall impact on the system in the sense that $\max_i [C\zeta^{\text{ss}}]_i$ is smallest.

Closely related is the linear program defining the minimal draining time

$$
\begin{aligned}
&\min \quad T \\
&\text{subject to} \quad \begin{aligned} x + Bz + \alpha T &= \boldsymbol{\theta}, \\ Cz &\leq T\mathbf{1}, \\ z &\geq \boldsymbol{\theta}\,, \end{aligned}
\end{aligned} \tag{2.12}
$$

where $x \in \mathbb{R}^\ell$ is given. The value of this linear program is equal to $T^*(x)$.

We let $W^*(x)$ denote the minimum time to drain the fluid model for an arrival-free model where $\alpha = \boldsymbol{\theta}$. The definition of load is thus motivated by considering the fluid model (2.5) without arrivals: on comparing (2.12) and (2.11) it is seen that $\rho = W^*(\alpha)$. Thus, if α units of material arrives at the network in one second, the *system load* is the amount of time required to clear this material, given that no other material arrives.

Alternative representations for the minimal emptying times are found through a representation of the velocity set V. Let V_0 denote the velocity set for the arrival-free model:

$$\mathsf{V}_0 := \{v - \alpha : v \in \mathsf{V}\} = \{B\zeta : \zeta \in \mathsf{U}\}\,. \tag{2.13}$$

THEOREM 2.1. *The sets V_0, V are described by the intersection of half-spaces: There exists a set of vectors $\{\xi^i : 1 \le i \le \ell_v\} \subset \mathbb{R}^{\ell_v}$, for some minimal integer $\ell_v \ge 1$, and binary values $\{b_i : 1 \le i \le \ell_v\} \subset \{0, 1\}$ such that the following hold:*

$$\mathsf{V}_0 = \{v : \langle \xi^i, v\rangle \ge -b_i,\ 1 \le i \le \ell_v\}, \tag{2.14}$$

$$\mathsf{V} = \{v : \langle \xi^i, v\rangle \ge -(b_i - \rho_i),\ 1 \le i \le \ell_v\}\,, \tag{2.15}$$

where in (2.15) *we set $\rho_i = \langle \xi^i, \alpha\rangle$ for $1 \le i \le \ell_v$.*

Proof. The representation V_0 in (2.14) follows from the fact that it is a polyhedral subset of $\mathbb{R}^\ell$ containing the origin. The representation for V then follows from the formula $\mathsf{V} = \{v + \alpha : v \in \mathsf{V}_0\}$ and the definition of $\{\rho_i\}$. ∎

The vector ξ^i is called a *workload vector* if $b_i \neq 0$. We denote by ℓ_r the number of distinct workload vectors.

For a given $\alpha \in \mathbb{R}^\ell_+$ we assume that the vectors $\{\xi^i\}$ are ordered so that $\rho_1 \ge \cdots \ge \rho_{\ell_v}$. Provided the linear program defining ρ is feasible, we see from Theorem 2.2(ii) that, under this ordering, the set of workload vectors is given by $\{\xi^i : 1 \le i \le \ell_r\}$ and that the system load defined in (2.11) is equal to ρ_1.

THEOREM 2.2. *The following hold for the model* (2.5)*, for any given $\alpha \in \mathbb{R}^\ell_+$, $x \in \mathsf{X}$:*

(i) *If $\langle \xi^i, x\rangle > 0$ for some $i > \ell_r$, then $W^*(x) = \infty$. Otherwise, the minimal emptying time for the arrival-free model is finite and given by*

$$W^*(x) = \max_{1 \le i \le \ell_r} \langle \xi^i, x\rangle\,.$$

(ii) *Suppose that the constraint set in the linear program* (2.11) *is nonempty. Then, $\rho_i \le 0$ for $i > \ell_r$, and the system load can be expressed as*

$$\rho = W^*(\alpha) = \max_{1 \le i \le \ell_r} \rho_i\,.$$

(iii) *If $\langle \xi^i, x\rangle > 0$ and $\rho_i \ge 0$ for some $i > \ell_r$, then $T^*(x) = \infty$. Otherwise, provided $\rho < 1$, the minimal emptying time T^* is finite and given by*

$$T^*(x) = \max_{1 \le i \le \ell_v} \frac{\langle \xi^i, x\rangle}{b_i - \rho_i}\,.$$

(iv) *The model is stabilizable if $\rho < 1$, and $\rho_i < 0$ for $i > \ell_r$. The second condition is automatic if the arrival-free model is stabilizable.*

Proof. Part (i) follows from Theorem 2.1: for $x \neq \boldsymbol{\theta}$, provided $W^*(x) < \infty$,

$$\begin{aligned} W^*(x) &= \min(T > 0 : -T^{-1}x \in \mathsf{V}_0) \\ &= \min(T > 0 : \langle \xi^i, -T^{-1}x \rangle \geq b_i,\ 1 \leq i \leq \ell_v) \\ &= \min(T > 0 : \langle \xi^i, x \rangle \leq b_i T,\ 1 \leq i \leq \ell_v). \end{aligned}$$

Recall that $b_i = 0$ for $i > \ell_r$. If for some such i we have $\langle \xi^i, x \rangle > 0$, then we see that the constraint set in the minimization is infeasible, and we conclude that $W^*(x)$ cannot be finite. Conversely, if $\langle \xi^i, x \rangle \leq 0$ for $i > \ell_r$, then the equation above gives the desired representation for W^*. This establishes (i), and (iii) follows similarly using the definition $\rho_i := \langle \xi^i, \alpha \rangle$.

The proof of (ii) follows from (i) and the representation $\rho = W^*(\alpha)$, and result (iv) follows directly from (iii). ∎

The workload vectors may be interpreted as Lagrange multipliers since they define sensitivity of the optimal draining time with respect to the initial condition x. The following results provide further interpretations. For a given $x \in \mathbb{R}^\ell$, consider the dual of the linear program (2.12)

$$\begin{aligned} &\max \quad \langle \xi, x \rangle \\ &\text{subject to} \quad \begin{aligned} B^{\scriptscriptstyle T}\xi + C^{\scriptscriptstyle T}\eta &\geq \boldsymbol{\theta}, \\ -\alpha^{\scriptscriptstyle T}\xi + \mathbf{1}^{\scriptscriptstyle T}\eta &\leq 1, \\ \eta &\geq \boldsymbol{\theta}. \end{aligned} \end{aligned} \tag{2.16}$$

On considering the extreme points of (2.16), we may express the value of this linear program as a piecewise linear function on the domain $\{x \in \mathbb{R}^\ell : T^*(x) < \infty\}$. Applying Theorem 2.2, we see that these correspond to the vectors $\{\xi^i : 1 \leq i \leq \ell_r\}$ used in the representation of the sets V and V_0.

In view of this solidarity we denote by $\{(\xi^i, \eta^i) : 1 \leq i \leq \ell_r\}$ the nonzero extreme points of the constraint set in (2.16) when $\alpha = \boldsymbol{\theta}$. For each i we have $\xi^i \in \mathbb{R}^\ell$, $\eta^i \in \mathbb{R}_+^{\ell_m}$. An interpretation of the vectors $\{\eta^i\}$ is provided in the following proposition.

PROPOSITION 2.3. *Consider the linear program* (2.16) *with* $\alpha = \boldsymbol{\theta}$. *If* (ξ, η) *is an extreme point in the constraint set satisfying* $\xi \neq \boldsymbol{\theta}$, *then* $\eta \in \mathbb{R}_+^{\ell_m}$ *satisfies* $\langle \mathbf{1}, \eta \rangle = 1$. *Consequently, for any* $1 \leq i \leq \ell_r$ *we may interpret the vector* η^i *as a probability distribution on the resources* $\{1, \ldots, \ell_m\}$.

Proof. Suppose that (ξ, η) is any feasible pair with $0 \leq \langle \mathbf{1}, \eta \rangle < 1$, and $\xi \neq \boldsymbol{\theta}$. Then $(\gamma\xi, \gamma\eta)$ is also feasible for any $0 < \gamma < \langle \mathbf{1}, \eta \rangle^{-1}$, which implies that (ξ, η) cannot be an extreme point. ∎

The *workload process* is defined on a fluid scale by

$$w(t; x) = \Xi q(t; x), \qquad t \geq 0, \quad x \in \mathsf{X}, \tag{2.17}$$

where Ξ denotes the $\ell_r \times \ell$ matrix whose ith row is given by ${\xi^i}^{\scriptscriptstyle T}$.

PROPOSITION 2.4. *The following lower bounds hold:*

$$\text{(i)} \qquad \langle \xi^i, B\zeta \rangle \geq -b_i, \qquad \zeta \in \mathsf{U}, \quad 1 \leq i \leq \ell_v;$$

$$\text{(ii)} \qquad \frac{d}{dt} w_i(t; x) \geq -(1 - \rho_i), \qquad t \geq 0, \quad 1 \leq i \leq \ell_r.$$

Proof. For (i), note that $v_0 := B\zeta$ is a generic element of V_0, so the result follows from the representation of V_0 in Theorem 2.1. As for (ii), observe that

$$\frac{d}{dt} w_i = \langle \xi^i, v \rangle \,,$$

where $v := B\zeta + \alpha$ is a generic element of V. This and Theorem 2.1 again imply the result since $b_i = 1$ for $1 \le i \le \ell_r$. ∎

We define the ith set of *pooled-resources* by

$$\mathcal{R}_i^\circ := \{j \le \ell_m : \eta_j^i > 0\}\,, \qquad 1 \le i \le \ell_r\,.$$

Resource j is called a *bottleneck* if $j \in \mathcal{R}_i^\circ$ for some $i \le \ell_r$, and $\rho_i = \rho$. The following result provides motivation for this terminology.

PROPOSITION 2.5. *For any $1 \le i \le \ell_r$, and any $\zeta \in \mathsf{U}$, the following are equivalent:*

(i) $\langle \xi^i, B\zeta\rangle = -1$,

(ii) $(C\zeta)_j = 1$ *for all* $j \in \mathcal{R}_i^\circ$, *and* ζ *satisfies the* complementary slackness *condition*

$$\zeta_j > 0 \Longrightarrow [B^{\scriptscriptstyle T}\xi^i + C^{\scriptscriptstyle T}\eta^i]_j = 0\,.$$

Proof. Suppose that (i) holds. Then we may multiply $\zeta^{\scriptscriptstyle T}$ times the constraint $B^{\scriptscriptstyle T}\xi + C^{\scriptscriptstyle T}\eta \ge \boldsymbol{\theta}$ in (2.16) to obtain the bound

$$-1 + \langle \eta^i, C\zeta\rangle = \langle \xi^i, B\zeta\rangle + \langle \eta^i, C\zeta\rangle = \langle \zeta, [B^{\scriptscriptstyle T}\xi^i + C^{\scriptscriptstyle T}\eta^i]\rangle \ge 0\,,$$

and it follows that $\langle \eta^i, C\zeta\rangle \ge 1$. Since the reverse inequality also holds when $\zeta \in \mathsf{U}$, we must have equality:

$$-1 + \langle \eta^i, C\zeta\rangle = \langle \zeta, [B^{\scriptscriptstyle T}\xi^i + C^{\scriptscriptstyle T}\eta^i]\rangle = 0\,. \tag{2.18}$$

In fact, since η^i is a probability distribution on $\{1, \dots, \ell_u\}$ and $C\zeta \le \mathbf{1}$, the equality (2.18) implies that $(C\zeta)_j = 1$ for all $j \in \mathcal{R}_i^\circ$. The equation (2.18) also implies the complementary slackness condition in (ii) since $[B^{\scriptscriptstyle T}\xi^i + C^{\scriptscriptstyle T}\eta^i] \ge \boldsymbol{\theta}$, and $\zeta \in \mathbb{R}_+^{\ell_m}$.

Conversely, if (ii) holds, then the complementary slackness condition implies the identity, $\langle \xi^i, B\zeta\rangle + \langle \eta^i, C\zeta\rangle = \langle \zeta, [B^{\scriptscriptstyle T}\xi^i + C^{\scriptscriptstyle T}\eta^i]\rangle = 0$. This combined with the assumption in (ii) that $(C\zeta)_j = 1$ whenever $j \in \mathcal{R}_i^\circ$ (equivalently $\langle \eta^i, C\zeta\rangle = 1$) gives (i) immediately. ∎

The workload vectors allow us to define "hot spots" in the network and give some intuition about the structure of good policies. Suppose that at time t the state takes the value $q(t; x) = y$. The ith pooled-resource is a *dynamic bottleneck* at time t if

$$T^*(y) = \langle \xi^i, y\rangle/(1 - \rho_i).$$

An ordinary resource j is called a *dynamic bottleneck* at time t if $j \in \mathcal{R}_i^\circ$ for some $1 \le i \le \ell_r$, and pooled-resource i is a dynamic bottleneck. We say that the ith pooled-resource is *working at capacity* at time t if $\langle \xi^i, B\zeta(t)\rangle = -1$.

The following is then immediate from Proposition 2.4, Proposition 2.5, and Theorem 2.2.

THEOREM 2.6. *Suppose that $\boldsymbol{q}$ is any solution to the fluid-model equations* (2.5) *starting at $x \in \mathsf{X}$.*

(i) *If $\boldsymbol{q}$ is time-optimal (so that $q(t; x) = \boldsymbol{\theta}$ for $t \ge T^*(x)$), then each dynamic-bottleneck pooled-resource is working at capacity for each $t < T^*(x)$.*

(ii) *If each dynamic-bottleneck pooled-resource works at capacity for $t < T^*(x)$, then the state trajectory $\boldsymbol{q}$ is time-optimal.* ∎

3. The relaxed control problem. We introduce here a relaxation of the optimal-control problem (2.6). The main idea is that, for the purposes of control, only a few of the workload vectors impose serious constraints. A much simpler optimal control problem is obtained by relaxing those constraints corresponding to relatively small load.

3.1. Almost-equivalent workload formulation. For arbitrary $1 \le n \le \ell_r$, the *nth relaxation* of (2.5) is defined as follows. As before, the state space X is taken as $\mathbb{R}^{\ell}_{+}$, but the velocity set is given by

$$\widehat{\mathsf{V}} = \left\{ v : \langle \xi^i, v \rangle \ge -(1-\rho_i), 1 \le i \le n \right\}.$$

An application of Theorem 2.1 establishes the inclusion $\mathsf{V} \subset \widehat{\mathsf{V}}$. It is assumed throughout that $\{\xi^i : 1 \le i \le n\}$ are linearly independent vectors.

We denote by $\widehat{q}$ any feasible state trajectory:

$$\widehat{q}(0;x) = x, \quad \widehat{q}(t;x) \in \mathsf{X}, \quad \textit{and} \quad \frac{q(t;x) - q(s;x)}{t-s} \in \widehat{\mathsf{V}}, \qquad 0 \le s < t. \tag{3.1}$$

The nth relaxation may also be described in a form analogous to (2.5):

$$\frac{d}{dt}\widehat{q}(t;x) = B\widehat{\zeta}(t;x) + \alpha\,, \quad t \ge 0\,, \qquad \widehat{q}(0;x) = x\,. \tag{3.2}$$

The allocation rates in (3.2) are subject to the constraints

$$\widehat{\zeta}(t;x) \in \widehat{\mathsf{U}} := \{\zeta \in \mathbb{R}^{\ell_u} : \widehat{C}\zeta \le \mathbf{1}\}\,,$$

where $\widehat{C} := -\widehat{\Xi}B$, and $\widehat{\Xi}$ denotes the $n \times \ell$ matrix

$$\widehat{\Xi} = [\xi^1 \mid \cdots \mid \xi^n]^{\mathrm{T}}. \tag{3.3}$$

The equivalence of the representations (3.1) and (3.2) follows from Propositions 2.4 and 2.5. The matrix $\widehat{C}$ may be viewed as a constituency matrix for the fluid model (3.2).

If $n \ll \ell_r$, then the behavior of this system may be entirely unnatural since so many constraints have been removed. We show in section 4 that this error can be bounded when considering optimal-control solutions for the fluid model. Related results are obtained for the stochastic model in section 4.2. Such solidarity requires that $n \ge 1$ be chosen sufficiently large, but in many examples this is significantly smaller than ℓ_r.

Our goal remains the same: We wish to minimize, over all feasible state trajectories, the infinite-horizon cost

$$\widehat{J}(x) = \int_0^\infty c(\widehat{q}(t;x))\,dt, \qquad x \in \mathsf{X}. \tag{3.4}$$

Procedures for *translation* of an optimal allocation $\widehat{\boldsymbol{z}}^*$ to both the original fluid model and to the stochastic model (2.1) are treated in sections 4.1 and 4.2, respectively.

In analogy with (2.17), the workload process for this model is given by

$$\widehat{w}(t;x) = \widehat{\Xi}\widehat{q}(t;x), \qquad t \ge 0.$$

For all $1 \le i \le n$ we retain the simple constraint

$$\frac{d}{dt}\widehat{w}_i(t;x) \ge -(1-\rho_i)\,, \qquad t \ge 0\,. \tag{3.5}$$

These constraints are *decoupled* under our assumption that the workload vectors are linearly independent. However, the workload process is also constrained to the set

$$\widehat{\mathsf{W}} := \{\widehat{\Xi}x : x \in \mathsf{X}\}\,. \tag{3.6}$$

The set $\widehat{\mathsf{W}} \subseteq \mathbb{R}^n$ is a convex cone since $\mathsf{X} = \mathbb{R}^\ell_+$. In general, $\widehat{\mathsf{W}} \not\subseteq \mathbb{R}^n_+$ since elements of a workload vector $w \in \widehat{\mathsf{W}}$ may have negative entries.

Two states $x,\ y \in \mathsf{X}$ are called *exchangeable* if $\widehat{\Xi}(x-y) = \boldsymbol{\theta}$. Letting $\widehat{T}^*(x,y)$ denote the optimal time to travel from x to y,

$$\widehat{T}^*(x,y) = \Big(\max_{1\le i\le n} \frac{\langle \xi^i, x-y\rangle}{1-\rho_i}\Big)^+,$$

we see that $\widehat{T}^*(x,y) = \widehat{T}^*(y,x) = 0$ when x and y are exchangeable.

If one is interested in optimal control, then of course one will never stay in a state x if there exists an exchangeable state y with lower cost. Hence an optimal trajectory $\widehat{\boldsymbol{q}}^*$ can always be chosen so that it takes values in

$$\widehat{\mathsf{X}} = \big\{x \in \mathsf{X} : c(x) \le c(y) \text{ whenever } \widehat{\Xi}x = \widehat{\Xi}y\big\}.$$

This is an example of *state space collapse* as described in the introduction.

This reasoning leads to the following definitions:

(i) The *effective cost* $\bar{c}\colon \widehat{\mathsf{W}} \to \mathbb{R}_+$ is defined for any $w \in \widehat{\mathsf{W}}$ as the value of the linear program

$$\begin{aligned} &\min \quad \gamma \\ &\text{subject to} \quad \begin{aligned} \gamma &\ge \langle c^i, x\rangle, \quad 1 \le i \le \ell_c, \\ \widehat{\Xi}x &= w, \\ x &\in \mathsf{X}, \end{aligned} \end{aligned} \tag{3.7}$$

where $\{c^i\}$ are the components of the cost function given in (2.8). The effective cost is piecewise linear:

$$\bar{c}(w) = \max_i \langle \bar{c}^i, w\rangle, \qquad w \in \widehat{\mathsf{W}}\,, \tag{3.8}$$

where $\{\bar{c}^i\} \in \mathbb{R}^n$ are the extreme points obtained in the dual of (3.7).

(ii) For any $w \in \widehat{\mathsf{W}}$, the *effective state* $\mathcal{X}^*(w)$ is defined to be the vector $x \in \widehat{\mathsf{X}}$ that minimizes the linear program (3.7):

$$\mathcal{X}^*(w) = \arg\min_{x\in\mathsf{X}} \Big(c(x) : \widehat{\Xi}x = w\Big). \tag{3.9}$$

(iii) For any $x \in \mathsf{X}$, the optimal exchangeable state $\mathcal{P}^*(x) \in \widehat{\mathsf{X}}$ is defined via

$$\mathcal{P}^*(x) = \mathcal{X}^*(\widehat{\Xi}x). \tag{3.10}$$

The function $\mathcal{X}^*$ may not be uniquely defined, but it is chosen to be a continuous map from $\widehat{\mathsf{W}}$ to $\widehat{\mathsf{X}}$. This is always possible by restricting to basic feasible solutions in (3.7) to obtain a piecewise linear function of x.

Let $\widehat{\mathsf{W}}^+ \subset \mathbb{R}^n$ denote the closed, positive cone

$$\widehat{\mathsf{W}}^+ = \{w \in \widehat{\mathsf{W}} : \bar{c}(w) \le \bar{c}(w') \quad \text{whenever } w' \ge w, w' \in \widehat{\mathsf{W}}\}. \tag{3.11}$$

The function $\bar{c}\colon \widehat{\mathsf{W}} \to \mathbb{R}_+$ is called *monotone* if $\widehat{\mathsf{W}}^+ = \widehat{\mathsf{W}}$ and $\widehat{\mathsf{W}} \subseteq \mathbb{R}^n_+$.

Let $\widehat{q}^*(\,\cdot\,;x)$ denote an optimal trajectory for the relaxed control problem with initial condition x, and let $\widehat{w}^*(\,\cdot\,;x)$ denote the corresponding workload process. By optimality we have the equivalence

$$c(\widehat{q}^*(t;x)) = \bar{c}(\widehat{w}^*(t;x)), \qquad t \ge 0.$$

PROPOSITION 3.1. *Suppose that the nth relaxation is stabilizable. Then, the optimal trajectory $\widehat{\boldsymbol{q}}^*$ can be chosen so that for any initial condition $x \in \mathsf{X}$,*

(i) *$c(\widehat{q}^*(t;x))$ is decreasing, convex, and piecewise linear,*

(ii) *both $\widehat{\boldsymbol{q}}^*$ and $\widehat{\boldsymbol{w}}^*$ are piecewise linear and continuous on $(0,\infty)$.*

Proof. The proof of (i) is identical to the result for the original network model (see [36, Proposition 5]).

To see (ii), consider first the workload process. By convexity, $\bar{c}(\widehat{w}^*(t;x))$ can be discontinuous only at $t = 0$. Moreover, we may assume that $\widehat{\boldsymbol{w}}^*$ is linear on each of the open intervals (T_i, T_{i+1}), $1 \le i \le m-1$, where $\{T_i\}$ denotes the times at which $\frac{d}{dt}c(\widehat{q}^*(t;x))$ is discontinuous, with $T_0 = 0$, $T_m = \infty$.

We now show that, without any loss of generality, the trajectory $\widehat{\boldsymbol{w}}^*$ can be taken to be continuous on $(0,\infty)$. Consider the second time-interval $[T_1, T_2]$. We consider the linear path on this interval given by

$$\widehat{w}^\circ(t) = \widehat{w}^*(T_1-;x) + \frac{t-T_1}{T_2-T_1}\Big[\widehat{w}^*(T_2-;x) - \widehat{w}^*(T_1-;x)\Big], \qquad T_1 < t < T_2.$$

The identity $\bar{c}(\widehat{w}^\circ(t)) = \bar{c}(\widehat{w}^*(t;x))$ holds on this interval since $\bar{c}(\widehat{w}^*(T_1-;x)) = \bar{c}(\widehat{w}^*(T_1+;x))$.

The trajectory $\widehat{\boldsymbol{w}}^\circ$ is feasible, and we can thus redefine $\widehat{\boldsymbol{w}}^*$ on $(0,T_2)$ so that it is continuous. This procedure can be continued on each interval to form an optimal solution that is continuous on $(0,\infty)$.

To show that $\widehat{\boldsymbol{q}}^*$ can also be taken as continuous, choose $\widehat{q}^*(t;x) = \mathcal{X}^*(\widehat{w}^*(t;x))$, $t > 0$. ∎

3.2. One-dimensional workload. The workload process for the relaxed control problem frequently admits an identifiable optimal solution, and in many instances this solution is pointwise optimal.

In the one-dimensional case the matrix $\widehat{\Xi}$ is a row vector, $\widehat{\Xi} = \xi^{1\mathsf{T}}$. Provided $\rho = \rho_1 < 1$, the minimal draining time is given by

$$\widehat{T}^*(x) = \frac{\max(0, \langle \xi^1, x\rangle)}{1-\rho_1}, \qquad x \in \mathsf{X}.$$

The following results follow from linearity of $\widehat{T}^*$ and radial homogeneity of the cost function.

PROPOSITION 3.2. *The following hold for the one-dimensional relaxation for any piecewise linear cost function:*

(i) *The velocity set* $\widehat{\mathsf{V}}$ *is the half space*

$$\widehat{\mathsf{V}} = \{v : \langle \xi^1, v\rangle \geq -(1-\rho)\}.$$

(ii) *The effective cost* $\bar{c}$ *and the lifting map* $\mathcal{X}^*$ *are linear functions of* w*, for* $w \geq 0$*. Hence, letting* $x^* = \mathcal{X}^*(1)$*, the following hold for any* $w \geq 0$ *and any* $x \in \mathsf{X}$ *satisfying* $\langle \xi^1, x\rangle \geq 0$*:*

$$\bar{c}(w) = wc(x^*), \quad \mathcal{X}^*(w) = wx^*, \quad \mathcal{P}^*(x) = \langle \xi^1, x\rangle x^* \,.$$

(iii) *For any* $x \in \mathsf{X}$ *satisfying* $\langle \xi^1, x\rangle \geq 0$*, an optimal state trajectory is given by*

$$\widehat{q}^*(t;x) = \mathcal{P}^*(x)\Big(\frac{\widehat{T}^*(x) - t}{\widehat{T}^*(x)}\Big), \qquad 0 < t \leq \widehat{T}^*(x).$$

(iv) *If the initial condition* $x \in \mathsf{X}$ *satisfies* $\langle \xi^1, x\rangle \leq 0$*, then an optimal solution is given by* $\widehat{q}^*(t;x) = \boldsymbol{\theta}$ *for* $t > 0$. ∎

Proposition 3.3 shows that the solution in (iii) is pointwise optimal.

PROPOSITION 3.3. *Consider the relaxed control problem with* $n = 1$*. For* any *monotone, convex cost function* $c\colon \mathsf{X} \to \mathbb{R}_+$ *and any initial condition, there exists a pointwise optimal allocation.*

Proof. Let $x \in \mathsf{X}$ be given. If $\langle \xi^1, x\rangle \leq 0$, then $\widehat{q}^*(t;x) = \boldsymbol{\theta}$ for all $t > 0$. This is a pointwise optimal solution.

The proof is by comparison when $\langle \xi^1, x\rangle > 0$. Let $x^*(t)$ be the solution to the nonlinear program

$$\begin{aligned} \min \quad & c(y) \\ \text{subject to} \quad & y = x + \widehat{v}t, \\ & \langle \xi^1, \widehat{v}\rangle \geq -(1-\rho), \\ & y \geq \boldsymbol{\theta}. \end{aligned}$$

Its value, $\widehat{c}^* = c(x^*(t))$, is a lower bound on $c(\widehat{q}(t;x))$ for any feasible state trajectory $\widehat{\boldsymbol{q}}$ since we are optimizing over all states attainable at time t. Moreover, because $\widehat{\mathsf{V}}$ is a half-space, the state trajectory $\widehat{q}^*(t;x) = x^*(t)$, $t > 0$, is feasible for the relaxed fluid model. ∎

When c is linear, the effective cost has the following specific form:

$$\bar{c}(w) = \Big(\frac{c_{i^*}}{\xi^1_{i^*}}\Big)w = \Big(\min \frac{c_i}{\xi^1_i} : \xi^1_i > 0\Big)w, \qquad w \geq 0\,,$$

and $x^* = (\xi^1_{i^*})^{-1}e^{i^*}$. In this case, Proposition 3.2(ii) may be viewed as a generalization of the $c\mu$-rule [6, 30].

The routing model shown in Figure 2 was used in [29] to illustrate a form of state space collapse for a stochastic model. We assume that the router with service rate μ_3 is fast, so that, in particular, $\mu_3 > \mu_1 + \mu_2$.

The fluid model is given by

$$B = \begin{bmatrix} -\mu_1 & 0 & \mu_3 & 0 \\ 0 & -\mu_2 & 0 & \mu_3 \\ 0 & 0 & -\mu_3 & -\mu_3 \end{bmatrix}, \quad \alpha = \begin{bmatrix} 0 \\ 0 \\ \alpha_3 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}. \tag{3.12}$$

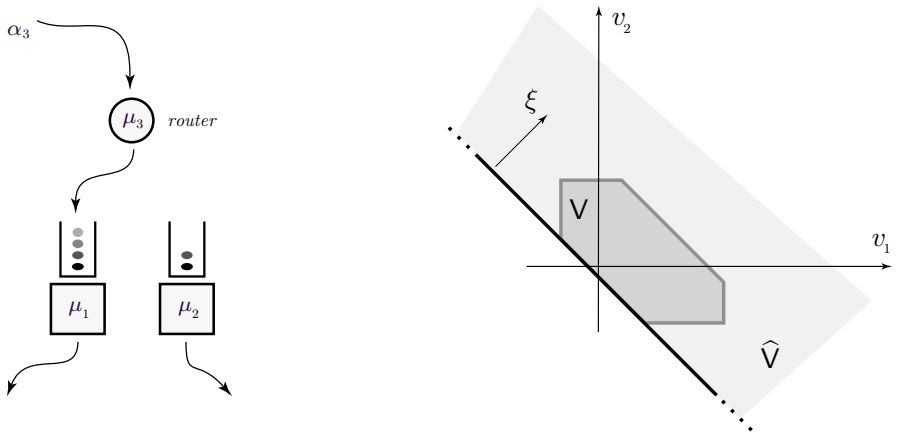


FIG. 2. *On the left is shown a simple routing model. At right is the velocity set* V*, and its one-dimensional relaxation, projected onto* $\{v \in \mathbb{R}^3 : v_3 = 0\}$.

We have four workload vectors,

$$\begin{aligned} \xi^1 &= (\mu_1+\mu_2)^{-1}(1,1,1)^{\scriptscriptstyle T}, & \eta^1 &= (\mu_1+\mu_2)^{-1}(\mu_1,\mu_2,0)^{\scriptscriptstyle T}, \\ \xi^2 &= (m_1,0,0)^{\scriptscriptstyle T}, & \eta^2 &= e^1, \\ \xi^3 &= (0,m_2,0)^{\scriptscriptstyle T}, & \eta^3 &= e^2, \\ \xi^4 &= (0,0,m_3)^{\scriptscriptstyle T}, & \eta^4 &= e^3, \end{aligned}$$

where $m_i = 1/\mu_i$. The vector ξ^1 defines the workload at the two downstream stations, pooled together to form a single resource.

The respective loads are given by $\rho_1 = \alpha_3/(\mu_1+\mu_2)$, $\rho_2 = \rho_3 = 0$, $\rho_4 = \alpha_3/\mu_3$. The system load is $\rho = \max(\rho_1,\rho_4) = \rho_1$ since we have assumed that $\mu_3 > \mu_1+\mu_2$. Using the formula given in Theorem 2.2 we can compute the minimum emptying time from an initial condition $x \in \mathsf{X} = \mathbb{R}^3_+$:

$$T^*(x) = \max\Big(\frac{1}{1-\rho_1}\frac{x_1+x_2+x_3}{\mu_1+\mu_2}, \frac{x_1}{\mu_1}, \frac{x_2}{\mu_2}, \frac{1}{1-\rho_4}\frac{x_3}{\mu_3}\Big).$$

Given the expression for ξ^1 we find that the velocity set for the first workload-relaxation is given by

$$\widehat{\mathsf{V}} = \{v : v_1+v_2+v_3 \ge -(\mu_1+\mu_2-\alpha_3)\}\,.$$

This set is compared to the entire velocity set V in Figure 2. Although both are defined to be a subset of $\mathbb{R}^3$, we can set $q_3 = v_3 \equiv 0$ to obtain the two-dimensional projection shown. We have $\widehat{\mathsf{W}} = \mathbb{R}_+$ in the first workload-relaxation, and if the cost is linear, $c(x) = c^{\scriptscriptstyle T}x$, $x \in \mathsf{X}$, then the effective cost is given by

$$c(w) = c_{i_*}(\mu_1+\mu_2)w\,, \qquad w \in \widehat{\mathsf{W}}\,,$$

where $c_{i_*} = \min(c_1,c_2,c_3)$.

3.3. Dimension two. Under certain conditions on the cost we can again be assured of a pointwise optimal solution even when $\widehat{\mathsf{V}}$ is not a half-space. We illustrate this in the two-dimensional case where

$$\begin{aligned} &\widehat{\Xi} = [\xi^1 \mid \xi^2]^{\scriptscriptstyle T}, \\ &\widehat{\mathsf{V}} = \{v : \langle \xi^i, v\rangle \ge -(1-\rho_i), \quad i = 1,2\}. \end{aligned}$$

The following result holds for any piecewise linear cost function. Recall the definition of the monotone set given in (3.11).

THEOREM 3.4. *Suppose that* $\widehat{\mathsf{W}} \subseteq \mathbb{R}^2_+$.

(i) *When the initial condition satisfies* $\widehat{w}(0) \in \widehat{\mathsf{W}}^+$*, then there exists a pointwise optimal solution.*

(ii) *If the vector* $(1-\rho_1, 1-\rho_2)^T$ *lies in* $\widehat{\mathsf{W}}^+$*, then there is a pointwise optimal solution from* any *initial condition.*

Proof. The proof follows from the rectangular geometry of the set of all states reachable from $\widehat{w}(0) = w$: If w^1 can be reached from w at time t using some allocation, then any $w^2 \in \widehat{\mathsf{W}}$ can also be reached, provided $w_i^2 \geq w_i^1$ for each i. Under the conditions imposed in (i), using the greedy policy we have $\widehat{w}^*(t;x) \in \widehat{\mathsf{W}}^+$ for all $t > 0$, and $\widehat{w}^*(t;x)$ is pointwise minimal within $\widehat{\mathsf{W}}^+$ in the sense that

$$\widehat{w}_i^*(t;x) \leq \widehat{w}_i(t;x), \qquad t \geq 0, \quad i = 1, 2,$$

for any other feasible trajectory $\widehat{\boldsymbol{w}}$ evolving in $\widehat{\mathsf{W}}^+$, starting at $w = \widehat{\Xi}x$. The result (ii) then follows from (i) since $\widehat{w}_i^*(t;x) \in \widehat{\mathsf{W}}^+$ for all $t > 0$ under the assumptions of (ii). ∎

Figure 3 shows the structure of the cost function, the set $\widehat{\mathsf{W}}^+$, and optimal state trajectories for a model that satisfies the assumptions of Theorem 3.4(ii).

Pathwise optimality cannot be expected in general. If the workload dimension is greater than one, and if the cost function c favors starvation of some dynamic bottleneck from some initial condition, then the greedy policy is not time-optimal and hence it cannot be pointwise optimal. Figure 4 illustrates one example with $(1-\rho_1, 1-\rho_2) \notin \widehat{\mathsf{W}}^+$ and $\widehat{w}(0) \notin \widehat{\mathsf{W}}^+$. The initial condition satisfies

$$\frac{d}{dw_2} c\left(\widehat{w}(0)\right) < 0.$$

From this initial condition it is advantageous in the short term to allow $\widehat{w}(0+) \in \partial\widehat{\mathsf{W}}^+$ since $\bar{c}$ is not monotone. This is the greedy, or myopic, policy, which is not time-optimal in this example. The paths shown minimize the infinite-horizon cost given in

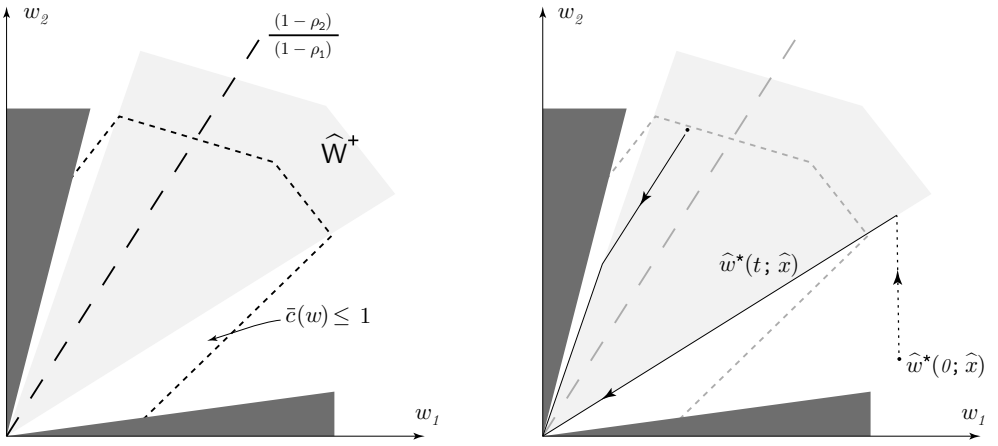


FIG. 3. *The figure at left shows a level set of the cost function* $\bar{c}$ *and the positive cone* $\widehat{\mathsf{W}}^+$ *on which the cost it is monotone. The plot at right shows three optimal state trajectories from varying initial conditions. The darkest region in each figure shows workload vectors* w *that are not feasible.*

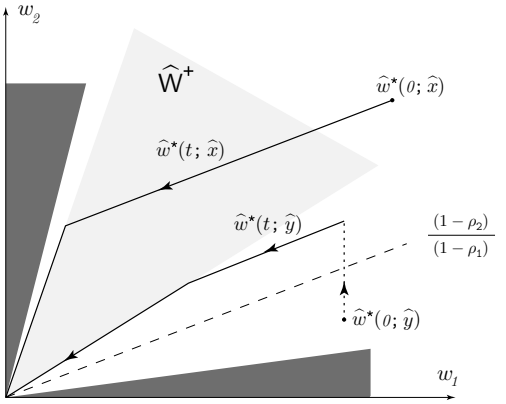


FIG. 4. *Shown are two optimal state trajectories from two different initial conditions. This example is exactly as before, except that ρ_2 is somewhat larger. A trade-off must be made in this case: An overly greedy decision at time 0+ will significantly extend the time to empty the network.*

(3.4), or equivalently $\widehat{J} = \int_0^\infty \bar{c}(\widehat{w}(t;x))\,dt$. An optimal allocation makes a trade-off between reducing the cost at time 0+ and preserving a fast draining time for the model, whenever $\widehat{w}(0) \notin \widehat{\mathsf{W}}^+$.

The three-buffer model shown in Figure 5 is described by the linear fluid model with parameters

$$B = \begin{bmatrix} -\mu_1 & 0 & 0 \\ \mu_1 & -\mu_2 & 0 \\ 0 & \mu_2 & -\mu_3 \end{bmatrix}, \quad \alpha = \begin{bmatrix} \alpha_1 \\ 0 \\ 0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}. \tag{3.13}$$

The load parameters and workload vectors are given by

$$\begin{aligned} \xi^1 &= (m_1 + m_3, m_3, m_3)^T, & \rho_1 &= \alpha_1(m_1 + m_3), \\ \xi^2 &= (m_2, m_2, 0)^T, & \rho_2 &= \alpha_1 m_2, \end{aligned}$$

where we have used $\boldsymbol{\rho} = \widehat{\Xi}\alpha$, with $\widehat{\Xi}$ given in (3.3) with $n = 2$, and $m_i = \mu_i^{-1}$.

Figure 6 shows the optimal solutions for the first and second workload-relaxations. In this numerical example we have taken $\rho_1 = \rho_2 = 9/10$ and $c = (1, 1, 1)^T$. The two plots are very different since the loads at stations one and two are equal.

In Figure 7 the optimal trajectory minimizing (2.6) is compared to the pointwise optimal solution for the two-dimensional relaxation. The triangular region shows the (apparent) error introduced by relaxing the original network optimization problem. We introduce a procedure in Theorem 4.1 below to translate the solution of the relaxed

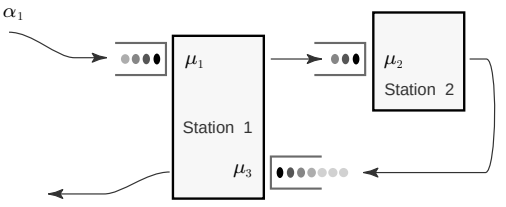


FIG. 5. *A two station scheduling problem*

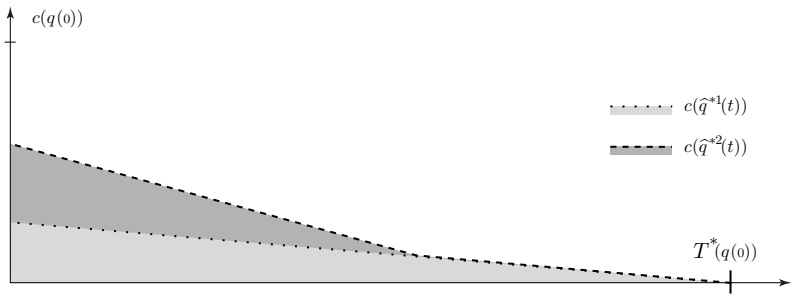


FIG. 6. *Optimal cost-curves for the first and second workload-relaxations.*

problem to the original network model. This yields precisely the optimal policy in this example.

Figure 1 shows a *pull model* in which four of the buffers are virtual. This example is analyzed in [14] under the assumption that the arrivals are *controlled*. An optimal policy will simultaneously determine sequencing and routing rules at each station and release rules for material to the network. Specific service rate values may be found in Chapter 3 of [14]. The cost c is linear, with a weighting of 10 for deficit and unity weighting at the two other virtual buffers and all real buffers.

Although the model is complex, the effective cost for the second workload-relaxation is very simple: as shown in Figure 8, it is defined by five linear functions $\{\bar{c}^i, i = 1, \ldots, 5\}$. Figure 8 shows that the set $\widehat{\mathsf{W}}^+$ contains the ray $\{w \in \mathbb{R}^2_+ : \bar{c}^1 w = \bar{c}^2 w\}$ but not much else, since both $\bar{c}^1$ and $\bar{c}^2$ have negative components. It follows that pointwise optimal trajectories exist for each initial condition only for a very small set of arrival-rate vectors α. (In this example, arrivals are interpreted as *demand* of material from the network.)

3.4. Higher workload dimension. The two-dimensional case is special because one can always find, for each initial w and each time t, a workload vector $\widehat{w}^*(t) \in \widehat{\mathsf{W}}^+$ that is pointwise minimal and reachable from w at time t. This geometry breaks down in three or more dimensions.

Consider first some positive results.

THEOREM 3.5. *Suppose that $\widehat{\mathsf{W}} \subseteq \mathbb{R}^n_+$ for the nth workload-relaxation. The following are then equivalent:*

(i) *A pointwise minimal solution $\widehat{\boldsymbol{w}}^*$ exists for any initial condition $x \in \mathsf{X}$ and any arrival-rate vector α.*

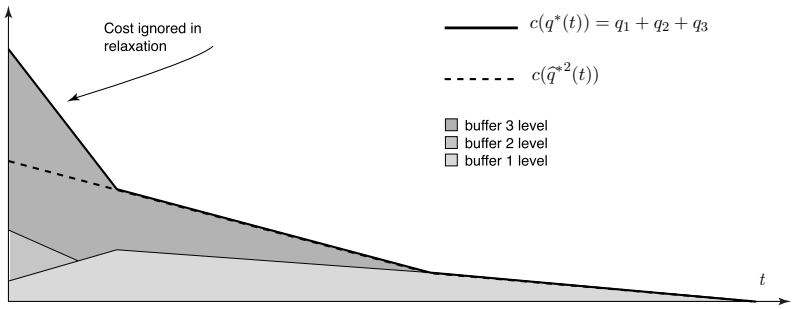


FIG. 7. *The dashed line shows the cost $c(\widehat{q}^*(t;x))$ for the optimized two-dimensional workload-relaxation. The actual optimal policy incurs a higher cost, but this error is bounded in ρ.*

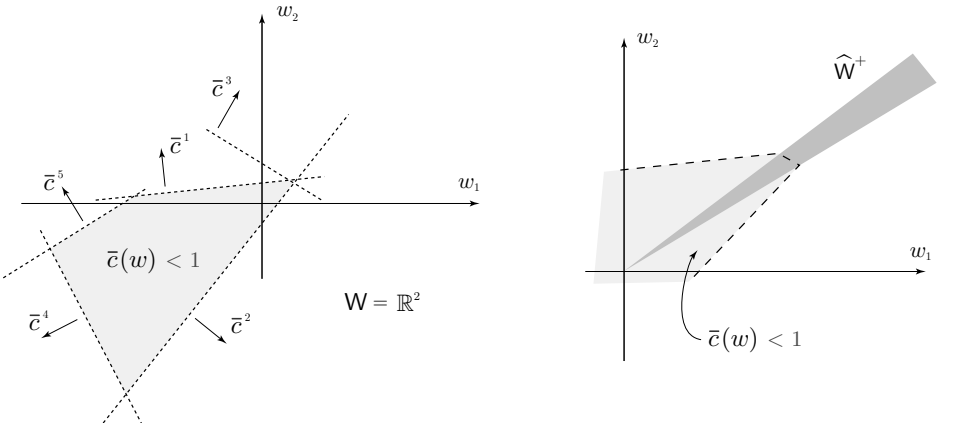


FIG. 8. *On the left is shown a sublevel set of the effective cost for one set of parameters in a two-dimensional relaxation of the network shown in* Figure 1. *The figure at right shows the set* $\widehat{\mathsf{W}}^+$ *together with a close-up of the sublevel set of* $\bar{c}$*, restricted to* $\mathbb{R}^2_+$*. The workload space* $\widehat{\mathsf{W}}$ *is equal to all of* $\mathbb{R}^2$ *in this example.*

(ii) *For each* $w \in \mathbb{R}^n$*, the set*

$$\widehat{\mathsf{W}}_w = \{\widehat{w} : \widehat{w} \in \widehat{\mathsf{W}}, \widehat{w}_i \geq w_i, \quad i = 1, \dots, n\}, \tag{3.14}$$

contains a pointwise minimal element.

If either of these equivalent conditions hold, then a pointwise minimal trajectory may be expressed,

$$\widehat{w}^*(t; x) = [\widehat{\Xi} x - (\mathbf{1} - \boldsymbol{\rho})t]_+, \tag{3.15}$$

where $[w]_+$*,* $w \in \mathbb{R}^n$*, is the projection of* w *onto the set* $\widehat{\mathsf{W}}_w$ *in the standard* ℓ_2 *norm.*

Proof. We first show that the pointwise minimal trajectory is given by (3.15) if (i) holds. Observe that for any t, x the inequality $\widehat{w}^*(t; x) \geq \widehat{\Xi} x - (\mathbf{1} - \boldsymbol{\rho})t$ holds, and $\widehat{w}^*(t; x) \in \widehat{\mathsf{W}}$. Since $\widehat{\boldsymbol{w}}^*$ is minimal, it serves as the projection as claimed.

This implication also shows that (i) $\Rightarrow$ (ii).

Conversely, if (ii) holds, then the trajectory given by $\widehat{w}^\circ(t; x) = [\widehat{\Xi} x - (\mathbf{1} - \boldsymbol{\rho})t]_+$, $t \geq 0$, is obviously pointwise minimal, and it is a piecewise linear function of t for each initial condition x. We show below that the semigroup property holds:

$$\widehat{w}^\circ(t + s; x) = [\widehat{w}^\circ(t; x) - (\mathbf{1} - \boldsymbol{\rho})s]_+, \qquad t, s \geq 0, \quad x \in \mathsf{X}. \tag{3.16}$$

This implies that $\frac{d}{dt}\widehat{w}^\circ(t; x) \geq -(\mathbf{1} - \boldsymbol{\rho})$ for all t, so that this trajectory is feasible for the relaxed fluid model, and hence (i) holds with $\widehat{\boldsymbol{w}}^* = \widehat{\boldsymbol{w}}^\circ$.

To establish (3.16), fix $s, t > 0$, let $T_1 = s + t$, and consider for comparison the trajectory

$$\widehat{w}(T; x) = \widehat{\Xi} x + \frac{T}{T_1}[\widehat{w}^\circ(t + s; x) - \widehat{\Xi} x], \qquad 0 \leq T \leq T_1.$$

This is feasible, and by minimality of $\widehat{w}^\circ(t; x)$ we have $\widehat{w}(t; x) \geq \widehat{w}^\circ(t; x)$. The following bounds then follow:

$$\begin{aligned} \widehat{w}^\circ(t + s; x) = \widehat{w}(T_1; x) &= [\widehat{w}(t; x) + (\widehat{w}(t + s; x) - \widehat{w}(t; x))]_+ \\ &\geq [\widehat{w}(t; x) - (\mathbf{1} - \boldsymbol{\rho})s]_+ \\ &\geq [\widehat{w}^\circ(t; x) - (\mathbf{1} - \boldsymbol{\rho})s]_+. \end{aligned}$$

To obtain an inequality in the reverse direction, note that $\widehat{w}^\circ(t;x) \ge \widehat{\Xi}x - (\mathbf{1}-\boldsymbol{\rho})t$, which implies that

$$\widehat{w}^\circ(t+s;x) := [\widehat{\Xi}x - (\mathbf{1}-\boldsymbol{\rho})(t+s)]_+ \le [\widehat{w}^\circ(t;x) - (\mathbf{1}-\boldsymbol{\rho})s]_+ .$$

We therefore obtain (3.16). □

Under these conditions there is some hope in finding a pointwise optimal solution to the relaxed optimal control problem.

COROLLARY 3.6. *Suppose that*

(i) *the effective cost $\bar{c}$ is monotone and*

(ii) *a pointwise minimal solution $\widehat{\boldsymbol{w}}^*$ exists for the nth workload-relaxation, for any initial condition $x \in \mathsf{X}$.*

Then for any $x \in \mathsf{X}$ there is a pointwise optimal solution for the nth workload-relaxation. □

Assumption (ii) fails in general. Consider the three-station network shown in Figure 9 (see [29, sections 6 and 7] for related examples of RBM networks). The arrival rates α_1, α_6 are equal, and all service rates are equal to unity. For any x, the vector $\widehat{w} = \widehat{\Xi}x \in \mathbb{R}^3$ can be written

$$\widehat{w}_1 = x_1 + x_2 + x_4 + x_6 , \quad \widehat{w}_2 = x_1 + x_3 + x_4 + x_6 , \quad \widehat{w}_3 = x_6 + x_1 + x_3 + x_5 .$$

For example, $\widehat{w}^3 := \widehat{\Xi}e^3 = [0,\ 1,\ 1]^{\mathrm{T}}$, and $\widehat{w}^4 := \widehat{\Xi}e^4 = [1,\ 1,\ 0]^{\mathrm{T}}$. The two states $\{e^3, e^4\}$ are not exchangeable for a three-dimensional relaxation since the workload vectors $\widehat{w}^3, \widehat{w}^4$ are different.

For simplicity consider the arrival-free model where $\alpha_1 = \alpha_6 = 0$ so that $\rho = 0$. The initial condition $x = e^3 + e^4$ has corresponding workload $\widehat{w}(0) = \widehat{\Xi}x = (1,2,1)^{\mathrm{T}}$. From this initial condition it is possible to reach either e^3 or e^4 in exactly one second. Any minimal workload vector $\widehat{w}^*$ must then satisfy $\widehat{w}^*(t;x) \le \widehat{w}^3$ and $\widehat{w}^*(t;x) \le \widehat{w}^4$ at $t = 1$, which implies that $\widehat{w}^*(1;x) \le (0,1,0)^{\mathrm{T}}$.

The only vector in $\widehat{\mathsf{W}}$ satisfying this inequality is $w = (0,0,0)^{\mathrm{T}}$. However, this state is not reachable in one second since the minimal draining time is $W^*(x) = T^*(x) = \max(\widehat{w}_1(0), \widehat{w}_2(0), \widehat{w}_3(0)) = 2$.

We now investigate the structure of pointwise optimal solutions under the conditions of Corollary 3.6.

The ith pooled-resource is said to be *satiated* at state x provided there exists $v \in \widehat{\mathsf{V}}$ satisfying $\langle \xi^i, v\rangle = -(1-\rho_i)$, and $v_i \ge 0$ whenever $x_i = 0$. A resource is said to be satiated if it is a component of a satiated pooled-resource.

Consider any $x \in \mathsf{X}$, and suppose $y \in \mathsf{X}$ with $\langle \xi^k, x\rangle > \langle \xi^k, y\rangle$ for some $1 \le k \le n$. Then the optimal time to travel from x to y is nonzero:

$$\widehat{T}^*(x,y) = \max_{1\le j\le n} \frac{\langle \xi^j, x-y\rangle}{1-\rho_j} > 0 .$$

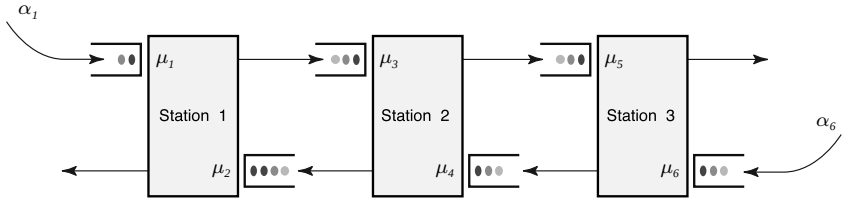


FIG. 9. *A three-station network.*

With $v = (y-x)/\widehat{T}^*(x,y) \in \widehat{\mathsf{V}}$, the trajectory below is both feasible and time-optimal:

$$\widehat{q}(t;x) = x + vt, \qquad 0 \le t \le \widehat{T}^*(x,y)\,.$$

Moreover, simple dynamic programming arguments ensure that

$$\frac{d}{dt}\widehat{T}^*(\widehat{q}(t;x),y) = -1, \qquad 0 < t < \widehat{T}^*(x,y).$$

Hence, whenever i is a maximizer, so that

$$\widehat{T}^*(x,y) = \frac{\langle \xi^i, x-y\rangle}{1-\rho_i},$$

we must have $\langle \xi^i, v\rangle = -(1-\rho_i)$. This implies that pooled-resource i is satiated by x and proves the following.

PROPOSITION 3.7. *Suppose that $\rho < 1$. Then $\widehat{T}^*(x,y) < \infty 0$, $x, y \in \mathsf{X}$, and if $\widehat{T}^*(x,y) > 0$, then*

$$\begin{aligned}\widehat{T}^*(x,y) &= \max\left\{\frac{\langle \xi^j, x-y\rangle}{1-\rho_j} : 1 \le j \le n\right\}\\ &= \max\left\{\frac{\langle \xi^j, x-y\rangle}{1-\rho_j} : j \text{ is satiated by } x\right\}.\end{aligned}$$ □

Satiated resources play a role analogous to dynamic bottlenecks in the construction of a time-optimal trajectory. The following result is the analogue of Theorem 2.6. It is an easy corollary to Proposition 3.7.

THEOREM 3.8. *Suppose that $\rho < 1$, and that the nth relaxation satisfies $\widehat{\mathsf{W}} \subseteq \mathbb{R}^n_+$. Let $\widehat{\boldsymbol{q}}$ be any solution to the nth workload-relaxation, starting at $x \in \mathsf{X}$, and let $\widehat{w}(t;x) = \widehat{\Xi}\widehat{q}(t;x)$, $t \ge 0$. We then have the following:*

(i) *If $\widehat{\boldsymbol{w}}$ is pointwise minimal, then each satiated pooled-resource is working at capacity for each $0 < t < \infty$. That is, if pooled-resource i is satiated at time t, then*

$$\frac{d}{dt}\widehat{w}_i(t) = -(1-\rho_i).$$

(ii) *If each satiated pooled-resource works at capacity for all t, then the resulting workload trajectory $\widehat{\boldsymbol{w}}$ is pointwise minimal.* □

3.5. Variability and continuity. We close this section with some continuity properties for pointwise minimal solutions. Our interest lies in the fluid model with *exogenous disturbance*, defined by

$$\widehat{q}(t;x) = B\widehat{z}(t;x) + \alpha t + d_0(t)\,, \quad t \ge 0\,, \qquad \widehat{q}(0;x) = x\,. \tag{3.17}$$

We assume as in (3.2) that the allocation is subject to the linear constraints

$$\widehat{C}[\widehat{z}(t;x) - \widehat{z}(s;x)] \le [t-s]\mathbf{1}, \qquad 0 \le s \le t\,, \tag{3.18}$$

where $\widehat{C} = -\widehat{\Xi}B$ is defined below (3.2), and we assume throughout that the disturbance $\boldsymbol{d}_0$ is of bounded variation.

Letting $\widehat{w}(t;x) := \widehat{\Xi}\widehat{q}(t;x)$, $d(t) = \widehat{\Xi}d_0(t)$, we obtain the corresponding workload model

$$\widehat{w}(t;x) = \widehat{\Xi}x - (\mathbf{1} - \boldsymbol{\rho})t + \iota(t) + d(t), \qquad t \geq 0, \tag{3.19}$$

where $\iota(t) := t\mathbf{1} - \widehat{C}\widehat{z}(t;x)$, $t \geq 0$. The idleness process $\boldsymbol{\iota}$ is nonnegative with nondecreasing components, and $\widehat{\boldsymbol{w}}$ evolves in $\widehat{\mathsf{W}}$.

Rather than define $\widehat{\boldsymbol{w}}$ through (3.17), for the purposes of optimization we may restrict attention to the simpler model (3.19). Given the current workload-value $\widehat{w} = \widehat{w}^*(t;x)$ we take $\widehat{z}^*(t;x)$ to be any optimizer of the linear program

$$\begin{aligned} &\min \quad \gamma \\ &\text{subject to} \quad \begin{aligned} \gamma &\geq \langle c^i, y\rangle, \quad 1 \leq i \leq \ell_c, \\ y &= x + Bz + \alpha t + d_0(t), \\ \widehat{\Xi}y &= \widehat{w}, \\ y &\in \mathsf{X}. \end{aligned} \end{aligned} \tag{3.20}$$

It follows from the definitions that the optimizer $\widehat{\boldsymbol{z}}^*$ satisfies the constraints given in (3.18).

If $\boldsymbol{d} \equiv \boldsymbol{\theta}$, then (3.19) is the linear workload model considered earlier. In this case, it follows from Theorem 3.5 and Assumption M below that (3.19) admits a pointwise minimal solution for any value of $\boldsymbol{\rho}$. This is generalized to arbitrary disturbances in Theorem 3.10.

Assumption M.

(i) The workload vectors for the nth relaxation are linearly independent and satisfy

$$\xi^i \in \mathbb{R}_+^\ell \text{ for } 1 \leq i \leq n.$$

(ii) For each $w \in \mathbb{R}^n$ the set $\widehat{\mathsf{W}}_w$ defined in (3.14) contains a pointwise minimal element denoted $[w]_+$.

Although the semigroup property (3.16) does not hold in general for a model with disturbances, we always have the lower bound.

LEMMA 3.9. *Under Assumption* M*, if $\widehat{\boldsymbol{w}}$ is a feasible state trajectory for the nth relaxation, then*

$$\widehat{w}(t;x) \geq [\widehat{w}(s;x) - (\mathbf{1} - \boldsymbol{\rho})(t-s) + d(t) - d(s)]_+, \qquad t \geq s \geq 0,\ x \in \mathsf{X}.$$

Proof. The lower bound $\widehat{w}(t;x) \geq (\widehat{w}(s;x) - (\mathbf{1} - \boldsymbol{\rho})(t-s) + d(t) - d(s))$ holds since the idleness process is nondecreasing. Hence the result follows from the definition of the projection, combined with Assumption M, which asserts that the projection can be taken to be pointwise minimal. ∎

Theorem 3.10 establishes existence of minimal solutions and some strong robustness properties. This existence question is closely related to the *generalized Skorokhod problem* [21, 26, 2, 15, 16, 18]. These results will facilitate the treatment of stochastic models in section 4.

THEOREM 3.10. *Under Assumption* M*, for any given disturbance $\boldsymbol{d}$ of bounded variation, the model* (3.19) *admits a solution $\widehat{\boldsymbol{w}}^*$ that is pointwise minimal. For two disturbances $(\boldsymbol{d}^1, \boldsymbol{d}^2)$ the respective minimal solutions $(\widehat{\boldsymbol{w}}^{*1}, \widehat{\boldsymbol{w}}^{*2})$ satisfy the following:*

(i) *Provided $d_0^1(t) \leq d_0^2(t)$, $t \geq 0$,*

$$\widehat{w}^{*2}(t;x) \leq \widehat{w}^{*1}(t;x) + d^2(t) - d^1(t), \qquad t \geq 0,\ x \in \mathsf{X}.$$

(ii) *Suppose that* $d_0^1(t) = d_0^2(t) - \varepsilon_0(t)$, $t \ge 0$, *with* $\varepsilon_0(\cdot)$ *a nonnegative and nondecreasing function from* $\mathbb{R}_+$ *to* $\mathbb{R}_+^\ell$. *Then,*

$$\widehat{w}^{*1}(t;x) \le \widehat{w}^{*2}(t;x), \qquad t \ge 0,\ x \in \mathsf{X}.$$

(iii) *For arbitrary disturbances* $\boldsymbol{d}_0^1$, $\boldsymbol{d}_0^2$,

$$\widehat{w}^{*1}(t;x) \le \widehat{w}^{*2}(t;x) + |\boldsymbol{d}^2 - \boldsymbol{d}^1|_\infty^t - [d^2(t) - d^1(t)], \qquad t \ge 0,\ x \in \mathsf{X},$$

where $(|\boldsymbol{f}|_\infty^t)_i := \sup_{0\le s\le t} |f_i(s)|$ *for any function* $\boldsymbol{f}\colon \mathbb{R}_+ \to \mathbb{R}^n$.

Proof. We first establish the three properties, given that minimal solutions exist.

To prove (i), observe that if the optimal allocation $\widehat{\boldsymbol{z}}^{*1}$ for the first system is applied to the second, then we have for all $t \ge 0$

$$\begin{aligned} \widehat{q}^{*1}(t;x) &= x + B\widehat{z}^{*1}(t) + \alpha t + d_0^1(t), \\ \text{and}\quad \widehat{q}^{2}(t;x) &= x + B\widehat{z}^{*1}(t) + \alpha t + d_0^2(t) \ge \widehat{q}^{*1}(t;x) \ge \boldsymbol{\theta}. \end{aligned} \tag{3.21}$$

Hence $\widehat{\boldsymbol{z}}^{*1}$ is feasible for the second disturbance, and consequently $\widehat{w}^{*2}(t;x) \le \widehat{w}^{2}(t;x)$, with $\widehat{w}^{2}(t;x) := \widehat{\Xi}\widehat{q}^{2}(t;x)$, by the assumed existence of a minimal process $\widehat{\boldsymbol{w}}^{*2}$. Moreover, (3.21) implies that $\widehat{w}^{2}(t;x) = \widehat{w}^{*1}(t;x) + d^2(t) - d^1(t)$, $t \ge 0$, which gives (i).

The proof of (ii) is similar: Define $\varepsilon(t) := \widehat{\Xi}\varepsilon_0(t)$, $t \ge 0$. Under Assumption M and the conditions imposed in (ii), this is nonnegative and nondecreasing. Let $\hat{\imath}^{*i}(t) = \mathbf{1}t - \widehat{C}\widehat{z}^{*i}(t)$, $i = 1, 2$, denote the optimal idleness, and set $\hat{\imath}^{1}(t) = \hat{\imath}^{*2} + \varepsilon(t)$, $t \ge 0$. We have $\frac{d}{dt}\hat{\imath}^{1}(t) \ge \boldsymbol{\theta}$, and we also have under this policy, applied to the first model, $\widehat{w}^{1}(t;x) = \widehat{w}^{*2}(t;x)$. This combined with minimality of $\widehat{\boldsymbol{w}}^{*1}$ proves (ii).

To prove (iii) let $\boldsymbol{d}_0^3$ denote the disturbance $d_0^3(t) = d_0^1(t) + |\boldsymbol{d}_0^2 - \boldsymbol{d}_0^1|_\infty^t$, and let $\widehat{\boldsymbol{w}}^{*3}$ denote the associated minimal solution. We have $d_0^3(t) \ge d_0^2(t)$, and $\varepsilon_0(t) := d_0^3(t) - d_0^1(t)$ is nonnegative and nondecreasing. Consequently, for any $t \ge 0$, $x \in \mathsf{X}$, we have

$$\begin{aligned} \widehat{w}^{*3}(t;x) &\le \widehat{w}^{2*}(t;x) + |\boldsymbol{d}^2 - \boldsymbol{d}^1|_\infty^t + [d^1(t) - d^2(t)] && \text{from (i)}, \\ \widehat{w}^{*1}(t;x) &\le \widehat{w}^{3*}(t;x) && \text{from (ii)}. \end{aligned}$$

Combining these bounds gives (iii).

We now establish existence. Consider first the special case in which all of the disturbances are continuous and piecewise linear. In this case we may construct a pointwise minimal trajectory $\widehat{\boldsymbol{w}}^*$ inductively by adapting the construction used in Theorem 3.5. Set $\widehat{w}^*(0;x) = \widehat{\Xi}x$, and

$$\widehat{w}^*(T_k + t; x) = [\widehat{w}^*(T_k;x) - (\mathbf{1} - \boldsymbol{\rho})t + m_k t]_+, \qquad 0 < t < T_{k+1} - T_k, \quad k \ge 0,$$

where $\{T_i\}$ are the times at which the slope of $\boldsymbol{d}$ changes, and m_k denotes the slope of $\boldsymbol{d}$ on the interval $[T_k, T_{k+1}]$. An application of Theorem 3.5 shows that this is the desired minimal solution on $[T_k, T_{k+1}]$ with initial condition $\widehat{w} = \widehat{w}^*(T_k;x)$, and by induction it follows that $\widehat{\boldsymbol{w}}^*$ is pointwise minimal.

For an arbitrary disturbance $\boldsymbol{d}$ of bounded variation we can construct a sequence of piecewise linear functions $\{\boldsymbol{d}^k\}$ such that $d^k(t) \downarrow d(t)$, $k \to \infty$. We let $\{\widehat{\boldsymbol{w}}^{*k}\}$ denote the respective optimal solutions and set $\widehat{w}^*(t;x) = \lim\inf_{k\to\infty} \widehat{w}^{*k}(t;x)$ for all t, x. Using property (i) for the $\{\widehat{\boldsymbol{w}}^{*k}\}$ we deduce that $\widehat{\boldsymbol{w}}^*$ is the desired pointwise minimal solution. ∎

We see that it is frequently possible to compute a pointwise optimal trajectory $\widehat{\boldsymbol{q}}^*$ for the relaxed control problem, with or without disturbances. What does this

then tell us about the original model of interest? The sharpest results are obtained by examining a model in heavy traffic, with $\rho \sim 1$.

4. Networks in heavy traffic. We consider here a sequence of networks, indexed by an integer $r \geq 1$, for which $\rho^r \uparrow 1$ as $r \to \infty$. It is in this heavily loaded regime that the time-scale separation developed in the previous section is most evident in the (unrelaxed) network model.

We assume that B and C are independent of r. Two arrival-rate vectors α^1, α^∞ are given, and for arbitrary $r \geq 1$ we set

$$\alpha^r := \alpha^\infty - \tfrac{1}{r}(\alpha^\infty - \alpha^1). \tag{4.1}$$

We impose the following assumptions throughout this section.

Assumption H.

(i) The model with arrival-rate vector α^1 is stabilizable. In particular,

$$\rho^1 := W^*(\alpha^1) < 1\,.$$

(ii) The arrival-rate vector α^∞ satisfies $\alpha^1 \leq \alpha^\infty$ and

$$\rho^\infty := W^*(\alpha^\infty) = 1.$$

We let $\mathcal{I}_b = \{i : \langle \xi^i, \alpha^\infty \rangle = 1\}$ denote the index set of bottleneck stations for the model with arrival rate α^∞. By reordering, we can assume, without loss of generality, that $\mathcal{I}_b = \{1, \ldots, \ell_b\}$ for some integer $\ell_b \geq 1$.

The choice of a perturbation in the arrival stream is for the sake of convenience since we can then take a fixed set of workload vectors. If we assume that V_r is a general, convergent sequence of polyhedra, then the theory below remains essentially unchanged.

Throughout this section we consider the nth workload-relaxation with $n = \ell_b$.

4.1. Fluid models. The rth network is defined on a fluid scale by

$$\frac{d}{dt} q(t;x) = B\zeta(t;x) + \alpha^r t, \qquad t \geq 0. \tag{4.2}$$

We let V_r denote the corresponding velocity space so that $\frac{d}{dt} q(t;x) \in \mathsf{V}_r$ for all t, x, r.

The following bound on ρ^r shows that this model is stabilizable for finite $r \geq 1$. The inequality is obtained using convexity of W^*:

$$\begin{aligned} \rho^r = W^*(\alpha^r) &= W^*\left((1 - \tfrac{1}{r})\alpha^\infty + \tfrac{1}{r}\alpha^1\right) \\ &\leq (1 - \tfrac{1}{r})\rho^\infty + \tfrac{1}{r}\rho^1 = 1 - \tfrac{1}{r}(1 - \rho^1) < 1. \end{aligned} \tag{4.3}$$

For finite r we have $\rho_i^r = 1 - r^{-1}\langle \xi^i, \alpha^\infty - \alpha^1 \rangle$, $i \in \mathcal{I}_b$.

Theorem 4.1 shows that little is lost when considering the ℓ_bth relaxation. Let J^*, $\widehat{J}^*$ denote the value functions for the infinite-horizon optimal control problems defined in (2.3), (3.4), respectively. We always have

$$\widehat{J}^*(x) \leq J^*(x), \qquad x \in \mathsf{X}.$$

We obtain a bound in the reverse direction in this section. The analysis is simplest when optimal trajectories are uniquely defined.

Assumption U.

(i) The linear program (3.9) that defines the effective state $\mathcal{X}^*(w)$ has a unique solution for each $w \in \widehat{\mathsf{W}}$.

(ii) For all $r \geq 1$ sufficiently large and each $T > 0$, $x \in \mathsf{X}$, the ℓ_bth workload-relaxation admits a solution $\widehat{\boldsymbol{q}}^{r*}$ that minimizes the total cost (2.6), and this solution is *unique*.

Consider for example the one-dimensional relaxation of the simple routing model shown in Figure 2. Assume that the cost is linear, so that $c(x) = \langle c, x\rangle$, with $c \in \mathbb{R}^3_+$. If $c_3 \geq c_2 > c_1$, then the above conditions hold. The greedy priority policy that prefers routing to buffer 1, whenever buffer 2 is nonempty, is the unique (pointwise) optimal solution.

Note that Assumption U(i) implies (ii) under Assumption M since in this case $\widehat{q}^*(t;x) = \mathcal{X}^*(\widehat{w}^*(t;x))$, and the pointwise minimal solution $\widehat{\boldsymbol{w}}^*$ is always uniquely defined when it exists.

Applying (3.5) and the form of the rate vector given in (4.1), we find that the constraints on the workload relaxation may be expressed as

$$\frac{d}{dt}\widehat{w}_i(t;x) \geq -\tfrac{1}{r}\delta_i, \qquad 1 \leq i \leq \ell_b,\ r \geq 1,\ t > 0\,,$$

where $\delta_i = \langle \xi^i, \alpha^\infty - \alpha^1\rangle$. Letting $\widehat{\boldsymbol{w}}^{1*}, \widehat{J}^{1*}$ denote the optimal trajectory and value function when $r = 1$, it follows that for any $r \geq 1$ the optimal solution is given by

$$\begin{aligned} \widehat{w}_i^*(t;x) &= \widehat{w}_i^{1*}(t/r;x), \qquad 1 \leq i \leq \ell_b,\ t > 0\,, \\ \widehat{J}^*(x) &= r\widehat{J}^{1*}(x), \qquad\qquad x \in \mathsf{X}\,. \end{aligned} \tag{4.4}$$

We define a policy for the unrelaxed model as follows. Applying Proposition 3.1 we are assured of the existence of a piecewise linear, optimal solution to the relaxed control problem, which we denote $[\widehat{q}^*(t;x), \widehat{\zeta}^*(t;x)]$. The allocation rate $\zeta(t;x)$ for the unrelaxed model is defined to be a function of $[\widehat{q}^*(t;x), \widehat{\zeta}^*(t;x), q(t;x)]$ for any initial condition x and any $t \geq 0$. Let $\mathcal{I}_c(x) = \{i : c(x) = \langle c^i, x\rangle\}$, and given the current states $y = q(t;x)$, $y^* = \widehat{q}^*(t;x)$, let $\zeta(t;x)$ be the optimizing value of the variable ζ in the linear program

$$\begin{aligned} &\min \quad \gamma \\ &\text{subject to} \quad \begin{aligned} \gamma &\geq \langle c^i, B\zeta\rangle, \quad i \in \mathcal{I}_c(y), \\ C\zeta &\leq \mathbf{1}, \\ \zeta &\geq \boldsymbol{\theta}, \\ (B\zeta + \alpha^r)_i &\geq 0 \quad \text{if } y_i = 0, \\ \langle \xi^i, (B\zeta + \alpha^r)\rangle &\leq \langle \xi^i, (B\widehat{\zeta}^* + \alpha^r)\rangle, \quad \text{whenever } i \leq \ell_b, \\ &\qquad\qquad\qquad\qquad\quad \text{and } \langle \xi^i, y\rangle = \langle \xi^i, y^*\rangle. \end{aligned} \end{aligned} \tag{4.5}$$

The last constraint ensures that $w_i(t;x) \leq \widehat{w}_i^*(t;x)$ for all $i \leq \ell_b$ and all t.

Assume that $q(t;x)$ is the resulting state trajectory using this policy for all t, and set

$$e^r(t;x) = q(t;x) - \widehat{q}^*(t;x),\ t > 0, \qquad \underline{T}_{r\circ}(x) = \min\{t : e^r(t;x) = \boldsymbol{\theta}\}.$$

The following result provides uniform bounds on $\underline{T}_{r\circ}$ and shows that this first hitting time is in fact a *coupling time*. It is possible to relax the uniqueness assumption in

Theorem 4.1, but one must redefine $\underline{T}_{r\circ}(x)$ as the first hitting time to *some* optimal $\widehat{q}^*(t;x)$. A proof is provided in Appendix A.

THEOREM 4.1. *Under Assumptions* H *and* U, *the following hold for the state trajectory* $\boldsymbol{q}$ *for any initial condition* $x \in \mathsf{X}$*:*

(i) $w(t;x) \le \widehat{w}^*(t;x)$, $t \ge 0$, *where the inequality is interpreted componentwise.*

(ii) *The time* $\underline{T}_{r\circ}$ *is uniformly bounded in* r*: For some* $b_0 < \infty$,

$$\underline{T}_{r\circ}(x) \le b_0 \|e^r(0+;x)\|\,, \qquad x \in \mathsf{X},\ r \ge 1.$$

(iii) $q(t;x) = \widehat{q}^*(t;x)$ *for all* $t \ge \underline{T}_{r\circ}(x)$.

(iv) *There is a constant* b_1 *such that*

$$\begin{aligned} J(x) := \int_0^\infty c(q(t;x)) &\le (1 + b_1/r)\widehat{J}^*(x) \\ &\le (1 + b_1/r)J^*(x)\,, \qquad x \in \mathsf{X},\ r \ge 1. \end{aligned}$$

(v) *Suppose that* $\widehat{\boldsymbol{q}}^*$ *is a pointwise optimal solution. Then*

$$c(q(t;x)) = \underline{c}^*(t;x), \qquad t \ge \underline{T}_{r\circ}(x),$$

where $\underline{c}^*$ *is given in* (2.9). □

4.2. Stochastic models. Although the workload-relaxation is in general a significant distortion of the original model, we have seen in Theorem 4.1 that this is negligible when the model is in heavy traffic. The workload constraints overwhelm all other constraints on the velocity vector field. In this section we establish similar solidarity for the stochastic model.

To obtain any such solidarity we must control modelling error, and we must understand when if ever a user can benefit from statistical information. Consider a $G/G/2$ queue, where the two servers are constrained so that only one can work at any given time-instance. The fluid model is given by the one-dimensional model

$$\frac{d}{dt}q(t;x) = -\mu_1\zeta_1(t;x) - \mu_2\zeta_2(t;x) + \alpha, \tag{4.6}$$

with the linear control constraint $\zeta_1 + \zeta_2 \le 1$. This can be viewed as an idealized two-armed bandit, where α is the rate at which a gambler is paying the casino, and $\mu_i\zeta_i$ is his rate of return on using the ith arm. The casino's reward at time t is a linear function of $q(t)$. If $\mu_1 = \mu_2 > \alpha$, then obviously any nonidling policy is optimal, from the gambler's point of view, for any monotone cost function.

For the stochastic model, however, the particular allocation chosen can have great impact since variability of service rates determines the steady-state queue length. For a priority policy in which server i is used exclusively, we obtain in steady-state an approximation of the form, for $\rho \sim 1$,

$$\mathsf{E}[Q(t)] = \tfrac{1}{2}\frac{\gamma^2}{1-\rho} + O(1).$$

The infinite-horizon optimal policy is precisely the priority policy that chooses the server with the smallest variability parameter γ^2.

This example is special because the optimal fluid policy is *not unique*. Typically, the optimal control problem for the fluid model may be solved uniquely since linear

programs generically have unique solutions. If this is the case, then we have fewer opportunities to successfully gamble.

In Theorem 4.3 we impose uniqueness through Assumption U, and an assumption that B is full-rank with $\ell \geq \ell_u$, so that $\boldsymbol{Z}$ is essentially determined by $\boldsymbol{Q}$. The latter assumption may be relaxed considerably by expanding the state space.

Take, for example, the routing model in which B is the 3×4 matrix given in (3.12). Consider the associated four-dimensional network model $\boldsymbol{Q}^{\mathrm{a}}$ on $\mathsf{X}:=\mathbb{R}^4_+$, in which the fourth component is the total-idleness at buffer 1, given by $Q^{\mathrm{a}}_4(t;x) = t - Z_1(t;x)$, $t \geq 0$. The associated matrix B^{a} is invertible, as seen by the explicit form

$$B^{\mathrm{a}} = \begin{bmatrix} -\mu_1 & 0 & \mu_3 & 0 \\ 0 & -\mu_2 & 0 & \mu_3 \\ 0 & 0 & -\mu_3 & -\mu_3 \\ -1 & 0 & 0 & 0 \end{bmatrix}, \qquad \alpha^{\mathrm{a}} = \begin{bmatrix} 0 \\ 0 \\ \alpha_3 \\ 1 \end{bmatrix}.$$

If the cost function on $\boldsymbol{Q}^{\mathrm{a}}$ is assumed linear, with $c_1 < c_2 < c_3$ and $c_4 > 0$, then Assumption U holds for the four-dimensional model.

For any network model one may augment the state space to include total-idleness, as well as total-allocation values. The cost may be similarly augmented to reflect the desire to maximize utilization of some resources, while minimizing utilization of others. The augmented model will satisfy assumptions (i)–(iii) of Theorem 4.3 for a very general class of network models and cost criteria.

How do we choose the allocation $\boldsymbol{Z}$ to maintain solidarity with an ideal fluid solution $[\widehat{\boldsymbol{q}}^*, \widehat{\boldsymbol{z}}^*]$? There are three issues that must be addressed:

(i) Suppose that for a given state x, a state $x^* \in \mathsf{X}$ is chosen as a target, with $\widehat{\Xi} x^* \in \widehat{\mathsf{W}}^+$. For the fluid model, even if the buffers are empty, an associated resource may be required to work at full capacity. This is not feasible for the discrete model: if a resource finds no work available, then it cannot work. This may be disastrous if the resource is a dynamic bottleneck since any idle time will rule out time-optimality.

(ii) To ensure feasibility we can impose small safety stocks, a well-motivated and standard technique in policy synthesis for manufacturing models [13, 20, 36]. We must ensure that these safety-stock levels can be maintained through a modification of the fluid-allocation without introducing idleness.

(iii) To ensure success we require bounds on the variability of the stochastic processes $(\boldsymbol{A}, \boldsymbol{R}, \boldsymbol{S})$ used in the stochastic model.

To simplify the statements of our assumptions we henceforth assume that the stochastic model (2.1) has the following specific form: For each $1 \leq i \leq \ell$ and $t \geq 0$,

$$Q^r_i(t;x) = x - \sum_{j=1}^{\ell_u} S_{ij}(Z_j(t)) + \sum_{j=1}^{\ell_u} R_{ij}(Z_j(t)) + A_i(t(1 - r^{-1}\delta^\alpha_i)),$$

where $\delta^\alpha_i := (\alpha^\infty_i - \alpha^1_i)/\alpha^\infty_i$ for each i, and the arrival-rate vectors α^1, α^∞ satisfy Assumption H. We assume that the stochastic model is consistent with the fluid model, in the sense that (2.4) holds with $\alpha = \alpha^\infty$. In particular, if $\alpha^\infty_i = 0$, then the process $\boldsymbol{A}_i$ is null. Assumption S formalizes our remaining probabilistic assumptions. Under this condition we can devise a policy that tracks any fluid idealization and simultaneously ensures that critical resources do not risk starvation.

Assumption S. For all $1 \leq i \leq \ell$ and $1 \leq k \leq \ell_u$, each of the stochastic processes $\{\boldsymbol{A}_i, \boldsymbol{R}_{ik}, \boldsymbol{S}_{ik},\ t \geq 0\}$ is either null or is an undelayed renewal process whose increment process possesses a moment generating function that is bounded in a neighborhood

of the origin. The stochastic processes $\{\boldsymbol{A}, \boldsymbol{R}, \boldsymbol{S}\}$ are adapted to a given filtration $\{\mathcal{H}_t : t \geq 0\}$.

We continue to assume that the allocation process $\boldsymbol{Z}$ satisfies the constraints (2.2), and we assume that any allocation $\boldsymbol{Z}$ is progressively measurable in the sense that

$$\sigma(Q^r(s), Z^r(s) : s \leq t) \subset \mathcal{H}_t, \qquad t \geq 0\,.$$

A relaxed model $[\widehat{\boldsymbol{Q}}, \widehat{\boldsymbol{Z}}]$ is defined in analogy with (3.2), in which the allocation constraint is relaxed to

$$\widehat{C}[\widehat{Z}(t;x) - \widehat{Z}(s;x)] \leq [t-s]\mathbf{1}\,, \qquad \widehat{C} := -\widehat{\Xi}^{\scriptscriptstyle T} B\,. \tag{4.7}$$

This is of course subject to the additional constraint that $\widehat{Q}(t;x)$ evolves in $\mathsf{X} := \mathbb{R}^{\ell}_{+}$. We assume that $\widehat{\boldsymbol{Z}}$ is of bounded variation, but unlike $\boldsymbol{Z}$, it is not subject to any statistical constraints.

For any feasible pair $[\widehat{\boldsymbol{Q}}, \widehat{\boldsymbol{Z}}]$ we define the *pseudodisturbance* $\boldsymbol{d}_0$ through the equation

$$\widehat{Q}(t;x) = x + B\widehat{Z}(t;x) + \alpha^{\scriptscriptstyle T} t + d_0(t)\,, \quad t \geq 0\,, \tag{4.8}$$

and we let $d(t) = \widehat{\Xi} d_0(t)$, $t \geq 0$. The associated workload process may be expressed in terms of $\boldsymbol{d}$ as follows: first define the idleness process by $\widehat{I}(t;x) := t\mathbf{1} - \widehat{C}\widehat{Z}(t;x)$, $t \geq 0$. This is vector-valued, and (4.7) implies that its components are nonnegative and nondecreasing. We then write

$$\widehat{W}(t;x) := \widehat{\Xi}\widehat{Q}(t;x) = -(\mathbf{1} - \boldsymbol{\rho}^r)t + \widehat{I}(t;x) + d(t)\,.$$

We consider below the optimal solution $[\widehat{\boldsymbol{q}}^*, \widehat{\boldsymbol{z}}^*]$ to the ℓ_bth fluid-model relaxation (3.17) with respect to the (random) pseudodisturbance $\boldsymbol{d}_0$. This of course depends upon $\widehat{\boldsymbol{Z}}$. These processes are used for comparison to obtain performance bounds. For example, under the conditions of Theorem 3.10 we obviously have the absolute lower bound, $\widehat{W}(t;x) \geq \widehat{w}^*(t;x) := \widehat{\Xi}\widehat{q}^*(t;x)$, $t \geq 0$. Perhaps surprisingly, the policies considered below almost achieve this lower bound, uniformly for the time-horizons considered.

4.3. Sensitivity and optimality. In the development that follows we construct a trajectory $[\boldsymbol{Q}^{r\circ}, \boldsymbol{Z}^{r\circ}]$ by attempting to mimic the flow of the optimal fluid trajectory. We begin with a list of desirable properties that $[\boldsymbol{Q}^{r\circ}, \boldsymbol{Z}^{r\circ}]$ should satisfy. In Theorem 4.3 we show that these general properties imply a strong form of approximate optimality.

Following this we provide a constructive procedure for policy synthesis to attain these properties. This requires some assumptions on the model that we illustrate first in one dimension in section 4.4 and then for general models in section 4.5.

The following result is central to all of the remaining analysis in this section and is essentially our only motivation for Assumption S. A proof may be found in Appendix B.

If $(\boldsymbol{X}, \boldsymbol{Y}) = \{(X_r, Y_r) : r \geq 1\}$ is a sequence of random variables, we write $\boldsymbol{X} \leq O(\boldsymbol{Y})$ if $X_r \leq b_\bullet Y_r$ for some fixed deterministic constant $b_\bullet$, $t \geq 0$, and we write $\boldsymbol{X} \leq o(\boldsymbol{Y})$ if $\lim_{r\to\infty} X_r/Y_r = 0$ a.s. The constant $b_\bullet$ is assumed fixed throughout.

PROPOSITION 4.2. *Let $\boldsymbol{X}$ be a real-valued i.i.d. process with common mean $m = \mathsf{E}[X_i] > 0$ and moment generating function bounded in a neighborhood of the origin.*

There exists $I_0 > \infty, \delta_0 > 0, B_0 < \infty$ such that for all $0 < \delta \le \delta_0$, we have the following:

(i) *For any $N \ge 1$, writing $S_T := \sum_{i=1}^T X_i$,*

$$\mathsf{P}\Big\{ \inf_{T\ge 1} (S_T - (m-\delta)T) \le -N \Big\} \le B_0 \exp(-I_0 \delta^2 N).$$

(ii) *Let $\boldsymbol{Y}$ be the undelayed renewal process with increment process $\boldsymbol{X}$. There exists $B_1 < \infty$ such that for $k_0 \ge 2$,*

$$\lim_{r\to\infty} \sup_{0\le s\le t\le r^{k_0}} \left(\frac{Y(t) - Y(s) - (t-s)(m^{-1}+\delta)}{\log(r)} \right) \le B_1 k_0 \delta^{-2} \qquad a.s. \qquad \square$$

Throughout this section we let $[\widehat{\boldsymbol{Q}}, \widehat{\boldsymbol{Z}}]$ denote any feasible trajectory for the relaxed stochastic model. It is defined on the same sample space through identical generating processes $(\boldsymbol{A}, \boldsymbol{R}, \boldsymbol{S})$. Our goal is to construct a policy for (2.1) that uniformly outperforms any such feasible trajectory on a time-window of the form $[\underline{T}_{r^\bullet}, T_{r^\bullet}]$, where

$$\underline{T}_{r^\bullet} = b_0[\|x - \mathcal{P}^*(x)\| + \log(r)], \qquad T_{r^\bullet} = r^3, \tag{4.9}$$

with $b_0 < \infty$ sufficiently large.

The following two uniform bounds will be established for the policies constructed below, and for the optimal policy. Property P1 appears to be desirable for any network and any cost function on X. However, Property P2 is desirable only when the effective cost is monotone.

Recall that $\widehat{\boldsymbol{w}}^*$ denotes the minimal solution to the workload relaxation (3.17), where the disturbance $\boldsymbol{d}_0$ is defined in (4.8).

Property P1 *(relative optimization).* For any $x \in \mathsf{X}$, $r \ge 1$,

$$\|\widehat{Q}(t;x) - \mathcal{P}^*(\widehat{Q}(t;x))\| \quad \le \quad O(\log(r)) + o(1), \qquad \underline{T}_{r^\bullet} \le t \le T_{r^\bullet}, \quad \text{a.s.}$$

Property P2 *(relative minimal workload).* For any $x \in \mathsf{X}$, $r \ge 1$,

$$\widehat{W}(t;x) - \widehat{w}^*(t;x) \quad \le \quad O(\log(r)) + o(1), \qquad \underline{T}_{r^\bullet} \le t \le T_{r^\bullet}, \quad \text{a.s.}$$

THEOREM 4.3. *Suppose that $\ell \ge \ell_u$ in* (2.1) *and the following additional assumptions hold:*

(i) *Assumption* M *holds with $n = \ell_b$, and the effective cost $\bar{c}$ for the ℓ_bth workload-relaxation is monotone.*

(ii) *Assumptions* H, S, *and* U *hold, and the matrix B has rank ℓ_u.*

(iii) *The pair $[\boldsymbol{Q}^{r\circ}, \boldsymbol{Z}^{r\circ}]$ satisfies conditons* P1 *and* P2.

Then, as $r \to \infty$,

$$\sup_{[\widehat{Q}^r, \widehat{Z}^r]} \left(\sup_{0\le T\le T_{r^\bullet}} \left(\frac{1}{T} \int_0^T [c(Q^{r\circ}(t;x)) - c(\widehat{Q}^r(t;x))]\, dt \right) \right) \le O(\log(r)) + o(1).$$

Proof. Given the allocation $\boldsymbol{Z}^{r\circ}$, and any other allocation $\widehat{\boldsymbol{Z}}^r$ satisfying (4.7), we can construct respective pseudodisturbances $\boldsymbol{d}_0^{r\circ}$, $\boldsymbol{d}_0^r$ via (4.8).

The proof of Theorem 4.3 is based upon a comparison of the respective optimal solutions to the ℓ_bth fluid-model relaxation (3.17), denoted $[\widehat{\boldsymbol{q}}^{r\circ^*}, \widehat{\boldsymbol{z}}^{r\circ^*}]$ and $[\widehat{\boldsymbol{q}}^{r^*}\, \widehat{\boldsymbol{z}}^{r^*}]$.

This comparison is made possible via the following "coupling property": For any $x \in \mathsf{X}$, and all small $\delta > 0$,

(4.10)
$$\|d_0^{r\circ}(t) - d_0^r(t)\| \le \delta \|Z^{r\circ}(t) - \widehat{Z}^r(t)\| + O(\delta^{-2}\log(r)) + o(1), \quad 0 \le t \le T_{r\bullet}, \text{ a.s.}$$

This bound follows directly from Proposition 4.2 and Assumption S.

Rather than a general allocation, for each r we consider a "near-optimal" solution $[\widehat{\boldsymbol{Q}}^r, \widehat{\boldsymbol{Z}}^r]$ defined as follows. We fix $0 < T_\bullet \le T_{r\bullet}$, and we assume that for any other solution $[\widehat{\boldsymbol{Q}}, \widehat{\boldsymbol{Z}}]$,

$$\frac{1}{T_\bullet}\int_0^{T_\bullet} c(\widehat{Q}^r(t;x))\,dt \le \frac{1}{T_\bullet}\int_0^{T_\bullet} c(\widehat{Q}(t;x))\,dt + O(\log(r)).$$

Recall that in this notation $O(\log(r)) \le b_\bullet \log(r)$ with $b_\bullet$ fixed throughout, so that the above bound is uniform in $\{\widehat{\boldsymbol{Q}}\}$. It is shown in Proposition B.2 in Appendix B that a solution can be chosen so that $[\widehat{\boldsymbol{Q}}^r, \widehat{\boldsymbol{Z}}^r]$ satisfies conditions P1 and P2.

Combining conditions P1 and P2 for $[\widehat{\boldsymbol{Q}}^{r\circ}, \widehat{\boldsymbol{Z}}^{r\circ}]$ and $[\widehat{\boldsymbol{Q}}^r, \widehat{\boldsymbol{Z}}^r]$ gives

$$\begin{aligned} \|Q^{r\circ}(t;x) - \widehat{q}^{r\circ^*}(t;x)\| &\le O(\log(r)) + o(1), \\ \|\widehat{Q}^r(t;x) - \widehat{q}^{r^*}(t;x)\| &\le O(\log(r)) + o(1) \quad \text{a.s.} \end{aligned} \tag{4.11}$$

Theorem 3.10 and Assumption U give

$$\|\widehat{q}^{r\circ^*}(t;x) - \widehat{q}^{r^*}(t;x)\| \le O(\|\boldsymbol{d}_0^{r\circ} - \boldsymbol{d}_0^r\|_\infty^t)\,. \tag{4.12}$$

Combining (4.11), (4.12) with (4.10) and the rank condition on B then gives

$$\begin{aligned} \|Z^{r\circ}(t;x) - \widehat{Z}^r(t;x)\| &\le O(\|Q^{r\circ}(t;x) - \widehat{Q}^r(t;x)\|) + O(\|d_0^{r\circ}(t) - d_0^r(t)\|) \\ &\le O(\|\boldsymbol{d}_0^{r\circ} - \boldsymbol{d}_0^r\|_\infty^t) + O(\log(r)) + o(1) \\ &\le \tfrac{1}{2}\|\boldsymbol{Z}^{r\circ} - \widehat{\boldsymbol{Z}}^r\|_\infty^t + O(\log(r)) + o(1) \\ &\le \tfrac{1}{2}\|\boldsymbol{Z}^{r\circ} - \widehat{\boldsymbol{Z}}^r\|_\infty^{T_\bullet} + O(\log(r)) + o(1) \end{aligned}$$

uniformly for $0 \le t \le T_\bullet$. It follows that $\|\boldsymbol{Z}^{r\circ} - \widehat{\boldsymbol{Z}}^r\|_\infty^{T_\bullet} = O(\log(r))$, and this easily implies the result. ∎

The proof of Theorem 4.3 hinges on uniqueness of $[\widehat{\boldsymbol{q}}^*, \widehat{\boldsymbol{z}}^*]$ for a given disturbance $\boldsymbol{d}$. Without uniqueness one can attempt to search over optimal fluid allocations whose associated translation $\boldsymbol{Z}^{r\circ}$ has minimal cost, as in the "two-armed bandit" (4.6). The monotonicity assumption is also critical and, as we have seen, often fails in complex network models when the workload dimension is taken larger than one. We return to this issue in section 4.5.

How then can we design a policy that satisfies conditions P1 and P2? We present here an approach based on a "discrete-review" structure, following [27, 33, 1]. Let $T_r > 0$, $\bar{x}^r \in \mathsf{X}$ denote, respectively, a planning horizon and safety-stock levels for the rth network. We take the explicit form

$$T_r = K_0 \log(r), \quad \bar{x}_i^r = K_1 \log(r)\bar{x}_i, \qquad r \ge 1,\ 1 \le i \le \ell, \tag{4.13}$$

where K_j, $j = 0, 1$, and $\bar{x}_i$, $i = 0, \dots, \ell$, are strictly positive constants. The ratio $\Delta_0 := K_1/K_0$ determines the likelihood of starvation.

In practice, taking a fixed safety-stock level is neither desirable nor practical—a fixed value $\bar{x}^r$ is chosen for convenience. A more desirable choice may be a "moving target," such as

$$\bar{x}^r = K_1 \log(\|x\| + 1)\bar{x}\,,$$

where x is the current state of the network. It is also not necessary to assume strict positivity of *every* element of $\bar{x}$: it is only necessary to assume that every pooled-resource, for $i \le \ell_r$, can work at capacity at time t if $q(t;x) \ge \bar{x}$. The results below can be extended to cover such generalizations.

We choose the allocation rates exactly as in the fluid-translation (4.5), except that we introduce safety-stock constraints that may be viewed as a shift of the origin. Let $\bar{w}^r = \widehat{\Xi}\bar{x}^r$, $r \ge 1$, and denote by $[w]^r_+$ the projection of w, in the standard ℓ_2 norm, onto the set $\widehat{\mathsf{W}}^r_w := \{\widehat{w} + \bar{w}^r : \widehat{w} \in \widehat{\mathsf{W}},\ \widehat{w} + \bar{w}^r \ge w\}$. This is equal to the pointwise minimal element of this set under Assumption M.

Fix $\delta_0 > 0$, and consider the following linear program:

$$\begin{aligned}
&\min \quad \gamma \\
&\text{subject to} \quad
\begin{aligned}
\gamma &\ge \langle c^i, y\rangle, \qquad 1 \le i \le \ell_c, \\
y &= x + Bz + \alpha^r T_r, \\
y_i &\ge (x_i + \delta_0 \bar{x}^r_i) \wedge \bar{x}^r_i, \qquad 1 \le i \le \ell, \\
\widehat{\Xi} y &\le [\widehat{\Xi} x - (\mathbf{1} - \boldsymbol{\rho})T_r]^r_+, \\
Cz &\le T_r \mathbf{1}, \\
z &\ge \boldsymbol{\theta}.
\end{aligned}
\end{aligned} \tag{4.14}$$

We assumed in Assumption M that the workload vectors satisfy $\{\xi^i : 1 \le i \le \ell_r\} \subset \mathbb{R}^\ell_+$. Under this condition, an application of Lemma A.1 implies that $\delta_0 > 0$ may be chosen sufficiently small so that this linear program is feasible for any $r \ge 1$.

Given a solution z^* to (4.14) we then set $\zeta^{r\circ} := z^*/T_r$, and $Z^{r\circ}(t;x) = t\zeta^{r\circ}$, $0 \le t \le T_r$. In practice, additional constraints on $\boldsymbol{Z}$ will force an approximation, but this will be negligible for large r. This can then be repeated for each interval $[kT_r, (k+1)T_r]$ for $k \ge 0$ to obtain $(Q^{r\circ}(t), Z^{r\circ}(t))$ for $t \ge 0$. On any time-interval $[kT_r, (k+1)T_r]$ the buffers behave like decoupled $G/G/1$ queues.

In addition to feasibility of the linear program (4.14), the definition of $\boldsymbol{Z}^{r\circ}$ requires *feasibility of the resulting state trajectory* so that $Q^{r\circ}(t;x) \in \mathsf{X}$ for all $t \ge 0$. Positivity of $\boldsymbol{Q}$ and approximate optimality follow from the large deviation bound in Proposition 4.2. We demonstrate this, and provide conditions under which P1–P2 also hold in the following two subsections.

4.4. One-dimensional workload. In this case there is a single set of pooled bottleneck-resources to be considered, and we set $\xi = \xi^1$, $\mathcal{R}^\circ = \mathcal{R}^\circ_1$. This case is special since the effective cost is always monotone, and the relaxed control problem admits a simple, pointwise optimal solution (see Proposition 3.2).

Recall that b_0 determines the times $\underline{T}_{r\bullet}$, and $\Delta_0 = K_1/K_0$ (see (4.13)).

THEOREM 4.4. *Suppose that the following assumptions hold:*

(i) *Assumption* M *holds with* $n = \ell_b = 1$.

(ii) *Assumptions* H, S, *and* U *hold.*

Then for all $\Delta_0 > 0$ *sufficiently large, there exists* $b_0 < \infty$ *such that Properties* P1 *and* P2 *hold for the policy defined via the linear program* (4.14).

Proof. In the one-dimensional case, provided $\langle \xi, x\rangle \ge 2\bar{w}^r := 2\langle \xi, \bar{x}^r\rangle$, the linear program (4.14) to obtain $\zeta^{r\circ}$ on $[0, T_r]$ reduces to

$$
\begin{aligned}
&\min \quad \gamma \\
&\text{subject to} \qquad
\begin{aligned}
\gamma &\ge \langle c^i, y\rangle, \qquad 1 \le i \le \ell_c, \\
y &= x + Bz + \alpha^r T_r, \\
y_i &\ge (x_i + \delta_0 \bar{x}_i^r) \wedge \bar{x}_i^r, \qquad 1 \le i \le \ell, \\
Cz &\le T_r \mathbf{1}, \\
z &\ge \boldsymbol{\theta}, \\
\langle \xi, Bz\rangle &= -T_r.
\end{aligned}
\end{aligned} \tag{4.15}
$$

Given a solution z^* to (4.15) we then set $\zeta^{r\circ} := T_r^{-1} z^*$. Let $y^{r\circ}$ denote the associated optimal state, starting from the initial condition x:

$$
y^{r\circ} = x + (B\zeta^{r\circ} + \alpha^r) T_r.
$$

As in the deterministic setting, we consider the error process

$$
E^{r\circ}(t; x) := Q^{r\circ}(t; x) - \mathcal{P}^*(Q^{r\circ}(t; x)). \tag{4.16}
$$

One can show as in Theorem 4.1 that for some fixed $\delta > 0$ independent of r, whenever $\|E^{r\circ}(0)\| = \|x - \mathcal{P}^*(x)\| \ge 2\|\bar{x}^r\|$, and $T^*(x) \ge T^*(\bar{x}^r) + T_r$,

$$
\begin{aligned}
\|y^{r\circ} - \mathcal{P}^*(y^{r\circ})\| &\le \|E^{r\circ}(0)\| - \delta T_r, \\
\mathsf{E}[\|E^{r\circ}((k+1)T_r)\| \mid \mathcal{H}_{kT_r}] &\le \|E^{r\circ}(kT_r)\| - \delta T_r.
\end{aligned} \tag{4.17}
$$

We will show that this implies P1, and that the constraint $\langle \xi, Bz\rangle = -T_r$ in (4.15) implies P2.

For $k \ge 1$ let $G_{r,k}$ denote the union of "good events,"

$$
\begin{aligned}
G_{r,k} = \Big\{ W^{r\circ}(kT_r) \le 2\bar{w}^r \Big\} \\
\bigcup \Big\{ \tfrac{1}{2}\delta_0 \bar{x}^r \le E^{r\circ}(t; x) \le \tfrac{3}{2}\bar{x}^r, \quad kT_r \le t \le (k+1)T_r, \\
\text{and} \quad \langle \xi, B[Z^{r\circ}((k+1)T_r) - Z^{r\circ}(kT_r)]\rangle = -T_r \Big\},
\end{aligned}
$$

and define for any r

$$
G_r = \bigcup_{\underline{T}_{r\bullet} \le k \le T_{r\bullet}/T_r} G_{r,k}.
$$

For sufficiently large b_0 and constants $B_2 < \infty$, $I_2 > 0$, Proposition 4.2 implies the bound $\mathsf{P}\{G_r^c\} \le r^3 B_2 \exp(-I_2 \Delta_0 \log(r))$, $r \ge 1$. For $\Delta_0 \ge 5 I_2^{-1}$ this is bounded by $B_2 r^{-2}$, and it then follows that

$$
\sum_{r=1}^{\infty} \mathsf{P}\{G_r^c\} < \infty.
$$

From the Borel–Cantelli lemma we can conclude that, with probability one, each state-allocation trajectory $[\boldsymbol{Q}^{r\circ}, \boldsymbol{Z}^{r\circ}]$ eventually satisfies G_r for large enough r. It follows that P1 and P2 also hold and that $\boldsymbol{Q}^{r\circ}$ evolves in X for all large r. ∎

4.5. Higher dimensions. For the general workload dimension, even if the fluid model admits a pointwise optimal solution, one cannot hope to obtain the strong sample-path optimality established in Theorem 4.4 for the stochastic model. Consider the workload processes

$$W^r(t;x) = \widehat{\Xi} Q^r(t;x), \qquad W^{r\circ}(t;x) = \widehat{\Xi} Q^{r\circ}(t;x).$$

In heavy traffic, any greedy policy would attempt to drive $W^r(t;x)$ into the set $\widehat{\mathsf{W}}^+$. This is illustrated in Figure 10. Initially, the trajectory $W^{r\circ}$ mimics the behavior of the fluid model. It is probable that a sample path will then drift throughout the region $\widehat{\mathsf{W}}^+$ if $\rho \sim 1$. For the sample path shown, initially $\bar{c}(W^r(t;x))$ is much greater than $\bar{c}(W^{r\circ}(t;x))$, but then the opposite is true following the upward drift of $W_2^{r\circ}$ shown in the figure.

This counterexample depends upon the specific geometry shown. Although a pointwise optimal solution exists for the fluid model, the effective cost $\bar{c}$ is not monotone. Consequently, property P2 is not desirable—the optimal workload trajectory $\widehat{\boldsymbol{w}}^*$ for the fluid model is not pointwise minimal.

Assuming that the effective cost is monotone, the arguments used in the proofs of Theorem 3.10 and Theorem 4.4 may be applied to establish the following consequences.

THEOREM 4.5. *Suppose that the following assumptions hold:*

(i) *Assumption* M *holds with* $n = \ell_b$*, and the effective cost* $\bar{c}$ *for the* ℓ_b*th workload-relaxation is monotone.*

(ii) *Assumptions* H, S, *and* U *hold.*

Then for all $\Delta_0 > 0$ *sufficiently large, there exists* $b_0 < \infty$ *such that Properties* P1 *and* P2 *hold for the policy defined via the linear program* (4.14).

Proof. The proof of condition P1, and positivity of the state trajectory $\boldsymbol{Q}^{r\circ}$, is identical to the proof in the one-dimensional case since (4.17) continues to hold for the associated error process given in (4.16).

To establish condition P2 we first note that the allocation rate $\zeta_k^{r\circ}$ on $[kT_r, (k+1)T_r]$ is designed to be *almost optimal* for a disturbance-free model on $[kT_r, (k+1)T_r]$,

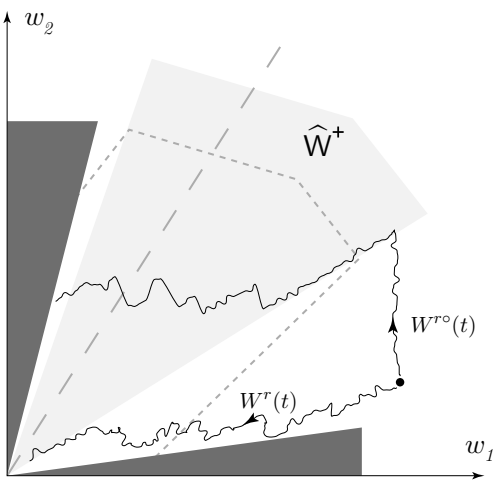


FIG. 10. *The figure shows two trajectories for the stochastic workloads* $W^r(t)$ *and* $W^{r\circ}(t)$*,* $t \geq 0$.

in the sense that

$$W^{r\circ}(kT_r;x) + \widehat{\Xi}[B\zeta_k^{r\circ}T_r + \alpha^r T_r] \le [W^{r\circ}(kT_r;x) - (\mathbf{1}-\boldsymbol{\rho}^r)T_r]_+^r \,. \tag{4.18}$$

Using (4.18) we verify the following restricted form of condition P2 by induction:

$$W^{r\circ}(kT_r;x) \le \widehat{w}^{r\circ^*}(kT_r;x) + 2\bar{w}^r \,, \qquad k \ge 0. \tag{4.19}$$

In fact, this bound combined with Proposition 4.2(ii) implies that P2 holds.

The inequality (4.19) is obvious for $k = 0$. Assuming it is valid for a given $k \ge 1$, denote $D(k+1) := d^{r\circ}((k+1)T_r) - d^{r\circ}(kT_r)$, and consider the following bounds:

$$\begin{aligned}
W^{r\circ}((k+1)T_r;x) &\le \big[W^{r\circ}(kT_r;x) - (\mathbf{1}-\boldsymbol{\rho}^r)T_r\big]_+^r + D(k+1) \\
&\qquad\qquad \text{by (4.18)} \\
&\le \big[\widehat{w}^{r\circ^*}(kT_r;x) + 2\bar{w}^r - (\mathbf{1}-\boldsymbol{\rho}^r)T_r\big]_+^r + D(k+1) \\
&\qquad\qquad \text{by induction} \\
&= \widehat{w}^{r\circ^*}(kT_r;x) + 2\bar{w}^r - (\mathbf{1}-\boldsymbol{\rho}^r)T_r + D(k+1) \\
&\qquad\qquad \text{for } r \text{ sufficiently large} \\
&\le \big[\widehat{w}^{r\circ^*}(kT_r;x) - (\mathbf{1}-\boldsymbol{\rho}^r)T_r + D(k+1)\big]_+ + 2\bar{w}^r \\
&\le \widehat{w}^{r\circ^*}((k+1)T_r;x) + 2\bar{w}^r \\
&\qquad\qquad \text{by Lemma 3.9.}
\end{aligned}$$

The equality in the third line follows since $\bar{w}^r$ is an interior point of $\widehat{\mathsf{W}}$, and $\boldsymbol{\rho}^r \to \mathbf{1}$ as $r \to \infty$. This completes the verification of the induction hypothesis. □

5. Conclusions. The results of this paper lead to practical control solutions for large networks. One must consider an appropriate relaxation for the fluid model, define idealized target states through this idealization, and use safety stocks and some regulation policy to attempt to meet these target values.

Consider for example the network illustrated in Figure 1. For certain parameters a two-dimensional workload-relaxation is justifiable, and a policy that is nearly optimal in heavy traffic can be computed by hand once the effective cost is found. When the cost c is linear and $\mathsf{X} = \mathbb{R}_+^\ell$, the effective cost $\bar{c}(x)$ is the value of the linear program

$$\begin{aligned}
&\min \quad \langle c, x\rangle \\
&\text{subject to} \quad \xi_i^T x = w_i, \quad i = 1,2, \\
&\qquad\qquad\quad x \ge \boldsymbol{\theta}.
\end{aligned} \tag{5.1}$$

It is amazing that optimal policy synthesis can be conceptualized so easily for such a complex model!

In practice, it may be difficult to summarize *every* goal in a single cost function. Optimization may be viewed as a method of generating a family of candidate *good policies*. One can then choose among these or formulate generalizations to satisfy various goals.

A complex network model such as that shown in Figure 1 illustrates the importance of taking a flexible viewpoint in modelling, and in control design. By restricting to a basic feasible solution of (5.1), one may assume that an optimal trajectory $\widehat{q}^*(t;x)$ is null, with the exception of at most two positive components when the cost function c is linear. This is born out in numerical experiments conducted in [14]. After a short

transient period, it is observed that in all but two of the buffers, all of the inventory vanishes in the optimal fluid state trajectory. Similar behavior is commonly seen in the heavy-traffic networks literature (see, e.g., [29, 25, 1]).

In practice, such behavior is rarely feasible because buffers are finite. One can add a state space constraint to both models: for an $\ell_s \times \ell$ *constraint matrix* C_s,

$$C_s Q(t;x) \le \mathbf{1}, \quad Q(t;x) \ge \boldsymbol{\theta}, \qquad t \ge 0. \tag{5.2}$$

The state space is then redefined via $\mathsf{X} = \{x \in \mathbb{R}^\ell : x \ge \boldsymbol{\theta},\ C_s x \le \mathbf{1}\}$, and in this case the set $\widehat{\mathsf{W}}$ given in (3.6) is no longer a simple positive cone. These additional constraints increase the complexity of optimal state trajectories so that work is distributed across a greater number of buffers. Alternatively, if a strictly convex cost function is used, rather than a linear one, then more reasonable optimal trajectories will be obtained.

Another question concerns uncertainty. In telecommunications applications one may know little about the arrival rates to the system, and in a manufacturing application *demand* may be uncertain. One approach is to define a set of *generalized Klimov indices*, as in Proposition 12 of [36]. Alternatively, given prior information regarding arrival rates, one can expand the definition of V. Suppose that A is a polyhedron that defines possible arrival rates. The corresponding worst-case emptying time is given by

$$\overline{T}^*(x) = \max_{\alpha \in \mathsf{A}} \max_{1 \le i \le \ell_r} \frac{\langle \xi^i, x \rangle}{1 - \langle \xi^i, \alpha \rangle}.$$

It is then straightforward to design efficient policies for the fluid model that drain the system before this time without knowledge of the exact value of α. Other approaches have been considered recently in [31, 40, 19].

It has been taken for granted in this paper that activities and buffers far outnumber resources. However, in communication applications, particularly in wireless models, one frequently finds that the set of possible allocation rates is a highly complex convex set (see, e.g., [41, 44]). In particular, it may not be a polyhedron. One interpretation is that in wireless models there are an infinite number of resources through time-division, frequency selection, multiple paths, or choices in coding schemes. Extensions of the methods developed here may be possible provided the rate set V is suitably smooth, and in this case a one-dimensional relaxation is suggested.

There are many questions left unanswered.

(i) Can one formulate a version of Theorem 4.5 when the fluid model admits pointwise optimal solutions, yet the effective cost $\bar{c}$ is not monotone? This question is interesting even in the case of a single bottleneck since sample-path optimality does not hold if ξ^1 has any negative components (see [7]).

(ii) What if a pointwise optimal allocation does not exist for the ℓ_bth workload-relaxation? Can one obtain a near-optimal policy in this case (in the infinite-horizon sense (2.3))?

(iii) The policies described in this paper are based on state-feedback, using a workload-based model. It is expected that RBM models will play a role in the determination of optimality in the mean and in a finer performance analysis.

(iv) Can efficient recursive algorithms, based on workload dimension, be constructed for policy synthesis on a fluid scale?

(v) Where are the sources of highest sensitivity in control design?

(vi) Do the results of this paper lead to improved methods for performance approximation via simulation, or through calculation, by exploiting the simplicity of the network model following state space collapse?

(vii) Finally, extensions of these algorithms have been formulated for sequencing and routing in the face of breakdowns or preventative maintenance. We are eager to see how these methods actually work in practice.

Topics (i)–(v) are considered in what follows [7, 28], but the story is far from complete.

Appendix A. Workload relaxations. The proof of Theorem 4.1 is based on the following lemma, which shows that, relative to the system load, exchangeable states for the ℓ_bth workload-relaxation are almost exchangeable for the original fluid model when r is large.

LEMMA A.1. *Suppose that Assumption* H *holds. There exists $b_0 < \infty$ such that for any $x, y \in \mathsf{X}$, and any $r \geq 1$, the time to reach y from x is uniformly bounded,*

$$T^*(x,y) \leq b_0 \|x - y\| \quad \textit{whenever} \quad \widehat{\Xi}(y - x) \geq \boldsymbol{\theta}.$$

Proof. If $\langle \xi^i, y - x \rangle \geq 0$ for $1 \leq i \leq \ell_b$, then it follows from the definition of T^* that

$$\begin{aligned} T^*(x,y) &= \max_{i \geq 1} \frac{\langle \xi^i, x - y \rangle}{1 - \langle \xi^i, \alpha^r \rangle} \\ &= \max_{i > \ell_b} \frac{\langle \xi^i, x - y \rangle}{1 - \langle \xi^i, \alpha^r \rangle}, \qquad 1 \leq r < \infty. \end{aligned}$$

The right-hand side is bounded in r by construction of α^r and the definition of ℓ_b. It also scales linearly on scaling $(x - y)$. This gives the required bound. ∎

Proof of Theorem 4.1. Fix $0 < t < \underline{T}_{r\circ}(x)$, and define

$$\widehat{v}^{\perp} = -\beta \frac{e^r(t;x)}{\|e^r(t;x)\|}.$$

The constant $\beta > 0$ is chosen so that $\widehat{v}^{\perp}$ is a boundary point of V_r. We have the explicit formula $\beta^{-1} = \frac{T^*(x^1,x^2)}{\|x^2 - x^1\|}$, with $x^1 = q(t;x)$, $x^2 = \widehat{q}^*(t;x)$.

We have already remarked that the constraints ensure that (i) holds so that $w_i(t;x) \leq \widehat{w}_i^*(t;x)$ for all $t \geq 0$, and $1 \leq i \leq \ell_b$. It follows that $\widehat{\Xi}\widehat{v}^{\perp} \geq \boldsymbol{\theta}$, and Lemma A.1 implies directly that $\beta = \beta(r)$ is uniformly bounded from below in r. Applying this and Assumption U(i), we conclude that there is some fixed $\Delta > 0$, independent of $x \in \mathsf{X}$ and $r \geq 1$, such that for all $0 \leq t < \underline{T}_{r\circ}(x)$ and sufficiently small $s > 0$,

$$c(q(t;x) + s\widehat{v}^{\perp}) - c(q(t;x)) \leq -\Delta s. \tag{A.1}$$

Now let

$$v = \widehat{v}^* + \Big(1 - \tfrac{1}{2}\frac{r_0}{r}\Big)\widehat{v}^{\perp}, \tag{A.2}$$

where $\widehat{v}^* = \frac{d}{dt}\widehat{q}^*(t;x)$ and r_0 is a constant. We show that this is in V_r for any $r \geq r_0$ provided r_0 is sufficiently large. For $1 \leq i \leq \ell_b$ we have $\langle \xi^i, \widehat{v}^{\perp} \rangle \geq 0$, and hence for $r \geq r_0$,

$$\langle \xi^i, v \rangle := \langle \xi^i, \widehat{v}^* + (1 - r_0/(2r))\widehat{v}^{\perp} \rangle \geq \langle \xi^i, \widehat{v}^* \rangle \geq -(1 - \rho_i^r).$$

For $i > \ell_b$ we can reason as follows: The identity (4.4) implies that $\|\frac{d}{dt}\widehat{w}^*(t)\| \le K_0/r$ for some $K_0 < \infty$ and all $t > 0$. Since $\mathcal{X}^*$ is continuous, we must have a similar bound for $\widehat{\boldsymbol{q}}^*$, so that $\|\widehat{v}^*\| \le K_1/r$ for some finite K_1. Then, for $i > \ell_b$ and $r \ge r_0$,

$$\begin{aligned}\langle \xi^i, v\rangle &\ge \langle \xi^i, \widehat{v}^*\rangle + \tfrac{1}{2}\frac{r_0}{r}(1-\rho_i^r) - (1-\rho_i^r)\\ &\ge \frac{1}{r}(\tfrac{1}{2}r_0(1-\rho_i^r) - K_1\|\xi^i\|) - (1-\rho_i^r).\end{aligned}$$

Hence, to ensure feasibility of v, it is enough to choose $r_0 > 2K_1\|\xi^i\|(1-\rho_i^\infty)^{-1}$ for all $i > \ell_b$.

Using (A.1), (A.2), and the minimality of $\langle \nabla c, \frac{d}{dt}q\rangle$, we obtain the bound, for any $0 < t < \underline{T}_{r\circ}$, $r \ge r_0$,

$$\frac{d}{dt}(c(q(t;x)) - c(\widehat{q}^*(t;x))) \le -\left(1 - \tfrac{1}{2}\frac{r_0}{r}\right)\Delta \le -\tfrac{1}{2}\Delta.$$

Let $g(t) = c(q(t;x)) - c(\widehat{q}^*(t;x))$, $t \ge 0$. This function is piecewise linear on $(0,\infty)$ and satisfies $g(0+) = c(x) - c(\mathcal{P}^*(x))$, where $\mathcal{P}^*$ is defined in (3.10). The previous bound then gives

$$g(t) \le g(0+) - \tfrac{1}{2}\Delta t, \qquad 0 < t < \underline{T}_{r\circ},$$

and since g is nonnegative, $g(t) = 0$ for $t \ge 2g(0+)/\Delta$. Assumption U(ii) then implies that $q(t;x) = \widehat{q}^*(t;x)$ for such t, and hence

$$\underline{T}_{r\circ}(x) < \frac{2}{\Delta}g(0+) = \frac{2}{\Delta}(c(x) - c(\mathcal{P}^*(x))).$$

This proves (ii) and (iii) since c is a norm, and results (iv) and (v) follow immediately. □

Appendix B. Stochastic models.

Proof of Proposition 4.2. In part (i) we are asking, *When can the graph* (T, S_T) *of the partial sums of* X_i *hit the line* $l(T) = (m-\delta)T - N$ *for some* $T \ge 0$*?* Hence, the bound in (i) follows easily from Cramer's theorem [12].

For (ii) we define, for any $i \ge 1$, the event

$$\mathcal{E}_i := \{Y(S_i + T) - Y(S_i) - m^{-1}T \ge \delta T + N \quad \text{some } T \ge 0\}.$$

Using the fact that $S_i := \sum_{j=1}^i X_j$, $i \ge 1$, is equal to the time of the ith jump of $\boldsymbol{Y}$, we obtain the identity

$$\mathcal{E}_i = \left\{ \sum_{j=i+1}^{(m^{-1}+\delta)T+N} X_j \le T, \quad \text{some } T \ge 0 \right\}.$$

Applying (i) gives a bound of the form $\mathsf{P}(\mathcal{E}_i) \le B_1 \exp(-I_1\delta^2 N)$ for some $B_1 < \infty$, $I_1 > 0$, and all i, N. Consequently, for any $r \ge 1$,

$$\begin{aligned}&\mathsf{P}\left\{\sup_{0\le s\le t\le r^{k_0}}\left(\frac{Y(t)-Y(s)-(t-s)(m^{-1}+\delta)}{N}\right) \ge 1\right\}\\ &\quad\le \mathsf{P}\{S_{r^{k_0+1}} \le r^{k_0}\} + \textstyle\sum_{i=1}^{r^{k_0+1}} \mathsf{P}\{\mathcal{E}_i\}\\ &\quad\le B_1\exp(-I_1\delta_0 r) + r^{k_0+1}B_1\exp(-I_1\delta^2 N).\end{aligned}$$

We now define N via $I_1\delta^2 N = \log(r^{k_0+3})$, so that the right-hand side is bounded by $2B_2 r^{-2}$. This is summable, and hence by the Borel–Cantelli lemma,

$$\limsup_{r\to\infty}\left\{\sup_{0\le s\le t\le r^{k_0}}\left(\frac{Y(t)-Y(s)-(t-s)(m^{-1}+\delta)}{(I_1\delta^2)^{-2}\log(r^{k_0+3})}\right)\right\}\le 1 \quad \text{a.s.} \qquad \square$$

We may now develop properties of the stochastic relaxation used to prove Theorem 4.3. Consider the value function for the relaxed stochastic network for a given $T>0$, $x\in\mathsf{X}$,

$$\widehat{\Gamma}^{r*}(T;x) = \min \frac{1}{T}\int_0^T c(\widehat{Q}(t))\,dt\,,$$

where the minimum is over all allocations $\widehat{\boldsymbol{Z}}$ satisfying the constraints (4.7), subject to the additional constraint that $\widehat{\boldsymbol{Q}}$ evolves in $\mathsf{X}=\mathbb{R}^{\ell}_+$.

Given the value of $\widehat{\boldsymbol{Z}}^*$ at time t, the associated idleness process is given by

$$\widehat{I}^{r*}(t;x) = t\mathbf{1} - \widehat{C}\widehat{Z}^{r*}(t;x), \qquad t\ge 0,\ x\in\mathsf{X}.$$

Conversely, one can determine the optimal allocation given the current value of the idleness. If $\widehat{I}^{r*}(t;x)=\widehat{I}$, then we take $\widehat{Z}^{r*}(t;x)$ to be the minimizer $\widehat{Z}^{r*}$ of the nonlinear optimization problem

$$\begin{aligned} \min c(y) \qquad \text{subject to} \qquad y &= x - S(\widehat{Z}) + R(\widehat{Z}) + A(t),\\ \widehat{C}\widehat{Z} &= \widehat{I},\\ \widehat{Z} &\in \mathbb{R}^{\ell_u},\\ y &\ge \boldsymbol{\theta}. \end{aligned} \tag{B.1}$$

This representation leads to the following conclusion.

PROPOSITION B.1. *Suppose that Assumptions* H *and* S *hold. Then, the optimal solution* $[\widehat{\boldsymbol{Q}}^{r*},\widehat{\boldsymbol{Z}}^{r*}]$ *satisfies condition* P1.

Proof. Fix any $t_0>0$, set $y=\widehat{Q}^{r*}(t_0)$, let $y^*=\mathcal{P}^*(y)$, and set ζ^+ as any solution to $B\zeta^+ = y^*-y$, so that $\widehat{C}\zeta^+=\boldsymbol{\theta}$. Under Assumption U(i) there exists $\kappa>0$ such that $c(y+sB\zeta^+)\le c(y)-s\|\zeta^+\|\kappa$ for $s\le 1$.

Consider the allocation $\widehat{\boldsymbol{Z}}^{*\Delta}$ given by $\widehat{Z}^{*\Delta}(t)=\widehat{Z}^{r*}(t)$ if $t\ne t_0$, and $\widehat{Z}^{r\Delta}(t_0)=\widehat{Z}^{r*}(t_0)+\Delta\zeta^+$. On the remainder of $\mathbb{R}_+$ we again suppose that this allocation is linear on each time-horizon. This is feasible for a range of $\Delta\ge 0$, and by Proposition 4.2 we have

$$c(\widehat{Q}^{r*}(t_0)) \ge c(\widehat{Q}^{r\Delta}(t_0)) - O(s\|\zeta^+\|\kappa) + O(\log(r)).$$

We must therefore have $\|\zeta^+\| = O(\log(r))$, so that P1 holds. $\square$

We may also establish a form of P2.

PROPOSITION B.2. *Suppose that Assumptions* H, M, U, *and* S *hold, where all bounds are with respect to the* ℓ_b*th workload-relaxation. Then for any* $r\ge 1$ *and any* $0<T_\bullet\le T_{r\bullet}$, *there exists a solution* $[\widehat{\boldsymbol{Q}}^{r},\widehat{\boldsymbol{Z}}^{r}]$ *that satisfies conditions* P1 *and* P2, *and*

$$\frac{1}{T_\bullet}\int_0^{T_\bullet} c(\widehat{Q}^{r}(t))\,dt \le \widehat{\Gamma}^{r*}(T_\bullet;x) + O(\log(r))\,. \tag{B.2}$$

Proof. The proof is again by comparison. We approximately retain the convention that $\widehat{\boldsymbol{Z}}^r$ is determined from its idleness process through the following restricted form of (B.1): For a given value $\widehat{I} = \widehat{I}^r(t;x)$ we take $\widehat{Z}^r(t;x)$ to be the minimizer $\widehat{Z}^*$ of

$$\begin{aligned} \min c(y) \qquad \text{subject to} \qquad y &= x - S(\widehat{Z}) + R(\widehat{Z}) + A(t), \\ \widehat{C}\widehat{Z} &= \widehat{I}, \\ \widehat{Z} &\in \mathbb{R}^{\ell_u}, \\ y &\geq \tfrac{1}{2}\bar{x}^r. \end{aligned} \tag{B.3}$$

Under this restriction, one can adapt the proof given in Proposition B.1 to show that $[\widehat{\boldsymbol{Q}}^r, \widehat{\boldsymbol{Z}}^r]$ satisfies Property P1.

We now show how $\widehat{\boldsymbol{I}}^r$ may be constructed so that P2 holds.

For a given $\Delta > 0$, let $\widehat{\boldsymbol{I}}^{r\Delta}$ denote the idleness process defined by $\widehat{I}^{r\Delta}(t;x) = \widehat{I}^{r*}(t;x) + \Delta\bar{w}^r$, $t \geq 0$, where $\bar{w}^r := \widehat{\Xi}\bar{x}^r$. An application of Proposition 4.2 implies that Δ can be chosen so large that the resulting state trajectory satisfies

$$\widehat{W}^{r*}(t;x) + 2\Delta\bar{w}^r + o(1) \geq \widehat{W}^{r\Delta}(t;x) \geq \bar{w}^r - o(1)\,, \quad 0 \leq t \leq T_\bullet\,.$$

This can be chosen independently of $r \geq 1$ and independently of $0 < T_\bullet \leq T_{r\bullet}$. It is obvious that (B.2) holds for the allocations $\{\widehat{\boldsymbol{I}}^{r\Delta}\}$.

We now refine this allocation to form an allocation $\widehat{\boldsymbol{I}}^r$ as follows. We first define this for $t = kT_r$ and then take it to be linear on each interval $[kT_r, (k+1)T_r]$ for each $k \geq 0$.

For $k = 0$ we set $\widehat{I}^r(0;x) := \widehat{I}^{r\Delta}(0;x) = \Delta\bar{w}^r$. For all $k \geq 1$, given that $\widehat{I}^r((k-1)T_r;x)$ has been specified, we define an upperbound $\bar{I} \in \mathbb{R}^{\ell_b}_+$ on the idleness rate on the interval $[(k-1)T_r, kT_r]$. This is given as the solution to

$$\widehat{W}^r((k-1)T_r;x) - (\mathbf{1} - \boldsymbol{\rho}^r)T_r + \bar{I}T_r = \big[\widehat{W}^r((k-1)T_r;x) - (\mathbf{1} - \boldsymbol{\rho}^r)T_r\big]^r_+. \tag{B.4}$$

We then define $\widehat{I}^r(kT_r;x)$ as

$$\widehat{I}^r(kT_r;x) := \min\Big([\widehat{I}^r((k-1)T_r;x) + \bar{I}T_r], \widehat{I}^{r\Delta}(kT_r;x)\Big),$$

where the minimum is interpreted componentwise. One can show that for sufficiently large $r \geq 1$, $\widehat{W}^r(t;x) - \frac{1}{2}\bar{w}^r \in \widehat{\mathsf{W}}$, $t \geq 0$. It follows that $\widehat{\boldsymbol{I}}^r$ defines a feasible allocation $\boldsymbol{Z}^r$, in the sense that the nonlinear program (B.3) is feasible when $\widehat{I} = \widehat{I}^r(t)$. Following familiar arguments we may conclude that the resulting state trajectory $\widehat{\boldsymbol{Q}}^r$ evolves in X for sufficiently large r.

Moreover, (B.4) implies that the resulting workload process satisfies a bound similar to (4.18): We have by construction

$$\widehat{W}^r(kT_r;x) - (\mathbf{1} - \boldsymbol{\rho}^r)T_r + \hat{\imath}^r_k T_r \leq [W^r(kT_r;x) - (\mathbf{1} - \boldsymbol{\rho}^r)T_r]^r_+\,,$$

where $\hat{\imath}^r_k := T_r^{-1}[\widehat{I}^r((k+1)T_r;x) - \widehat{I}^r(kT_r;x)]$ denotes the idleness on the kth interval. The proof of P2 is then identical to the proof of P2 for $\widehat{\boldsymbol{Q}}^{r\circ}$ given in Theorem 4.5.

Finally, since $\widehat{I}^r(kT_r;x) \leq \widehat{I}^{r\Delta}(kT_r;x)$ for all k and $\widehat{\boldsymbol{I}}^{r\Delta}$ satisfies (B.2) by construction, we may establish (B.2) for $\widehat{\boldsymbol{I}}^r$ by an application of Proposition 4.2 as in the proof of Proposition B.1. ∎

Acknowledgments. The author would like to express his thanks to Michael Chen, Michael Harrison, and Shane Henderson for many useful comments on an earlier

draft of this manuscript. The anonymous referee also provided valuable input. Of course, the author takes full responsibility for any remaining errors. Thanks are also due to Maury Bramson, Kavita Ramanan, and Ruth Williams for sharing their unpublished work, and for many fruitful discussions.

REFERENCES

[1] S. L. Bell and R. J. Williams, *Dynamic scheduling of a system with two parallel servers: Asymptotic policy in heavy traffic*, in Proceedings of the 38th IEEE Conference on Decision and Control, Vol. 3, IEEE Press, Piscataway, NJ, 1999, pp. 2255–2260.

[2] A. Bernard and A. el Kharroubi, *Régulations déterministes et stochastiques dans le premier "orthant" de R^n*, Stochastics Stochastics Rep., 34 (1991), pp. 149–167.

[3] D. Bertsekas and R. Gallager, *Data Networks*, Prentice–Hall, Englewood Cliffs, NJ, 1992.

[4] M. Bramson, *State space collapse with application to heavy traffic limits for multiclass queueing networks*, Queueing Systems Theory Appl., 30 (1998), pp. 89–148.

[5] C. Chen, Z. Jia, and P. Varaiya, *Causes and cures of highway congestion*, IEEE Control Systems Magazine, 21 (2001), pp. 26–32.

[6] H. Chen and D. D. Yao, *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*, Appl. Math. 46, Springer–Verlag, New York, 2001.

[7] M. Chen, C. Pandit, and S. P. Meyn, *In search of sensitivity in network optimization*, Queueing Systems Theory Appl., to appear.

[8] J. G. Dai, *On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models*, Ann. Appl. Probab., 5 (1995), pp. 49–77.

[9] J. G. Dai, *A fluid-limit model criterion for instability of multiclass queueing networks*, Ann. Appl. Probab., 6 (1996), pp. 751–757.

[10] J. G. Dai and S. P. Meyn, *Stability and convergence of moments for multiclass queueing networks via fluid limit models*, IEEE Trans. Automat. Control, 40 (1995), pp. 1889–1904.

[11] J. G. Dai and G. Weiss, *A fluid heuristic for minimizing makespan in job-shops*, Oper. Res., 50 (2002), pp. 692–707.

[12] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed., Springer–Verlag, New York, 1998.

[13] B. T. Doshi, *Optimal control of the service rate in an M/G/1 queueing system*, Adv. Appl. Probab., 10 (1978), pp. 682–701.

[14] M. Chen, R. Dubrawski, and S. P. Meyn, *Management of demand-driven production systems*, IEEE Trans. Automat. Control., submitted.

[15] P. Dupuis and K. Ramanan, *Convex duality and the Skorokhod problem*. I, Probab. Theory Related Fields, 115 (1999), pp. 153–195.

[16] P. Dupuis and K. Ramanan, *Convex duality and the Skorokhod problem*. II, Probab. Theory Related Fields, 115 (1999), pp. 197–236.

[17] P. Dupuis and K. Ramanan, *An explicit formula for the solution of certain optimal control problems on domains with corners*, Teor. Ĭmovīr. Mat. Stat., 63 (2000), pp. 32–48.

[18] P. Dupuis and K. Ramanan, *A multiclass feedback queueing network with a regular Skorokhod problem*, Queueing System Theory Appl., 36 (2000), pp. 327–349.

[19] A. Eryilmaz, R. Srikant, and J. R. Perkins, *Stable scheduling policies for broadcast channels*, in Proceedings of the 2002 IEEE Symposium on Information Theory, IEEE Press, Piscataway, NJ, 2002, p. 382.

[20] S. B. Gershwin, *Manufacturing Systems Engineering*, Prentice–Hall, Englewood Cliffs, NJ, 1993.

[21] Ĭ. Ī. Gīhman and A. V. Skorohod, *Stochastic Differential Equations*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 72, Kenneth Wickwire, trans., Springer–Verlag, New York, 1972.

[22] J. M. Harrison, *Brownian models of queueing networks with heterogeneous customer populations*, in Stochastic Differential Systems, Stochastic Control Theory and Applications (Minneapolis, Minn., 1986), Springer–Verlag, New York, 1988, pp. 147–186.

[23] J. M. Harrison, *Brownian models of open processing networks: Canonical representations of workload*, Ann. Appl. Probab., 10 (2000), pp. 75–103.

[24] J. M. Harrison, *Stochastic networks and activity analysis*, in Analytic Methods in Applied Probability: In Memory of Fridrikh Karpelevich, Amer. Math. Soc. Transl. Ser. 2 207, Yu. M. Suhov, ed., AMS, Providence, RI, 2002.

[25] J. M. Harrison and J. A. Van Mieghem, *Dynamic control of Brownian networks: State space*

collapse and equivalent workload formulations, Ann. Appl. Probab., 7 (1997), pp. 747–771.

[26] J. M. Harrison and R. J. Williams, *Brownian models of open queueing networks with homogeneous customer populations*, Stochastics, 22 (1987), pp. 77–115.

[27] J. M. Harrison, *The BIGSTEP approach to flow management in stochastic processing networks*, in Stochastic Networks Theory and Applications, F. P. Kelly, S. Zachary, and I. Ziedins, eds., Clarendon Press, Oxford, UK, 1996, pp. 57–89.

[28] S. G. Henderson, S. P. Meyn, and V. Tadic, *Performance evaluation and policy selection in multiclass networks*, Discrete Event Dyn. Syst., 13 (2003), pp. 149–189.

[29] F. C. Kelly and C. N. Laws, *Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling*, Queueing Systems Theory Appl., 13 (1993), pp. 47–86.

[30] H. J. Kushner, *Heavy Traffic Analysis of Controlled Queueing and Communication Networks*, Springer–Verlag, New York, 2001.

[31] R. Leelahakriengkrai and R. Agrawal, *Scheduling in multimedia wireless networks*, in Proceedings of the 17th International Teletraffic Congress, Brazil, 2001.

[32] X. Luo and D. Bertsimas, *A new algorithm for state-constrained separated continuous linear programs*, SIAM J. Control Optim., 37 (1998), pp. 177–210.

[33] C. Maglaras, *Design of dynamic control policies for stochastic processing networks via fluid models*, in Proceedings of the 36th IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 1997, pp. 1208–1213.

[34] S. P. Meyn, *Transience of multiclass queueing networks via fluid limit models*, Ann. Appl. Probab., 5 (1995), pp. 946–957.

[35] S. P. Meyn, *Stability and optimization of queueing networks and their fluid models*, in Mathematics of Stochastic Manufacturing Systems (Williamsburg, VA, 1996), American Mathematical Society, Providence, RI, 1997, pp. 175–199.

[36] S. P. Meyn, *Sequencing and routing in multiclass queueing networks part* I: *Feedback regulation*, SIAM J. Control Optim., 40 (2001), pp. 741–776.

[37] M. C. Pullan, *Forms of optimal solutions for separated continuous linear programs*, SIAM J. Control Optim., 33 (1995), pp. 1952–1977.

[38] M. I. Reiman, *Open queueing networks in heavy traffic*, Math. Oper. Res., 9 (1984), pp. 441–458.

[39] M. I. Reiman, *A multiclass queue in heavy traffic*, Adv. Appl. Probab., 20 (1988), pp. 179–207.

[40] S. Shakkottai and A. Stolyar, *Scheduling for multiple flows sharing a time-varying channel: The exponential rule*, in Analytic Methods in Applied Probability: In Memory of Fridrikh Karpelevich, Amer. Math. Soc. Transl. Ser. 2 207, Yu. M. Suhov, ed., AMS, Providence, RI, 2002.

[41] D. N. C. Tse and S. V. Hanly, *Multiaccess fading channels*. I. *Polymatroid structure, optimal resource allocation and throughput capacities*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2796–2815.

[42] G. Weiss, *Optimal draining of fluid re-entrant lines*, in Stochastic Networks: Theory and Applications, Roy. Statist. Soc. Lecture Note Ser. 4, F. P. Kelly, S. Zachary, and I. Ziedlins, eds., Oxford University Press, Oxford, 1996, pp. 19–34.

[43] R. J. Williams, *Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse*, Queueing Systems Theory Appl., 30 (1998), pp. 27–88.

[44] L. Xiao, M. Johansson, H. Hindi, S. Boyd, and A. Goldsmith, *Joint optimization of communication rates and linear systems*, in Proceedings of the 40th IEEE Conference on Decision and Control, Vol. 3, IEEE Press, Piscataway, NJ, 2001, pp. 2321–2326.

[45] L. Zuckerman and M. L. Wald, Gridlock in the skies: A special report; *Crisis for air traffic system: More passengers, more delays*, New York Times, 5 September, 2000, sec. A, p. 1, late edition.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 218–238

NEEDLE VARIATIONS THAT CANNOT BE SUMMED*

ROSA-MARIA BIANCHINI† AND MATTHIAS KAWSKI‡

Abstract. This article analyzes sets of higher order tangent vectors to reachable sets of analytic control systems (affine in the control). Both small-time local *output* controllability and small-time local controllability about a nonstationary reference trajectory are considered. In a series of purposefully constructed examples it is shown that the *cones* generated by *needle variations* may fail to be convex. The examples demonstrate that the usual technical condition that *needle variations* must be *movable* is essential to guarantee desirable convexity properties. Moreover, new doubts are cast on the structural stability of controllability properties, as apparently higher order perturbations can reverse the (lack of) controllability of *lower order* nilpotent approximating systems, thereby providing new insights about the ultimate question of whether controllability is finitely determined.

Key words. nonlinear controllability, optimality, control variations, tangent cones

AMS subject classifications. 93B05, 49K15

PII. S0363012902402876

1. Introduction.

A fundamental dichotomy for a control system with a reference trajectory is whether the trajectory is *optimal* or whether the system is *controllable* about this trajectory. Topologically, this translates into the reference trajectory lying either on the boundary or in the interior of the *funnel of reachable sets*. Derivatives of the *endpoint map* with respect to the controls serve as the main analytic tool: If the derivative has full rank, then the trajectory lies in the interior and the system is *controllable*. Conversely, a necessary condition for *optimality* is that the derivative does not have full rank, which leads to the Pontryagin maximum principle [19] (which is a *first derivative* test). See [23] for the current state of the art unifying nonsmooth with differential geometric approaches.

Since commonly the first derivative does not have full rank (at the reference control), there have been many efforts to obtain higher order conditions for controllability and optimality, with Krener's high order maximal principle [17], Stefani's high order conditions for optimality [20], and Sussmann's general theorem for controllability [22] being a few prominent classical ones. See [23] and the references therein for the current state of the art. As opposed to purely theoretical statements in terms of high order Frechet derivatives, the main interest is in *effectively computable* conditions, leading, e.g., to algebraic rank conditions in terms of iterated Lie brackets of the system vector fields evaluated at a point as, e.g., in [20, 22]. Underlying such conditions are notions of higher order tangent vectors together with open mapping theorems. The higher order tangent vectors are basically high order directional derivatives of the endpoint

*Received by the editors February 21, 2002; accepted for publication (in revised form) October 10, 2002; published electronically March 26, 2003. A preliminary announcement of selected results of this paper has been presented as *Lack of convexity for tangent cones of needle variations*, 41st IEEE Conference on Decision and Control, Las Vegas, NV, 2002.

http://www.siam.org/journals/sicon/42-1/40287.html

†Dipartimento di Matematica, "Ulisse Dini," Università degli Studi di Firenze, Viale Morgagni 67a, 50134 Firenze, Italy (bianchin@math.unifi.it).

‡Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287-1804 (kawski@asu.edu). This author gratefully acknowledges the hospitality of the Universita degli Studi in Firenze and the support of the Italian CNR for this work. He also acknowledges the support of the NSF through grant DMS 00-72369.

map, obtained from *families of control variations* (i.e., curves in the space of admissible controls). Open mapping theorems guarantee that if a *cone of tangent vectors* is the whole tangent space, then the reference trajectory lies in the interior. Of course, depending on the precise technical notion of the tangent object, one needs dedicated open mapping theorems—see, e.g., Lee and Markus [18] for such classical statements. A notable effort in the 1980s by Frankowska employed *nonsmooth analysis* to obtain a general purpose open mapping theorem [9, 10], relying on high order variations of set-valued maps.

The basic trade-off is between narrowly defined tangent objects (with straightforward open mapping theorems) and *large* tangent objects, which may lack nice convexity and approximation properties. Recall that classical calculus of variations primarily employed families of variations that are parameterized by their amplitude. In optimal control theory several alternative approaches have been analyzed. These include both *curves* $s \mapsto u_s(\cdot)$, $s \geq 0$ (with $u_0 = u^*$ the *reference control*) such that $\|u_s - u^*\|_\infty \searrow 0$, curves such that $\|u_s - u^*\|_1 \searrow 0$ as $s \searrow 0$ (compare with [8]), and the so-called *needle variations*: Loosely speaking, these are such that the *variation* u_s agrees with the reference control u^* everywhere except on a (finite number of) interval(s) whose (combined) length goes to zero as $s \searrow 0$. The main attraction of such needle variations is that they are conceptually easy to combine for the purpose of generating convex combinations of tangent vectors, resulting in tangent cones with nice convexity properties. Of course, such schemes hinge on the intervals on which the variations differ from the reference control, eventually (for sufficiently small $s > 0$) becoming disjoint. Alternatively, one might require that every needle variation be *movable*, i.e., one must be able to *move* the variation by a small time-amount and still be able to generate the same tangent vector. Precise technical specification of such conditions can be quite involved.

This article is not overly concerned about specific ways of combining needle variations. The main results are independent of such technical intricacies because we show that certain directions can be generated by the most basic needle variations, while certain other points cannot be reached by any control variations at all. Of course, this implies that one cannot reach those points or generate those tangent vectors by any kind of composition of needle variations either, no matter how these are defined.

For analytic, affine control systems with a *stationary* reference trajectory, no such conditions are needed, and a wealth of necessary and sufficient conditions for controllability and optimality have been found in recent decades; see, e.g., [22]. On the other hand, distinctive features of the existing literature for nonstationary reference trajectories are much more careful and provide more delicate notions of variations, tangent objects, and open mapping theorems; see, e.g., [1, 4, 5, 6, 16, 23]. A major open question has been whether such technical conditions are truly essential, or whether they are merely artifacts of a still imperfect knowledge level (e.g., future work might show that they are automatically satisfied for reasonable classes of systems). This article provides a definitive answer: Loosely speaking, even for the most benign classes of nonlinear systems (polynomial cascades, affine in the control, convex compact control values), the tangent *cones* generated by needle variations can be nonconvex. Moreover, even if their convex hull is the entire space the reference trajectory may lie on the boundary of the reachable sets. Both may happen even for arbitrary small positive times.

This article is organized as follows: After this introduction, we give and review technical definitions of controllability, needle variations, tangent cones, and open map-

ping theorems. The subsequent three sections analyze a sequence of increasingly more delicate examples. First we consider the *output controllability* of a simple polynomial cascade system about an equilibrium (stationary reference trajectory). Most of the innovative constructions and critical inequalities may be found here. The next section slightly modifies the system so that it falls into the class of standard small-time local controllability about a nonstationary reference trajectory. Finally, we add some terms which at first sight appear to be higher order perturbations, but we show that these actually destroy the delicate controllability properties. The final section reflects and speculates on the possible implications of such a lack of robustness of the notion of controllability.

The constructions in the proofs fall into two categories: On one side we provide very explicit constructions of *needle variations* and demonstrate that they generate certain tangent vectors. Similarly, in the absence of applicable open mapping theorems, we explicitly construct controls that steer the system to any given target point. On the other side, we show that certain (regions of) points cannot be reached by any control whatsoever, and thus not by any needle variations either. Such arguments naturally involve delicate arguments with integral inequalities.

2. Control variations and approximating cones. The main thrust of this article is *counterexamples*. Thus for positive controllability results we use constructions that rely on small sets of tools, e.g., very narrow notions of (needle) variations. Such results clearly hold a fortiori if the sets of admissible variations etc. are broadened. Conversely, negative results (lack of convexity or controllability) are usually stated in the form that even for very large sets of controls something is impossible.

The systems under consideration are finite-dimensional, deterministic, analytic systems (possibly including an analytic *output* function) that are affine in the control.

$$\dot{x} = f_0(x) + \sum_{i=1}^{\ell} u_i(t) f_i(x), \qquad x(0) = 0, \qquad y = \varphi(x). \tag{2.1}$$

Here $x \in \mathbb{R}^n$, $f_i\colon \mathbb{R}^n \mapsto \mathrm{T}\mathbb{R}^n$ are analytic vector fields; the controls $u\colon [0,T] \mapsto U \subseteq \mathbb{R}^\ell$ are measurable functions defined on finite intervals and take values in a compact convex set U, usually $U = [-1,1]^\ell$. The output map $\varphi\colon \mathbb{R}^n \mapsto \mathbb{R}^p$ is analytic. In this article the vector fields are actually always of a polynomial cascade form, the controls in all constructions are piecewise constant, and the output map is either the identity or a projection onto a subspace. The solution curves of (2.1) corresponding to a control u, starting from $x(0) = 0$, are denoted by $x(t,u)$ or, when no confusion arises, by $x(t)$. Their images under the output map are denoted $y(t,u) = \varphi(x(t,u))$ or simply $y(t)$. Throughout this article we also conveniently identify the tangent spaces $\mathrm{T}_p\mathbf{R}^n$ with $\mathbf{R}^n$.

The reachable sets $\mathcal{R}(t)$ consist of all points $x(t,u)$ reached in time t by trajectories of (2.1) from $x(0) = 0$ by means of admissible controls (measurable, and with values in U almost everywhere). Given a reference trajectory $x^*(t) = x(t,u^*)$, controllability is defined as follows.

Definition 2.1. *The system* (2.1) *is small-time locally controllable* (STLC) *about* x^* *if* $x^*(t) \in \mathrm{int}\mathcal{R}(t)$ *for all* $t > 0$. *The system* (2.1) *is small-time locally output controllable* (STLOC) *about* $y^* = \varphi(x^*)$ *if* $y^*(t) \in \mathrm{int}(\varphi(\mathcal{R}(t)))$ *for all* $t > 0$.

The latter notion is related to Kaskosz's g-controllability [13] but is not yet standard. In general one needs to distinguish STLOC about a reference output y^* with a fixed initial condition $x(0) = 0$ from a notion that requires only $x(0) \in \varphi^{-1}(y_0)$. In

this article this distinction will not play a role.

Throughout this article we will use the norm $\|x\|_\infty \stackrel{\text{def}}{=} \max_{k=1}^n |x_k|$ for $x \in \mathbb{R}^n$. In particular, the open ball of *radius* $r \geq 0$ about $p \in \mathbb{R}^n$ is $B_p^\infty(r) = \{x \in \mathbb{R}^k : \|x - p\|_\infty < r\}$. Since $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$, this does not change the quality of any statements but allows us to dispense with lots of additional factors of $\sqrt{n}$ throughout our constructions and statements of results.

DEFINITION 2.2. *A one-parameter family of control variations of a reference control* $u^* : [0,T] \mapsto U$ *is a curve* $s \mapsto u_s \in \mathcal{U} = \{u : [0,T] \mapsto U \text{ measurable}\}$ *defined for* $s \in [0, s_0]$ *for some* $s_0 > 0$ *and with* $u_0 = u^*$. *Such a family is called a family of control variations* at zero *if* $u_s(t) = u^*(t)$ *for all* $t \in (s,T]$.

In the latter case it is convenient to identify u_s with its restriction to $[0,s]$. Conversely we generally do not distinguish between a control $u_s : [0,s] \mapsto U$ and its extension to $[0,T]$ defined by setting $u_s(t) = u^*(t)$ if $t \in (s,T]$. Using either kind of curves one may define derivatives to the family of reachable sets. Denote by $\Phi : (t,p) \mapsto \Phi_t(p) = x(t,u^*;p)$ the flow corresponding to the reference control u^*. In all systems analyzed in this article, this flow will be analytic, and hence there are no subtleties about transporting tangent vectors via the tangent maps Φ_{t*} from one tangent space to another. To facilitate comparisons, we move all tangent vectors back to the tangent space at the initial point. In the following definitions let $u^* : [0,T] \mapsto U$ be an admissible control generating a (reference) trajectory $x^* : [0,T] \mapsto \mathbf{R}^n$.

DEFINITION 2.3. *A vector* $v \in \mathbb{R}^n$ *is an* m*th order tangent vector to the family of reachable sets* $\{\mathcal{R}(T)\}_{T \geq 0}$ *at (time) zero, written* $v \in \mathcal{K}_0^m$, *if there exists a family of control variations at zero* $u_s : [0,s] \mapsto U$ *of* u^* *such that*

$$x(s,u_s) = x(s,u^*) + s^m \Phi_{s*}(v) + o(s^m). \tag{2.2}$$

DEFINITION 2.4. *A vector* $v \in \mathbb{R}^p$ *is called an* m*th order tangent vector to the output-reachable sets of the system* (2.1) *with stationary reference trajectory* $x^* \equiv 0$, *written* $v \in \mathcal{K}_\varphi^m$, *if there exists a family of control variations* $u_s : [0,s] \mapsto U$ *such that*

$$\varphi(x(s,u_s)) = \varphi(x(0)) + s^m v + o(s^m). \tag{2.3}$$

Most of the statements in this article involve these approximating sets to the *family* of reachable sets at the initial time. However, in a few places (e.g., system (2.9) and Corollary 5.8) it is convenient for precise statements to use the following notion of tangent vectors to a fixed reachable set (but transported back to the initial time for easy comparisons). In order to allow rudimentary comparisons with the above notions of higher order tangent vectors, the following definition uses the L^1-norm of the differences $(u_s - u^*)$.

DEFINITION 2.5. *A vector* $v \in \mathbb{R}^n$ *is an* m*th order tangent vector to the reachable set* $\mathcal{R}(T)$ *at* $x^*(T)$, *written* $v \in \mathcal{K}_T^m$, *if there exists a family of control variations* $u_s : [0,T] \mapsto U$ *of* u^* *such that*

$$x(T,u_s) = x(T,u^*) + \|u_s - u^*\|_1^m \Phi_{T*}(v) + o(\|u_s - u^*\|_1^m). \tag{2.4}$$

Note, that in all three definitions no regularity whatsoever is assumed or required for the maps $s \mapsto u_s$.

If $\{u_s\}_{s>0}$ is a family of control variations at zero (and thus $u_s(t) = u^*(t)$ for $s < t \leq T$), it is clear that $\|u_s - u^*\|_1^m \leq c^m s^m$, where $c = \max_{w \in U} |w|$ is finite (due to compactness of U). Thus, basically from the definition,

$$\mathcal{K}_0^m \subseteq \mathcal{K}_T^m \quad \text{for all } m \geq 0 \text{ and all } T > 0. \tag{2.5}$$

However, the reverse inequality need not hold as, e.g., $\|u_s - u^*\|_1^m \searrow 0$ might go to zero faster than s^m as $s \searrow 0$. Nonetheless, for typical families of control variations at zero, as constructed in what follows, one often can readily conclude that they generate corresponding tangent vectors of the same order m, lying in $\mathcal{K}_0^m$ and $\mathcal{K}_T^m$. Write $\overline{\mathcal{K}}_T$, $\overline{\mathcal{K}}_0^m$, and $\overline{\mathcal{K}}_\varphi^m$ for the cones $\overline{\mathcal{K}}_0^m = \{\lambda v : v \in \mathcal{K}_0^m,\ \lambda \geq 0\}$ etc. generated by $\mathcal{K}_T$, $\mathcal{K}_0^m$, and $\mathcal{K}_\varphi^m$, respectively, and also refer to the elements of the cones $\overline{\mathcal{K}}^m$ as *tangent vectors.* In the case of a stationary reference trajectory $x^* \equiv x(0)$ together with full-state output $\varphi \equiv \mathrm{id}_{\mathbf{R}^n}$, the definitions of $\mathcal{K}_0^m$ and $\mathcal{K}_\varphi^m$ agree, and they are subsets of adjacent tangent cones and, moreover, have nice convexity and approximation properties [9].

The sets $\overline{\mathcal{K}}_\varphi^m$ of tangent vectors to output-reachable sets in general may have more complicated structures. In particular, they generally are not simply the images of the corresponding sets $\overline{\mathcal{K}}_0^m$ under the tangent map φ_* as illustrated in the basic example

$$\begin{cases} \dot{x}_1 = u, & x(0) = 0, \\ \dot{x}_2 = x_1^2, & |u(\cdot)| \leq 1, \\ \dot{x}_3 = x_1^3, & \varphi(x) = (x_1, x_3) \end{cases} \tag{2.6}$$

with reference control $u^* \equiv 0$. Here $\overline{\mathcal{K}}_0^m = \mathbf{R} \times \{(0,0)\}$ for $m = 1, 2$, and $\overline{\mathcal{K}}_0^m = \mathbf{R} \times [0, \infty) \times \{0\}$ for $m \geq 3$, while $\overline{\mathcal{K}}_\varphi^m = \mathbf{R}^2$ for $m \geq 4$. In particular, $\overline{\mathcal{K}}_\varphi^4 \neq \varphi_* \overline{\mathcal{K}}_0^4$.

A key step to rendering the sets of tangent vectors useful for obtaining optimality results is to establish their convexity. For a wide variety of different settings, results that are similar to the following may be found in, e.g., [4, 5, 6, 8, 9, 10, 14, 18].

PROPOSITION 2.1. *For systems of form* (2.1) *with a stationary reference trajectory the following hold:*

If $\lambda \in [0,1]$ *and* $v \in \mathcal{K}_0^m$, *then* $\lambda v \in \mathcal{K}_0^m$.

If $m \leq m'$, *then* $\mathcal{K}_0^m \subseteq \mathcal{K}_0^{m'}$.

If v_1, $v_2 \in \mathcal{K}_0^m$ *and* $\lambda \in [0,1]$, *then* $\lambda^m v_1 + (1-\lambda)^m v_2 \in \mathcal{K}_0^m$.

THEOREM 2.2 (see [9, 10, 11, 14]). *Suppose* $x^* \equiv 0$ *is the stationary reference trajectory of a system of form* (2.1). *Then for every closed convex cone* K *that is strictly contained in* $\{0\} \cup \mathrm{int}\overline{\mathcal{K}}_0^m$, *there exist* $C > 0$ *and* $T > 0$ *such that*

$$K \cap B_0^\infty(Ct^m) \subseteq \mathcal{R}(t) \text{ for all } 0 \leq t \leq T. \tag{2.7}$$

Moreover, if $\overline{\mathcal{K}}_0^m = \mathbb{R}^n$, *then the system* (2.1) *is STLC about* $x^* = 0$, *and the minimum time function* $V(q) = \min\{t \geq 0 : \exists u \quad \text{such that} \quad x(t,u) = q\}$ *is Hölder continuous of order* $\frac{1}{m}$ *at* $x = 0$.

If the reference trajectory is nonstationary, then the cones $\overline{\mathcal{K}}_0^m$ need not have similarly nice convexity and approximation properties. Thus much work in recent decades has focused on developing more refined notions of tangent objects for that setting. The classical open mapping theorems (compare, e.g., [18, 19]) use topological arguments that rely on continuity of the multiparameter families of control variations that naturally generalize the (single-parameter) curves introduced above. Typically, one might have constructed n curves $s \mapsto u_s^{(i)} \in \mathcal{U}$, which generate the tangent vectors $v^{(i)} \in \mathcal{K}_0^m$, $i = 1, \ldots, n$. The critical next step is to somehow combine these variations in order to show that the convex combinations $(c_1 v_1 + \cdots + c_n v_n)$ with $c_1 + \cdots + c_n = 1$ are mth order tangent vectors, i.e., are contained in $\overline{\mathcal{K}}_0^m$, too. A natural first try is to consider convex combinations of suitable reparameterizations of the original curves

$$s \mapsto u_{s,c_1,\ldots,c_n} = c_1 u^{(1)}_{\alpha_1(c,s)} + c_2 u^{(2)}_{\alpha_2(c,s)} + \cdots + c_n u^{(n)}_{\alpha_n(c,s)} \tag{2.8}$$

and analyze the corresponding curves $s \mapsto x(s, u_{s,c})$ of endpoints for each value c. Variations that lend themselves especially well to generating such convex combinations are needle variations. (Alternatively, compare with, e.g., [8] for variations such that $\|u_s - u^*\|_1 \searrow 0$.) Basically, each such control u_s agrees with u^* on all of $[0,T]$ except on an interval whose length is of order s for $s \searrow 0$. Two such families of variations are easily combined unless the intervals on which they disagree from u^* overlap for all small $s > 0$. In that case, one might be tempted to simply *move* the interval in one of the families. The continuous dependence of solution curves on the data suggests that after such small (or vanishing, as $s \searrow 0$) moves, the combined variations might still generate the desired tangent vectors.

Before giving a technical (broad) definition of needle variations, we discuss a very simple example that illustrates the problem that some (narrowly defined) *needle variations* might *not be able to be moved*:

$$\begin{cases} \dot{x}_0 = 1, & |u(\cdot)| \le 1, \\ \dot{x}_1 = u, & x(0) = 0, \\ \dot{x}_2 = (x_0 - 1)x_1^2, & x^*(t) = (t,0,0,0), \\ \dot{x}_3 = x_1^7, & t \in [0,2] = [0,T]. \end{cases} \tag{2.9}$$

In this case the family of *needle variations* $u_s^{\pm}\colon [0,2] \mapsto [-1,1]$ defined by $u_s^{\pm}(t) = \pm 1$ if $1 - \frac{1}{2}s \le t < 1$, $u_s^{\pm}(t) = \mp 1$ if $1 \le t < 1 + \frac{1}{2}s$, and $u_s^{\pm}(t) = 0$ else, generates the curves $x(2, u_s^{\pm}) = (2,0,0,\pm 2^{-10}s^8) \in \mathcal{R}_0(2)$, and thus $v = (0,0,0,\pm 1) \in \overline{\mathcal{K}}_2^8$ as 8th order tangent vectors at $T = 2$ (note that $\|u_s - u^*\|_1 = s$).

In general, suppose $u_s\colon [0,2] \mapsto [-1,1]$ is any family of control variations that agrees with $u_s(t) = u^*(t) = 0$ for all $t \in [0,T] \setminus I_s$, where $I_s \subseteq [0,T]$ is an interval of any of the forms $[a, a+s]$, $[a-s, a]$, or $[a - \frac{s}{2}, a + \frac{s}{2}]$ for some fixed $a \in [0,2]$. Then $x(T, u_s) = (2,0,0,Cs^8) + o(s^8)$ is only possible if I_s is of the third form, and in addition $a = 1$. This means that the only *single* family of *needle variations* that can generate $v = (0,0,0,\pm 1)$ as tangent vector *"cannot be moved."*

In this simple example it is clear that either v can easily be generated by control variations whose support (technically the support of $(u_s - u^*)$) consists of two disjoint intervals, e.g.,

$$u_s(t) = \begin{cases} 1 & \text{if } \ \pm\frac{1}{2} - \frac{1}{4}s \le t < \pm\frac{1}{2}, \\ -1 & \text{if } \ \pm\frac{1}{2} \le t < \pm\frac{1}{2} + \frac{1}{4}s, \\ 0 & \text{else.} \end{cases} \tag{2.10}$$

This demonstrates that if one considers only control variations that are supported in a single interval shrinking to a point, then it may be possible that certain tangent vectors can be generated *only* at a specific point t_0 that cannot be moved.

The example appears contrived and the notion of needle variation unnecessarily constrained. It is customary to broaden the notion of needle variations to allow the support of $(u_s - u^*)$ to be a finite union of intervals of total length of order s.

DEFINITION 2.6. *A family of control variations $s \mapsto u_s\colon [0,T] \mapsto U$ defined for $s \in [0, s_0]$ with $s_0 > 0$ is a (family of)* needle variations *of the reference control $u^* = u_0$ if there exist a constant $C > 0$ and a finite number of pairs of increasing functions $s \mapsto a_s^{(k)}$ and decreasing functions $s \mapsto b_s^{(k)}$ defining intervals $[a_s^{(k)}, b_s^{(k)}] \subseteq [0,T]$, $k = 1, \dots, N$, such that*

$$\operatorname{supp}(u_s - u^*) \subseteq \bigcup_{k=1}^{N} [a_s^{(k)}, b_s^{(k)}] \ \text{ and } \ \sum_{k=1}^{N} (b_s^{(k)} - a_s^{(k)}) \le Cs \ \text{ for all } s \le s_0. \tag{2.11}$$

Note that a family of variations *at zero* as defined earlier automatically qualifies as a family of needle variations as one may take $N = 1$ and intervals defined by $a_s^{(1)} = 0$ and $b_s^{(1)} = s$. Definition 2.6 also includes other common constructions with $N = 1$ and $[a_s^{(1)}, b_s^{(1)}] = [\alpha, \alpha + s]$ (*right* variations at $t = \alpha$) or $[a_s^{(1)}, b_s^{(1)}] = [\beta - s, \beta]$ (*left* variations at $t = \beta$). The constructions in this article use mainly (right) needle variations at zero, combined with some (left) needle variations at the final time T. But the main claim of *"needle variations that cannot be summed"* holds even if the combinations (or sums) are allowed to lie in a more general class as in Definition 2.6.

3. The main construction. This section analyzes a custom-designed polynomial cascade system (affine in the control) with output, which forms the heart of the promised counterexample in section 5. We demonstrate that the growth rates of its reachable sets are very sensitive to reflections about the origin. More specifically, the approximating cones of tangent vectors are described not only by intersections, but also by unions of half spaces. The basic system with *output* φ is

$$\begin{cases} \dot{x}_1 = u_1, & |u_1(\cdot)| \le 1, \\ \dot{x}_2 = u_2, & |u_2(\cdot)| \le 1, \\ \dot{x}_3 = x_1^2, & x(0) = 0, \\ \dot{x}_4 = x_2^2, & \varphi(x) = (x_1, x_2, x_5, x_6), \\ \dot{x}_5 = x_4 x_1^2 - x_1^7, & \\ \dot{x}_6 = x_3 x_2^2 - x_2^7. & \end{cases} \tag{3.1}$$

Figure 3.1 pictorially summarizes key properties of this system, showing cross-sections of approximating cones of the image $\varphi(\mathcal{R}(t))$ of the reachable set under the output map. The remainder of this section proves technical statements that make this illustration precise.

THEOREM 3.1. *For all $T > 0$ (sufficiently small)*

$$\varphi(R_0(T)) \supseteq B_0^\infty(2^{-18}T^8) \cap \{y \in \mathbb{R}^4 \colon (y_3 \ge 0 \textbf{ or } \ y_4 \ge 0\,) \textbf{ and } y_1 = y_2 = 0\}. \tag{3.2}$$

The proof shows in particular that points of the forms $x = (0, 0, *, *, -2^{-18}s^8, 0)$ and $x = (0, 0, *, *, 0, -2^{-18}s^8)$ can be reached in time s, i.e., by control variations at zero. Thus the tangent vectors $(0, 0, -1, 0), (0, 0, 0, -1) \in \overline{\mathcal{K}}_\varphi^8$ are generated by *needle variations* of the zero reference control.

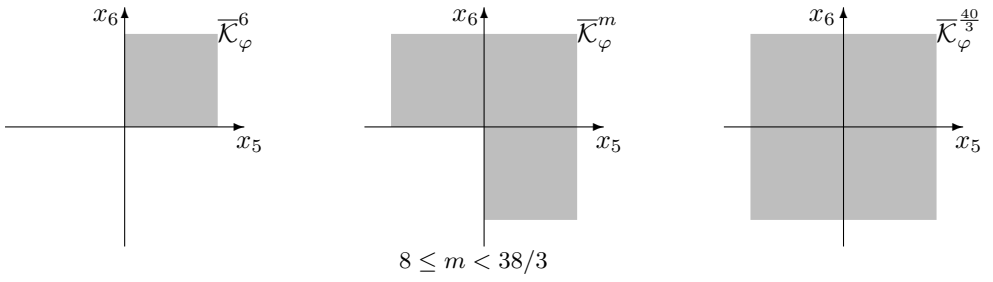


FIG. 3.1. *Cross-sections of approximating cones of tangent vectors for system* (3.1).

Proof. Suppose $T > 0$ and $\eta \in \mathbb{R}^4$ are given, with $\|\eta\|_\infty < 2^{-18}T^8$, $\eta_3 < 0$, and $\eta_4 \geq 0$. (The case of $\eta_4 < 0$ and $\eta_3 \geq 0$ is analogous.) Set $t_1 = (-4\eta_3)^{1/8} > 0$. Define the control $u\colon [0,T] \mapsto [-1,1]^2$ (with $t_2 \geq 0$ to be determined later) by

$$u(t) = \begin{cases} (\ 1,\ \ 0) & \text{if} \quad 0 \leq t \leq t_1, \\ (-1,\ \ 0) & \text{if} \quad t_1 \leq t \leq 2t_1, \\ (\ 0,\ -1) & \text{if} \quad 2t_1 \leq t \leq 2t_1 + t_2, \\ (\ 0,\ \ 1) & \text{if} \quad 2t_1 + t_2 \leq t \leq 2t_1 + 2t_2, \\ (\ 0,\ \ 0) & \text{else.} \end{cases} \tag{3.3}$$

From $u_2(t) = 0$ for $t \leq 2t_1$ it is easily seen that $x(2t_1, u) = (0, 0, \frac{2}{3}(-4\eta_3)^{3/8}, 0, \eta_3, 0)$ and thus $\varphi(x(2t_1, u)) = (0, 0, \eta_3, 0)$. Then $\varphi(x(2t_1 + 2t_2, u)) = (0, 0, \eta_3, \xi(t_2))$, where

$$\xi\colon t_2 \mapsto x_6(2t_1 + 2t_2) = 2\int_0^{t_2} \Big(\tfrac{2}{3}(-4\eta_3)^{3/8}t^2 + t^7\Big)\, dt = \tfrac{4}{9}(-4\eta_3)^{3/8}t_2^3 + \tfrac{1}{4}t_2^8 \tag{3.4}$$

is a strictly increasing function. Thus there exists a unique $0 \leq t_2^* \leq (4\eta_4)^{1/8}$ such that $\xi(t_2^*) = \eta_4$. One easily verifies that the total time satisfies

$$(2t_1 + 2t_2^*) \leq 2(-4\eta_3)^{1/8} + 2(4\eta_4)^{1/8} \leq 4 \cdot \|4\eta\|_\infty^{1/8} = \big(2^{18}\|4\eta\|_\infty\big)^{1/8} < T. \tag{3.5}$$

Points $\eta = (0, 0, \eta_3, \eta_4)$ with $\eta_3 > 0$ and $\eta_4 > 0$ are even easier to reach using again piecewise constant controls, but now both u_1 and u_2 start with value $+1$, followed by -1 and zero. Note that in the directions of this positive quadrant the reachable set grown even faster, of order T^6. ∎

Several stronger statements appear possible, e.g., characterizing the image $\varphi(\mathcal{R}_0(T))$ of the reachable set without the condition $y_1 = y_2 = 0$. However, in what follows we have no need for this, and it is quite cumbersome to find the precise complete boundary of the reachable sets.

One consequence of the next statement is that there does not exist a constant $C > 0$ such that one can reach points $x = (0, 0, *, *, -CT^8, -CT^8)$ in the fourth quadrant for all sufficiently small times $T > 0$ using *any* controls. Thus, a fortiori, one cannot reach such points by any combination of needle variations in time $T > 0$, and $(0, 0, -1, -1) \notin \overline{\mathcal{K}}_\varphi^8$.

THEOREM 3.2. *Suppose* $0 \leq T < 1$, *and* $x_1(T, u) = x_2(T, u) = 0$. *Define* $C = 3^{1/5} \cdot 2^{-35/3} > 0$. *If* $x_5(T, u) < 0$, *then* $x_6(T, u) > -CT^{38/3}$, *and if* $x_6(T, u) < 0$, *then* $x_5(T, u) > -CT^{38/3}$.

The following useful restatements and consequences are immediate corollaries from the theorem (or from its proof given in what follows).

COROLLARY 3.3. *If* $C, m > 0$ *are such that the images of reachable set of system* (3.1) *contain the open balls* $B_0^\infty(CT^m) \subseteq \varphi(\mathcal{R}_0(T))$ *for all* $T > 0$ *(sufficiently small), then* $m \geq \frac{38}{3}$.

COROLLARY 3.4. *For system* (3.1), $(0, 0, a, b) \in \overline{\mathcal{K}}_\varphi^8$ *if and only if* $a \geq 0$ *or* $b \geq 0$.

Proof of Corollary 3.4. The *"if"* part is clear from Theorem (3.1). For the *"only if"* part, suppose $(0, 0, a, b) \in \mathcal{K}_\varphi^8$ for some $a, b \in \mathbb{R}$. By definition of $\mathcal{K}_\varphi$ there exist $0 < s_0 < 1$ and a family of control variations $u_s\colon [0, s] \mapsto [-1, 1]^2$, $0 \leq s \leq s_0$, such that for some $\xi\colon [0, s_0] \mapsto \mathbb{R}^6$ with $\|\xi(s)\|_\infty = o(s^8)$ the endpoints have the form

$$x(s, u_s) = s^8 \cdot (0, 0, *, *, a, b) + \xi(s) \in \mathcal{R}(s). \tag{3.6}$$

Define a new family of control variations $v_{2s}\colon [0,2s] \mapsto [-1,1]^2$, $0 \le 2s \le s_0$, by

$$v_{2s}(t) = \begin{cases} u_s(t) & \text{if } 0 \le t \le s, \\ -\frac{1}{s} \cdot (\xi_1(s), \xi_2(s)) & \text{if } s < t \le 2s. \end{cases} \tag{3.7}$$

Then $x(t, v_{2s}) = x(t, u_s)$ for $0 \le t \le s$, and in particular $x(s, v_{2s}) = x(s, u_s)$. For $t \in [s, 2s]$ in the second half of the domain of v_{2s} the first two components $x_i(t, v_{2s}) = (2s-t)\xi_i(s)$, $i = 1, 2$, remain of size $o(s^8)$. Consequently also $x_3(t, v_{2s})$ and $x_4(t, v_{2s})$ remain of size $o(s^8)$, and the change in the last two components

$$\|x_i(2s, v_{2s}) - x_i(s, v_{2s})\|_\infty = o(s^{24}) \quad \text{for } i = 5, 6 \tag{3.8}$$

is again of higher order. The endpoints are of the form

$$x(2s, v_{2s}) = (0, 0, *, *, s^8 a + o(s^8), s^8 b + o(s^8)), \tag{3.9}$$

and in particular they lie in the plane $x_1 = x_2 = 0$, satisfying the hypotheses of Theorem 3.2. Thus $a \ge 0$ or $b \ge 0$. ∎

COROLLARY 3.5. *For $8 < m < 38/3$ the approximating cones $\overline{\mathcal{K}}^m_\varphi$ of the output-reachable set are not convex.*

To streamline the proof of Theorem 3.2 we first establish two technical lemmata. Consider the following simple system about its equilibrium point $x = 0$:

$$\begin{cases} \dot{x}_1 = u, & |u(\cdot)| \le 1, \\ \dot{x}_2 = c x_1^2 - x_1^7, & x(0) = 0. \end{cases} \tag{3.10}$$

LEMMA 3.6. *If $c \neq 0$, then the system (3.10) is not STLC about $x = 0$. If $T > 2\left(\frac{8}{3}|c|\right)^{\frac{1}{5}}$, then $0 \in \operatorname{int} \mathcal{R}(T)$.*

Proof. W.l.o.g. assume $c > 0$ (else consider the coordinate change $(x_1, x_2) \mapsto (x_1, -x_2)$). The first assertion follows from $x_2(T) = \int_0^T x_1^2(t)(c - x_1^5(t))\, dt \ge 0$ since $|x_1^5(t)| \le t^5 \le T^5 \le c$ if $T \le c^{\frac{1}{5}}$.

Conversely, suppose $T > 2b_0$, where $b_0 = (\frac{8}{3}c)^{\frac{1}{5}}$ is the first positive zero of $t \mapsto \frac{1}{3}ct^3 - \frac{1}{8}t^8$.

Consider the 2-parameter family of controls

$$u_{a,b}(t) = \begin{cases} 0 & \text{if} \quad 0 \le t \le T - 2b_0 - 2b - a, \\ 1 & \text{if} \quad T - 2b_0 - 2b - a < t \le T - b_0 - b - a, \\ -1 & \text{if} \quad T - b_0 - b - a < t \le T. \end{cases} \tag{3.11}$$

Since $T - 2b_0 > 0$, these controls are defined for (a, b) in some neighborhood of $(0, 0)$. (To be precise, they are defined for $-b_0 \le b$, $-b_0 \le a + b$, and $a + 2b \le T - 2b_0$.) It is straightforward to verify that $x(T, u_{0,0}) = 0$ and that the Jacobian (for some polynomial p in two variables)

$$\frac{\partial\, x(T, u_{a,b})}{\partial(a, b)} = \begin{pmatrix} -1 & 0 \\ ca^2 + a^7 & -\frac{20}{9}\left(54c^7\right)^{\frac{1}{5}} + b \cdot p(b, c^{\frac{1}{5}}) \end{pmatrix} \tag{3.12}$$

of the map $(a, b) \mapsto x(T, u_{a,b})$ has full rank at $(a, b) = (0, 0)$ when $(c \neq 0)$. By the inverse mapping theorem, $0 \in \operatorname{int} \mathcal{R}(T)$, and the system (3.6) is STLC. ∎

LEMMA 3.7. *Suppose $f\colon [0, T] \mapsto \mathbb{R}$ is absolutely continuous with $|f'(t)| \le \varepsilon$ a.e. and $f(0) = f(T) = 0$. Then for any even integer $m = 2k \ge 2$*

$$\int_0^T f^m(t)\, dt \ge \tfrac{2}{(m+1)\varepsilon} \|f\|_\infty^{m+1}. \tag{3.13}$$

Proof. Since f is continuous there exists $t_0 \in [0,T]$ such that $f(t_0) = \|f(t)\|_\infty$. W.l.o.g. assume $f(t_0) > 0$. Since $|f'(t)| \le \varepsilon$ a.e., use the lower bound $f(t_0 \pm \tau) \ge f(t_0) - \varepsilon\tau$ for $0 \le \tau \le \tau_0 = f(t_0)/\varepsilon$ to obtain the estimate

(3.14)
$$\int_0^T f^{2k}(t)\,dt \ge \int_{t_0-\tau_0}^{t_0+\tau_0} f^{2k}(t)dt \ge 2\cdot\int_0^{\tau_0} (\varepsilon t)^{2k}\,dt = \frac{2\varepsilon^{2k}\tau_0^{2k+1}}{2k+1} = \frac{2\,\|f\|_\infty^{2k+1}}{(2k+1)\varepsilon}. \qquad \square$$

Proof of Theorem 3.2. Before going into technical details we lay out some heuristic ideas that motivate the strategy.

The presence of the definite terms in system (3.1) suggests using controls u_1 and u_2 with disjoint support. If one first keeps $u_2 \equiv 0$, thus also $x_4 \equiv 0$, and using a simple bang-bang control u_1 one easily reaches a point of the form $(0,0,q_3,0,-q_5,0)$ with $q_3, q_5 > 0$. Next one holds $u_1 \equiv 0$ and, using a suitable control u_2, tries to reach a point $(0,0,q_3,q_4,-q_5,-q_6)$ with all $q_i > 0$. But according to Lemma 3.6 this second step requires that the *available time* is sufficiently large (as determined by q_3). The subsequent estimates quantify this delicate trade-off, with the conclusion that, roughly speaking, if $x_6(T) < 0$, then $|x_5(T)| < CT^{38/3}$ for all T sufficiently small. The proof shows that this is true for all controls, not only for whatever clever combinations of the needle variations are used in the proof of Theorem 3.1. One of the key obstacles to making this heuristics into a general argument is that in such setting of *all possible* controls, there does not exist a clear notion of which of the two directions is obtained first. Thus the first objective is to identify an abstraction of *"which of the directions x_5 and x_6 is generated first."* Since

$$\int_0^T\!\!\left(\int_0^t x_1^2(s)\,ds\right)x_2^2(t)\,dt + \int_0^T\!\! x_1^2(t)\left(\int_0^t x_2^2(s)\,ds\right)dt = \left(\int_0^T\!\! x_1^2(t)\,dt\right)\cdot\left(\int_0^T\!\! x_2^2(t)\,dt\right)$$

and the system is invariant under the index permutation $(x_1,x_2,x_3,x_4,x_5,x_6) \mapsto (x_2,x_1,x_4,x_3,x_6,x_5)$, we may assume w.l.o.g. that

$$\int_0^T\left(\int_0^t x_1^2(s)\,ds\right)x_2^2(t)\,dt \ge \tfrac{1}{2}\left(\int_0^T x_1^2(t)\,dt\right)\left(\int_0^T x_2^2(t)\,dt\right), \tag{3.15}$$

which may be rewritten in the most useful form

$$\int_0^T\left(\int_0^t x_1^2(s)\,ds\right)x_2^2(t)\,dt \ge \tfrac{1}{2}\int_0^T\left(\int_0^T x_1^2(s)\,ds\right)x_2^2(t)\,dt. \tag{3.16}$$

The key advantage of this form is that the weight against which x_2^2 is integrated is now a constant, so that Lemma 3.6 applies. If $x_6(T) \ge 0$, nothing needs to be shown. Thus assume $x_6(T) < 0$. Then

$$0 > x_6(T) = \int_0^T \big(x_3^2(t)x_2^2(t) - x_2^7(t)\big)dt \ge \int_0^T x_2^2(t)\,\left(\tfrac{1}{2}\int_0^T x_1^2(s)\,ds - x_2^5(t)\right)dt \tag{3.17}$$

and thus

$$\xi_2 \stackrel{\text{def}}{=} \max_{0\le t\le T} x_2(t) > \left(\tfrac{1}{2}\int_0^T x_1^2(s)\,ds\right)^{\frac{1}{5}}. \tag{3.18}$$

The term on the right-hand side may be considered a *cost*, or *energy*, that is required to move $x_6(0) = 0$ to $x_6(T) < 0$. The next step is to find a lower bound for this energy in terms of the displacement. Using Lemma 3.7 with $\varepsilon = 1$ and $f = x_1$ gives the crude estimate

$$\int_0^T x_1^2(s)\,ds \geq 2 \cdot \left(\tfrac{1}{3}\right) \cdot \xi_1^3 \stackrel{\text{def}}{=} \tfrac{2}{3} \max_{0\leq t\leq T} x_1(t). \tag{3.19}$$

On the other hand, using Minkowski's inequality (backwards), we get

$$-x_5(T) = \int_0^T (x_1^7(t) - x_4(t)x_1^2(t))dt \leq \int_0^T |x_1^7(t)|dt \leq T \cdot \xi_1^7. \tag{3.20}$$

Combining (3.18), (3.19), and (3.20) yields

$$\xi_2 > \left(\tfrac{1}{2}\int_0^T x_1^2(s)\,ds\right)^{\frac{1}{5}} \geq \left(\tfrac{1}{2}\cdot\tfrac{2}{3}\,\xi_1^3\right)^{\frac{1}{5}} \geq \left(\tfrac{1}{3}\left(\frac{-x_5(T)}{T}\right)^{\frac{3}{7}}\right)^{\frac{1}{5}}. \tag{3.21}$$

Using that $\xi_2 \leq \frac{1}{2}T$ (from $u_2(\cdot)| \leq 1$ together with $x_2(0) = x_2(T) = 0$) implies

$$-x_5(T) \leq 3^{\frac{1}{5}} \cdot \xi_2^{\frac{35}{3}} \cdot T \leq 3^{\frac{1}{5}} \cdot 2^{-\frac{35}{3}} \cdot T^{\frac{38}{3}}. \tag{3.22}$$

Note that, in particular, if $x(T) = (0, 0, *, *, -\rho, -\rho) \in \mathcal{R}_0(T)$ with $\rho > 0$, then $\rho < CT^{38/3}$ (with $C = 3^{1/5} \cdot 2^{-35/3}$). □

THEOREM 3.8. *There exists a constant $C > 0$ such that for all sufficiently small $T > 0$ the image $\varphi(\mathcal{R}(T))$ of the reachable set of system* (3.1) *contains the open ball $B_0^\infty(CT^{40/3})$.*

COROLLARY 3.9. *The system* (3.1) *is STLOC about $y = 0$.*

COROLLARY 3.10. *The approximating cone $\overline{\mathcal{K}}_\varphi^{\frac{40}{3}}$ is the whole tangent space $T_0\mathbb{R}^4$.*

Main elements of the construction. These statements all follow directly from Theorem 4.3 and we refer to its proof for the full technical details. At this point we illustrate the critical elements in the construction showing a slightly weaker result. The general proof given in section 4 requires no major innovation but mainly involves keeping track of many more quantities.

The objective in this construction is to suitably combine the families of control variations

$$u_s^{(1)}(t) = \begin{cases}(+1,0) & \text{if } 0 \leq t < \frac{s}{2},\\ (-1,0) & \text{if } \frac{s}{2} \leq t < s\end{cases} \text{ and } u_s^{(2)}(t) = \begin{cases}(0,+1) & \text{if } 0 \leq t < \frac{s}{2},\\ (0,-1) & \text{if } \frac{s}{2} \leq t < s,\end{cases} \tag{3.23}$$

which steer to points $y(s, u_s^{(1)}) = (0, 0, -2^{-7}s^8, 0)$ and $y(s, u_s^{(2)}) = (0, 0, 0, -2^{-7}s^8)$ on the negative y_3 and y_4 axes to obtain families of control variations that steer to points in the third quadrant $y_3 < 0$ and $y_4 < 0$ in the plane $y_1 = y_2 = 0$. Thus consider the control

$$u(t) = \begin{cases}(+1,0) & \text{if} \quad 0 \leq t < t_1 = c\varepsilon^r,\\ (-1,0) & \text{if} \quad t_1 \leq t < t_2 = 2c\varepsilon^r,\\ (0,+1) & \text{if} \quad t_2 \leq t < t_3 = 2c\varepsilon^r + \varepsilon + \mu\varepsilon^m,\\ (0,-1) & \text{if} \quad t_3 \leq t < T = 2c\varepsilon^r + 2\varepsilon + 2\mu\varepsilon^m,\end{cases} \tag{3.24}$$

with $r, m \geq 1$ and constants c, μ to be determined (first think of $r = 1$ and $\mu = 0$).

Verify that at selected switching times the trajectory passes through

$$
\begin{aligned}
x(t_1) &= (c\varepsilon^r, 0, \tfrac{1}{3}c^3\varepsilon^{3r}, 0, -\tfrac{1}{8}c^8\varepsilon^{8r}, 0),\\
x(t_2) &= (\ 0\ , 0, \tfrac{2}{3}c^3\varepsilon^{3r}, 0, -\tfrac{1}{4}c^8\varepsilon^{8r}, 0),\\
x(t_3) &= (0, \varepsilon+\mu\varepsilon^m, \tfrac{2}{3}c^3\varepsilon^{3r}, \tfrac{1}{3}(\varepsilon+\mu\varepsilon^m)^3, -\tfrac{1}{4}c^8\varepsilon^{8r}\tfrac{2}{9}c^3\varepsilon^{3+3r} - \tfrac{1}{8}\varepsilon^8 + p_{36}\cdot\mu\varepsilon^m),\\
x(T) &= (\ 0\ ,\ 0\ , \tfrac{2}{3}c^3\varepsilon^{3r}, \tfrac{2}{3}(\varepsilon+\mu\varepsilon^m)^3, -\tfrac{1}{4}c^8\varepsilon^{8r}, \tfrac{4}{9}c^3\varepsilon^{3+3r} - \tfrac{1}{4}\varepsilon^8 + p_{46}\cdot\mu\varepsilon^m).
\end{aligned}
\tag{3.25}
$$

Here p_{36} and p_{46} are polynomial expressions in $(c, \varepsilon, \varepsilon^r, \mu\varepsilon^m)$. In particular,

$$
p_{46} = \left(\tfrac{4}{3}c^3\varepsilon^{3r+2} - 2\varepsilon^7\right) + q_{46}\cdot\mu\varepsilon^m, \tag{3.26}
$$

where q_{46} is a polynomial expression in $(c, \varepsilon, \varepsilon^r, \mu\varepsilon^m)$. First consider the case with $\mu = 0$ (no *perturbation*). If $r = 1$, then $x_6(2c\varepsilon^r + 2\varepsilon) > 0$ for all sufficiently small $\varepsilon > 0$. However, (still with $\mu = 0$) if one chooses $r \geq \frac{5}{3}$, then one may obtain $y_4(T) < 0$ for suitable choices of c. Choosing the critical value $r = \frac{5}{3}$ yields equality of the exponents in ε^8 and ε^{3+3r}, and thus by varying the parameter c one reaches the points $y(T) = \varepsilon^{\frac{40}{3}}\ (0, 0, -\frac{1}{4}c^8, \frac{4}{9}c^3 - \frac{1}{4})$ in the third quadrant of the plane $y_1 = y_2 = 0$.

Alternatively, fix the value $c^* = (9/16)^{1/3}$, which yields $y_4(T) = 0$ when $\mu = 0$. To reach a given point $\eta = (0, 0, \eta_3, \eta_4)$ with $\eta_3, \eta_4 < 0$ first choose $\varepsilon = (16/9)^{1/5}\cdot(-4\eta_3)^{3/40}$. Then, using that $p_{46}|_{\mu=0} \neq 0$, the implicit function theorem guarantees that (for $-\eta_4 > 0$ sufficiently small) one can solve $\eta_4 = \mu\varepsilon^m \cdot p_{46}(c, \varepsilon, \varepsilon^r, \mu\varepsilon^m)$ (with $m = \frac{40}{3}$) for $\mu > 0$.

Note that the total time $T = 2\varepsilon + 2c\varepsilon^{\frac{5}{3}} + 2\mu\varepsilon^{\frac{40}{3}} \stackrel{!}{=} s$ is of order ε (for $\varepsilon < 1$), and it is straightforward to reparameterize the controls choosing a constant multiple $s = a\varepsilon$ so that they qualify as a family of control variations at zero. Such linear reparameterization clearly does not affect the exponents such as $\frac{40}{3}$, thus yielding the statements in Theorem 3.8 and its corollaries. The construction of controls steering to any point $\eta \in \mathbb{R}^4$ with η_1, η_2 not necessarily zero requires no major innovation, just much book-keeping, and it is completely analogous to the proof of Theorem 4.3 given below. □

Remark 3.11. Recall that one commonly *linearly* reparameterizes needle variations when constructing convex combinations of tangent vectors as tangent vectors *of the same order*. What is new here is that one must consider nonlinear reparameterizations (here basically replace s by a power s^r with r strictly larger than one). As a consequence one may still obtain the convex combinations of the original tangent vectors here but not as tangent vectors of the same order. (Here the combination of eighth order tangent vectors yields at best tangent vectors of order (40/3) as shown above.) Further, much more severe implications for controllability properties in general are analyzed in section 5.

In summary—using the traditional language of STLC—the analysis of system (3.1) *quantified* the delicate interplay of simultaneous *neutralization* of two *obstructions* to controllability by (what initially might appear to be) *higher order* terms: The terms $\int x_4 x_1^2$ and $\int x_3 x_2^2$, which are always nonnegative, are 6th order in time, whereas the indefinite terms $\int x_2^7$ and $\int x_1^7$ are 8th order in time. The constructions in this section showed that it is indeed possible to reach points $p = (p_1, \ldots, p_6)$ with either $p_5 < 0$ or $p_6 < 0$ from $x(0) = 0$ in a time of order $\|p\|_\infty^{1/8}$, and even points p with both $p_5 < 0$ and $p_6 < 0$, but only in time of order at least $\|p\|_\infty^{3/38}$—as a result we have the nonconvex approximating cones for $8 \leq m < \frac{40}{3}$ sketched in Figure 3.1. The critical value obtained in this section is the exponent $\frac{5}{3} > 1$, which relates the maximal distance the variation can be translated to the *length* of the needle variation.

These phenomena never played a role in the classical study STLC as the *lower order obstructions* $\int x_1^2$ and $\int x_2^2$ cannot be *neutralized*, making any further study of the system in view of STLC irrelevant. Nonetheless, not only does this example demonstrate a noteworthy feature from the point of view of STLOC, but it also serves as the foundation of constructions in the next sections, which basically replace the role of the output map $\varphi\colon x \mapsto (x_1, x_2, x_5, x_6)$ by a nonstationary reference trajectory in order to obtain unexpected controllability properties.

4. A nonstationary reference trajectory. A slight modification of this system with output yields a corresponding result for controllability about (and optimality of) a *nonstationary reference trajectory*. It might well be possible to construct a similar example with fewer controls. We chose this implementation with four controls for its structural simplicity and because it allows one to easily build on the results of the previous section.

$$\begin{cases} \dot{x}_1 = u_1, & |u_1(\cdot)| \le 1, \\ \dot{x}_2 = u_2, & |u_2(\cdot)| \le 1, \\ \dot{x}_3 = x_1^2 + (1 + u_{01}), & |u_{01}(\cdot)| \le 1, \\ \dot{x}_4 = x_2^2 + (1 + u_{02}), & |u_{02}(\cdot)| \le 1, \\ \dot{x}_5 = x_4 x_1^2 - x_1^7, & x(0) = 0, \\ \dot{x}_6 = x_3 x_2^2 - x_2^7, & x^*(t) = (\,0, 0, t, t, 0, 0). \end{cases} \tag{4.1}$$

The key in this construction is that the *lower order* (i.e., apparently dominant) definite components $\int_0^T x_1^2(t)dt$ and $\int_0^T x_2^2(t)dt$ are aligned with the direction of the nonstationary reference trajectory and that the *zero speed* $\|\dot{x}\| = 0$ is on the *boundary* of the set of admissible *velocities* (on the hyperplane $x_1 = x_2 = 0$, which includes the reference trajectory). With the fixed boundary velocity $u_{01} = u_{02} = -1$ the system may be considered a two-input system (with controls (u_1, u_2)) about the *stationary reference trajectory* $x \equiv 0$ (basically system (3.1)). This two-input system is not STLC as obviously the (x_3, x_4)-directions are uncontrollable. (It is impossible to reach points x with $x_3 < 0$ or $x_4 < 0$ from $x = 0$.) But in (4.1) these uncontrollable directions are aligned with (as opposed to transversal to) the nonstationary reference trajectory, and thus controllability involves comparison with $\dot{x}_3 = \dot{x}_4 \equiv 1$ (as opposed to $\dot{x}_3 = \dot{x}_4 \equiv 0$).

With this design in mind, it is natural to combine standard techniques and results about STLC and convexity of approximating cones that apply to systems about an equilibrium point (i.e., first holding $u_{01} = u_{02} = -1$ fixed) with constructions unique to nonstationary reference trajectories. In particular, after having *generated* desired tangent vectors working in the vicinity of the *equilibrium point*, it is straightforward to *catch up* with the prescribed nonstationary reference trajectory by using $u_{01} > 0$ and $u_{02} > 0$; compare Figure 4.1 for a schematic illustration.

THEOREM 4.1. *For* $8 \le m < \frac{38}{3}$ *the approximating cones* $\overline{\mathcal{K}}_0^m$ *of system* (4.1) *are not convex as they have* inward corners *in the sense that*

$$(0, 0, 0, 0, v_5, v_6) \in \overline{\mathcal{K}}_0^m \iff (\, v_5 \ge 0 \ \text{ or } \ v_6 \ge 0\,). \tag{4.2}$$

Proof. First note that if $\xi(\cdot, u)$ and $x(\cdot, u)$ denote the solution curves of systems (3.1) and (4.1), respectively (for the same control), then clearly $x_5(t, u) \ge \xi_5(t, u)$ and $x_6(t, u) \ge \xi_6(t, u)$ for all $t \ge 0$. Thus, it is a direct consequence of Theorem 3.2 that if $m < \frac{38}{3}$, $v_5 < 0$, *and* $v_6 < 0$, then $(0, 0, 0, 0, v_5, v_6) \notin \mathcal{K}_0^m$.

In the other direction, we show that every $v = (0, 0, v_3, v_4, v_5, v_6) \in \mathbb{R}^6$ with $v_5 \ge 0$ or $v_6 \ge 0$ is contained in the cones $\overline{\mathcal{K}}_0^m$ for $m \ge 8$. (More general conclusions

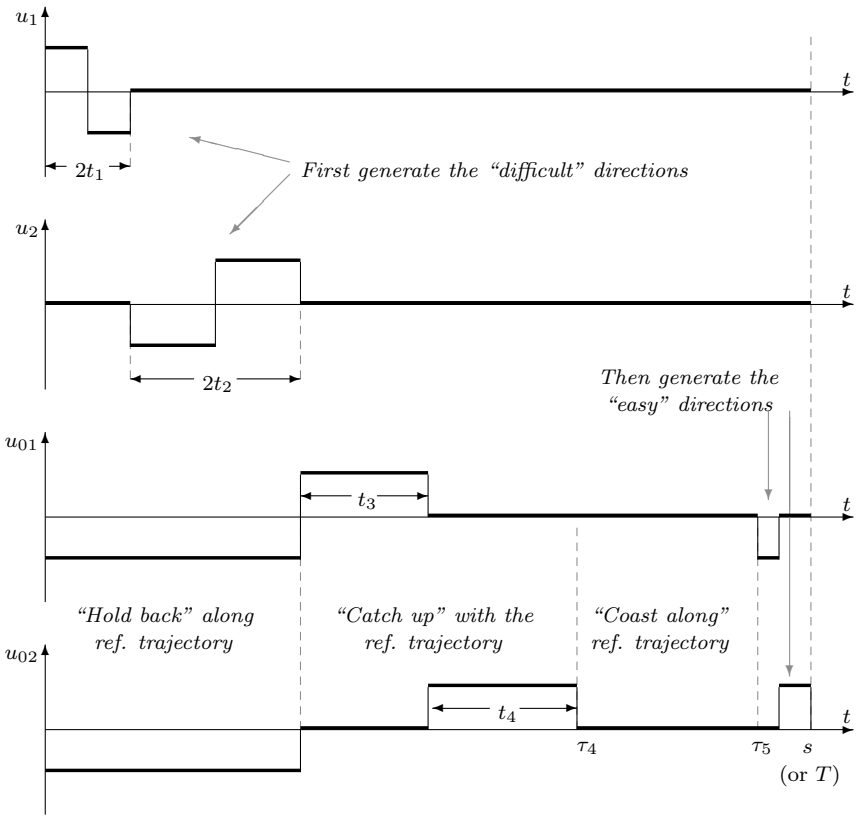


FIG. 4.1. *Schematic design of the needle variations for the proof of Theorem* 4.1.

allowing $v_1, v_2 \neq 0$ may be obtained using Theorem 4.3.) Suppose $0 < s < 1$ is given and $q = (0, 0, q_3, q_4, q_5, q_6) \in R^6$ is such that $\|q\|_\infty < (1/26)^8$ and w.l.o.g. $q_6 \geq 0$. We exhibit a family of control variations at zero, $u_s = u_{s,q}\colon [0, s] \mapsto [-1, 1]^4$, $0 \leq s \leq 1$, such that $x(s, u_s) = x^*(s) + s^8 q$. Write $t_1 = s\,(4|q_5|)^{1/8}$, and let $t_2 \geq 0$ be the unique nonnegative number such that $s^8 q_6 = \frac{2}{9} t_1^3 t_2^3 + \frac{1}{8} t_2^8$. Moreover, let $t_3 = 2t_1 + 2t_2 - \frac{2}{3} t_1^3$, and $t_4 = 2t_1 + 2t_2 - \frac{2}{3} t_2^3$. Write $\tau_4 = 2t_1 + 2t_2 + t_3 + t_4$ and $\tau_5 = s - s^8(|q_3| + |q_4|)$.

It is critical that $\tau_4 \leq \tau_5$. To see this, note that $t_i \leq 2s\|q\|_\infty^{1/8}$ for both $i = 1, 2$. Next $t_i \leq 8s\|q\|_\infty^{1/8}$ for both $i = 3, 4$, and thus $\tau_4 \leq 24s\|q\|_\infty^{1/8}$. On the other hand, $\tau_5 \geq s - 2s\|q\|_\infty^{1/8}$ (since both $s < 1$ and $\|q\|_\infty < 1$). Thus $\tau_5 - \tau_4 \geq s(1 - 26\|q\|_\infty^{1/8}) \geq 0$ follows from the assumption that $\|q\|_\infty < (1/26)^8$. Consequently, the following is a well-defined one-parameter family of control variations of $u^* \equiv (0, 0, 0, 0)$:

$$
u(t) = \begin{cases}
(\operatorname{sign}(q_5), 0, -1, -1) & \text{if} \quad 0 \leq t < t_1, \\
(-\operatorname{sign}(q_5), 0, -1, -1) & \text{if} \quad t_1 \leq t < 2t_1, \\
(0, 1, -1, -1) & \text{if} \quad 2t_1 \leq t < 2t_1 + t_2, \\
(0, -1, -1, -1) & \text{if} \quad 2t_1 + t_2 \leq t < 2t_1 + 2t_2, \\
(0, 0, 1, 0) & \text{if} \quad 2t_1 + 2t_2 \leq t < 2t_1 + 2t_2 + t_3, \\
(0, 0, 0, 1) & \text{if} \quad 2t_1 + 2t_2 + t_3 \leq t < \tau_4, \\
(0, 0, 0, 0) & \text{if} \quad \tau_4 \leq t < \tau_5, \\
(0, 0, \operatorname{sign}(q_3), 0) & \text{if} \quad \tau_5 \leq t < s - s^8|q_4|, \\
(0, 0, 0, \operatorname{sign}(q_4)) & \text{if} \quad s - s^8|q_4| \leq t < s.
\end{cases} \tag{4.3}
$$

Note that as a family of control variations *at zero* this is a family of needle variations. One easily verifies that at selected switching times the trajectory passes through

$$\begin{aligned} x(2t_1) &= (0,0,\tfrac{2}{3}t_1^3,0,s^8q_5,0),\\ x(2t_1+2t_2) &= (0,0,\tfrac{2}{3}t_1^3,\tfrac{2}{3}t_2^3,s^8q_5,s^8q_6),\\ x(\tau_4) &= (0,0,\tau_4,\tau_4,s^8q_5,s^8q_6),\\ x(\tau_5) &= (0,0,\tau_5,\tau_5,s^8q_5,s^8q_6),\\ x(s) &= (0,0,s,s,0,0)+s^8q. \end{aligned} \tag{4.4}$$

The flow Φ corresponding to the control $u^* \equiv 0$ satisfies $\Phi_{-s}(0,0,x_3,x_4,x_5,x_6) = (0,0,x_3-s,x_4-s,x_5,x_6)$, and thus $x(s,u_s) = x(s,u^*)+s^8\Phi_s(0,0,q_3,q_4,q_5,q_6)$. Consequently $(0,0,\lambda q_3,\lambda q_4,\lambda q_5,\lambda q_6) \in \overline{\mathcal{K}}_0^8$ for all $\lambda \geq 0$. □

The proof showed, in particular, that the vectors $(0,0,0,0,-1,0)$ and $(0,0,0,0,0,-1)$ are members of $\overline{\mathcal{K}}_0^8$ that are generated by *needle variations*. In contrast, according to Theorem 3.2, the convex combination $\frac{1}{2}(0,0,0,0,-1,-1)$ cannot at all be generated as an eighth order tangent vector. In particular, it cannot be generated by any (however clever) combination of needle variations. Nonetheless, in analogy to Corollary 3.10, in this particular system, this convex combination can still be generated as an even higher order tangent vector as shown below. First we establish a basic estimate needed later.

LEMMA 4.2. *If $0 < a < M^5$ and $0 < b < M^8$, then the unique positive root t_0 of the scalar function $f(t) = b + at^3 - \frac{1}{8}t^8$ satisfies $t_0 < 2M$.*

Proof. The existence of a unique positive root is elementary. Set $t_{00} = (8a)^{1/5}$, which is the first positive root of $t \mapsto at^3 - \frac{1}{8}t^8$. Since $f''(t) < 0$ for all $t \geq t_{00}$, the linear estimate $0 \overset{!}{=} f(t_0) \leq f(t_{00}) + f'(t_{00})(t_0 - t_{00}) = b - 5 \cdot 8^{2/5}a^{7/5}(t_0 - (8a)^{1/5})$ yields the bound

$$t_{00} \leq t_0 \leq t_{00} + \frac{b}{10 \cdot 2^{1/5}a^{7/5}} = \frac{10 \cdot 2^{4/5}a^{8/5} + b}{10 \cdot 2^{1/5}a^{7/5}}. \tag{4.5}$$

In the case that $b < a^{8/5}$, this immediately yields the desired estimate $t_0 < 2a^{1/5} \leq 2M$. Alternatively, if a is too small, i.e., if $b \geq a^{8/5}$, we find directly that $f(2b^{1/8}) \leq b + b^{5/8}(2b^{1/8})^3 - \frac{1}{8}(2b^{1/8})^8 = -23b < 0$, and hence the first positive zero t_0 of f again satisfies $x_0 \leq 2b^{1/8} < 2M$. □

THEOREM 4.3. *For every $T > 0$ sufficiently small, the reachable set $\mathcal{R}(T)$ of system (4.1) contains the open ball $B^\infty_{x^*(T)}((2^{-5}T)^{40/3})$ centered at $x^*(T)$.*

COROLLARY 4.4. *The system (4.1) is STLC about the reference trajectory $x^*(t) = (0,0,t,t,0,0)$.*

COROLLARY 4.5. *The approximating cone $\overline{\mathcal{K}}_0^{40/3}$ of system (4.1) is the whole tangent space $\overline{\mathcal{K}}_0^{40/3} = T_0\mathbb{R}^6$.*

Proof. Similar to the proof of Theorem 4.1 we explicitly construct for any given time $T > 0$ and any given point q sufficiently close to $x^*(T)$ a control $u = u_q\colon [0,T] \mapsto [-1,1]^4$ which steers the system to q. The main differences to the prior work is that now we need to also consider the case when $q_1 \neq 0$ and $q_2 \neq 0$. Note that since suitable open mapping principles for approximating cones for this setting are not available, we cannot rely on just generating tangent vectors.

Suppose $0 < T < 1$ and $q \in B_0^\infty(2^{-40}T^{40/3})$ are given. We consider only the case that both $q_5 < 0$ and $q_6 < 0$. All other cases can be handled analogously, albeit

with considerably simpler constructions. The basic construction for moving into the *difficult* directions $(0,0,0,0,*,*)$ is the same as in earlier proofs. These are followed by simple control actions that move into the *easy* directions. However, these later moves also have some effect on the difficult directions. Thus we first calculate these side effects and modify the initial controls to *overshoot* the targets in the difficult directions. The required corrections are calculated to cancel the integrals that result from (4.11) (also compare (4.12)).

$$\Delta q_{31} = q_1^2|q_2| = \int_0^{|q_2|} q_1^2\,dt, \qquad \Delta q_3 = \tfrac{1}{3}|q_1|^3 + |q_2|q_1^2 = \Delta q_{31} + \int_0^{|q_1|} (\sigma_1 t)^2 dt, \tag{4.6}$$

$$\Delta q_4 = \tfrac{1}{3}|q_2|^3 = \int_0^{|q_2|} t^2 dt, \tag{4.7}$$

$$\begin{aligned}\Delta q_{51} &= q_1^2|q_2|(T+q_4) - \tfrac{1}{2}q_1^2q_2^2 - \tfrac{1}{4}q_1^2q_2^4 - q_1^7|q_2| \\ &= \int_0^{|q_2|} \left((\tau_8 + q_4 - \Delta q_4 + (t + \tfrac{1}{3}t^3))q_1^2 - q_1^7\right)\,dt,\end{aligned} \tag{4.8}$$

$$\begin{aligned}\Delta q_5 &= \Delta q_{51} + \tfrac{1}{3}|q_1|^3(T+q_4) - \tfrac{1}{3}|q_1|^3|q_2| - \tfrac{1}{9}|q_1|^3|q_2|^3 - \tfrac{1}{12}q_1^4 - \tfrac{1}{8}|q_1|q_1^7 \\ &= \Delta q_{51} + \int_0^{|q_1|} \left((\tau_7 + t + q_4 - \Delta q_4)(\sigma_1 t)^2 + (\sigma_1 t)^7\right)\,dt,\end{aligned} \tag{4.9}$$

$$\begin{aligned}\Delta q_6 &= \tfrac{1}{3}|q_2|^3 T + \tfrac{1}{3}|q_2|^3 q_3 - \tfrac{1}{12}q_1^2q_2^4 - \tfrac{1}{12}q_2^4 - \tfrac{1}{8}|q_2|q_2^7 \\ &= \int_0^{|q_2|} \left((\tau_8 + q_3 - \Delta q_{31} + (q_1^2+1)t)(\sigma_2 t)^2 - (\sigma_2 t)^7\right)\,dt.\end{aligned} \tag{4.10}$$

What really matters in the subsequent construction is the size of these correction terms—clearly they are all of order at least 3 in $\|q\|_\infty$. Next let $t_1 = (-4(q_5 - \Delta q_5))^{\frac{1}{8}}$, $t_2 \geq 0$, be the smallest nonnegative number such that $(q_6 - \Delta q_6) = \frac{2}{9}t_1^3t_2^3 - \frac{1}{8}t_2^8$ (which is unique in the challenging case of $(q_6 - \Delta q_6) < 0$). Let $t_3 = 2t_1 + 2t_2 - \frac{2}{3}t_1^3$ and $t_4 = 2t_1 + 2t_2 - \frac{2}{3}t_2^3$. Write $\tau_3 = 2t_1 + 2t_2 + t_3$, $\tau_4 = 2t_1 + 2t_2 + t_3 + t_4$ and working backwards $\tau_8 = T - |q_2|$, $\tau_7 = \tau_8 - |q_1|$, $\tau_6 = \tau_7 - |q_4 - \Delta q_4|$, $\tau_5 = \tau_6 - |q_3 - \Delta q_3|$. Also use the abbreviations $\sigma_i = \operatorname{sign}(q_i)$ for $i = 1,2$ and $\sigma_i = \operatorname{sign}(q_i - \Delta q_i)$ for $i = 3,4$. Define the control $u = u_{q,T}\colon [0,T] \mapsto [-1,1]^4$ by

$$u(t) = \begin{cases} (1, & 0, & -1, & -1) & \text{if} & 0 \leq t < t_1, \\ (-1, & 0, & -1, & -1) & \text{if} & t_1 \leq t < 2t_1, \\ (0, & 1, & -1, & -1) & \text{if} & 2t_1 \leq t < 2t_1 + t_2, \\ (0, & -1, & -1, & -1) & \text{if} & 2t_1 + t_2 \leq t < 2t_1 + 2t_2, \\ (0, & 0, & 1, & 0) & \text{if} & 2t_1 + 2t_2 \leq t < \tau_3, \\ (0, & 0, & 0, & 1) & \text{if} & \tau_3 \leq t < \tau_4, \\ (0, & 0, & 0, & 0) & \text{if} & \tau_4 \leq t < \tau_5, \\ (0, & 0, & \sigma_3, & 0) & \text{if} & \tau_5 \leq t < \tau_6, \\ (0, & 0, & 0, & \sigma_4) & \text{if} & \tau_6 \leq t < \tau_7, \\ (\sigma_1, & 0, & 0, & 0) & \text{if} & \tau_7 \leq t < \tau_8, \\ (0, & \sigma_2, & 0, & 0) & \text{if} & \tau_8 \leq t < T. \end{cases} \tag{4.11}$$

It is essential to verify that $\tau_4 \le \tau_5$. Start with the crude estimate $|\Delta q_i| \le 2\|q\|_\infty^3 \le 2\|q\|_\infty$ for $i = 1, 2, 5, 6$. Next $t_1 \le 2\|q\|_\infty^{1/8}$. The critical estimate is for t_2—note the sign reversal in the defining equation in this construction compared to the one in the proof of Theorem 4.1: The earlier construction allowed one to estimate $s^8 q_6 = \frac{2}{9}t_1^3 t_2^3 + \frac{1}{8}t_2^8 \overset{!}{\ge} \frac{1}{8}t_2^8$ yielding $t_2 \le (8s^8 q\_6)^{1/8}$. Now in the interesting case $0 \ge (q_6 - \Delta q_6) = \frac{2}{9}t_1^3 t_2^3 - \frac{1}{8}t_2^8$ Lemma 4.2 applies with $a = \frac{2}{9}t_1^3 \le \frac{4}{9}\|q\|_\infty^{3/8}$, $b = (\Delta q_6 - q_6) \le 3\|q\|_\infty$, and $M = \|q\|_\infty^{3/40}$, yielding $t_2 \le 2\|q\|_\infty^{3/40}$. Easily $t_i \le 8\|q\|_\infty^{3/40}$ for $i = 3, 4$, and thus $\tau_4 \le 24\|q\|_\infty^{3/40}$.

On the other hand, $T - \tau_5 \le 4\|q\|_\infty + |\Delta q_5| + |\Delta q_6| \le 8\|q\|_\infty \le 8\|q\|_\infty^{3/40}$. Together with $0 < T < 1$ and $\|q\|_\infty < (2^{-5}T)^{\frac{40}{3}}$ this yields $\tau_5 - \tau_4 \ge T - 32\|q\|_\infty^{3/40} \ge 0$.

Note that $x_1(t) = x_2(t) = 0$ for all $t \in [2t_1 + 2t_2, \tau_7]$. Thus $x_5(\cdot)$ and $x_6(\cdot)$ are constant on this interval. It is straightforward to verify that at selected switching times the trajectory passes through

$$
\begin{aligned}
x(t_1) &= (t_1, 0, \tfrac{1}{3}t_1^3, 0, -\tfrac{1}{8}t_1^8, 0),\\
x(2t_1) &= (0, 0, \tfrac{2}{3}t_1^3, 0, q_5 - \Delta q_5, 0),\\
x(2t_1 + t_2) &= (0, t_2, \tfrac{2}{3}t_1^3, \tfrac{1}{3}t_2^3, q_5 - \Delta q_5, -\tfrac{1}{8}t_2^8 + \tfrac{2}{9}t_1^3 t_2^3),\\
x(2t_1 + 2t_2) &= (0, 0, \tfrac{2}{3}t_1^3, \tfrac{2}{3}t_2^3, q_5 - \Delta q_5, q_6 - \Delta q_6),\\
x(\tau_3) &= (0, 0, \tau_3, t_3 + \tfrac{2}{3}t_2^3, q_5 - \Delta q_5, q_6 - \Delta q_6),\\
x(\tau_4) &= (0, 0, \tau_4, \tau_4, q_5 - \Delta q_5, q_6 - \Delta q_6),\\
x(\tau_5) &= (0, 0, \tau_5, \tau_5, q_5 - \Delta q_5, q_6 - \Delta q_6),\\
x(\tau_6) &= (0, 0, \tau_6 + q_3 - \Delta q_3, \tau_6, q_5 - \Delta q_5, q_6 - \Delta q_6),\\
x(\tau_7) &= (0, 0, \tau_7 + q_3 - \Delta q_3, \tau_7 + q_4 - \Delta q_4, q_5 - \Delta q_5, q_6 - \Delta q_6),\\
x(\tau_8) &= (q_1, 0, \tau_8 + q_3 - \Delta q_{31}, \tau_8 + q_4 - \Delta q_4, q_5 - \Delta q_{51}, q_6 - \Delta q_6),\\
x(T) &= (q_1, q_2, T + q_3, T + q_4, q_5, q_6). \qquad \square
\end{aligned}
\tag{4.12}
$$

Note that the construction above was given for a fixed terminal time T. But it is easily modified to yield families of control variations $u_s\colon [0, s] \mapsto U$ at zero by replacing T by s and q by $s^{40/3}q$. This yields the curve of endpoints $x(s, u_s) = (0, 0, s, s, 0, 0) + s^{40/3}q$ yielding $q \in \overline{\mathcal{K}}_0^{40/3}$.

In summary, this section modified the system (3.1) so that its STLOC properties translate into even less expected properties of STLC about a nonstationary reference trajectory. The key step in the construction is to align the *obstructions* $\int x_1^2$ and $\int x_2^2$ with the reference trajectory and to utilize control values that lie on the boundary of the control set U. As a result, the *difficult* directions (negative in x_5 and negative in x_6) can only be generated by *needle variations* at $t = 0$, which moreover can only be moved by time intervals whose length is a power of the length of the variation that is strictly larger than 1 (here 5/3). The tangent cones again exhibit loss of convexity for $8 \le m < \frac{38}{3}$, as illustrated in Figure 3.1.

5. Delicate structural instability. This section analyzes small modifications of the systems considered in the preceding sections. The main result is that what at first appear to be *"at a higher order"* perturbations can nonetheless destroy the delicate controllability exhibited in the earlier sections. More specifically, we exhibit a polynomial cascade system which has nonconvex approximating cones generated by needle variations and which is not controllable.

Returning to the discussion of the system (4.1), we recall that cones $\overline{\mathcal{K}}_0^8$ are not convex. While the vectors $(0,0,0,0,-1,0) \in \overline{\mathcal{K}}_0^8$ and $(0,0,0,0,0,-1) \in \overline{\mathcal{K}}_0^8$ could be generated by needle variations at zero as 8th order tangent vectors, it was shown that it is impossible to obtain convex combinations of these as tangent vectors of the same order. However, such convex combinations can be generated as even higher order tangent vectors, e.g., $(0,0,0,0,-\frac{1}{2},-\frac{1}{2}) \in \overline{\mathcal{K}}_0^{40/3}$ by fairly delicate constructions. As we illustrate next, such constructions can be rendered impossible through quite innocent-looking perturbations. Similar to section 3, we start again with an affine system with output and a stationary reference trajectory:

$$\text{(5.1)} \qquad \begin{cases} \dot z_1 = u_1, & |u_1(\cdot)| \le 1, \\ \dot z_2 = u_2, & |u_2(\cdot)| \le 1, \\ \dot z_3 = z_1^2, & z(0) = 0, \\ \dot z_4 = z_2^2, & \varphi(z) = (z_1, z_2, z_5, z_6), \\ \dot z_5 = z_4 z_1^2 - z_1^7 + z_1^{10} + z_2^{10}, & \\ \dot z_6 = z_3 z_2^2 - z_2^7 + z_1^{10} + z_2^{10}. & \end{cases}$$

It is clear that the added 10th order terms do not at all affect the construction given in the proof of Theorem 3.2 and Corollary 3.4, and thus it is still true that the 8th order cone is nonconvex.

PROPOSITION 5.1. *For the system* (5.1), $(0,0,a,b) \in \overline{\mathcal{K}}_\varphi^8$ *if and only if* $a \ge 0$ *or* $b \ge 0$.

However, we will show that the perturbations destroy the controllability, and the tangent cones of all orders $m \ge 8$ are not convex.

THEOREM 5.2. *For* $0 \le T < 1$ *the image* $\varphi(\mathcal{R}(T))$ *of the reachable set of system* (5.1) *does not contain any points of the form* $(0,0,\eta_3,\eta_4)$ *with* $\eta_3 < 0$ *and* $\eta_4 < 0$.

COROLLARY 5.3. *The system* (5.1) *is not* STLOC *about* $y = 0$.

COROLLARY 5.4. *For the system* (5.1) *and* $m \ge 8$, *both* $(0,0,-1,0) \in \overline{\mathcal{K}}_\varphi^m$ *and* $(0,0,0,-1) \in \overline{\mathcal{K}}_\varphi^m$ *are generated by needle variations, but if* $\eta_3, \eta_4 < 0$, *then* $(0,0,\eta_3,\eta_4) \notin \overline{\mathcal{K}}_\varphi^m$ *for all* $m > 0$. *In particular, for all* $m > 0$, *the convex combination* $(0,0,-\frac{1}{2},-\frac{1}{2}) \notin \overline{\mathcal{K}}_\varphi^m$.

Proof. We write z and x for the corresponding solutions of the perturbed and unperturbed systems (3.1) and (5.1); i.e., $z_i(t,u) = x_i(t,u)$ for $i = 1,2,3,4$, while for $i = 5,6$

$$\text{(5.2)} \qquad z_i(t,u) = x_i(t,u) + \int_0^t \big(x_1^{10}(\tau,u) + x_2^{10}(\tau,u)\big)\, d\tau \quad \text{for all } \ 0 \le t \le T.$$

W.l.o.g. again assume that (3.15) holds (else permute in the following arguments the indices $(1,2,3,4,5,6) \mapsto (2,1,4,3,6,5)$). If $x_5(T,u) \ge 0$ (or $x_6(T,u) \ge 0$), then also $z_5(T,u) \ge 0$ (or $z_6(T,u) \ge 0$, respectively), and nothing needs to be shown. Thus suppose $x_5(T,u) < 0$ and $x_6(T) < 0$. Then use Lemma 3.7 and the key estimate (3.21) to obtain

$$\text{(5.3)} \qquad \int_0^T x_2^{10}(t)\, dt = \int_0^T |x_2(t)|^{10} dt \ge \tfrac{2}{11}\, \xi_2^{11} \ge \tfrac{2}{11} \left(\left(\frac{-x_5(T)}{T} \right)^{\frac{3}{35}} \right)^{11} \ge |x_5(T)| > 0.$$

If $T < 1$, clearly $|x_5(T)| < 1$. Together with $\frac{33}{35} < 1$ this shows that $z_5(T) = x_5(T) + \int_0^T |x_2(t)|^{10} dt \ge 0$ if $z_6(T) < 0$. ∎

For corresponding results for loss of controllability about a nonstationary reference trajectory, consider the system

$$
\begin{cases}
\dot{z}_1 = u_1, & |u_1(\cdot)| \le 1,\\
\dot{z}_2 = u_2, & |u_2(\cdot)| \le 1,\\
\dot{z}_3 = z_1^2 + (1+u_{01}), & |u_{01}(\cdot)| \le 1,\\
\dot{z}_4 = z_2^2 + (1+u_{02}), & |u_{02}(\cdot)| \le 1,\\
\dot{z}_5 = z_4 z_1^2 - z_1^7 + z_1^{10} + z_2^{10}, & z(0) = 0,\\
\dot{z}_6 = z_3 z_2^2 - z_2^7 + z_1^{10} + z_2^{10}, & z^*(t) = (\,0,0,t,t,0,0).
\end{cases}
\tag{5.4}
$$

By combining the arguments from the previous sections, one readily obtains the following statements. In particular, the 10th order terms in (5.4) do not affect the constructions of the 8th order tangent vectors; compare the proof of Theorem 3.2.

PROPOSITION 5.5. *For the system* (5.4), $(0,0,0,0,a,b) \in \overline{\mathcal{K}}_0^8$ *if and only if* $a \ge 0$ *or* $b \ge 0$.

On the other hand the estimate (5.3) in the proof of Theorem 5.2 applies also to the system with nonstationary reference trajectory.

THEOREM 5.6. *For* $0 \le T < 1$ *the reachable set* $\mathcal{R}(T)$ *of system* (5.4) *does not contain any points of the form* $(0,0,0,0,q_5,q_6)$ *with* $q_5 < 0$ *and* $q_6 < 0$.

COROLLARY 5.7. *The system* (5.4) *is not* STLC *about* $z^*(t) = (0,0,t,t,0,0)$.

The control variations constructed to generate the tangent vectors $(0,0,0,0,-1,0)$ and $(0,0,0,0,0,-1)$ satisfy the requirements of Definition 2.5. Together with Theorem 5.6 this allows us to state the last corollary for both notions of tangent vectors $\mathcal{K}_0^m$ and $\mathcal{K}_T^m$.

COROLLARY 5.8. *For the system* (5.4), $m \ge 8$, $0 \le T < 1$, *none of the cones* $\overline{\mathcal{K}}_T^m$ *is convex. In particular,* $(0,0,0,0,-1,0) \in \overline{\mathcal{K}}_T^m$ *and* $(0,0,0,0,0,-1) \in \overline{\mathcal{K}}_T^m$ *are generated by needle variations at zero, but* $(0,0,0,0,-\frac{1}{2},-\frac{1}{2}) \notin \overline{\mathcal{K}}_T^m$ *for all* $m > 0$ *and* $T \ge 0$.

In summary, this section exploited the lack of convexity of tangent cones of order $8 \le m < \frac{38}{3}$ to *destroy* the controllability properties of systems (3.1) and (4.1) by adding some *perturbations*, which at first view may appear to be of even higher order. While in the systems in the previous sections needle variations could still be combined to generate convex combinations of tangent vectors, albeit at a higher order, these *perturbed* systems make it entirely impossible to obtain convex combinations (at any order) of tangent vectors that are generated by needle variations.

6. Conclusion and further outlook. Summarizing, the analysis of the custom-designed systems in this article has yielded two major insights. The immediate result is a counterexample for a long-standing natural conjecture about needle variations.

- The most general notion of a tangent vector (which works well for stationary reference trajectories) does not necessarily yield convex cones either for nonstationary reference trajectories or for systems with output.
- Narrower notions of tangent vectors that yield convex cones, e.g., by requiring explicitly that needle variations must be *movable* by fixed times, may miss the controllability of some systems. (Recall in section 4 that the systems were controllable, but this cannot be detected with needle variations that require standard *movability*.)
- The common *technical* conditions on *needle variations* are *not* stated only because of *not yet discovered* stronger theorems, but they are essential even for some of the most benign systems. In a nutshell, the *time(s)* at which needle variations are generated must be movable—though not necessarily by a fixed distance, but by a distance of the same order of magnitude as the

duration of the needle variation. Movability by a smaller amount, e.g., some power larger than one, is not sufficient; compare Remark 3.11. Otherwise, one should expect a lack of convexity of the *cones* of tangent vectors and a consequent failure to satisfy the hypotheses for open mapping theorems that are essential for making the cones into meaningful *"approximating cones."*

On a deeper level these examples cast further doubt on the structural stability and robustness of the STLC property. It is well known that simple Taylor approximations do not preserve controllability (see, e.g., [3]). For example, the system $\dot{x}_1 = u$, $\dot{x}_2 = x_1$, $\dot{x}_3 = x_2^2 + x_1^3$ [21] is STLC about $x(0) = 0$ and $|u(\cdot)| \leq 1$, but its quadratic Taylor approximation $\dot{y}_1 = u$, $\dot{y}_2 = y_1$, $\dot{y}_3 = y_2^2$ is not STLC. Instead, many controllability results of the past two decades rely on nilpotent approximating systems [4, 12] which are based on *graded structures*: For a fixed set of local coordinates $(x_1, \ldots, x_n)$ on $\mathbb{R}^n$ and weights $r_1, \ldots, r_n \geq 1$, consider the family of dilations $\Delta: \mathbb{R}^+ \times \mathbb{R}^n \mapsto \mathbb{R}$ defined by $\Delta_s(x) = (s^{r_1}x_1, \ldots, s^{r_n}x_n)$. A function $\varphi: \mathbb{R}^n \mapsto \mathbb{R}^n$ is called homogeneous of order m with respect to Δ if $\varphi \circ \Delta_s = s^m \varphi$. A vector field f is called homogeneous of order k if (considered as a first order partial differential operator) it maps (smooth) homogeneous functions of any order m to homogeneous functions of order $m + k$. To *approximate* the system (2.1) by a *nilpotent* system of the same form, replace each vector field f_i by the *principal part* g_i of the expansion of f_i into a (generally infinite) series of homogeneous vector fields. If all vector fields g_i have negative degrees of homogeneity, then all are polynomial vector fields of *cascade form*, and they generate a *nilpotent* Lie algebra; see the survey [12] for details. Roughly speaking, the fundamental result is that if such *homogeneous* nilpotent approximating system with vector fields g_i is STLC, then the original system is *STLC*; see, e.g., [4, 22]. By varying the construction of the local coordinates and dilation exponents r_i, several major results for STLC follow; see, e.g., [4, 21, 22]. On the other hand, it is known that there exist polynomial systems that are STLC, but for them any such homogeneous nilpotent approximating system is not STLC [15].

Returning to the systems (3.1) and (4.1) discussed in this article which are STLOC and STLC, respectively, these are similar to the system in [15], as any of the standard nilpotent approximation schemes will simply delete the terms x_1^7 and x_2^7 on the right-hand side, yielding homogeneous nilpotent approximating systems that are not STLOC nor STLC. Since the example constructed in [15], one has been searching for more general approximating schemes that would *recognize* the need to preserve terms such as x_1^7 and x_2^7 in the above systems, which are essential for their controllability and which would only truncate *truly higher order* terms that are not essential for controllability. However, the analysis in the preceding section of the systems (5.1) and (5.4) raises new questions as the inclusion of the even higher order terms x_1^{10} and x_2^{10} again causes loss of controllability. One naturally wonders whether it is possible to construct such an infinite chain, which could possibly lead to a negative answer to the long-standing question of whether STLC (of analytic systems) is *finitely determined* [2].

Acknowledgments. The authors thank the anonymous referees and H. Frankowska who made numerous suggestions to help improve the article.

REFERENCES

[1] A. Agrachev, *Newton diagrams and tangent cones to attainable sets*, in Analysis of Controlled Dynamical Systems, Progr. Systems Control Theory 8, B. Bonnard, B. Bride, J. P. Gauthier, and I. Kupka, eds., Birkhäuser Boston, Boston, MA, 1991, pp. 11–20.

[2] A. Agrachev, *Is it possible to recognize local controllability in a finite number of differentiations?*, in Open Problems in Mathematical Systems and Control Theory, V. Blondel, E. Sontag, M. Vidyasagar, and J. Willems, eds., Springer, London, 1999, pp. 15–18.
[3] R. M. Bianchini and G. Stefani, *Normal local controllability of order one*, Internat. J. Control, 39 (1984), pp. 701–714.
[4] R. M. Bianchini and G. Stefani, *Graded approximations and controllability along a trajectory*, SIAM J. Control Optim., 28 (1990), pp. 903–924.
[5] R. M. Bianchini and G. Stefani, *Controllability along a trajectory: A variational approach*, SIAM J. Control Optim., 31 (1993), pp. 900–927.
[6] R. M. Bianchini, *Good needle-like variations*, in Proc. Sympos. Pure Math. 64, AMS, Providence, RI, 1999, pp. 91–101.
[7] R.-M. Bianchini and M. Kawski, *Lack of convexity for tangent cones of needle variations*, in Proceedings of the 41st IEEE Conference on Decision and Control, Vol. 2, IEEE Press, Piscataway, NJ, 2002, pp. 1916–1921.
[8] A. Bressan, *A high-order test for optimality of bang-bang controls*, SIAM J. Control Optim., 23 (1985), pp. 38–48.
[9] H. Frankowska, *An open mapping principle for set-valued maps*, J. Math. Anal. Appl., 127 (1987), pp. 172–180.
[10] H. Frankowska, *Local controllability of control systems with feedback*, J. Optim. Theory Appl., 60 (1989), pp. 277–296.
[11] H. Frankowska, *A conical open mapping principle for set-valued maps*, Bull. Austral. Math. Soc., 45 (1992), pp. 53–60.
[12] H. Hermes, *Nilpotent and high-order approximations of vector field systems*, SIAM Rev., 33 (1991), pp. 238–264.
[13] B. Kaskosz, *Abundant subsets of generalized control systems*, in Systems Modelling and Optimization, Chapman & Hall/CRC. Res. Notes. Math. 396, Boca Raton, FL, 1999, pp. 108–116.
[14] M. Kawski, *An angular open mapping theorem*, in Analysis and Optimization of Systems, Lect. Notes in Control and Inform. Sci. 111, A. Bensoussan and J. L. Lions, eds., Springer, Berlin, 1988, pp. 361–371.
[15] M. Kawski, *Control variations with an increasing number of switchings*, Bull. Amer. Math. Soc., 18 (1988), pp. 149–152.
[16] H. Knobloch, *Higher Order Necessary Conditions in Optimal Control Theory*, Lect. Notes in Control and Inform. Sci. 34, Springer–Verlag, Berlin-New York, 1981.
[17] A. J. Krener, *The high order maximal principle and its application to singular extremals*, SIAM J. Control Optim., 15 (1977), pp. 256–293.
[18] E. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[19] L. Pontryagin, V. Boltyanskii, R. Gamkrelidze, and E. Mischenko, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.
[20] G. Stefani, *Polynomial approximations to control systems and local controllability*, in Proceedings of the 25th IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 1985, pp. 33–38.
[21] H. J. Sussmann, *Lie brackets and local controllability: A sufficient condition for scalar-input systems*, SIAM J. Control Optim., 21 (1983), pp. 686–713.
[22] H. J. Sussmann, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.
[23] H. Sussmann, *Needle variations and almost lower semicontinuous differential inclusions*, Set-Valued Anal., to appear.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 239–259

BOUNDARY FEEDBACK STABILIZATION OF SHALLOW SHELLS*

SHUGEN CHAI†, YUXIA GUO‡, AND PENG-FEI YAO§

Abstract. We consider the stabilization of the shallow shell by boundary feedbacks where the model has a middle surface of any shape. First, we put the shallow shell model in a suitable semigroup scheme. The existence, the uniqueness, and the properties of solutions to the shallow shell are then treated by the semigroup approach and the regularity of elliptic boundary value problems. Finally, we establish the uniform energy decay rate for the shallow shell under some checkable geometric conditions on the middle surface.

Key words. shallow shell, regularity, boundary stabilization

AMS subject classifications. 35A, 35L, 35Q, 49A, 49B, 49E

PII. S0363012901397156

1. Introduction.

We are concerned with the stabilization of the shallow shell by boundary feedbacks. This issue has been analyzed a great deal for the wave equation and plates; see Lagnese [8], Lagnese and Lions [9], Lasiecka and Triggiani [10], [11], and many others. For thin shells, we know very little about this problem. A circular cylindrical shell is considered by Chen, Coleman, and Liu [3] and a spherical shell by Lasiecka, Triggiani, and Valente [12], and Triggiani [17]. In the above cases the models are expressed in terms of special coordinates and all the work takes place in those coordinates.

We study the shallow shell model where the tensor of change of curvature is given by the Hessian of the normal displacement; see Ciarlet [4], Mason [13], Niordson [14], or Koiter [7]. The model is written into a coordinate free form by using the global geometry analysis in Yao [20]. This is one of the simplest thin shell models. For other models, for example, the Koiter model where the change of the curvature tensor is much more complicated, the control problems seem to be even more difficult; see Chai and Yao [2].

We shall carry out the control scheme, which is given in Lagnese [8] for the boundary stabilization of thin plates, to study the boundary stabilization of the shallow shell and we obtain the exponential stabilization under very weak geometrical conditions. There are some difficulties we have to overcome.

One of the key problems in getting the uniform energy decay rate is obtaining the regularity of solutions to the shallow shell. By using some ideas in Lagnese [8] and the geometry approach, we address the resulting closed-loop system of the shallow shell after exerting the boundary feedback controls in an appropriate semigroup scheme so that the regularity in the time variable follows from the semigroup theory; see

*Received by the editors October 29, 2001; accepted for publication (in revised form) September 10, 2002; published electronically March 26, 2003. This work is supported by the NSF of China grant 60074006.

http://www.siam.org/journals/sicon/42-1/39715.html

†Department of Mathematics, Shanxi University, Taiyuan 030006, China; Institute of Systems Science, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing 100080, China (sgchai@china.com).

‡Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China (yguo@math.tsinghua.edu.cn).

§Institute of Systems Science, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing 100080, China; Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, Exhibition Road, London SW7 2BT, UK (p.yao@ic.ac.uk).

Pazy [15]. In addition, the regularity we need in the spatial variables is obtained by using the elliptic boundary value theory; see Agmon, Douglis, and Nirenberg [1]. We mention that, for the static problem, the existence, uniqueness, and regularity (i.e., the ellipticity) have been thoroughly treated by Ciarlet [4].

Another problem is that we have to develop some trace estimates for solutions of the closed-loop system which permit certain boundary traces to be expressed in terms of other traces modulo lower-order interior terms. We use Horn [6] to obtain the trace estimates on the tangential component of solutions of the shallow shell and Lasiecka and Triggiani [11] on the bending component. Those results allow us to have the stabilization under very weak geometrical conditions. We mention that trace estimates for the wave equation were developed in Lasiecka and Triggiani [10] to eliminate the geometric constraints.

1.1. Some notation. We introduce some notations in preparation for the shallow shell.

Denote the usual inner product in $\mathbb{R}^3$ by $\langle \cdot, \cdot \rangle$, i.e., the dot product. Let M be a surface in $\mathbb{R}^3$. For simplicity, M is assumed to be smooth. Surface M produces a natural Riemannian manifold of dimension 2 with the induced metric in $\mathbb{R}^3$. We denote this induced metric on surface M by g or by $\langle \cdot, \cdot \rangle$, as is convenient. For each $x \in M$, M_x is the tangential space of M at x. It is assumed that surface M is orientable with the unit normal field N on M. Denote the set of all vector fields on M by $\mathcal{X}(M)$. Denote the set of all k-order tensor fields and the set of all k-forms on M by $T^k(M)$ and $\Lambda^k(M)$, respectively, where k is a nonnegative integer. Then

$$\Lambda^k(M) \subset T^k(M).$$

In particular, $\Lambda^0(M) = T^0(M) = C^\infty(M)$ is the set of all C^∞ functions on M and

$$T^1(M) = T(M) = \Lambda(M) = \mathcal{X}(M),$$

where $\Lambda(M) = \mathcal{X}(M)$ is in the following isomorphism: for $X \in \mathcal{X}(M)$ given, the equation

$$U(Y) = \langle Y, X \rangle \qquad \forall\, Y \in \mathcal{X}(M)$$

determines a unique $U \in \Lambda(M)$.

It is well known that, for each $x \in M$, k-order tensor space T_x^k on M_x is an inner product space defined as follows. Let e_1, e_2 be an orthonormal basis of M_x. For any α, $\beta \in T_x^k$, $x \in M$, the inner product is given by

$$\langle \alpha, \beta \rangle_{T_x^k} = \sum_{i_1,\dots,i_k=1}^{2} \alpha(e_{i_1}, \dots, e_{i_k}) \beta(e_{i_1}, \dots, e_{i_k}) \qquad \text{at } x. \tag{1.1}$$

In particular, for $k = 1$ definition (1.1) becomes

$$g(\alpha, \beta) = \langle \alpha, \beta \rangle_{T_x} = \langle \alpha, \beta \rangle \qquad \forall\, \alpha, \beta \in M_x,$$

that is, the induced inner product of M_x in $\mathbb{R}^3$.

Let Ω be a bounded region of surface M with a regular boundary Γ or without boundary (when Γ is empty). From (1.1), $T^k(\Omega)$ are then inner product spaces in the following sense:

$$(T_1, T_2)_{T^k(\Omega)} = \int_\Omega \langle T_1, T_2 \rangle_{T_x^k}\, dx \qquad \forall\, T_1, T_2 \in T^k(\Omega), \tag{1.2}$$

where dx is the volume element of surface M in its Riemaniann metric g.

The completions of $T^k(\Omega)$ in inner products (1.2) are denoted by $L^2(\Omega, T^k)$. In particular, $L^2(\Omega, \Lambda) = L^2(\Omega, T)$. $L^2(\Omega)$ is the completion of $C^\infty(\Omega)$ in the following inner product:

$$(f, h)_{L^2(\Omega)} = \int_\Omega f(x)h(x)\, dx \qquad \forall\, f,\, h \in C^\infty(\Omega).$$

Let D be the Levi–Civita connection on M in the induced metric g of surface M. For $U \in \mathcal{X}(M)$, DU is the covariant differential of U which is a 2-order covariant tensor field in the following sense:

$$DU(X, Y) = D_Y U(X) = \langle D_Y U,\, X\rangle \qquad \forall\, X,\, Y \in M_x,\, x \in M. \tag{1.3}$$

We also define $D^*U \in T^2(M)$ by

$$D^*U(X, Y) = DU(Y, X) \qquad \forall\, X,\, Y \in M_x,\, x \in M, \tag{1.4}$$

that is, $D^*U \in T^2(M)$ is the transpose of DU. For any $T \in T^2(M)$, the trace of T at $x \in M$ is defined by

$$\mathrm{tr}T = \sum_{i=1}^{2} T(e_i, e_i),$$

where $e_1,\ e_2$ is an orthonormal basis of M_x. It is obvious that $\mathrm{tr}T \in C^\infty(M)$ if $T \in T^2(M)$.

For $T \in T^k(M)$ and $X \in \mathcal{X}(M)$, we define $l_X T \in T^{k-1}(M)$ by

$$l_X T(X_1, \ldots, X_{k-1}) = T(X, X_1, \ldots, X_{k-1}) \qquad \forall\, X_1,\, \ldots,\, X_{k-1} \in \mathcal{X}(M).$$

The Sobolev space $H^k(\Omega)$ is the completion of $C^\infty(\Omega)$ with respect to the norm

$$\|f\|^2_{H^k(\Omega)} = \sum_{i=1}^{k} \|D^i f\|^2_{L^2(\Omega, T^i)} + \|f\|^2_{L^2(\Omega)}, \qquad f \in C^\infty(\Omega), \tag{1.5}$$

where $D^i f$ is the ith covariant differential of f in the induced metric g of M which is an i-order tensor field on Ω, and $\|\cdot\|_{L^2(\Omega, T^i)}$ and $\|\cdot\|_{L^2(\Omega)}$ are the induced norms in inner products (1.1)–(1.2), respectively. For details on Sobolev spaces on Riemannian manifolds, we refer to Hebey [5] or Taylor [16].

Another important Sobolev space for us is $H^k(\Omega, \Lambda)$, defined by

$$H^k(\Omega, \Lambda) = \{\, U \,|\, U \in L^2(\Omega, \Lambda),\ D^i U \in L^2(\Omega, T^{i+1}),\ 1 \le i \le k \,\}$$

with inner product

$$(U, V)_{H^k(\Omega,\Lambda)} = \sum_{i=0}^{k} (D^i U, D^i V)_{L^2(\Omega, T^{i+1})} \qquad \forall\, U,\ V \in H^k(\Omega, \Lambda);$$

for example, see Wu [18]. In particular, $H^0(\Omega, \Lambda) = L^2(\Omega, \Lambda)$.

For $\hat{\Gamma} \subset \Gamma$, set

$$H^1_{\hat{\Gamma}}(\Omega, \Lambda) = \{\, W \,|\, W \in H^1(\Omega, \Lambda),\ W|_{\hat{\Gamma}} = 0 \,\}, \tag{1.6}$$

$$H^2_{\hat{\Gamma}}(\Omega) = \left\{ w \,|\, w \in H^2(\Omega),\ w\Big|_{\hat{\Gamma}} = \frac{\partial w}{\partial n}\Big|_{\hat{\Gamma}} = 0 \right\}. \tag{1.7}$$

In particular, $H^1_0(\Omega, \Lambda) = H^1_\Gamma(\Omega, \Lambda)$ and $H^2_0(\Omega) = H^2_\Gamma(\Omega)$.

1.2. Model. We assume that the middle surface of the shell is a bounded region Ω of surface M in $\mathbb{R}^3$ before the deformation takes place. The shell, a body in $\mathbb{R}^3$, is defined by

$$\mathcal{S} = \{\, p \,|\, p = x + zN(x),\ x \in \Omega,\ -h/2 < z < h/2 \,\},$$

where h is the thickness of the shell, small.

Denote by $\eta(x)$ the displacement vector of point x of the middle surface. We decompose the displacement vector η into a sum

$$\eta(x) = W(x) + w(x)N(x), \qquad x \in \Omega,\ W(x) \in M_x, \tag{1.8}$$

i.e., W and w are components of η on the tangent plane and on the normal of the undeformed middle surface Ω, respectively. The linearized strain tensor and the change of curvature tensor of the middle surface Ω are given by

$$\Upsilon(\eta) = \frac{1}{2}(DW + D^*W) + w\Pi \tag{1.9}$$

and

$$\rho(\eta) = -D^2 w \tag{1.10}$$

in a coordinate free form, respectively, where Π is the second fundamental form of surface M and D^2w the Hessian of w, which are justified for a shallow shell. For (1.9) and (1.10), we refer to Ciarlet [4], Niordson [14], Mason [13], or to Koiter [7].

Remark 1.1. If we express the two tensors (1.9) and (1.10) by a coordinate, they look complicated. Let the middle surface of the shell be given by a coordinate

$$\varphi = (\varphi_1(x_1, x_2),\ \varphi_2(x_1, x_2),\ \varphi_3(x_1, x_2)), \qquad (x_1, x_2) \in \mathbb{R}^2.$$

Set

$$\mathbf{a}_\alpha = \left(\frac{\partial \varphi_1}{\partial x_\alpha},\ \frac{\partial \varphi_2}{\partial x_\alpha},\ \frac{\partial \varphi_3}{\partial x_\alpha}\right), \qquad W = w_1\mathbf{a}_1 + w_2\mathbf{a}_2.$$

Then the tensors (1.9) and (1.10) become

$$\Upsilon_{\alpha\beta} = \frac{1}{2}(w_{\alpha|\beta} + w_{\beta|\alpha}) - b_{\alpha\beta}w,$$

$$\rho_{\alpha\beta} = -w|_{\alpha\beta},$$

where $1 \le \alpha,\ \beta \le 2$, $b_{\alpha\beta} = -\partial_{\mathbf{a}_\alpha} N \cdot \partial_{\mathbf{a}_\beta}$ is the second fundamental form, and

$$w_{\alpha|\beta} = \partial_{\mathbf{a}_\beta} w_\alpha - \Gamma^\lambda_{\alpha\beta} w_\lambda, \qquad w|_{\alpha\beta} = \partial_{\mathbf{a}_\beta}\partial_{\mathbf{a}_\alpha} w - \Gamma^\lambda_{\alpha\beta} w.$$

The shell strain energy associated with a displacement field η of the middle surface Ω can be written as

$$\mathcal{B}_1(\eta, \eta) = \frac{Eh}{1-\mu^2} \int_\Omega B(\eta, \eta)\, dx, \tag{1.11}$$

where

$$B(\eta, \eta) = a(\Upsilon(\eta), \Upsilon(\eta)) + \gamma a(\rho(\eta), \rho(\eta)), \qquad \gamma = h^2/12, \tag{1.12}$$

$$a(T_1, T_1) = (1-\mu)\langle T_1, T_1\rangle_{T_x^2} + \mu(\mathrm{tr}T_1)^2, \qquad T_1 \in T^2(\overline{\Omega}), \tag{1.13}$$

for $x \in \Omega$, where E, μ, respectively, denote Young's modulus and Poisson's coefficient of the material.

Thus, with expression (1.11) we are able to associate the following symmetric bilinear form, directly defined on the middle surface Ω:

$$\mathcal{B}(\eta, \zeta) = \int_\Omega B(\eta, \zeta)\, dx, \tag{1.14}$$

where η is given in (1.8) and

$$\zeta = U + uN, \qquad U(x) \in M_x, \quad x \in \Omega.$$

Denote by H and by k the mean curvature and the Gauss curvature of surface M, respectively. From Yao [20], we have the following Green's formula for the shallow shell.

Formula I. Let the bilinear form $\mathcal{B}(\cdot,\cdot)$ be given in (1.14). For all sufficiently smooth $\eta = (W, w)$ and $\varsigma = (U, u)$, we have

$$\mathcal{B}(\eta, \varsigma) = (\mathcal{A}\eta, \varsigma)_{L^2(\Omega,\Lambda)\times L^2(\Omega)} + \int_\Gamma \partial(\mathcal{A}\eta, \varsigma) d\Gamma, \tag{1.15}$$

where

$$\partial(\mathcal{A}\eta, \varsigma) = v_1(\eta)\langle U, n\rangle + v_2(\eta)\langle U, \tau\rangle + v_3(\eta)\frac{\partial u}{\partial n} + v_4(\eta)u, \tag{1.16}$$

n, τ are the normal and the tangential along curve Γ, respectively,

$$\mathcal{A}\eta = \begin{pmatrix} -\Delta_\mu W - (1-\mu)kW - \mathcal{F}(w) \\ \gamma[\Delta^2 w - (1-\mu)\delta(kdw)] + (H^2 - 2(1-\mu)k)w + \mathcal{G}(W) \end{pmatrix}, \tag{1.17}$$

Δ_μ is of the Hodge-Laplacian type, applied to the 1-form (or equivalently vector fields), defined by

$$\Delta_\mu = -\left(\frac{1-\mu}{2}\delta d + d\delta\right), \tag{1.18}$$

d the exterior differential, δ the formal adjoint of d, Δ the Laplacian on manifold M,

$$\begin{cases} \mathcal{F}(w) &= (1-\mu)l_{dw}\Pi + \mu H dw + w dH, \\ \mathcal{G}(W) &= (1-\mu)\langle DW, \Pi\rangle_{T_x^2} - \mu H \delta W, \end{cases} \tag{1.19}$$

and

$$\begin{cases} v_1(\eta) &= (1-\mu)\Upsilon(\eta)(n,n) + \mu(wH - \delta W), \\ v_2(\eta) &= (1-\mu)\Upsilon(\eta)(n,\tau), \\ v_3(\eta) &= \gamma[\Delta w - (1-\mu)D^2 w(\tau,\tau)], \\ v_4(\eta) &= -\gamma\left\{\dfrac{\partial \Delta w}{\partial n} + (1-\mu)\left[\dfrac{\partial}{\partial \tau}(D^2 w(\tau, n)) + k(x)\dfrac{\partial w}{\partial n}\right]\right\}. \end{cases} \tag{1.20}$$

By the "principle of virtual work" and Formula I, we obtain the following displacement equations for a shallow shell (see Yao [20]) after changing t to t/λ with $\lambda^2 E/(1-\mu^2) = 1$.

Formula II. We assume that there are no external loads on the shell and the shell is clamped along a portion Γ_0 of Γ and free on Γ_1, where $\Gamma_0 \cup \Gamma_1 = \Gamma$ and $\Gamma_0 \cap \Gamma_1 = \emptyset$. Then the displacement vector $\eta = (W, w)$ satisfies the following boundary value problem:

$$\begin{cases} W_{tt} - [\Delta_\mu W + (1-\mu)kW + \mathcal{F}(w)] = 0, \\ w_{tt} - \gamma\Delta w_{tt} + \gamma\left(\Delta^2 w - (1-\mu)\delta(kdw)\right) \\ \qquad\qquad +(H^2 - 2(1-\mu)k)w + \mathcal{G}(W) = 0, \\ \eta(0) = \eta^0, \quad \eta_t(0) = \eta^1, \end{cases} \quad \text{in } Q_\infty, \tag{1.21}$$

$$\begin{cases} W = 0, \\ w = \dfrac{\partial w}{\partial n} = 0, \end{cases} \quad \text{on } \Sigma_{0\infty}, \tag{1.22}$$

$$v_1(\eta) = v_2(\eta) = v_3(\eta) = 0 \quad \text{and} \quad v_4(\eta) + \gamma\frac{\partial w_{tt}}{\partial n} = 0 \quad \text{on } \Sigma_{1\infty}, \tag{1.23}$$

where

$$Q_\infty = \Omega \times (0,\infty), \quad \Sigma_{0\infty} = \Gamma_0 \times (0,\infty), \quad \Sigma_{1\infty} = \Gamma_1 \times (0,\infty). \tag{1.24}$$

Remark 1.2. If the shell is flat, a plate, the equations in (1.21) are uncoupled. The equation on the component w is the same as in Lagnese [8]—a Kirchhoff plate (see Yao [20]).

1.3. Uniform stabilization. We write (1.21) as

$$\eta_{tt} - \gamma(0, \Delta w_{tt}) + \mathcal{A}\eta = 0 \tag{1.25}$$

and define the total energy of shell by

$$E(t) = \frac{1}{2}[\|W_t\|^2_{L^2(\Omega,\Lambda)} + \|w_t\|^2_{L^2(\Omega)} + \gamma\|Dw_t\|^2_{L^2(\Omega,\Lambda)} + \mathcal{B}(\eta,\eta)]. \tag{1.26}$$

By Green's formula (1.15), the equations (1.25), and the boundary conditions (1.22) we obtain

$$\begin{aligned} &\frac{d}{dt}E(t) \\ &= \frac{d}{dt}\left\{\frac{1}{2}[\|W_t\|^2_{L^2(\Omega,\Lambda)} + \|w_t\|^2_{L^2(\Omega)} + \gamma\|Dw_t\|^2_{L^2(\Omega,\Lambda)} + \mathcal{B}(\eta,\eta)]\right\} \\ &= \mathcal{B}(\eta,\eta_t) + \int_\Omega [\langle W_{tt}, W_t\rangle + w_{tt}w_t + \gamma\langle Dw_{tt}, Dw_t\rangle]dx \\ &= \int_\Omega [\langle W_{tt}, W_t\rangle + w_{tt}w_t - \gamma\Delta w_{tt}w_t]dx + (\mathcal{A}\eta, \eta_t)_{L^2(\Omega,\Lambda)\times L^2(\Omega)} \\ &\quad + \int_\Gamma \left[\partial(\mathcal{A}\eta,\eta_t) + \gamma\frac{\partial w_{tt}}{\partial n}w_t\right]d\Gamma \\ &= \int_\Omega \langle \eta_{tt} - \gamma(0,\Delta w_{tt}) + \mathcal{A}\eta, \eta_t\rangle dx + \int_\Gamma \left(\partial(\mathcal{A}\eta,\eta_t) + \gamma\frac{\partial w_{tt}}{\partial n}w_t\right)d\Gamma \\ &= \int_{\Gamma_1} \left\{v_1(\eta)\langle W_t, n\rangle + v_2(\eta)\langle W_t, \tau\rangle + v_3(\eta)\frac{\partial w_t}{\partial n} + \left[v_4(\eta) + \gamma\frac{\partial w_{tt}}{\partial n}\right]w_t\right\}d\Gamma. \end{aligned} \tag{1.27}$$

For simplicity, we set

$$\breve{\zeta} = \left(\langle U, n \rangle, \langle U, \tau \rangle, \frac{\partial u}{\partial n}, \frac{\partial u}{\partial \tau}, u \right) \tag{1.28}$$

for any $\zeta = (U, u)$. In this paper, we shall consider feedback laws to be defined by

$$\begin{cases} v_i(\eta) = \mathcal{J}_i(\eta_t), & i = 1, 2, 3, \\ v_4(\eta) + \gamma \dfrac{\partial w_{tt}}{\partial n} = \mathcal{J}_4(\eta_t), \end{cases} \tag{1.29}$$

where the feedback operators $\mathcal{J}_i$ are given by

$$\begin{cases} \mathcal{J}_i(\zeta) = -\breve{\zeta} F_i^\tau, & i = 1, 2, 3, \\ \mathcal{J}_4(\zeta) = -\breve{\zeta} F_5^\tau + \dfrac{\partial}{\partial \tau}(\breve{\zeta} F_4^\tau), \end{cases} \tag{1.30}$$

$F_i = (f_{i1}, f_{i2}, f_{i3}, f_{i4}, f_{i5})$ for $1 \le i \le 5$, and $\zeta = (U, u)$. In the formula (1.30) the superscript τ denotes a transpose, f_{ij} are real $L^\infty(\Gamma_1)$ functions, and the matrix $F = (F_1^\tau, F_2^\tau, F_3^\tau, F_4^\tau, F_5^\tau)$ satisfies

$$\text{F is symmetric and positive semidefinite on } \Gamma_1. \tag{1.31}$$

If we put the feedback laws of the formulas (1.29) and (1.30) into the formula (1.27), by the assumption (1.31) we obtain

$$\begin{aligned} \frac{d}{dt} E(t) &= \int_{\Gamma_1} \left[-\breve{\eta}_t F \breve{\eta}_t{}^\tau + \frac{\partial w_t}{\partial \tau} F_4 \breve{\eta}_t{}^\tau + w_t \frac{\partial}{\partial \tau} F_4 \breve{\eta}_t{}^\tau \right] d\Gamma \\ &= - \int_{\Gamma_1} \breve{\eta}_t F \breve{\eta}_t{}^\tau d\Gamma \le 0 \end{aligned} \tag{1.32}$$

so that the resulting closed-loop system under the feedback laws of (1.29) and (1.30) is dissipative in the sense that $E(t)$ is nonincreasing.

Remark 1.3. When the tangent component $W = 0$, the feedback laws of (1.29) and (1.30) are what Lagnese [8] presented for the uniform stabilization of the Kirchhoff plate.

We now set up some geometric conditions on the middle surface of the shallow shell which are necessary to get the energy decay.

Assumption (H.1). There is a constant λ_0 such that

$$\lambda_0 \mathcal{B}(\eta, \eta) \ge \|DW\|^2_{L^2(\Omega, T^2)} + \gamma \|D^2 w\|^2_{L^2(\Omega, T^2)} \tag{1.33}$$

for $\eta = (W, w) \in H^1(\Omega, \Lambda) \times H^2(\Omega)$.

Assumption (H.2). There is a vector field $V \in \mathcal{X}(M)$ such that

$$DV(X, X) = b(x)|X|^2, \quad X \in M_x, \quad x \in \overline{\Omega}, \tag{1.34}$$

where b is a function on $\overline{\Omega}$. Set

$$a(x) = \frac{1}{2} \langle DV, \mathcal{E} \rangle_{T^2_x}, \qquad x \in \overline{\Omega},$$

where $\mathcal{E}$ is the volume element of M. Moreover, suppose that b and a meet inequality

$$2\min_{x\in\overline{\Omega}} b(x) > \lambda_0(1+\mu)\max_{x\in\overline{\Omega}}|a(x)|. \tag{1.35}$$

Assumption (H.3). Γ_0 and Γ_1 satisfy the following conditions:

$$\overline{\Gamma}_0 \neq \emptyset, \quad \overline{\Gamma}_0 \cap \overline{\Gamma}_1 = \emptyset, \quad \text{and } V(x)\cdot n(x) \leq 0 \quad \text{on } \Gamma_0. \tag{1.36}$$

Assumption (H.4). $F \in C^1(\Gamma_1)$ is a positive definite matrix.

Remark 1.4. The assumptions (H.1)–(H.3) are geometric conditions on the middle surface of the shell, while the assumption (H.4) is on the feedback. For a plate the assumptions (H.1)–(H.2) automatically satisfy, where we set $V = x - x_0$. For the general case, the assumptions (H.1)–(H.2) can be verified by the geometry method; see, for example, Yao [21]. Here the geometric assumption (H.3) is, generally, considered to be much weaker than the following:

$$V(x)\cdot n(x) \leq 0 \quad \text{on} \quad \Gamma_0 \quad \text{and} \quad V(x)\cdot n(x) > 0 \quad \text{on} \quad \Gamma_1,$$

which is used to avoid the complex trace estimates.

We are now in a position to state our main results.

THEOREM 1.1. *Assume that the assumptions* (H.1)–(H.4) *hold. Let the energy* $E(t)$ *be defined by* (1.26) *for the closed-loop system* (1.21), (1.22), *and* (1.29). *Then there are positive constants* K *and* ω *such that*

$$E(t) \leq Ke^{-\omega t}E(0), \qquad t \geq 0, \tag{1.37}$$

for any $\eta^0 \in H^1_{\Gamma_0}(\Omega,\Lambda)\times H^2_{\Gamma_0}(\Omega)$ *and any* $\eta^1 \in L^2(\Omega,\Lambda)\times H^1_{\Gamma_0}(\Omega)$.

2. Existence, uniqueness, and properties of solutions. In this section, we follow the ideas in Lagnese [8] for the Kirchhoff plate to put the shallow shell problem into a semigroup frame. Then the regularity of solutions we need for the stabilization is worked out by Agmon, Douglis, and Nirenberg [1].

2.1. Variational formulation. We shall set

$$\mathcal{W} = H^1_{\Gamma_0}(\Omega,\Lambda)\times H^2_{\Gamma_0}(\Omega), \quad \mathcal{V} = L^2(\Omega,\Lambda)\times H^1_{\Gamma_0}(\Omega), \quad \text{and} \quad \mathcal{L} = L^2(\Omega,\Lambda)\times L^2(\Omega). \tag{2.1}$$

Introduce the forms

$$a_0(\eta,\zeta) = \int_\Omega [\langle \eta,\zeta\rangle + \gamma\langle Dw, Du\rangle]dx \tag{2.2}$$

and

$$a_1(\eta,\zeta) = \int_{\Gamma_1} \breve{\eta} F \breve{\zeta}^\tau d\Gamma \tag{2.3}$$

for $\eta = (W,w)$ and $\zeta = (U,u)$. It follows from Green's formula (1.15) that an appropriate variational formulation of the systems (1.21), (1.22), and (1.29) is as follows: Find a vector field $\eta \in C([0,\infty);\mathcal{W})\cap C^1([0,\infty);\mathcal{V})$ such that

$$\begin{cases} \dfrac{d}{dt}[a_0(\eta_t,\zeta) + a_1(\eta,\zeta)] + \mathcal{B}(\eta,\zeta) = 0 \qquad \forall\, \zeta \in \mathcal{W}, \\ \eta(0) = \eta^0 \in \mathcal{W}, \quad \eta_t(0) = \eta^1 \in \mathcal{V}. \end{cases} \tag{2.4}$$

2.2. Well-posedness of (1.21), (1.22), and (1.29). The bilinear forms $a_0(\cdot,\cdot)$, $a_1(\cdot,\cdot)$, and $\mathcal{B}(\cdot,\cdot)$ are continuous, symmetric, and nonnegative on $\mathcal{V}$ and $\mathcal{W}$, respectively, and if we set

$$a_0(\eta) = a_0(\eta,\eta) \quad \text{and} \quad a_1(\eta) = a_1(\eta,\eta), \tag{2.5}$$

then we have

$$a_0(\eta) \geq \gamma \|\eta\|^2_{L^2(\Omega,\Lambda)\times H^1(\Omega)} \quad \text{and} \quad a_1(\eta) \geq 0. \tag{2.6}$$

The form $a_0(\cdot,\cdot)$ defines a scalar product on $\mathcal{V}$ and so does $\mathcal{B}(\cdot,\cdot)$ on $\mathcal{W}$ because of the ellipticity (1.33). Those scalar products are equivalent to the ones previously introduced in those spaces. We identify $\mathcal{L}$ with its dual $\mathcal{L}'$ so that we have the dense and continuous embeddings

$$\mathcal{W} \subset \mathcal{V} \subset \mathcal{L} \subset \mathcal{V}' \subset \mathcal{W}'. \tag{2.7}$$

Let A_0 (respectively, P) denote the canonical isomorphism of $\mathcal{V}$ (respectively, $\mathcal{W}$) endowed with the scalar product $a_0(\cdot,\cdot)$ (respectively, $\mathcal{B}(\cdot,\cdot)$) onto $\mathcal{V}'$ (respectively, $\mathcal{W}'$). Then

$$a_0(\eta,\zeta) = \langle A_0\eta,\zeta\rangle \qquad \forall \eta, \quad \zeta \in \mathcal{V},$$

$$\mathcal{B}(\eta,\zeta) = \langle P\eta,\zeta\rangle \qquad \forall \eta, \quad \zeta \in \mathcal{W},$$

where $\langle\cdot,\cdot\rangle$ refers to $(\cdot,\cdot)_{L^2(\Omega,\Lambda)\times L^2(\Omega)}$. Furthermore, there is a nonnegative operator $A_1 \in \mathcal{B}(\mathcal{W},\mathcal{W}')$ such that

$$a_1(\eta,\zeta) = \langle A_1\eta,\zeta\rangle \qquad \forall \eta, \quad \zeta \in \mathcal{W}.$$

We write (2.4) as

$$\frac{d}{dt}(A_0\eta_t + A_1\eta) + P\eta = 0 \quad \text{in} \quad \mathcal{W}'. \tag{2.8}$$

Let us formally rewrite (2.4) as the system

$$\begin{pmatrix} P & 0 \\ 0 & A_0 \end{pmatrix} \begin{pmatrix} \eta \\ \eta_t \end{pmatrix}' + \begin{pmatrix} 0 & -P \\ P & A_1 \end{pmatrix} \begin{pmatrix} \eta \\ \eta_t \end{pmatrix} = 0$$

or

$$\mathbb{C}Y' + \mathcal{Q}Y = 0, \qquad t \geq 0, \tag{2.9}$$

where

$$\mathbb{C} = \begin{pmatrix} P & 0 \\ 0 & A_0 \end{pmatrix}, \quad \mathcal{Q} = \begin{pmatrix} 0 & -P \\ P & A_1 \end{pmatrix}, \quad \text{and} \quad Y = \begin{pmatrix} \eta \\ \eta_t \end{pmatrix}.$$

We wish to solve (2.9) in the space $\mathcal{W}\times\mathcal{V}$. In order to make sense of (2.9) in that space it is natural to introduce

$$\mathcal{D}(\mathcal{Q}) = \{(\eta,\zeta) | \eta \in \mathcal{W}, \zeta \in \mathcal{W}, P\eta + A_1\zeta \in \mathcal{V}'\}.$$

Then $\mathcal{Q} : \mathcal{D}(\mathcal{Q}) \to \mathcal{W}' \times \mathcal{V}'$. Since $\mathbb{C}$ is the canonical isomorphism of $\mathcal{W} \times \mathcal{V}$ onto $\mathcal{W}' \times \mathcal{V}'$, we rewrite (2.9) in the form

$$Y' + \mathbb{C}^{-1}\mathcal{Q}Y = 0 \quad \text{in } \mathcal{W} \times \mathcal{V}. \tag{2.10}$$

Solutions of the system (1.21), (1.22), and (1.29) are therefore defined via (2.10).

THEOREM 2.1. *$-\mathbb{C}^{-1}\mathcal{Q}$ is the infinitesimal generator of a C_0-semigroup of contraction on $\mathcal{W} \times \mathcal{V}$.*

Proof. (i) $\mathcal{D}(\mathcal{Q})$ is dense in $\mathcal{W} \times \mathcal{V}$.

By the definition of P and A_1, for $\varsigma = (U, u) \in \mathcal{W}$, we obtain

$$\begin{aligned}
\langle P\eta + A_1\zeta, \varsigma\rangle &= \mathcal{B}(\eta, \varsigma) + a_1(\zeta, \varsigma) \\
&= -\int_\Omega \langle \Delta_\mu W + (1-\mu)kW + \mathcal{F}(w),\ U\rangle dx - \gamma \int_\Omega \langle d\Delta w, du\rangle dx \\
&\quad + \int_\Omega [-(1-\mu)\delta(kdw) + \mathcal{G}(W) + (H^2 - 2(1-\mu)k)w]u dx \\
&\quad + \int_{\Gamma_1} \left[v_1(\eta)\langle U, n\rangle + v_2(\eta)\langle U, \tau\rangle + v_3(\eta)\frac{\partial u}{\partial n}\right] d\Gamma \\
&\quad - (1-\mu)\gamma \int_{\Gamma_1} \left[\frac{\partial}{\partial \tau} D^2 w(n, \tau) + k\frac{\partial w}{\partial n}\right] u\, d\Gamma \\
&\quad - \int_{\Gamma_1} \left[\mathcal{J}_1(\zeta)\langle U, n\rangle + \mathcal{J}_2(\zeta)\langle U, \tau\rangle + \mathcal{J}_3(\zeta)\frac{\partial u}{\partial n} + \mathcal{J}_4(\zeta)u\right] d\Gamma.
\end{aligned} \tag{2.11}$$

The expression on the right-hand side of the formula (2.11) implies the relation

$$\begin{aligned}
\mathcal{D}(\mathcal{Q}) \supset \mathcal{D}_0 = \{\, (\eta, \zeta) \,|\, \eta \in \mathcal{W} \cap H^2(\Omega, \Lambda) \times H^4(\Omega),\ \zeta \in \mathcal{W},\ v_1(\eta) = \mathcal{J}_1(\zeta), \\
v_2(\eta) = \mathcal{J}_2(\zeta),\ \text{and } v_3(\eta) = \mathcal{J}_3(\zeta) \text{ on } \Gamma_1 \,\}.
\end{aligned}$$

Indeed, if $(\eta, \zeta) \in \mathcal{D}_0$, then

$$\begin{aligned}
|\langle P\eta + A_1\zeta, \varsigma\rangle| &\le C(\|W\|_{H^2(\Omega,\Lambda)} + \|w\|_{H^3(\Omega)})(\|U\|_{L^2(\Omega,\Lambda)} + \|u\|_{H^1(\Omega)}) \\
&\quad + C\left\| \mathcal{J}_4(\zeta) + (1-\mu)\gamma \left[\frac{\partial}{\partial \tau} D^2 w(n, \tau) + k\frac{\partial w}{\partial n}\right] \right\|_{H^{-\frac{1}{2}}(\Gamma_1)} \|u\|_{H^{\frac{1}{2}}(\Gamma_1)} \\
&\le C_{\eta,\zeta}\|\varsigma\|_{\mathcal{V}},
\end{aligned}$$

that is,

$$P\eta + A_1\zeta \in \mathcal{V}'.$$

We mention that in the above inequality the following result is used: $\eta \in \mathcal{W} \cap H^2(\Omega, \Lambda) \times H^4(\Omega)$ and $\zeta \in \mathcal{W}$ imply $\mathcal{J}_4(\zeta) + (1-\mu)\gamma[\frac{\partial}{\partial \tau} D^2 w(n, \tau) + k\frac{\partial w}{\partial n}] \in H^{-\frac{1}{2}}(\Gamma_1)$.

$\mathcal{D}(\mathcal{Q})$ is then dense in $\mathcal{W} \times \mathcal{V}$ since $\mathcal{D}_0$ is in $\mathcal{W} \times \mathcal{V}$.

(ii) $-\mathbb{C}^{-1}\mathcal{Q}$ is dissipative. This is shown by

$$\begin{aligned}
\langle \mathbb{C}^{-1}\mathcal{Q}(\eta, \zeta), (\eta, \zeta)\rangle &= \langle (-\zeta, A_0^{-1}(P\eta + A_1\zeta)), (\eta, \zeta)\rangle \\
&= -\mathcal{B}(\zeta, \eta) + \mathcal{B}(\eta, \zeta) + a_1(\zeta, \zeta) \\
&= a_1(\zeta, \zeta) \ge 0
\end{aligned} \tag{2.12}$$

for $(\eta, \zeta) \in \mathcal{D}(\mathcal{Q})$.

(iii) We also have $\text{Range}(\lambda I + \mathbb{C}^{-1}\mathcal{Q}) = \mathcal{W} \times \mathcal{V}$, for $\lambda > 0$, $(\eta, \zeta) \in \mathcal{D}(\mathcal{Q})$. In fact, this is equivalent to

$$\text{Range}(\lambda^2 A_0 + \lambda A_1 + P) = \mathcal{V}'.$$

But, by the Lax–Milgram theorem, it is actually true. □

As a consequence of Theorem 2.1, we have the following result.

THEOREM 2.2. *Assume that* (1.29) *and* (1.31) *hold and*

$$\eta^0 \in \mathcal{W}, \quad \eta^1 \in \mathcal{W}, \quad P\eta^0 + A_1\eta^1 \in \mathcal{V}'. \tag{2.13}$$

Then the problem (2.4) *admits a unique solution with*

$$\eta \in C^1([0,\infty); \mathcal{W}) \cap C^2([0,\infty); \mathcal{V}),$$

$$\eta_{tt} \in C([0,\infty); \mathcal{V}),$$

$$A_0\eta_{tt} + A_1\eta_t + P\eta = 0, \qquad t \geq 0, \tag{2.14}$$

$$\eta(0) = \eta^0, \qquad \eta_t(0) = \eta^1.$$

2.3. Regularity of solutions. In the following, we shall use local coordinate systems to obtain the regularity of variational solutions to the system (1.21), (1.22), and (1.29). Let us consider the system

$$\begin{cases} \eta \in \mathcal{W}, \quad \mathcal{A}\eta \in \mathcal{V}', \\ v_i(\eta) \in H^{\frac{1}{2}}(\Gamma_1), \qquad i = 1,2,3, \\ v_4(\eta) \in H^{-\frac{1}{2}}(\Gamma_1). \end{cases} \tag{2.15}$$

First, we have the following lemma.

LEMMA 2.3. *Let* η *satisfy the problem* (2.15). *Then*

$$\eta \in H^2(\Omega, \Lambda) \times H^3(\Omega) \cap \mathcal{W}. \tag{2.16}$$

Proof. First, we prove that, for any $\varphi \in C^\infty(\overline{\Omega})$, $\varphi\eta$ still satisfies the problem (2.15), so our analysis of η on $\overline{\Omega}$ can be localized.

Note that

$$\mathcal{A}(\varphi\eta) = \varphi\mathcal{A}\eta + [\mathcal{A}, \varphi]\eta, \tag{2.17}$$

where the commutator $[\mathcal{A}, \varphi]$ is a first-order differential operator on the component W and a third-order differential operator on the component w. Then the hypothesis $\mathcal{A}\eta \in \mathcal{V}'$, together with $\eta \in \mathcal{W}$, gives $\mathcal{A}(\varphi\eta) \in \mathcal{V}'$. Similarly, it is easy to check from the formulas in (1.20) that when η satisfies the problem (2.15) all the boundary conditions in the problem (2.15) for $\varphi\eta$ are still true.

So suppose η, satisfying (2.15), is supported on a coordinate chart $(\mathcal{U}, \psi)$ where $(\mathcal{U}, \psi)$ is chosen in such a way that there exists a positive smooth function Θ on $\mathcal{U}$ to meet

$$g = \Theta(dx_1^2 + dx_2^2) \quad \text{on} \quad \mathcal{U}, \tag{2.18}$$

where g is the induced metric of the Riemannian manifold M; see Wu [19]. It is noticeable that the expression in (2.18) does not hold in general when the dimension of the manifold is larger than 2.

In addition, the formulas in (1.19) and the hypothesis $\eta \in \mathcal{W}$ imply

$$\mathcal{F}(w) \in H^1_{\Gamma_0}(\Omega) \quad \text{and} \quad \mathcal{G}(W) \in L^2(\Omega, \Lambda). \tag{2.19}$$

Set

$$W = u_1 \frac{\partial}{\partial x_1} + u_2 \frac{\partial}{\partial x_2}. \tag{2.20}$$

By the relations (2.18)–(2.20) and through a computation, we separate the problem (2.15) into the two following ones:

$$\begin{cases} (u_1, u_2) \in H^1_{\partial O_0}(O), \\ \dfrac{\partial^2 u_1}{\partial x_1^2} + \dfrac{1-\mu}{2}\dfrac{\partial^2 u_1}{\partial x_2^2} + \dfrac{1+\mu}{2}\dfrac{\partial^2 u_2}{\partial x_1 \partial x_2} \in L^2(O), \\ \dfrac{1-\mu}{2}\dfrac{\partial^2 u_2}{\partial x_1^2} + \dfrac{\partial^2 u_2}{\partial x_2^2} + \dfrac{1+\mu}{2}\dfrac{\partial^2 u_1}{\partial x_1 \partial x_2} \in L^2(O), \\ n_1 \dfrac{\partial u_1}{\partial n} + n_2 \dfrac{\partial u_2}{\partial n} - \mu n_2 \dfrac{\partial u_1}{\partial \tau} + \mu n_1 \dfrac{\partial u_2}{\partial \tau} \in H^{\frac{1}{2}}(\partial O_1), \\ -n_2 \dfrac{\partial u_1}{\partial n} + n_1 \dfrac{\partial u_2}{\partial n} + n_1 \dfrac{\partial u_1}{\partial \tau} + n_2 \dfrac{\partial u_2}{\partial \tau} \in H^{\frac{1}{2}}(\partial O_1), \end{cases} \tag{2.21}$$

and

$$\begin{cases} w \in H^2_{\partial O_0}(O), \\ \Delta_0^2 w \in H^{-1}(O), \\ \Delta_0 w - (1-\mu)\Theta \dfrac{\partial^2 w}{\partial \tau^2} \in H^{\frac{1}{2}}(\partial O_1), \\ \dfrac{\partial \Delta_0 w}{\partial n} + (1-\mu)\Theta \left[k \dfrac{\partial w}{\partial n} + \dfrac{\partial^3 w}{\partial \tau^2 \partial n} \right] \in H^{-\frac{1}{2}}(\partial O_1), \end{cases} \tag{2.22}$$

where

$$O = \psi(\mathcal{U} \cap \Omega), \quad \partial O_1 = \psi(\Gamma_1 \cap \mathcal{U}), \quad \partial O_0 = \partial O / \partial O_1,$$

$$n = n_1 \frac{\partial}{\partial x_1} + n_2 \frac{\partial}{\partial x_2}, \quad \tau = -n_2 \frac{\partial}{\partial x_1} + n_1 \frac{\partial}{\partial x_2}, \quad \Delta_0 = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}.$$

It is clear that the problem (2.22) is a classical elliptic boundary value problem since $\overline{\Gamma}_0 \cap \overline{\Gamma}_1 = \emptyset$ and Γ is smooth enough. We therefore obtain

$$w \in H^3(O) \cap H^2_{\partial O_0}(O).$$

Next, since the determinant of coefficients of $\{\partial u_1/\partial n, \partial u_1/\partial n\}$ in the boundary conditions of the problem (2.21) is $1/\Theta > 0$, the classical theory of Agmon, Douglis, and Nirenberg [1] yields

$$(u_1, u_2) \in \left(H^2(O)\right)^2.$$

Finally, the partition of unity subject to a coordinate cover of $\overline{\Omega}$ completes the proof. □

In order to apply Green's formula to solutions of the system (1.21), (1.22), and (1.29), we have to get more regularity than what is given by Theorem 2.2. To this end, we need to require more regularity of the initial data than what is supposed in (2.13). The requisite assumptions are obtained as follows. We introduce

$$\tilde{A}_0 = A_0|_{\mathcal{W}}, \quad \mathcal{H} = \text{Range}\tilde{A}_0, \tag{2.23}$$

and we assume

$$\eta^0 \in \mathcal{W}, \quad \eta^1 \in \mathcal{W} \cap H^2(\Omega,\Lambda) \times H^3(\Omega), \quad \text{and} \quad A_1\eta^1 + P\eta^0 \in \mathcal{H}. \tag{2.24}$$

From (2.14) we have $A_0\eta_{tt}(0) = -(A_1\eta^1 + P\eta^0) \in \mathcal{H}$ and, therefore,

$$\eta_{tt}(0) = -\tilde{A}_0^{-1}(A_1\eta^1 + P\eta^0) \in \mathcal{W}. \tag{2.25}$$

We further assume

$$v_i(\eta^1) = \mathcal{J}_i(\eta_{tt}(0)), \qquad i = 1,2,3, \quad \text{on} \quad \Gamma_1. \tag{2.26}$$

Then we conclude that

$$A_1\eta_{tt}(0) + P\eta^1 \in \mathcal{V}'. \tag{2.27}$$

Indeed, for $\varsigma = (U,u) \in \mathcal{V}$, from (2.3), (1.31), and (2.26) we have

$$a_1(\eta_{tt}(0),\varsigma) = -\int_{\Gamma_1}\left[v_1(\eta^1)\langle U,n\rangle + v_2(\eta^1)\langle U,\tau\rangle + v_3(\eta^1)\frac{\partial u}{\partial n} + \mathcal{J}_4(\eta_{tt}(0))u\right]d\Gamma \tag{2.28}$$

and, therefore, by Green's formula (1.15),

$$\begin{aligned}
&\langle A_1\eta_{tt}(0) + P\eta^1,\, \varsigma\rangle = a_1(\eta_{tt}(0),\varsigma) + \mathcal{B}(\eta^1,\varsigma)\\
&= -\int_\Omega \langle \Delta_\mu W^1 + (1-\mu)kW^1 + \mathcal{F}(w^1),\ U\rangle\, dx - \gamma\int_\Omega \langle d\Delta w^1,\, du\rangle\, dx\\
&\quad + \int_\Omega \gamma[-(1-\mu)\delta(kdw^1) + (H^2 - 2(1-\mu)k)w^1 + \mathcal{G}(W^1)]u\, dx\\
&\quad - \int_{\Gamma_1}\left\{\mathcal{J}_4(\eta_{tt}(0)) + \gamma(1-\mu)\left[\frac{\partial}{\partial\tau}D^2w^1(n,\tau) + k\frac{\partial w^1}{\partial n}\right]\right\}u\, d\Gamma.
\end{aligned} \tag{2.29}$$

Now the expression (2.29) produces

$$\begin{aligned}
&|\langle A_1\eta_{tt}(0) + P\eta^1,\, \varsigma\rangle|\\
&\le C\left(\|W^1\|_{H^2(\Omega,\Lambda)} + \|w^1\|_{H^3(\Omega)}\right.\\
&\quad \left. + \left\|\mathcal{J}_4(\eta_{tt}(0)) + \gamma(1-\mu)\left[\frac{\partial}{\partial\tau}D^2w^1(n,\tau) + k\frac{\partial w^1}{\partial n}\right]\right\|_{H^{-\frac{1}{2}}(\Gamma_1)}\right)\|\varsigma\|_{\mathcal{V}}.
\end{aligned} \tag{2.30}$$

Consequently, (2.27) holds as claimed.

We have shown that if (2.24) and (2.26) hold, then

$$\eta^1 \in \mathcal{W}, \quad \eta_{tt}(0) \in \mathcal{W}, \quad A_1\eta_{tt}(0) + P\eta^1 \in \mathcal{V}', \tag{2.31}$$

that is, $\{\eta^1, \eta_{tt}(0)\} \in \mathcal{D}(\mathbb{C}^{-1}Q)$. It follows from Theorem 2.2 that η satisfies

$$\eta_t \in C^1([0,\infty); \mathcal{W}) \quad \text{and} \quad \eta_{tt} \in C([0,\infty); \mathcal{W}). \tag{2.32}$$

Then, as a consequence of $A_1\eta_t + P\eta = -A_0\eta_{tt}$ we obtain that η satisfies

$$\begin{cases} \eta \in C^2([0,\infty); \mathcal{W}), \\ \mathcal{A}\eta = -(\eta_{tt} - \gamma(0, \Delta w_{tt})) \in C([0,\infty); L^2(\Omega,\Lambda) \times L^2(\Omega)), \\ v_i(\eta) = \mathcal{J}_i(\eta_t) \in H^{\frac{1}{2}}(\Gamma_1), \qquad i = 1,2, \\ v_3(\eta) = \mathcal{J}_3(\eta_t) \in H^{\frac{3}{2}}(\Gamma_1), \\ v_4(\eta) + \gamma \dfrac{\partial w_{tt}}{\partial n} = \mathcal{J}_4(\eta_t) \in H^{\frac{1}{2}}(\Gamma_1). \end{cases} \tag{2.33}$$

Elliptic theory then yields

$$w \in H^4(\Omega).$$

We now write the above analysis into the following result.

THEOREM 2.4. *Assume that* Γ *is smooth enough,* $\overline{\Gamma}_1 \cap \overline{\Gamma}_0 = \emptyset$*, and conditions* (2.24) *and* (2.26) *hold. Then variational solutions of the system* (1.21), (1.22), *and* (1.29) *satisfy*

$$\eta \in C([0,\infty); H^2(\Omega,\Lambda) \times H^4(\Omega) \cap \mathcal{W}) \cap C^1([0,\infty); H^1(\Omega,\Lambda) \times H^3(\Omega) \cap \mathcal{V}). \tag{2.34}$$

Remark 2.1. If $\eta^0 \in H^3(\Omega,\Lambda) \times H^4(\Omega) \cap \mathcal{W}$ and $\eta^1 \in H^2(\Omega,\Lambda) \times H^3(\Omega) \cap \mathcal{W}$, then conditions (2.24) and (2.26) hold.

3. Proof of Theorem 1.1. We assume that the initial data satisfy the assumptions of Theorem 2.4 and, therefore, the solution $\eta = (W, w)$ of the system (1.21), (1.22), and (1.29) meets the regularity of (2.34).

Let a vector field V be given to satisfy the assumption (H.2). Set

$$\eta_1 = (W, 0), \quad \eta_2 = (0, w), \quad m(\eta) = (D_V W, V(w)),$$

$$L(t) = \|W\|^2_{L^2(\Omega,\Lambda)} + \|w\|^2_{L^2(\Omega)} + \gamma\|w_t\|^2_{L^2(\Omega)} + \|Dw\|^2_{L^2(\Omega,\Lambda)},$$

$$\sigma_0 = \max_{x\in\overline{\Omega}} |V|, \qquad \sigma_1 = \min_{x\in\overline{\Omega}} b(x) - \frac{\lambda_0(1+\mu)}{2} \max_{x\in\overline{\Omega}} |a(x)|,$$

$$Q^T = (0,T) \times \Omega, \quad \Sigma^T = (0,T) \times \Gamma, \quad \Sigma_0^T = (0,T) \times \Gamma_0, \quad \text{and} \quad \Sigma_1^T = (0,T) \times \Gamma_1,$$

where $T > 0$ is given.

LEMMA 3.1. *Let the assumptions* (H.1) *and* (H.2) *hold. If we denote*

$$p(\eta) = \left(b - \frac{\sigma_1}{2}\right)\eta_1 - \frac{b}{2}\eta_2, \tag{3.1}$$

then we have

$$\begin{aligned}\sigma_1 \int_0^T E(t)dt \le \frac{1}{2}\int_{\Sigma_1^T} [|\eta_t|^2 + \gamma|Dw_t|^2 - B(\eta,\eta)]\langle V, n\rangle \, d\Sigma \\ + \frac{1}{2}\int_{\Sigma_0^T} B(\eta,\eta)\langle V, n\rangle \, d\Sigma - \int_0^T a_1(\eta_t,\, m(\eta) + p(\eta))\, dt \\ + C_T\left[E(0) + E(T) + \int_0^T L(t)\, dt\right]. \end{aligned} \tag{3.2}$$

Proof. By the embedding theorem there is $C_T > 0$ such that

$$L(0) \le C_T E(0) \quad \text{and} \quad L(T) \le C_T E(T). \tag{3.3}$$

Then, from (3.3), Theorem 1.1 of Yao [21] gives the following inequality:

$$\sigma_1 \int_0^T E(t)\, dt \le (SB)_1|_{\Sigma^T} + (SB)_2|_{\Sigma^T} + C_T\left[E(0) + E(T) + \int_0^T L(t)\, dt\right], \tag{3.4}$$

where

$$(SB)_1|_{\Sigma^T} = \frac{1}{2}\int_{\Sigma^T} [|\eta_t|^2 + \gamma|Dw_t|^2 - B(\eta,\eta)]\langle V, n\rangle \, d\Sigma, \tag{3.5}$$

$$(SB)_2|_{\Sigma^T} = \int_{\Sigma^T}\left[\partial\left(\mathcal{A}\eta,\, m(\eta) + p(\eta)\right) + \gamma\left(V(w) - \frac{1}{2}bw\right)\frac{\partial w_{tt}}{\partial n}\right] d\Sigma. \tag{3.6}$$

Let us examine the integrals over Σ_0^T in (3.5) and (3.6). By the boundary conditions of (1.22) on Γ_0 we have, from Proposition 2.12(ii) of Yao [21],

$$(SB)_1|_{\Sigma_0^T} = -\frac{1}{2}\int_{\Sigma_0^T} B(\eta,\eta)\langle V, n\rangle \, d\Sigma \quad \text{and} \quad (SB)_2|_{\Sigma_0^T} = \int_{\Sigma_0^T} B(\eta,\eta,)\langle V, n\rangle \, d\Sigma. \tag{3.7}$$

First, we consider the calculation of $(SB)_2|_{\Sigma_1^T}$.

From the formula (1.16) and the feedback law of (1.29) and (1.30) we obtain, for any $\varsigma = (U, u) \in L^2(\Omega, \Lambda) \times H^1(\Omega)$,

$$\begin{aligned}\int_{\Gamma_1}\left[\partial(\mathcal{A}\eta,\, \varsigma) + \gamma u\frac{\partial w_{tt}}{\partial n}\right] d\Gamma &= -a_1(\eta_t,\, \varsigma) + \int_{\Gamma_1}\left[u\frac{\partial}{\partial\tau}(\breve{\eta}_t F_4^\tau) + \breve{\eta}_t F_4^\tau\frac{\partial u}{\partial\tau}\right] d\Gamma \\ &= -a_1(\eta_t,\, \varsigma)\end{aligned} \tag{3.8}$$

since $\overline{\Gamma}_0 \cap \overline{\Gamma}_1 = \emptyset$. By applying the formula (3.8) to the expression (3.6) with $\varsigma = m(\eta) + p(\eta)$, one finds out

$$(SB)_2|_{\Sigma_1^T} = -\int_0^T a_1(\eta_t,\, m(\eta) + p(\eta))\, dt. \tag{3.9}$$

Substituting (3.7) and (3.9) into (3.4) yields the inequality (3.2). □

Next, we consider a trace result which is a consequence of Horn [6]. Let U be a vector field on $\overline{\Omega}$. Set

$$S(U) = \frac{1}{2}(DU + D^*U), \tag{3.10}$$

a two-order tensor on $\overline{\Omega}$.

LEMMA 3.2. *Let $T > 0$ and $1/2 > \epsilon > 0$ be given. Suppose that a vector field U satisfies the problem*

$$\begin{cases} U \in L^2([0,T]; H^{\frac{1}{2}+\epsilon}(\Omega,\Lambda)), \quad U_t \in L^2(\Sigma^T,\Lambda), \\ U_{tt} - \Delta_\mu U \in H^{-\frac{1}{2}}(Q^T,\Lambda), \\ (1-\mu)S(U)(n,n) - \mu\delta U \in L^2(\Sigma^T,\Lambda), \\ S(U)(n,\tau) \in L^2(\Sigma^T,\Lambda). \end{cases} \tag{3.11}$$

Then, for $T/2 > \alpha > 0$, there is $C_{T,\alpha,\epsilon} > 0$ such that

$$\begin{aligned} &\|D_\tau U\|^2_{L^2([\alpha,\,T-\alpha];L^2(\Gamma,\Lambda))} \\ &\le C_{T,\alpha,\epsilon}(\|U_t\|^2_{L^2(\Sigma^T,\Lambda)} + \|(1-\mu)S(U)(n,n) - \mu\delta U\|^2_{L^2(\Sigma^T)} + \|S(U)(n,\tau)\|^2_{L^2(\Sigma^T)}) \\ &\quad + C_{T,\alpha,\epsilon}(\|W_{tt} - \Delta_\mu W\|^2_{H^{-1/2}(Q^T,\Lambda)} + \|U\|^2_{L^2([0,T];H^{1/2+\epsilon}(\Omega,\Lambda))}). \end{aligned} \tag{3.12}$$

Proof. It is easy to check that, if U satisfies the problem (3.11), then, for any $\varphi \in C^\infty(\overline{\Omega})$, φU still does, which means the problem (3.11) can be localized. Suppose that U is supported in a coordinate chart $(\mathcal{U},\phi)$ with the metric g of (2.18). Set

$$U = u_1\frac{\partial}{\partial x_1} + u_2\frac{\partial}{\partial x_2}.$$

On the chart $(\mathcal{U},\phi)$, the problem (3.11) changes into the problem

$$\begin{cases} u_i \in L^2([0,T]; H^{1/2+\epsilon}(O)), \quad u_{it} \in L^2((0,T)\times\partial O), \qquad i = 1,2, \\ u_{1tt} - \left(\dfrac{\partial^2 u_1}{\partial x_1^2} + \dfrac{1-\mu}{2}\dfrac{\partial^2 u_1}{\partial x_2^2} + \dfrac{1+\mu}{2}\dfrac{\partial^2 u_2}{\partial x_1\partial x_2}\right) \in H^{-1/2}((0,T)\times O), \\ u_{2tt} - \left(\dfrac{1-\mu}{2}\dfrac{\partial^2 u_2}{\partial x_1^2} + \dfrac{\partial^2 u_2}{\partial x_2^2} + \dfrac{1+\mu}{2}\dfrac{\partial^2 u_1}{\partial x_1\partial x_2}\right) \in H^{-1/2}((0,T)\times O), \\ f_1, f_2 \in L^2([0,T]\times\partial O), \end{cases} \tag{3.13}$$

where

$$\begin{cases} f_1 = n_1\dfrac{\partial u_1}{\partial n} + n_2\dfrac{\partial u_2}{\partial n} - \mu n_2\dfrac{\partial u_1}{\partial\tau} + \mu n_1\dfrac{\partial u_2}{\partial\tau}, \\ f_2 = -n_2\dfrac{\partial u_1}{\partial n} + n_1\dfrac{\partial u_2}{\partial n} + n_1\dfrac{\partial u_1}{\partial\tau} + n_2\dfrac{\partial u_2}{\partial\tau}, \end{cases} \tag{3.14}$$

$O = \phi(\mathcal{U})$, and $\partial O = \partial\phi(\mathcal{U})$.

Next we want to show that the problem (3.13) is a special case of some three dimensional dynamic elasticity, so Theorem 1.2 of Horn [6] can be applied. To this end, we set

$$u = (u_1, u_2, 0) \quad \text{and} \quad O' = O\times(0,1).$$

If we let the Lamé coefficients λ and μ in Horn [6] be μ and $\frac{1-\mu}{2}$, respectively, we have

$$\sigma_{ij} = \mu(\mathrm{div}u)\delta_{ij} + \frac{1-\mu}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \tag{3.15}$$

$$\frac{\partial u_i}{\partial x_3} = 0 \tag{3.16}$$

for any $1 \le i, j \le 3$. By the notation of Horn [6], we have

$$\sigma(u)n' = \begin{pmatrix} n_1 & n_2 & 0 \\ -n_2 & n_1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ 0 \end{pmatrix} \tag{3.17}$$

and the problem (3.13) is then equivalent to the three dimensional problem

$$\begin{cases} u \in L^2([0,T];(H^{1/2+\epsilon}(O'))^3), \quad u_t \in (L^2([0,T]\times\partial O'))^3, \\ u_{tt} - \nabla\cdot\sigma(u) \in (H^{-1/2}([0,T]\times O'))^3, \\ \sigma(u)\cdot n' \in (L^2([0,T]\times\partial O'))^3, \end{cases} \tag{3.18}$$

where $n' = (n_1, n_2, 0)$ and $\sigma(u) = (\sigma_{ij})_{3\times 3}$.

Applying Theorem 1.2 of Horn [6] locally and using the partition of unity complete our proof. □

By using the same ideas as in Lemma 3.2 and applying Theorem 2.1 of Lasiecka and Triggiani [11], we get the following lemma.

LEMMA 3.3. *Let $T/2 > \alpha > 0$ and $1/2 > \epsilon > 0$ and $1/2 > s_0$. Suppose that w satisfies the problem*

$$\begin{cases} w_{tt} - \gamma\Delta w_{tt} + \gamma\Delta^2 w \in H^{-s_0}(Q^T), \\ w = \dfrac{\partial w}{\partial n} = 0 \quad on \quad \Sigma_0^T, \\ \Delta w - (1-\mu)D^2 w(\tau,\tau) \in L^2(\Sigma_1^T), \\ \dfrac{\partial \Delta w}{\partial n} + (1-\mu)\left[\dfrac{\partial}{\partial\tau}(D^2 w(\tau,n)) + k(x)\dfrac{\partial w}{\partial n} - \dfrac{\partial w_{tt}}{\partial n}\right] \in H^{-1}(\Sigma_1^T). \end{cases} \tag{3.19}$$

Then there is $C_{\alpha,\epsilon,T}$ such that

$$\begin{aligned} &\|D^2 w\|^2_{L^2([\alpha,T-\alpha];L^2(\Gamma_1,T^2))} \\ &\le C_{\alpha,\epsilon,T}\left\{\|w_{tt} - \gamma\Delta w_{tt} + \gamma\Delta^2 w\|^2_{H^{-s_0}(Q^T)} + \|\Delta w - (1-\mu)D^2 w(\tau,\tau)\|^2_{L^2(\Sigma_1^T)}\right. \\ &\quad + \left\|\frac{\partial\Delta w}{\partial n} + (1-\mu)\left[\frac{\partial}{\partial\tau}(D^2 w(\tau,n)) + k(x)\frac{\partial w}{\partial n} - \frac{\partial w_{tt}}{\partial n}\right]\right\|^2_{H^{-1}(\Sigma_1^T)} + \|Dw_t\|^2_{L^2(\Sigma_1^T,\Lambda)} \\ &\quad \left. + \|w\|^2_{L^2([0,T];H^{3/2+\epsilon}(\Omega))} + \|w_t\|^2_{L^2(\Sigma_1^T)}\right\}. \end{aligned} \tag{3.20}$$

LEMMA 3.4. *Let η be a solution of the system* (1.21), (1.22), *and* (1.29) *with the regularity* (2.34). *Let $T/2 > \alpha > 0$ and $1/2 > \epsilon > 0$ and $1/2 > s_0 > 0$ be given. Then*

$$\int_\alpha^{T-\alpha}\int_{\Gamma_1}(|DW|^2 + |D^2 w|^2)\,d\Sigma \le C_{\alpha,T,s_0,\epsilon}\int_{\Sigma_1^T}(|\eta_t|^2 + |Dw_t|^2)\,d\Sigma + lot(\eta). \tag{3.21}$$

Proof. From the formulas in (1.20), (1.29), and (1.30),

$$\begin{cases} (1-\mu)\Upsilon(\eta)(n,n) + \mu(wH - \delta W) = -\breve{\eta}_t F_1^\tau \quad \text{on} \quad \Gamma_1, \\ (1-\mu)\Upsilon(\eta)(n,\tau) = -\breve{\eta}_t F_2^\tau \quad \text{on} \quad \Gamma_1, \end{cases} \tag{3.22}$$

and we obtain

$$\begin{cases} DW(n,n) = -\mu DW(\tau,\tau) - w[(1-\mu)\Pi(n,n) + \mu H] - \breve{\eta}_t F_1^\tau, \\ DW(\tau,n) = -DW(n,\tau) - 2w\Pi(n,\tau) - \dfrac{2}{1-\mu}\breve{\eta}_t F_2^\tau. \end{cases} \tag{3.23}$$

It follows from (3.23) that

$$\begin{aligned} |DW|^2 &= |D_\tau W|^2 + [DW(n,n)]^2 + [DW(\tau,n)]^2 \\ &\le C(|D_\tau W|^2 + |\eta_t|^2 + |Dw_t|^2 + w^2) \quad \text{on} \quad \Gamma_1. \end{aligned} \tag{3.24}$$

Next, note $\Upsilon(\eta) = S(W) + w\Pi$ and the formulas (3.22) and, after applying the inequality (3.12) with $U = W$, we have

$$\begin{aligned} &\|D_\tau W\|^2_{L^2([\alpha, T-\alpha];L^2(\Gamma,\Lambda))} \\ &\le C_{T,\alpha,\epsilon}(\|W_t\|^2_{L^2(\Sigma^T,\Lambda)} + \|w_t\|^2_{L^2(\Sigma_1^T)} + \|Dw_t\|^2_{L^2(\Sigma^T,\Lambda)}) + lot(\eta), \end{aligned} \tag{3.25}$$

where the following inequality is used:

$$\begin{aligned} \|W_{tt} - \Delta_\mu W\|^2_{H^{-1/2}(Q^T,\Lambda)} &= \|(1-\mu)kW + \mathcal{F}(w)\|^2_{H^{-1/2}(Q^T,\Lambda)} \\ &\le C\|(1-\mu)kW + \mathcal{F}(w)\|^2_{L^2(Q^T,\Lambda)} = lot(\eta). \end{aligned} \tag{3.26}$$

Let us consider the component w. We now prove that there is a constant $C > 0$ such that

$$\|D^2 w\|^2_{H^{-s_0}(Q^T,T^2)} \le C\int_0^T \|w\|^2_{H^{2-s_0}(\Omega)}\,dt \qquad \forall\ w \in H^{-s_0}(Q^T). \tag{3.27}$$

For simplicity, we assume that Ω is a coordinate $\mathcal{U}$ with the coordinate system $x = (x_1, x_2)$. Then $Q^T = (0,T) \times \mathcal{U}$. Denote the Fourier transform variable of (t,x) by (ζ_0, ζ). By definition,

$$H^{-s_0}(Q^T) = (H_0^{s_0}(Q))^*, \quad H_0^{s_0}(Q^T) = \{\, w \,|\, w \in H^{s_0}(\mathbb{R}^3),\ \operatorname{supp} w \subset Q^T \,\}.$$

For $u \in H^{-s_0}(Q^T)$ given, we then have

$$\begin{aligned} \|u\|^2_{H^{-s_0}(Q^T)} &= \|(1+|\zeta_0|^2+|\zeta|^2)^{-s_0/2}\hat{u}\|^2_{L^2(\mathbb{R}^3)} \\ &\le \|(1+|\zeta|^2)^{-s_0/2}\hat{u}\|^2_{L^2(\mathbb{R}^3)} \\ &= \int_{\mathbb{R}^2}(1+|\zeta|^2)^{-s_0}\left(\int_{\mathbb{R}}|\hat{u}|^2 d\zeta_0\right)d\zeta_1\zeta_2 \\ &= \int_{\mathbb{R}^2}(1+|\zeta|^2)^{-s_0}\left(\int_0^T|\hat{u}^x|^2 dt\right)d\zeta_1\zeta_2 \\ &= \int_0^T \|u\|^2_{H^{-s_0}(\Omega)}dt \end{aligned} \tag{3.28}$$

since $\operatorname{supp} u \subset Q^T$, where $\hat{u}^x$ denotes the Fourier transform of u with respect to the variable x. In addition, for $i = 1, 2$,

$$\begin{aligned} \left\|\frac{\partial^2 w}{\partial x_i \partial x_j}\right\|^2_{H^{-s_0}(\Omega)} &= \left\|(1+|\zeta|^2)^{-\frac{s_0}{2}}\frac{\partial^2 \hat{w}}{\partial x_i \partial x_j}\right\|^2 \\ &\le \|(1+|\zeta|^2)^{\frac{2-s_0}{2}}\hat{w}\|^2 = \|w\|^2_{H^{2-s_0}(\Omega)}. \end{aligned} \tag{3.29}$$

The inequality (3.27) follows from the inequalities (3.28) and (3.29). The same argument leads us to the following:

$$\|DW\|^2_{H^{-s_0}(Q^T,T^2)} \le C\int_0^T \|W\|^2_{H^{1-s_0}(\Omega,\Lambda)}\,dt, \tag{3.30}$$

and

$$\left\|\frac{\partial}{\partial\tau}(\breve{\eta}_t F_5^\tau)\right\|^2_{H^{-1}(Q^T)} \le C\|\breve{\eta}_t F_5^\tau\|^2_{L^2(Q^T)}. \tag{3.31}$$

Apply the inequality (3.20), together with the inequalities (3.27), (3.30), and (3.31), to obtain

$$\|D^2 w\|^2_{L^2([\alpha,T-\alpha];L^2(\Gamma_1,T^2))} \le C_{\alpha,\epsilon,T}\int_{\Sigma_1^T}(|\eta_t|^2+|Dw_t|^2)\,d\Sigma + lot(\eta). \tag{3.32}$$

The inequality (3.21) follows from the inequalities (3.24), (3.25), and (3.32). □

Proof of Theorem 1.1. First, we make some estimates for the terms in the right-hand side of the inequality (3.2).

Let s_1, $s_2 > 0$ be such that

$$s_1|\nu|^2 \le \nu F\nu^\tau \le s_2|\nu|^2 \qquad \forall\ \nu\in\mathbb{R}^5. \tag{3.33}$$

Then the definition (2.3) of $a_1(\cdot,\cdot)$ gives, for $\varsigma = (U,u)$,

$$s_1\int_{\Gamma_1}(|\varsigma|^2+|Du|^2)\,d\Gamma \le a_1(\varsigma,\varsigma) \le s_2\int_{\Gamma_1}(|\varsigma|^2+|Du|^2)\,d\Gamma. \tag{3.34}$$

Use of the right-hand side of the inequality (3.34), therefore, yields

$$\begin{aligned}
&|a_1(\eta_t,\, m(\eta)+p(\eta))|\\
&\le [a_1(\eta_t,\eta_t)]^{1/2}[a_1(m(\eta)+p(\eta),\ m(\eta)+p(\eta))]^{1/2}\\
&\le C\int_{\Gamma_1}(|\eta_t|^2+|Dw_t|^2+|m(\eta)+p(\eta)|^2+|D(V(w)-b/2w)|^2)\,d\Gamma\\
&\le C\int_{\Gamma_1}[|\eta_t|^2+|Dw_t|^2+|DW|^2+|W|^2+|D^2w|^2+|Dw|^2+w^2]\,\Gamma\\
&= C\int_{\Gamma_1}(|\eta_t|^2+|Dw_t|^2+|DW|^2+|D^2w|^2)\,d\Gamma + l(\eta).
\end{aligned} \tag{3.35}$$

In addition, we have

$$B(\eta,\eta) \le C(|DW|^2+|D^2w|^2+w^2) \quad \text{on} \quad \Gamma_1 \tag{3.36}$$

and, from the geometrical condition (H.3),

$$\int_{\Sigma_0^T} B(\eta,\eta)\langle V,n\rangle\,d\Sigma \le 0. \tag{3.37}$$

Now we substitute the inequalities (3.35)–(3.37) into the inequality (3.2) to obtain

$$\begin{aligned}
\sigma_1\int_0^T E(t)\,dt \le C\int_{\Sigma_1^T}(|\eta_t|^2+|Dw_t|^2+|DW|^2+|D^2w|^2)\,d\Sigma\\
+ C_T(E(0)+E(T)) + lot(\eta).
\end{aligned} \tag{3.38}$$

Next, change the integral domain Σ_1^T into $[\alpha, T-\alpha]\times\Gamma_1$ in both sides of the inequalities (3.38) and use the inequality (3.21) to give

(3.39)
$$\sigma_1\int_\alpha^{T-\alpha} E(t)\,dt \le C_T\left\{E(\alpha)+E(T-\alpha)+\int_{\Sigma_1^T}(|\eta_t|^2+|Dw_t|^2)\,d\Sigma\right\}+lot(\eta).$$

Note the relation $E'(t)=-a_1(\eta_t,\eta_t)$ and the right-hand side of the inequality (3.34) again, and we find, for any $T>\beta>0$,

$$E(\beta)=E(T)+\int_\beta^T a_1(\eta_t,\eta_t)\,dt \le E(T)+C\int_{\Sigma_1^T}(|\eta_t|^2+|Dw_t|^2)\,d\Sigma. \tag{3.40}$$

Using the inequality (3.40) in the inequality (3.39), we obtain, for $T>0$ suitably large,

$$E(T)\le C_T\int_{\Sigma^T}(|\eta_t|^2+|Dw_t|^2)\,d\Sigma+lot(\eta). \tag{3.41}$$

By the compactness and uniqueness (Proposition 2.13 of Yao [21]) approach, we now have

$$E(T)\le C_T\int_{\Sigma_1^T}(|\eta_t|^2+|Dw_t|^2)\,d\Sigma. \tag{3.42}$$

Finally, using the inequality (3.42) and the left-hand side of the inequality (3.34) leads to

$$E(T)\le C_T\int_0^T a_1(\eta_t,\eta_t)\,dt = C_T(E(0)-E(T)), \tag{3.43}$$

that is,

$$E(T)\le \frac{C_T}{1+C_T}E(0). \tag{3.44}$$

Theorem 1.1 follows from the inequality (3.44). □

Acknowledgments. The authors would like to thank Professors I. Lasiecka and R. Triggiani for picking up two mistakes in the first version of this paper and for some advice.

REFERENCES

[1] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for the solutions of elliptic partial equations satisfying general boundary conditions* II, Comm. Pure Appl. Math., 17 (1964), pp. 35–92.

[2] S. Chai and P. F. Yao, *The observability inequalities for the thin shell*, Sci. China Ser. A, 32 (2002), pp. 1006–1017 (in Chinese).

[3] G. Chen, M. P. Coleman, and K. Liu, *Boundary stabilization of Donell's shallow circular cylindrical shell*, J. Sound Vibration, 209 (1998), pp. 265–298.

[4] P. G. Ciarlet, *Mathematical Elasticity, Vol.* II, Stud. Math. Appl. 27, North–Holland, Amsterdam, 1997.

[5] E. Hebey, *Sobolev Spaces on Riemannian Manifolds*, Lecture Notes in Math. 1635, Springer–Verlag, Berlin, Heidelberg, 1996.

[6] M. A. Horn, *Sharp trace regularity for the solutions of the equations of dynamic elasticity*, J. Math. Systems Estimation Control, 8 (1998), pp. 1–11.
[7] W. T. Koiter, *A consistent first approximation in the general theory of thin elastic shells*, in Proceedings of the IUTAM Symposium on the Theory of Thin Shells, Delft (August 1959), North–Holland, Amsterdam, 1960, pp. 12–33.
[8] J. E. Lagnese, *Boundary Stabilization of Thin Plates*, SIAM Stud. Appl. Math. 10, SIAM, Philadelphia, 1989.
[9] J. Lagnese and J. L. Lions, *Modelling Analysis and Control of Thin Plates*, Rech. Math. Appl. 6, Masson, Paris, 1988.
[10] I. Lasiecka and R. Triggiani, *Uniform stabilization of the wave equation with Dirichlet or Neumann feedback control without geometric conditions*, Appl. Math. Optim., 25 (1992), pp. 189–224.
[11] I. Lasiecka and R. Triggiani, *Sharp trace estimates of solutions to Kichhoff and Euler-Bernoulli equations*, Appl. Math. Optim., 28 (1993), pp. 277–306.
[12] I. Lasiecka, R. Triggiani, and V. Valente, *Uniform stabilization of spherical shells by boundary dissipation*, Adv. Differential Equations, 1 (1996), pp. 635–674.
[13] J. Mason, *Variational, Incremental and Energy Methods in Solid Mechanics and Shell Theory*, Stud. Appl. Mech. 4, Elsevier, Amsterdam, 1980.
[14] F. I. Niordson, *Shell Theory*, North-Holland Series in Applied Mathematics and Mechanics 29, North–Holland, Amsterdam, 1985.
[15] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer–Verlag, New York, Berlin, 1983.
[16] M. E. Taylor, *Partial Differential Equations* I, Springer–Verlag, New York, 1996.
[17] R. Triggiani, *Regularity theory, exact controllability and optimal quadratic cost problem for spherical shells with physical boundary controls*, Control Cybernet., 25 (1996), pp. 553–568.
[18] H. Wu, *Selected Lecture in Riemannian Geometry*, University of Beijing, Beijing, 1981 (in Chinese).
[19] H. Wu, C. L. Shen, and Y. L. Yu, *An Introduction to Riemannian Geometry*, University of Beijing, Beijing, 1989 (in Chinese).
[20] P. F. Yao, *On the Shallow Shell Equation*, preprint, Institute of Systems Science, Chinese Academy of Sciences, Beijing, 1997.
[21] P.-F. Yao, *Observability inequalities for shallow shells*, SIAM J. Control Optim., 38 (2000), pp. 1729–1756.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 260–265

EXISTENCE OF SOLUTIONS TO DIFFERENTIAL INCLUSIONS AND TO TIME OPTIMAL CONTROL PROBLEMS IN THE AUTONOMOUS CASE*

ARRIGO CELLINA† AND ANTÓNIO ORNELAS‡

Abstract. We prove existence of solutions to upper semicontinuous differential inclusions and to time optimal control problems under conditions that are strictly weaker than the usual assumption of convexity.

Key words. differential inclusions, time optimal control problems

AMS subject classifications. 49J15, 34A34, 34A36

PII. S0363012902408046

1. Introduction. The condition of convexity with respect to the variable gradient has been of universal use in the calculus of variations, in optimal control, and in differential inclusions to prove the existence of solutions. In fact, convexity is the property required in order to pass to a weak limit along a sequence, be it a minimizing sequence or a sequence of successive approximations, preserving the properties that are needed. This approach, however, because of its generality, need not always provide the best results, since it does not take into account possible additional information such as, for instance, the presence of symmetries in the problem. One is led to think that, by suitably exploiting these symmetries, the convexity condition could be substantially reduced. The purpose of the present paper is to show that, for the simplest of such symmetries, the time invariance in the problem of the existence of solutions to upper semicontinuous differential inclusions, convexity can be replaced by a strictly weaker condition, our almost convexity, below. Moreover, we show that, in the case of autonomous control systems of the form

$$x'(t) = f(x(t), u(t)), \quad u(t) \in U(x(t))$$

for the existence of a time optimal solution, Filippov's classical assumption of convexity of the images of the map $F(x) = f(x, U(x))$ can be replaced by the weaker assumption of almost convexity of the same images. As will be shown, our assumption does not imply that the set of solutions to the differential inclusion is closed in the space of continuous functions with uniform convergence, as happens in the case of the assumption of convexity, but only that the sections of this set of solutions are closed. This property is sufficient to establish the existence of time optimal solutions.

2. Main results. The following is our assumption of almost convexity.

DEFINITION 1. *Let X be a vector space. A set $K \subset X$ is called* almost convex *if for every $\xi \in coK$ there exist λ_1 and λ_2, $0 \le \lambda_1 \le 1 \le \lambda_2$, such that $\lambda_1 \xi \in K, \lambda_2 \xi \in K$.*

*Received by the editors May 22, 2002; accepted for publication (in revised form) October 23, 2002; published electronically April 17, 2003.

http://www.siam.org/journals/sicon/42-1/40804.html

†Dipartimento di Matematica e Applicazioni, Università di Milano Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy (cellina@matapp.unimib.it).

‡Cima-ue, rua Romão Ramalho 59, P-7000-671 Évora, Portugal (ornelas@uevora.pt). This author was financially supported by INVOTAN from September, 2001, through February, 2002.

Every convex set is almost convex. If a set K is almost convex and $0 \in co(K)$, then $0 \in K$. Typical cases of almost convex sets are $K = \partial C$, with C a convex set not containing the origin, or $K = \{0\} \cup \partial C$, C a convex set containing the origin.

It is our purpose to prove the following theorem.

THEOREM 1. *Let $\Omega \subset R^N$ be open, and let F, from Ω to the nonempty subsets of R^N, be upper semicontinuous with bounded, closed, and almost convex values. Then the Cauchy problem*

$$y'(s) \in F(y(s)), \quad y(0) = x_0 \in \Omega$$

admits a solution defined on some interval $[-\delta, \delta]$, $\delta > 0$. Moreover, for every $\tau \in [-\delta, \delta]$, the attainable set at τ, $A_{x_0}(\tau)$, is closed and coincides with $A^{co}_{x_0}(\tau)$, the attainable set at τ of the convexified problem

$$y'(s) \in coF(y(s)), \quad y(0) = x_0.$$

Remarks. (1) The upper semicontinuous map F, from R to the closed subsets of R, defined by $F(x) = -\text{sign}(x)$ for $x \neq 0$, $F(0) = \{-1, +1\}$, is not almost convex at $x = 0$, and the corresponding Cauchy problem with the condition $y(0) = 0$ admits no local solution.

(2) Under the condition of almost convexity, the attainable sets are closed for every (small) τ, but the set of solutions need not be closed in $C(I)$, unlike in the convex case.

The following corollary to Theorem 1, to be compared with Theorem 1 of Filippov [3], shows that, in the case of autonomous control systems, for the existence of a time optimal solution, Filippov's assumption that the set $f(x, U(x))$ is convex can be replaced by the weaker assumption that the same set is almost convex.

COROLLARY 1. *Let $f(x, u)$ be continuous for $x \in \Omega$ and $u \in U(x)$, and let the set valued map $U(x)$, from Ω to the nonempty compact subsets of R^N, be upper semicontinuous. Moreover, assume that the set*

$$F(x) = f(x, U(x))$$

is almost convex for every $x \in \Omega$. Let x_0 and x_1 be given in Ω, and assume that for some $\tilde{t} \geq 0$, $x_1 \in A_{x_0}(\tilde{t})$. Then the problem of reaching x_1 from x_0 in minimum time admits a solution.

For the proof of Theorem 1 we shall need the following preliminary result.

THEOREM 2. *Let F be upper semicontinuous. Let $x : [a, b] \to R^n$ be a solution to*

$$y'(t) \in co(F(y(t))), \quad y(a) = x_a.$$

Assume that there are two integrable functions $\lambda_1(.), \lambda_2(.)$, from $[a, b]$ to R, satisfying $0 \leq \lambda_1(t) \leq 1 \leq \lambda_2(t)$ and such that, for almost every $t \in [a, b]$, we have $\lambda_1(t)x'(t) \in F(x(t))$ and $\lambda_2(t)x'(t) \in F(x(t))$. Then there exists $t = t(s)$, a nondecreasing absolutely continuous map of the interval $[a, b]$ onto itself, such that the map $\tilde{x}(s) = x(t(s))$ is a solution to

$$y'(s) \in F(y(s)), \quad y(a) = x_a.$$

Moreover, $\tilde{x}(a) = x(a)$ and $\tilde{x}(b) = x(b)$.

Proof. (a) When $x'(t) = 0$, we shall assume, without loss of generality, that $\lambda_2(t) = 1$. Consider the set

$$C = \{t \in I : 0 \in F(x(t))\}.$$

From the continuity of x and the upper semicontinuity of F, we obtain that C is closed. Without loss of generality we shall assume that, for t in C, $\lambda_1(t)x'(t) = 0$.

(b) Let $[\alpha, \beta]$ be an interval, and assume that, on this interval, there exist two functions $\lambda_1(.), \lambda_2(.)$ with the properties stated above. In addition, assume that $\lambda_1(t) > 0$ a.e. We claim that there exist two measurable subsets of $[\alpha, \beta]$, having characteristic functions χ_1 and χ_2 such that $\sum \chi_i = \chi_{[\alpha,\beta]}$, and an absolutely continuous function $s = s(t)$ on $[\alpha, \beta]$, $s(\alpha) - s(\beta) = \alpha - \beta$, such that

$$s'(t) = \chi_1(t)\lambda_1(t) + \chi_2(t)\lambda_2(t).$$

This concludes the proof of this claim. □

Redefine $\lambda_1(t)$ and $\lambda_2(t)$ on a set of measure zero to have both functions positive for every $t \in [\alpha, \beta]$. Set $p(.)$ to be $\frac{1}{2}$ when $\lambda_1(t) = \lambda_2(t) = 1$, to be $\frac{\lambda_2 - 1}{\lambda_2 - \lambda_1}$ otherwise. With this definition we have that $0 \le p(t) \le 1$ and that both equalities

$$1 = p(t) + (1 - p(t))$$

and

$$1 = p(t)\lambda_1(t) + (1 - p(t))\lambda_2(t)$$

hold true. In particular, we have

$$\int_\alpha^\beta 1dt = \int_\alpha^\beta [p(t) + (1 - p(t))]dt = \int_\alpha^\beta \left[\frac{p(t)\lambda_1(t)}{\lambda_1(t)} + \frac{(1 - p(t))\lambda_2(t)}{\lambda_2(t)}\right] dt.$$

We wish to apply Liapunov's theorem on the range of measures, to infer the existence of two measurable subsets having characteristic functions $\chi_1(.), \chi_2(.)$ such that $\sum \chi_i = \chi_{[\alpha,\beta]}$ and with the property that

$$\int_\alpha^\beta 1dt = \int_\alpha^\beta \left[\chi_1(t)\frac{1}{\lambda_1(t)} + \chi_2(t)\frac{1}{\lambda_2(t)}\right] dt.$$

However, it is not obvious that the function $\frac{1}{\lambda_1(t)}$ is integrable, and thus the results of [2] need not be applicable. For this purpose we shall use a device already used in [1]. Consider the sequence of disjoint sets

$$E^n = \left\{ t \in [\alpha, \beta] : n < \frac{1}{\lambda_1(t)} \le n + 1 \right\}.$$

We have that $\cup E^n = [\alpha, \beta]$. Applying Liapunov's theorem to each E^n, we infer the existence of two sequences of measurable subsets E_1^n, E_2^n, having characteristic functions χ_1^n, χ_2^n, such that for every n,

$$\int_{E^n} 1dt = \int_{E^n} \left[\chi_1^n(t)\frac{1}{\lambda_1(t)} + \chi_2^n(t)\frac{1}{\lambda_2(t)}\right] dt.$$

Set $\cup E_1^n = E_1$, $\cup E_2^n = E_2$ and $\chi_1 = \sum \chi_1^n$, $\chi_2 = \sum \chi_2^n$. For each m, the function

$$\sigma^m(t) = \sum_{n=0}^{m} \left[\chi_1^n(t)\frac{1}{\lambda_1(t)} + \chi_2^n(t)\frac{1}{\lambda_2(t)}\right]$$

is positive, and the sequence converges pointwise monotonically to

$$\sigma(t) = \chi_1(t)\frac{1}{\lambda_1(t)} + \chi_2(t)\frac{1}{\lambda_2(t)}.$$

Moreover, the sequence of sets $V^m = (\cup_{n=0}^m E^n)_m$ is monotonically increasing to $[a, b]$, so that $\int_\alpha^\beta 1dt = \lim_m \int_{V^m} 1dt$. Hence

$$\int_\alpha^\beta 1dt = \lim_m \int s^m(t)dt = \int \lim_m s^m(t)dt,$$

so that we obtain

$$\int s(t)dt = \int \left[\chi_1(t)\frac{1}{\lambda_1(t)} + \chi_2(t)\frac{1}{\lambda_2(t)}\right] dt = \int_\alpha^\beta 1dt.$$

Define $s'(t) = \sigma(t)$. Then $\int_\alpha^\beta s'(t)dt = \beta - \alpha$. This proves the claim.

(c) Consider the case in which C is empty. In this case, it cannot be that $\lambda_1(t) = 0$ on a set of positive measure, and the previous point (b) can be applied to the interval $[a, b]$. Set $s(t) = a + \int_a^t s'(\tau)d\tau$. By the previous point (b), s is strictly monotonic increasing and maps $[a, b]$ onto itself. Let $t = t(s)$ be its inverse, so that, in particular, $t(a) = a$: we have that $1 = s'(t(s))t'(s)$. Consider the map $\tilde{x}(s) = x(t(s))$. We have

$$\begin{aligned}
\frac{d}{ds}\tilde{x}(s) &= x'(t(s))t'(s) = x'(t(s))\frac{1}{s'(t(s))} \\
&= x'(t(s))\frac{1}{s'(t(s))}\chi_1(t(s)) + x'(t(s))\frac{1}{s'(t(s))}\chi_2(t(s)) \\
&= x'(t(s))\lambda_1(t(s))\chi_1(t(s)) + x'(t(s))\lambda_2(t(s))\chi_2(t(s)) \\
&\in F(x(t(s)) = F(\tilde{x}(s)).
\end{aligned}$$

Hence the theorem is proved in this case.

(d) From now on we shall assume that C is nonempty. Set $c = \sup C$, so that $c \in C$. The complement of C is open relative to $[a, b]$; it consists of at most countably many nonoverlapping open intervals (a_i, b_i), with the possible exception of one of the form $[a_{i_i}, b_{i_i})$ with $a_{i_i} = a$, and one $(a_{i_f}, b_{i_f}]$ with $a_{i_f} = c$. For each i apply point (b) to the interval (a_i, b_i) to infer the existence of K_1^i and K_2^i, two subsets of (a_i, b_i) with characteristic functions $\chi_1^i(t)$ and $\chi_2^i(t)$ such that $\chi_1^i(t) + \chi_2^i(t) = \chi_{(a_i,b_i)}$ and such that, setting

$$s'(t) = \chi_1^i(t)\frac{1}{\lambda_1(t)} + \chi_2^i(t)\frac{1}{\lambda_2(t)},$$

we obtain

$$\int_{a_i}^{b_i} s'(\tau)d\tau = b_i - a_i.$$

(e) On $[a, c]$ set

$$s'(t) = \frac{1}{\lambda_2(t)}\chi_C(t) + \sum\left(\chi_1^i(t)\frac{1}{\lambda_1(t)} + \chi_2^i(t)\frac{1}{\lambda_2(t)}\right),$$

where the sum is over all intervals contained in $[a, c]$, i.e., with the exception of $(c, b]$. We have that

$$\int_a^c s'(\tau)d\tau = \kappa \leq c - a$$

since $\lambda_2(t) \geq 1$ and $\int_{a_i}^{b_i} s'(\tau)d\tau = b_i - a_i$. Setting $s(t) = a + \int_a^t s'(\tau)d\tau$, we obtain that s is an invertible map from $[a, c]$ to $[a, \kappa]$.

(f) Define $t = t(s)$ from $[a, \kappa]$ to $[a, c]$ to be the inverse of $s(.)$. Extend $t(.)$ as an absolutely continuous map $\tilde{t}(.)$ on $[a, c]$, setting

$$\tilde{t}'(s) = 0$$

for $s \in [\kappa, c]$. We claim that the function $\tilde{x}(s) = x(\tilde{t}(s))$ is a solution to

$$y'(s) \in F(y(s)), \quad y(a) = x_a$$

on the interval $[a, c]$. Moreover, we claim that it satisfies $\tilde{x}(c) = x(c)$.

To prove the claim, notice that, as in (c), we have that for s in $[a, \kappa]$, $\tilde{t}(s) = t(s)$ is invertible and

$$\begin{aligned}\frac{d}{ds}\tilde{x}(s) &= x'(\tilde{t}(s))\frac{1}{s'(\tilde{t}(s))} = x'(t(s))\frac{1}{s'(t(s))} \\ &= x'(\tilde{t}(s))\left[\lambda_2(t(s))\chi_C(t(s)) + \sum\left(\chi_1^i(t(s))\lambda_1(t(s)) + \chi_2^i(t(s))\lambda_2(t(s))\right)\right] \\ &\in F(x(t(s)) = F(\tilde{x}(s)).\end{aligned}$$

In particular, from $t(\kappa) = c$ we obtain $\tilde{x}(\kappa) = x(c)$. On $[\kappa, c]$, $\tilde{x}$ is constant and we have

$$\frac{d}{ds}\tilde{x}(s) = 0 \in F(x(c)) = F(\tilde{x}(\kappa)) = F(\tilde{x}(s)).$$

This proves the claim.

(g) It is left to define the solution on $[c, b]$. On it, $\lambda_1(t) > 0$ and the construction of points (b) and (c) can be repeated to find a solution to

$$y'(s) \in F(y(s)), \quad y(c) = x(c)$$

on $[c, b]$. This completes the proof. □

Proof of Theorem 1. By the upper semicontinuity of F, there exist a ball $B = B[x_a, \rho]$ and a positive real M such that F is bounded by M on $B = B[x_a, \rho]$; set $\delta = \frac{\rho}{M}$ and consider the interval $I = [a - \delta, a + \delta]$. On I a solution x to the Cauchy problem

$$y'(s) \in coF(y(s)), \quad y(a) = x_a$$

exists.

Fix any $\tau \in I$. Since the attainable set at τ, $A_{x_0}(\tau)$, is contained in the attainable set for the solutions to the convexified problem, $A_{x_0}^{co}(\tau)$, it is enough to show that $A_{x_0}^{co}(\tau) \subset A_{x_0}(\tau)$.

By assumption, for every x, $F(x)$ is almost convex; hence for almost every $t \in I$ there exist two nonempty sets $\Lambda_1(t)$ and $\Lambda_2(t)$ such that $\lambda_1 x'(t)) \in F(x(t))$ for

$\lambda_1 \in \Lambda_1(t)$, $\lambda_2 x'(t)) \in F(x(t))$ for $\lambda_2 \in \Lambda_2(t)$, and $0 \leq \lambda_1 \leq 1$, $1 \leq \lambda_2$. Set $Z = \{t : x'(t) = 0\}$: there is no loss of generality in assuming that, for $t \in Z$, $\Lambda_1(t) = \Lambda_2(t) = \{1\}$. We claim that the set valued map $t \to \Lambda_1(t)$ is measurable. Applying Lusin's theorem to x', write $I \setminus Z$ as $(\cup K_i) \cup N$, where the measure of N is 0, each K_i is compact, and the restriction to K_i of x' is continuous on K_i. Then, by the continuity on K_i of x' and of x and the upper semicontinuity of F, it follows that the map Λ_1 has a closed graph on $K_i \times R^N$; since, in addition, its values are closed subsets of $[0, 1]$, it is upper semicontinuous. It follows then that Λ_1 is measurable on I. The proof that $t \to \Lambda_2(t)$ is measurable is similar, with the difference that the values of Λ_2 need not be bounded. In this case, write $I \setminus Z$ as the countable union of the sets $M_n = \{t : \|x'(t)\| \geq 1/n\}$. On each M_n, Λ_2 has an upper bound, since F is bounded, and the same reasoning as in the previous point can be applied.

Hence, by standard arguments, there exist measurable selections $\lambda_1(.)$ and $\lambda_2(.)$ from the maps $\Lambda_1(.)$ and $\Lambda_2(.)$.

Apply Theorem 2 to the interval $[a, \tau]$ or $[\tau, a]$ to prove Theorem 1. □

Proof of Corollary 1. Let $t^* = \inf\{t \in [0, \tilde{t}] : x_1 \in A_{x_0}(t)\}$. Let (t_n) be decreasing to t_* and let x_n be solutions to the differential inclusion

$$x'(t) \in F(x(t))$$

such that $x_n(0) = x_0$ and $x_n(t_n) = x_1$. A subsequence of this sequence converges uniformly to x_*, and it is known that x_* is a solution to

$$x'(t) \in coF(x(t)).$$

Then, $x_*(t_*) \in A^{co}_{x_0}(t_*)$, and by Theorem 1 this set coincides with $A_{x_0}(t_*)$. Hence, x_* is the solution to the minimum time problem, and t_* is the minimum time required. □

REFERENCES

[1] A. Cellina and G. Colombo, *On a classical problem of the calculus of variations without convexity assumptions*, Ann. Inst. H. Poincaré, Anal. Non Linéaire, 7 (1990), pp. 97–106.

[2] L. Cesari, *Optimization—Theory and Applications*, Springer-Verlag, New York, 1983.

[3] A. F. Filippov, *On certain questions in the theory of optimal control*, Vestnik Moskov. Univ., Ser. Mat. Mech. Astr., 2 (1959), pp. 25–32; translated in SIAM J. Control, 1 (1962), pp. 76–84.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 266–299

THE FRACTIONAL REPRESENTATION APPROACH TO SYNTHESIS PROBLEMS: AN ALGEBRAIC ANALYSIS VIEWPOINT PART I: (WEAKLY) DOUBLY COPRIME FACTORIZATIONS*

A. QUADRAT†

Abstract. In this paper, we show how to reformulate the fractional representation approach to analysis and synthesis problems within an algebraic analysis framework. In terms of modules, we give necessary and sufficient conditions so that a system admits (weakly) left/right/doubly coprime factorizations. Moreover, we explicitly characterize the integral domains A such that every plant—defined by means of a transfer matrix whose entries belong to the quotient field of A—admits (weakly) doubly coprime factorizations. Finally, we show that this algebraic analysis approach allows us to recover, on the one hand, the approach developed in [M. C. Smith, *IEEE Trans. Automat. Control*, 34 (1989), pp. 1005–1007] and, on the other hand, the ones developed in [K. Mori and K. Abe, *SIAM J. Control Optim.*, 39 (2001), pp. 1952–1973; V. R. Sule, *SIAM J. Control Optim.*, 32 (1994), pp. 1675–1695 and 36 (1998), pp. 2194–2195; M. Vidyasagar, H. Schneider, and B. A. Francis, *IEEE Trans. Automat. Control*, 27 (1982), pp. 880–894; M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985].

Key words. fractional representation approach to synthesis problems, (weakly) left/right/doubly coprime factorizations, coherent rings and modules, coherent Sylvester domains, $H_\infty(\mathbb{C}_+)$, Bézout domains, algebraic analysis, module theory, homological algebra

AMS subject classifications. 93C05, 93D25, 93B25, 93C20, 16D40, 16P70, 16E60, 30D55

PII. S0363012902417127

Introduction. In the seventies, Vidyasagar and others introduced the idea of representing a class of transfer functions as the quotient field of a certain integral domain A of proper and stable transfer functions. Examples of such integral domains A, usually encountered in the literature, are the Banach algebra $H_\infty(\mathbb{C}_+)$ of bounded analytic functions in the open right half-plane $\mathbb{C}_+ = \{s \in \mathbb{C} \,|\, \mathrm{Re}\, s > 0\}$ [8], the algebra $RH_\infty = \mathbb{R}(s) \cap H_\infty(\mathbb{C}_+)$ of proper stable real rational functions [49], and the Wiener algebras $\mathcal{A}$, $\hat{\mathcal{A}}$ [3, 8], and $l_1(\mathbb{Z}_+)$ [49]. In the early eighties, this idea naturally led to the *fractional representation approach to synthesis problems*, principally developed in [3, 9, 48, 49]. The main outcome of this point of view is a reformulation of various questions of feedback stabilization of systems in terms of algebraic properties of some matrices whose entries belong to A (e.g., internal/strong/simultaneous/robust/optimal stabilization, parametrization of all the stabilizing controllers, graph topology, etc.).

Unfortunately, questions seem to remain for some classes of (infinite-dimensional) systems, in particular, the following:

1. Do necessary and sufficient conditions exist for internal stabilizability?
2. Is it possible to characterize all the integral domains A such that every plant—defined by means of a transfer matrix whose entries belong to the quotient field of A—is internally stabilizable?
3. What are the links between internal stabilizability and the existence of a doubly coprime factorization for the transfer matrix?

*Received by the editors November 15, 1999; accepted for publication (in revised form) November 4, 2002; published electronically April 17, 2003. This work was supported by grant HPMF-CT-1999-95 during the author's stay at the University of Leeds (United Kingdom).
http://www.siam.org/journals/sicon/42-1/41712.html
†INRIA Sophia Antipolis, CAFE Project (Computer Algebra and Functional Equations), 2004 Route des Lucioles, BP 93, 06902 Sophia Antipolis cedex, France (Alban.Quadrat@sophia.inria.fr).

4. Is it always possible to parametrize all the stabilizing controllers of a stabilizing plant by means of the Youla–Kučera parametrization?

In order to solve certain of these open questions, the authors of [40] tried to revisit the fractional representation approach to stabilization problems of single-input single-output (SISO) systems using a more intrinsic framework than the one used in [3, 9, 48, 49]. A module approach has recently been developed in [44] and continued in [22]. In the same years as that last work, another approach was developed in [30, 31, 32] using the ideas of *algebraic analysis* (see [27, 28, 29] and the references therein). The purpose of this paper is to present this new mathematical framework and to explain how certain of the previous open questions can be solved using this algebraic analysis point of view.

In this paper, we first introduce the concepts of *weakly left/right/doubly coprime factorizations*, give necessary and sufficient conditions in terms of modules so that a transfer matrix admits such factorizations, and characterize all the integral domains A over which every transfer matrix admits weakly doubly coprime factorizations (namely, *coherent Sylvester domains*, e.g., $H_\infty(\mathbb{C}_+)$). Moreover, we also give necessary and sufficient conditions so that a transfer matrix admits left/right/doubly coprime factorizations, and we recover a result of Vidyasagar [49] describing all the integral domains A over which every transfer matrix admits doubly coprime factorizations (namely, *Bézout domains*). In particular, we recover and generalize some standard results of [22, 43, 44, 49]. In the second part of the paper [33], we shall use the same mathematical framework and the previous results to develop necessary and sufficient conditions for *internal stabilizability* [9, 48, 49]. Moreover, we shall characterize all the integral domains A over which every plant—defined by means of a transfer matrix whose entries belong to the quotient field of A—is internally stabilizable (namely, *Prüfer domains*). Hence, the algebraic analysis framework seems to solve the first three questions listed above. We refer the reader to [34] for a general answer to the fourth one. Let us note that all these results use the techniques of module theory and homological algebra, and they seem difficult to obtain using only a matrix approach.

If we want to develop some general algorithms (i.e., valid for a general integral domain A) that check the existence of (weakly) left/right/doubly coprime factorizations and compute them, we then have to overcome the difficulty arising from the fact that most of the integral domains of SISO stable plants are *Banach algebras* (e.g., $H_\infty(\mathbb{C}_+)$, $\mathcal{A}$, $\hat{\mathcal{A}}$, $l_1(\mathbb{Z}_+)$). Indeed, a result proves that *noetherian* Banach algebras are only finite-dimensional [41], and thus, most of the Banach algebras used in systems theory are not noetherian. Therefore, it seems that we cannot use the standard techniques of commutative algebra, module theory, and homological algebra developed for noetherian rings to study general (infinite-dimensional) linear systems. (Some modules may fail to be finitely generated.) We show that the only possibility for coping with this difficulty seems to require that the Banach algebras be *coherent rings*. This result could explain why coherent Sylvester domains, Prüfer and Bézout domains, which play important roles in the fractional representation approach (see above), are all coherent. Using the fact that a system is defined by means of a transfer matrix, we prove that, if A is a coherent domain, then every system defines a *coherent A-module*. Now, the (category of) coherent A-modules over a coherent ring A (is) are invariant under all the elementary algebraic manipulations (e.g., intersection, sum, quotient, tensor product, duality, etc.). Therefore, we can use homological algebra to develop general algorithms which check the existence of (weakly) left/right/doubly coprime factorizations (or internal stabilizability in [33]) of (infinite-dimensional) linear systems defined over a coherent domain A.

Plan. In the first part of this paper (section 1), we describe the framework of the fractional representation approach to analysis and synthesis problems and explain why and how it is possible to use module theory. We recall some definitions of module theory and homological algebra that will be constantly used in the rest of the paper and in [33]. The second part (section 2) is related to factorization problems. We first introduce the concept of *a weakly doubly coprime factorization* and show that it corresponds to the weakest coprimeness for transfer matrices. We give necessary and sufficient conditions so that a transfer matrix admits a weakly left/right/doubly coprime factorization. In the third part (section 3), we introduce the concept of coherent rings and modules. We prove that every transfer matrix, with entries in the quotient field of A, admits a weakly doubly coprime factorization iff A is a *coherent Sylvester domain*. We show that $H_\infty(\mathbb{C}_+)$ is a coherent Sylvester domain. Finally, in the last part (section 4), we give necessary and sufficient conditions so that a transfer matrix admits left/right/doubly coprime factorizations.

Notation. In the course of the text, A denotes a commutative integral domain ($a\,b = 0, a \neq 0 \Rightarrow b = 0$) with a unit, $M_{q\times p}(A)$ (resp., $M_p(A)$) the set of $q \times p$ (resp., $p\times p$) matrices whose entries belong to A, and I_p the identity matrix. If $R \in M_{q\times p}(A)$, then R^T is the transposed matrix. *By convention, every vector with entries in A is a row vector.* The positive integers p, $q \in \mathbb{Z}_+$ will always satisfy $p \geq q$. If M and N are two A-modules, then $M \cong N$ means that M and N are isomorphic as A-modules, $\hom_A(M, N)$ is the A-module of the A-morphisms (i.e., A-linear maps) from M to N, and $M^* = \hom_A(M, A)$. Finally, $(a_1, \ldots, a_n)$ denotes the ideal $A\,a_1 + \cdots + A\,a_n$, and $\triangleq$ means "by definition."

1. The fractional representation approach to synthesis problems.

1.1. Introduction. Following ideas of Zames [51], a class of transfer functions needs to have the structure of a ring if we want to connect two systems in cascade (product) or in parallel (sum). In the fractional representation approach to analysis and synthesis problems, we start with an integral domain A of *SISO stable plants* [3, 8, 9, 48, 49]. Classical examples of integral domains of SISO stable plants are the Banach algebra $H_\infty(\mathbb{C}_+)$ of the bounded analytic functions on the open right half-plane $\mathbb{C}_+ = \{s \in \mathbb{C} \mid \mathrm{Re}\, s > 0\}$ [8], the ring RH_∞ of proper stable real rational functions [49], or the Wiener algebras $\mathcal{A}$, $\hat{\mathcal{A}}$, $l_1(\mathbb{Z}_+)$ [8, 49]. Then, the class of (unstable) SISO plants considered is defined by the field of fractions of A:

$$K = Q(A) = \{n/d \mid 0 \neq d,\, n \in A\}. \tag{1.1}$$

Example 1.1. Let us give some examples.

- Let us consider $A = RH_\infty = \{p = n/d \mid \deg n \leq \deg d,\, d(\underline{s}) = 0 \Rightarrow \mathrm{Re}\,\underline{s} < 0\}$ the integral domain of proper stable real rational functions. The transfer function $p = 1/(s-1)$ (resp., $p = s$) does not belong to A because p has the unstable pole 1 in $\mathbb{C}_+$ (resp., p is not proper) but belongs to $K = Q(A) = \mathbb{R}(s)$ because p can be represented as $p = n/d$ with $n = 1/(s+1) \in A$ and $d = (s-1)/(s+1) \in A$ (resp., $n = s/(s+1) \in A$ and $d = 1/(s+1) \in A$).
- Let us consider the following Wiener algebra [3, 8]:

$$\mathcal{A} = \left\{ f(t) + \sum_{i=0}^{\infty} a_i\, \delta(t - t_i) \mid f \in L_1(\mathbb{R}_+), (a_i)_{i\geq 0} \in l_1(\mathbb{Z}_+),\ 0 = t_0 \leq t_1 \ldots \right\},$$

with the two operations + and the convolution $\star$ and the Dirac distribution δ as the unit. Endowed with the topology defined by the norm

$$\| g \|_{\mathcal{A}} = \| f \|_{L_1(\mathbb{R}_+)} + \| (a_i)_{i\geq 0} \|_{l_1(\mathbb{Z}_+)},$$

$\mathcal{A}$ becomes a *Banach algebra* and an integral domain [3, 8, 48, 49]. The same properties hold for $\hat{\mathcal{A}} = \{\hat{f} \mid f \in \mathcal{A}\}$ ($\hat{\cdot}$ is the Laplace transform) with the norm $\| \hat{f} \|_{\hat{\mathcal{A}}} = \| f \|_{\mathcal{A}}$. For instance, an example of a transfer function which belongs to $K = Q(\mathcal{A})$ is the following:

$$p = e^t\, Y(t) \star \delta(t-1) = (\delta(t) - 2\, e^{-t}\, Y(t))^{-1} \star (e^{-t}\, Y(t) \star \delta(t-1)),$$

where $Y(t)$ denotes the Heaviside distribution (i.e., 1 for $t \geq 0$, and 0 otherwise) and $\delta(t) - 2\, e^{-t}\, Y(t)$, $e^{-t}\, Y(t) \star \delta(t-1) \in \mathcal{A}$. Equivalently, in the frequency domain, the same plant is defined by the following transfer function:

$$p = \frac{e^{-s}}{s-1} = \frac{\left(\frac{e^{-s}}{s+1}\right)}{\left(\frac{s-1}{s+1}\right)} \in Q(\hat{\mathcal{A}}), \qquad \frac{e^{-s}}{s+1},\ \frac{s-1}{s+1} \in \hat{\mathcal{A}}.$$

If $P \in M_{q\times(p-q)}(K)$, then it is always possible to write it as $P = D^{-1}\, N = \tilde{N}\, \tilde{D}^{-1}$, where $D \in M_q(A)$ and $\tilde{D} \in M_{p-q}(A)$ are two invertible matrices and $N \in M_{q\times(p-q)}(A)$, $\tilde{N} \in M_{q\times(p-q)}(A)$, i.e., all the entries of these four matrices are stable. For example, we can use $D = d\, I_q$ and $N = d\, P$, where $0 \neq d \in A$ is the product of the denominators of all the entries of P, and similarly for $\tilde{D} = d\, I_{p-q}$ and $\tilde{N} = P\, d$.

Example 1.2. Let $A = H_\infty(\mathbb{C}_+)$, and let us consider the plant defined by

$$P = \begin{pmatrix} \frac{e^{-s}}{s+1} & \frac{s-1}{s+1} \\ 0 & \frac{1}{s-1} \end{pmatrix} \in M_2(K), \qquad K = Q(A). \tag{1.2}$$

Then, P can be written as $P = D^{-1}\, N$ with

$$D = \begin{pmatrix} \frac{s-1}{s+1} & 0 \\ 0 & \frac{s-1}{s+1} \end{pmatrix} \in M_2(A), \qquad N = \begin{pmatrix} \frac{(s-1)\, e^{-s}}{(s+1)^2} & \left(\frac{s-1}{s+1}\right)^2 \\ 0 & \frac{1}{s+1} \end{pmatrix} \in M_2(A). \tag{1.3}$$

Thus, instead of representing a plant by $y = P\, u$ with $P \in M_{q\times(p-q)}(K)$, the fractional representation approach studies the following two systems:

$$(D : -N) \begin{pmatrix} y \\ u \end{pmatrix} = 0, \qquad \begin{pmatrix} y \\ u \end{pmatrix} = \begin{pmatrix} \tilde{N} \\ \tilde{D} \end{pmatrix} z,$$

with $R = (D : -N) \in M_{q\times p}(A)$ and $\tilde{R} = (\tilde{N}^T : \tilde{D}^T)^T \in M_{p\times(p-q)}(A)$. Then, using linear algebra over the ring A, it is possible to study the structural properties of P by looking at the algebraic properties of the matrices R and $\tilde{R}$ (left/right/doubly factorizations). See [3, 8, 9, 48, 49] for more information.

For linear finite-dimensional systems [49], the fractional representation approach gives necessary and sufficient conditions for internal stabilizability, existence of doubly coprime factorizations, or Youla–Kučera parametrizations of all the stabilizing controllers, etc. [49]. The possibility of generalizing these results to linear infinite-dimensional

systems (delay systems, partial differential equations such as the wave/heat/Euler–Bernoulli equations) has naturally been asked from theoretical and practical points of view [3, 8, 9, 48, 49]. In this framework, classes of linear infinite-dimensional systems are generally modelized by means of Banach algebras such as $H_\infty(\mathbb{C}_+)$, $\hat{\mathcal{A}}$, $l_1(\mathbb{Z}_+)$. These rings are algebraically and topologically more complex than the ring RH_∞ used for finite-dimensional systems. Hence, some questions, such as the ones given in the introduction, are still open for some classes of infinite-dimensional systems [8, 48, 49]. As we described it in the introduction, the purpose of this paper is to show that we can solve certain of these problems if we adopt an *algebraic analysis* framework rather than a matricial one. Here, we call "algebraic analysis" a mathematical framework which uses commutative algebra, module theory, and homological algebra combined with functional analysis (Banach algebras). This idea could seem natural if we notice that, in order to understand the structural properties of a plant, defined by the transfer matrix $P \in M_{q\times(p-q)}(K)$, we need to study the matrices $R \in M_{q\times p}(A)$ and $\tilde{R} \in M_{p\times(p-q)}(A)$, whose entries belong to a certain algebra of functions (e.g., a Banach algebra), and linear algebra over a ring is just a part of module theory.

1.2. Definitions. In this section, we give some definitions that we shall need to characterize intrinsically the structural properties of systems.

Let $R \in M_{q\times p}(A)$, and let us define the A-morphism (i.e., an A-linear map) $.R$ by

$$\begin{array}{rccc} .R: & A^q & \longrightarrow & A^p, \\ & (a_1 : \cdots : a_q) & \longrightarrow & (a_1 : \cdots : a_q)R. \end{array}$$

Then, the image $\operatorname{im} .R$ is the A-module generated by the A-linear combinations of the rows of R. This A-module is usually used in control theory [22, 44]. In *algebraic analysis* [27, 28, 29], one usually prefers to use the A-module $M = \operatorname{coker} .R = A^p/A^q R$.

Definition 1.1. *We have the following definitions (see* [1, 2, 39]*):*

- *A* complex *is a sequence of A-modules F_i and A-morphisms d_i, denoted by* $\cdots \longrightarrow F_{i+1} \xrightarrow{d_{i+1}} F_i \xrightarrow{d_i} F_{i-1} \longrightarrow \cdots$, *such that* $d_i \circ d_{i+1} = 0$, *i.e.,*

$$\operatorname{im} d_{i+1} \subseteq \ker d_i.$$

- *The ith* A-module of homology *of a complex is defined by*

$$H(F_i) = \frac{\ker d_i}{\operatorname{im} d_{i+1}}.$$

- *A sequence is said to be* exact *at F_i if $H(F_i) = 0$, i.e.,* $\ker d_i = \operatorname{im} d_{i+1}$, *and* exact *if we have $H(F_i) = 0$ for all i.*

Example 1.3. For instance, the exact sequence $0 \longrightarrow M' \xrightarrow{f} M$ means that f is injective, whereas the exact sequence $M \xrightarrow{g} M'' \longrightarrow 0$ means that g is surjective.

Let π be the A-morphism which associates to every row vector of A^p its class in the quotient A-module $M = A^p/A^q R$. We have the following exact sequence:

$$A^q \xrightarrow{.R} A^p \xrightarrow{\pi} M \longrightarrow 0. \tag{1.4}$$

Let $\{e_1, \ldots, e_p\}$ be the canonical basis of A^p, and $\{f_1, \ldots, f_q\}$ that of A^q. We define

$z_i = \pi(e_i),\ i = 1, \dots, p$. Then, we have for $i = 1, \dots, q$

$$
\begin{aligned}
f_i\, R = (R_{i1} : \cdots : R_{ip}) &= \sum_{j=1}^{p} R_{ij}\ e_j \in A^q\, R \\
\Rightarrow \pi(f_i\, R) &= \sum_{j=1}^{p} R_{ij}\ \pi(e_j) = \sum_{j=1}^{p} R_{ij}\ z_j = 0.
\end{aligned} \tag{1.5}
$$

Hence, M is defined by the equations $R\, z = 0$, where $z = (z_1 : \cdots : z_p)^T$, and their A-linear combinations. Moreover, for all $m \in M$, there exists $(a_1 : \cdots : a_p) \in A^p$ such that

$$
m = \pi((a_1 : \cdots : a_p)) = \pi\left(\sum_{i=1}^{p} a_i\, e_i\right) = \sum_{i=1}^{p} a_i\, \pi(e_i) = \sum_{i=1}^{p} a_i\, z_i.
$$

This means that every element m of M is an A-linear combination of the elements $z_1, \dots, z_p$, and the A-module M is said to be *finitely generated*. In fact, the module is defined by a finite number of equations (q equations) with a finite number of unknowns (p unknowns). In this case, we say that M is *finitely presented*, a fact which is equivalent to the existence of the exact sequence (1.4).

Example 1.4. Let us reconsider Example 1.2. We have $A = H_\infty(\mathbb{C}_+)$, and the matrix $R = (D : -N) \in M_{2\times 4}(A)$ is defined by

$$
R = \begin{pmatrix} \frac{s-1}{s+1} & 0 & -\frac{(s-1)\, e^{-s}}{(s+1)^2} & -\left(\frac{s-1}{s+1}\right)^2 \\ 0 & \frac{s-1}{s+1} & 0 & -\frac{1}{s+1} \end{pmatrix}. \tag{1.6}
$$

Then, the A-morphism $.R$ is defined by

$$
\begin{aligned}
A^2 &\longrightarrow A^4, \\
(a_1 : a_2) &\longrightarrow \left(a_1\, \left(\tfrac{s-1}{s+1}\right) :\ a_2\, \left(\tfrac{s-1}{s+1}\right) :\ -a_1\, \left(\tfrac{(s-1)\, e^{-s}}{(s+1)^2}\right) :\ -a_1\, \left(\tfrac{s-1}{s+1}\right)^2 - a_2\, \tfrac{1}{(s+1)}\right).
\end{aligned}
$$

Therefore, the A-module $M = A^4/A^2\, R$ is defined by the equations

$$
\begin{cases} \frac{(s-1)}{(s+1)}\, y_1 - \frac{(s-1)\, e^{-s}}{(s+1)^2}\, u_1 - \left(\frac{s-1}{s+1}\right)^2 u_2 = 0, \\ \frac{(s-1)}{(s+1)}\, y_2 - \frac{1}{(s+1)}\, u_2 = 0 \end{cases}
$$

and their A-linear combinations, where $y_i = \pi(e_i),\ u_i = \pi(e_{i+2}),\ i = 1, 2$.

Definition 1.2. *We have the following definitions (see* [1, 39]*):*

- *An A-module M is* finitely generated *if there exists an A-module of the form $F_0 \cong A^{r_0}$, $r_0 \in \mathbb{Z}_+$, such that we have the following exact sequence:*

$$
F_0 \xrightarrow{d_0} M \longrightarrow 0. \tag{1.7}
$$

- *An A-module M is* finitely presented *if there exist two A-modules $F_i \cong A^{r_i}$, $r_i \in \mathbb{Z}_+$, $i = 0, 1$, such that we have the following exact sequence:*

$$
F_1 \xrightarrow{d_1} F_0 \xrightarrow{d_0} M \longrightarrow 0. \tag{1.8}
$$

One of the main interests of associating an A-module with a matrix $R \in M_{q\times p}(A)$ is that we can use the natural classification of the properties of the module to understand the structural properties of the system $R\,z = 0$. The purpose of the next sections and [33] is to illustrate how some concepts of modules play interesting roles in characterizing the existence of a (weakly) left/right/doubly coprime factorization and internal stabilizability.

DEFINITION 1.3 (see [2, 39]). *If M is a finitely generated A-module, then*

- *M is* free *if $M \cong A^r$ for $r \in \mathbb{Z}_+$.*
- *M is* stably free *if there exist $r, s \in \mathbb{Z}_+$ such that $M \oplus A^s \cong A^r$.*
- *M is* projective *if there exist an A-module N and $r \in \mathbb{Z}_+$ such that*

$$M \oplus N \cong A^r.$$

- *M is* reflexive *if the A-morphism $\epsilon : M \longrightarrow \hom_A(\hom_A(M, A), A)$—defined by $\epsilon(m)(f) = f(m)$ for all $m \in M$, for all $f \in \hom_A(M, A)$—is an isomorphism.*
- *M is* torsion-free *if its* torsion submodule

$$t(M) = \{m \in M \mid \exists\, a \in A \setminus 0 :\ a\,m = 0\}$$

is reduced to 0. $m \in t(M)$ is a torsion element. *We have the exact sequence:*

$$0 \longrightarrow t(M) \longrightarrow M \longrightarrow M/t(M) \longrightarrow 0. \tag{1.9}$$

- *M is a* torsion *A-module if $t(M) = M$.*
- *M is a* flat *A-module if, for every relation $\sum_{i=1}^{q} a_i\, m_i = 0,\ a_i \in A,\ m_i \in M$, there exist $n_i \in M$ and $b_{ij} \in A$ such that*

$$\begin{cases} m_i = \displaystyle\sum_{j=1}^{r} b_{ij}\, n_j, & 1 \le i \le q, \\ \displaystyle\sum_{i=1}^{q} a_i\, b_{ij} = 0, & 1 \le j \le r. \end{cases}$$

PROPOSITION 1.4 (see [1, 2, 39]). *We have the following relations:*

1. *Free $\Rightarrow$ stably free $\Rightarrow$ projective $\Rightarrow$ reflexive $\Rightarrow$ torsion-free.*
2. *Projective $\Rightarrow$ flat $\Rightarrow$ torsion-free.*
3. *A finitely generated flat A-module is projective.*
4. *If A is a Bézout domain—namely, a domain A such that every finitely generated ideal I of A has the form $I = (a)$ for a certain $a \in A$—then every finitely generated torsion-free A-module is free and $M \cong t(M) \oplus M/t(M)$.*

DEFINITION 1.5. *We have the following definitions (see* [6, 20]*):*

- *A is a* Hermite ring *if every finitely generated stably free A-module is free, or, equivalently, if every* unimodular row $(a_1 : \cdots : a_n) \in A^n$ *—namely a row such that there exists $(b_1 : \cdots : b_n)^T$ satisfying $\sum_{i=1}^{n} a_i\, b_i = 1$—can be completed to a* unimodular matrix, *i.e., to a matrix of $GL_n(A)$.*
- *A is a* projective-free ring *if every finitely generated projective A-module is free.*

In particular, a projective-free ring (e.g., a Bézout domain) is a Hermite ring. A difficult result, namely, the Quillen–Suslin theorem [20, 39], proves that the ring of polynomials $k[\chi_1, \ldots, \chi_n]$ in χ_i, with coefficients in a field k, is projective-free.

THEOREM 1.6. *RH_∞ and $k[s]$, where k is a field, are principal ideal domains (namely, a domain such that every ideal I of A is principal, i.e., has the form $I = (a)$ for a certain $a \in A$; see* [49]*). The domain of entire functions with coefficients in $k = \mathbb{R}, \mathbb{C}$,*

$$E(k) = \left\{ f(s) = \sum_{n=0}^{+\infty} a_n \, s^n \mid a_n \in k, \lim_{n \to +\infty} |a_n|^{1/n} = 0 \right\},$$

and $\mathcal{E} = E(\mathbb{R}) \cap \mathbb{R}(s)\,[e^{-s}]$ are Bézout domains (see [18, 21]*). Thus, all these rings are projective-free.*

We introduce the concept of *localization* of modules. In the following sections, we shall show why this concept is interesting for the study of the links between a transfer matrix $y = P\,u$, $P = D^{-1}\,N \in M_{q\times(p-q)}(K)$ and the system $D\,y = N\,u$, where $R = (D : -N) \in M_{q\times p}(A)$.

DEFINITION 1.7 (see [1, 39]). *We have the following definitions:*

- *A* multiplicative set *S of A is a subset of A such that for all $a, b \in S \Rightarrow a\,b \in S$ and $1 \in S$.*
- *Let M be an A-module. We define an equivalence relation $\sim$ on $S \times M$ by $(s, m) \sim (s', m')$ if there exists $t \in S$ such that $t(sm' - s'm = 0)$. The* localization *of the A-module M with respect to S is the $S^{-1}A = \{a/s \mid (s, a) \in S \times A\}$-module*

$$S^{-1}M = (S \times M)/\sim .$$

If we denote by m/s the equivalence class of (s, m) in $S^{-1}M$, then we have

$$S^{-1}M = \{(a/s)\,m \mid (s, a) \in S \times A,\, m \in M\}\,.$$

The localization of a module is just a way to extend the scalars of the A-module M from A to $S^{-1}A$. Moreover, we have the following canonical A-morphism:

$$\begin{array}{rcl} i_S : M & \longrightarrow & S^{-1}M, \\ m & \longrightarrow & m/1, \end{array}$$

from which we obtain $\ker i_S = \{m \in M \mid \exists\, 0 \neq a \in S,\ a\,m = 0\}$.

DEFINITION 1.8. *If $S = A \setminus 0$, then S is a multiplicative set of A, and the* field of fractions *of A is defined by $S^{-1}A = Q(A) = \{a/b \mid 0 \neq b,\, a \in A\}$. We have*

$$\ker\, i_S = t(M) = \{m \in M \mid \exists\, 0 \neq a \in A : a\,m = 0\}. \tag{1.10}$$

In the course of the paper, we shall denote by K the field of fractions $Q(A)$ of A.

Example 1.5. Let us reconsider the A-module $M = A^4/A^2\,R$ defined in Example 1.4. We can check that the element $z = y_1 - \frac{e^{-s}}{(s+1)}\,u_1 - \frac{(s-1)}{(s+1)}\,u_2$ satisfies $\frac{(s-1)}{(s+1)}\,z = 0$, i.e., z is a torsion element of M. (See Example 3.4 for an explicit computation of the torsion elements of M.) If $S = A \setminus 0$, then, in the $K = Q(A)$-module $S^{-1}\,M$, we have

$$\frac{(s-1)}{(s+1)}\left(\frac{z}{1}\right) = \frac{\frac{(s-1)}{(s+1)}\,z}{1} = \frac{0}{1} \Rightarrow \frac{(s+1)}{(s-1)}\,\frac{(s-1)}{(s+1)}\left(\frac{z}{1}\right) = \frac{z}{1} = \frac{0}{1} \Rightarrow i_S(z) = \frac{0}{1}.$$

Moreover, we have the following isomorphism (see [1, 39]):

$$\begin{array}{rcl} S^{-1}M & \cong & S^{-1}A \otimes_A M, \\ (a/b)\,m & \longleftrightarrow & a/b \otimes m, \end{array} \tag{1.11}$$

which shows that the localization corresponds to the *tensor product* $S^{-1}A\otimes_A$.

DEFINITION 1.9 (see [1, 39]). *The* rank *of an A-module is defined by*

$$\operatorname{rank}_A(M) = \dim_K(K \otimes_A M),$$

where $K \otimes_A M$ is a K-vector space and $\dim_K(K \otimes_A M)$ is its dimension over K.

PROPOSITION 1.10 (see [1, 2, 39]). *If S is a multiplicative set of A, then $S^{-1}A$ is a flat A-module, and, for every exact sequence $0 \longrightarrow M' \longrightarrow M \longrightarrow M'' \longrightarrow 0$, we have the following exact sequence:*

$$0 \longrightarrow S^{-1}A \otimes_A M' \longrightarrow S^{-1}A \otimes_A M \longrightarrow S^{-1}A \otimes_A M'' \longrightarrow 0.$$

Moreover, if M', M, and M'' are finitely generated A-modules, then we have

$$\operatorname{rank}_A(M) = \operatorname{rank}_A(M') + \operatorname{rank}_A(M''). \tag{1.12}$$

2. Weakly doubly coprime factorizations. Since the work of Rosenbrock [38] on coprime factorizations of rational matrices, this concept has played an increasing role in analysis and synthesis problems. This technique has been popularized by the book of Vidyasagar [49]. However, contrary to finite-dimensional systems, the transfer matrices of more general systems (delay systems, partial differential equations) do not generally admit doubly coprime factorizations. Intuitively, this comes from the fact that the algebraic properties of rings such as $H_\infty(\mathbb{C}_+)$, $\mathcal{A}$, $\hat{\mathcal{A}}$, and $l_1(\mathbb{Z}_+)$ are more complex than the ones of RH_∞ or $k[s]$ (with k a field), which are used for finite-dimensional systems. In the next section, we shall give a mathematical formulation of the term "more complex." In order to achieve this goal, we shall need to introduce the concept of weakly doubly coprime factorizations of a transfer matrix.

2.1. Weak primeness and torsion-free modules. Let us introduce the concept of a weakly left-prime matrix.

DEFINITION 2.1. *Let A be an integral domain and $K = Q(A)$ its field of fractions. The matrix $R \in M_{q\times p}(A)$ is* weakly left-prime *if it satisfies*

$$K^q R \cap A^p = A^q R,$$

namely, if, for every $\lambda \in A^p$ satisfying $\lambda = \mu R$ for a certain $\mu \in K^q$, there exists $\nu \in A^q$ such that $\lambda = \nu R$. The concept of weak right-primeness *can be defined similarly. Let us notice that R is weakly right-prime iff R^T is weakly left-prime.*

If $R \in M_{q\times p}(A)$ has *full row rank*, namely, if the q rows of R are A-linearly independent, then R is weakly left-prime iff

$$\mu \in K^q, \quad \mu R \in A^p \Rightarrow \mu \in A^q.$$

Example 2.1. Let us consider the full row rank matrix R defined by (1.6). This matrix R is not weakly left-prime because $(\frac{s+1}{s-1} : 0) \notin A^2$ and we have

$$\left(\tfrac{s+1}{s-1} : 0\right) \begin{pmatrix} \frac{s-1}{s+1} & 0 & -\frac{(s-1)\,e^{-s}}{(s+1)^2} & -\left(\frac{s-1}{s+1}\right)^2 \\ 0 & \frac{s-1}{s+1} & 0 & -\frac{1}{s+1} \end{pmatrix} = \left(1 : 0 : -\tfrac{e^{-s}}{s+1} : -\tfrac{s-1}{s+1}\right) \in A^4.$$

PROPOSITION 2.2 (see [43]). *If A is* a greatest common divisor domain *(GCDD), namely, every couple of elements of A has a greatest common divisor, then a full row*

rank matrix $R \in M_{q\times p}(A)$ is weakly left-prime iff R is irreducible *(or* minor left-prime [27]*), namely,* 1 *is the greatest common divisor of the $q \times q$ minors of R.*

Let us notice that if A is no longer a GCDD, then the previous result fails to be true [43]. In particular, it is not known whether or not $\mathcal{A}$ or $\hat{\mathcal{A}}$ are GCDD.

LEMMA 2.3. *Let A be an integral domain, $K = Q(A)$ its field of fractions, and $R \in M_{q\times p}(A)$. Then, we have*

$$K^q R \cap A^p = \overline{A^q R},$$

where $\overline{A^q R} = \{\lambda \in A^p \mid \exists\, 0 \neq a \in A : a\lambda \in A^q R\}$ is called the A-closure of the submodule $A^q R$ in A^p *(see* [7]*).*

Proof. Let $\lambda \in K^q R \cap A^p$; then $\lambda \in A^p$, and there exists $\mu \in K^q$ such that $\lambda = \mu R$. Let us write $\mu = d^{-1}\nu$ with $\nu \in A^q$ and $0 \neq d \in A$. Then, we have $d\lambda = \nu R$, i.e., $\lambda \in \overline{A^q R}$. Conversely, let $\lambda \in \overline{A^q R}$; then $\lambda \in A^p$, and there exists $0 \neq d \in A$ such that $d\lambda \in A^q R$. Thus, there exists $\nu \in A^q$ such that $d\lambda = \nu R$; i.e., $\lambda = (d^{-1}\nu) R \in K^q R$, i.e., $\lambda \in K^q R \cap A^p$. □

PROPOSITION 2.4. *Let A be an integral domain, $K = Q(A)$ its field of fractions, $R \in M_{q\times p}(A)$, and $M = A^p/A^q R$. Then, we have*

$$\begin{cases} t(M) = (K^q R \cap A^p)/A^q R, \\ M/t(M) = A^p/(K^q R \cap A^p). \end{cases}$$

Proof. Let us note that we have $A^q R \subseteq K^q R \cap A^p$. Therefore, we have the following commutative exact diagram,

$$\begin{array}{ccccccccc}
 & & 0 & & 0 & & & & \\
 & & \downarrow & & \downarrow & & & & \\
0 & \longrightarrow & A^q R & \longrightarrow & A^p & \longrightarrow & M & \longrightarrow & 0 \\
 & & \downarrow & & \| & & \downarrow & & \\
0 & \longrightarrow & K^q R \cap A^p & \longrightarrow & A^p & \longrightarrow & A^p/(K^q R \cap A^p) & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
 & & (K^q R \cap A^p)/A^q R & & 0 & & 0 & & \\
 & & \downarrow & & & & & & \\
 & & 0 & & & & & &
\end{array}$$

from which we deduce the following exact sequence (snake lemma [2, 39]):

$$0 \longrightarrow (K^q R \cap A^p)/A^q R \longrightarrow M \longrightarrow A^p/(K^q R \cap A^p) \longrightarrow 0. \tag{2.1}$$

Using Lemma 2.3, we obtain

$$(K^q R \cap A^p)/A^q R = \overline{A^q R}/A^q R = \{m \in M \mid \exists\, 0 \neq a \in A : a\, m = 0\} = t(M).$$

Then, we have $A^p/(K^q R \cap A^p) = M/t(M)$ (see (1.9)), which proves the proposition. □

A direct consequence of Proposition 2.4 is the following corollary, which gives a module interpretation of the weak left-primeness.

COROLLARY 2.5. *Let A be an integral domain and $K = Q(A)$ its field of fractions, $R \in M_{q\times p}(A)$, and the A-module $M = A^p/A^q R$. Then, we have the equivalences*

1. *R is weakly left-prime, i.e., $\overline{A^q R} = K^q R \cap A^p = A^q R$;*
2. *M is a torsion-free A-module, i.e., $t(M) = 0$.*

Example 2.2. From Example 2.1, we know that the matrix R defined by (1.6) is not weakly left-prime. By Corollary 2.5, we deduce that the A-module $M = A^4/A^2\,R$ is not torsion-free. A torsion element is obtained by taking the class of the vector $(1:\ 0:\ -\frac{e^{-s}}{s+1}:\ -\frac{s-1}{s+1})$ (see Example 2.1) in M to obtain $z = y_1 - \frac{e^{-s}}{(s+1)}\,u_1 - \frac{(s-1)}{(s+1)}\,u_2$. We recover the torsion element z obtained in Example 1.5. It satisfies $\frac{(s-1)}{(s+1)}\,z = 0$.

Dually, we can prove that $\tilde{R} \in M_{p\times(p-q)}(A)$ is weakly right-prime iff the A-module $A^p/A^{p-q}\,\tilde{R}^T$ is torsion-free.

2.2. Transfer matrices. The following lemma shows that if a transfer matrix $P \in M_{q\times(p-q)}(K)$ is such that $P = D^{-1}\,N$, then the A-module $\overline{A^q\,R}$ depends only on P and not on $R = (D:\ -N) \in M_{q\times p}(A)$.

LEMMA 2.6. *Let A be an integral domain, $K = Q(A)$ its field of fractions, and $P \in M_{q\times(p-q)}(K)$ a transfer matrix. If P can be written as $P = D_1^{-1}\,N_1 = D_2^{-1}\,N_2$, where $R_1 = (D_1:\ -N_1) \in M_{q\times p}(A)$ and $R_2 = (D_2:\ -N_2) \in M_{q\times p}(A)$, then we have*

$$\begin{cases} A^q\,R_1 \subseteq \overline{A^q\,R_2}, \\ \\ A^q\,R_2 \subseteq \overline{A^q\,R_1}, \end{cases} \Rightarrow \overline{A^q\,R_1} = \overline{A^q\,R_2},$$

and thus, $\overline{A^q\,R_i}$ and $M_i/t(M_i) = A^p/\overline{A^q\,R_i}$ depend only on P and not on R_i, where $M_i = A^p/A^q\,R_i$. In particular, if $A^q\,R_1$ (resp., $A^q\,R_2$) is A-closed, then we have $\overline{A^q\,R_2} = A^q\,R_1$ (resp., $\overline{A^q\,R_1} = A^q\,R_2$). The same result holds for $P = \tilde{N}_i\,\tilde{D}_i^{-1}$.

Proof. Clearing the denominators of P, we have $P = d^{-1}\,H = H\,d^{-1}$, where $0 \neq d \in A$, and $H \in M_{q\times(p-q)}(A)$. Let $R = (d\,I_q:\ -H) \in M_{q\times p}(A)$. Then, we have

$$\begin{cases} D_i\,H = d\,N_i, \\ \\ (\det D_i)\,H = (D_i^c\,d)\,N_i, \end{cases} \Rightarrow \begin{cases} D_i\,R = d\,R_i, \\ \\ (\det D_i)\,R = (D_i^c\,d)\,R_i, \end{cases} \quad i = 1, 2,$$

where D_i^c is the cofactors matrix of D_i, i.e., it satisfies $D_i^c\,D_i = (\det D_i)\,I_q$. Let $\lambda \in A^q\,R_i$; then there exists $\mu \in A^q$ such that $\lambda = \mu\,R_i$, and thus,

$$d\,\lambda = \mu\,(d\,R_i) = \mu\,(D_i\,R) = (\mu\,D_i)\,R \Rightarrow \lambda \in \overline{A^q\,R} \Rightarrow A^q\,R_i \subseteq \overline{A^q\,R}.$$

Conversely, let $\lambda \in A^q\,R$; then there exists $\mu \in A^q$ such that $\lambda = \mu\,R$. Thus,

$$(\det\,D_i)\,\lambda = \mu\,((\det\,D_i)\,R) = \mu\,(D_i^c\,d\,R_i) = (\mu\,D_i^c\,d)\,R_i \Rightarrow \lambda \in \overline{A^q\,R_i} \Rightarrow A^q\,R \subseteq \overline{A^q\,R_i}.$$

Using the fact that $X \subseteq Y \Rightarrow \overline{X} \subseteq \overline{Y}$ for two submodules X and Y of a free A-module, we obtain

$$A^q\,R_i \subseteq \overline{A^q\,R} \subseteq \overline{A^q\,R_j} \quad \Rightarrow \quad \overline{A^q R_i} = \overline{A^q R_j}, \qquad i, j = 1, 2.$$

Now, if $A^q\,R_i$ is A-closed, then

$$A^q\,R_i \subseteq \overline{A^q\,R_j} \subseteq \overline{A^q\,R_i} = A^q\,R_i \Rightarrow \overline{A^q\,R_j} = A^q\,R_i. \qquad \square$$

LEMMA 2.7. *If $R \in M_{q\times p}(A)$ has full row rank and F is a free submodule of $\ker.R^T$ of rank $p-q$, then $\overline{F} = \ker.R^T$, where $\overline{F}$ is the A-closure of F in A^p.*

Proof. Let us note $N \triangleq \operatorname{coker}.R^T$. We have the following exact sequence:

$$0 \longleftarrow N \longleftarrow A^q \xleftarrow{.R^T} A^p \longleftarrow \ker.R^T \longleftarrow 0.$$

The A-module N is defined by the A-linear combinations of the equations $R^T z = 0$, where z_i is the class of the ith vector of the canonical basis of A^q in $N = A^q/A^p R^T$ (see (1.5)). Using the fact that R has full row rank, then there exist q equations of $R^T z = 0$ which are A-linearly independent. If we denote by $R_0^T \in M_q(A)$ the full rank matrix corresponding to these q A-linearly independent equations, then we have $R_0^T z = 0$, and thus, by multiplying R_0^T by its cofactors matrix, we obtain $(\det R_0^T)\, z = 0$ with $0 \neq \det R_0^T \in A$. This equation shows that N is a torsion A-module. Now, let us notice that we have $K \otimes_A N = 0$ because N is a torsion A-module: for all $n \in N$, there exists $0 \neq a \in A :\ a\, n = 0$, and thus, $1 \otimes n = (a/a) \otimes n = (1/a) \otimes a\, n = 0$. Using the fact that $K = Q(A)$ is a flat A-module (see Proposition 1.10), we have the following exact sequence:

$$0 = K \otimes_A N \longleftarrow K^q \xleftarrow{.R^T} K^p \longleftarrow K \otimes_A \ker .R^T \longleftarrow 0.$$

Here $K \otimes_A \ker .R^T$ is a subvector space of K^p of dimension $p - q$. As F is a free submodule of $\ker .R^T$ of rank $p-q$, we have $K \otimes_A F = K \otimes_A \ker .R^T \subset K^p$, and thus,

$$\overline{F} = (K \otimes_A F) \cap A^p = (K \otimes_A \ker .R^T) \cap A^p = \overline{\ker .R^T} = \ker .R^T$$

because $\ker .R^T$ is an A-closed submodule of A^p. Indeed, using the fact that A is an integral domain, we obtain

$$\begin{aligned} \lambda \in \overline{\ker .R^T} &\Rightarrow \exists\, 0 \neq a \in A :\ a\, \lambda \in \ker .R^T \Rightarrow a\, (\lambda\, R^T) = 0 \Rightarrow \lambda\, R^T = 0 \\ &\Rightarrow \lambda \in \ker .R^T. \qquad \square \end{aligned}$$

Proposition 2.8. *Let $P \in M_{q\times(p-q)}(K)$ be such that $P = D^{-1} N = \tilde{N}\, \tilde{D}^{-1}$, where $R = (D :\ -N) \in M_{q\times p}(A)$ and $\tilde{R} = (\tilde{N}^T :\ \tilde{D}^T)^T \in M_{p\times(p-q)}(A)$. If we define the A-modules $M = A^p/A^q\, R$ and $\tilde{M} = A^p/A^{p-q}\, \tilde{R}^T$, then we have*

$$\begin{cases} \ker .R^T = \overline{A^{p-q}\, \tilde{R}^T}, \\ \ker .\tilde{R} = \overline{A^q\, R}, \end{cases} \qquad \begin{cases} \tilde{M}/t(\tilde{M}) \cong A^p\, R^T, \\ M/t(M) \cong A^p\, \tilde{R}. \end{cases}$$

Proof. Using the fact that $R\, \tilde{R} = 0$, we obtain the following two complexes:

$$\begin{aligned} 0 \longrightarrow \quad & A^q \xrightarrow{.R} A^p \xrightarrow{.\tilde{R}} A^{p-q}, \\ & A^q \xleftarrow{.R^T} A^p \xleftarrow{.\tilde{R}^T} A^{p-q} \longleftarrow 0. \end{aligned}$$

Thus, $A^{p-q}\, \tilde{R}^T$ (resp., $A^q\, R$) is a free submodule of $\ker .R^T$ (resp., $\ker .\tilde{R}$) of rank $p - q$ (resp., q). By Lemma 2.7, Proposition 2.4, and Lemma 2.3, we obtain that

$$\begin{cases} \ker .R^T = \overline{A^{p-q}\, \tilde{R}^T}, \\ \ker .\tilde{R} = \overline{A^q\, R}, \end{cases} \Rightarrow \begin{cases} A^p\, R^T \cong A^p/\ker .R^T = A^p/\overline{A^{p-q}\, \tilde{R}^T} = \tilde{M}/t(\tilde{M}), \\ A^p \tilde{R} \cong A^p/\ker .\tilde{R} = A^p/\overline{A^q\, R} = M/t(M). \qquad \square \end{cases}$$

Let us notice that Proposition 2.8 is close in its spirit to some results obtained in [26] for linear multidimensional systems in the behavioral approach.

From Proposition 2.8 and Lemma 2.6, we obtain that the A-modules $A^p\, \tilde{R}$ and $A^p\, R^T$ depend only, up to an isomorphism, on the transfer matrix P. This result was proved in [22] in a different way (without any references to torsion-free A-modules). Using the fact that the structural properties of P do not depend on the choice of the fractional representation of P, we obtain the following corollary.

COROLLARY 2.9. *Let $P \in M_{q\times(p-q)}(K)$ be such that $P = D^{-1}\,N = \tilde{N}\,\tilde{D}^{-1}$, where $R = (D:\ -N) \in M_{q\times p}(A)$ and $\tilde{R} = (\tilde{N}^T:\ \tilde{D}^T)^T \in M_{p\times(p-q)}(A)$. Then, the structural (intrinsic) properties of P depend only on the A-modules $\overline{A^q\,R}$ and $\overline{A^{p-q}\,\tilde{R}^T}$ or, up to an isomorphism, on the A-modules $A^p\,\tilde{R}$ and $A^p\,R^T$.*

2.3. Weakly doubly coprime factorizations. Let us introduce the concepts of weakly left/right/doubly coprime factorizations.

DEFINITION 2.10. *Let A be an integral domain and $K = Q(A)$.*

- *A transfer matrix $P \in M_{q\times(p-q)}(K)$ admits a* weakly left-coprime factorization *if there exists a weakly left-prime matrix $R = (D:\ -N) \in M_{q\times p}(A)$, with $\det D \neq 0$, such that $P = D^{-1}\,N$.*
- *A transfer matrix $P \in M_{q\times(p-q)}(K)$ admits a* weakly right-coprime factorization *if there exists a weakly right-prime matrix $\tilde{R} = (\tilde{N}^T:\ \tilde{D}^T)^T \in M_{p\times(p-q)}(A)$, with $\det\tilde{D} \neq 0$, such that $P = \tilde{N}\,\tilde{D}^{-1}$.*
- *A transfer matrix P admits a* weakly doubly coprime factorization *if P admits weakly left- and right-coprime factorizations $P = D^{-1}\,N = \tilde{N}\,\tilde{D}^{-1}$.*

THEOREM 2.11. *Let A be an integral domain, $K = Q(A)$ its quotient field, $P = D^{-1}\,N = \tilde{N}\,\tilde{D}^{-1} \in M_{q\times(p-q)}(K)$ a transfer matrix, $R = (D:\ -N) \in M_{q\times p}(A)$, and $\tilde{R} = (\tilde{N}^T:\ \tilde{D}^T)^T \in M_{p\times(p-q)}(A)$. Then, $P = D^{-1}\,N$ admits a weakly left-coprime factorization (resp., weakly right-coprime factorization) iff $\overline{A^q\,R}$ (resp., $\overline{A^{p-q}\,\tilde{R}^T}$) is a free A-module of rank q (resp., $p-q$).*

Proof. $\Rightarrow$ If P admits a weakly left-coprime factorization, then there exists a weakly left-prime matrix $R' = (D':\ -N') \in M_{q\times p}(A)$, with $\det D' \neq 0$, such that we have $P = D'^{-1}\,N'$. Using Lemma 2.6, we deduce that $\overline{A^q\,R} = A^q\,R'$. Moreover, $A^q\,R' \cong A^q$ because R' has full row rank, and thus, $\overline{A^q\,R} \cong A^q$.

$\Leftarrow$ If $\overline{A^q\,R}$ is a finitely generated free A-module of rank q, then, choosing a basis for $\overline{A^q\,R}$, we obtain a full row rank matrix $R' \in M_{q\times p}(A)$ such that $\overline{A^q\,R} = A^q\,R'$, and thus, R' is weakly left-prime. If R_i is the ith row of R, then $R_i \in \overline{A^q\,R} = A^q\,R'$ because $R_i \in A^q\,R \subseteq \overline{A^q\,R}$. Therefore, there exists $R''_i \in A^q$ such that $R_i = R''_i\,R'$, and then, there exists $R'' \in M_q(A)$ such that $R = R''\,R'$. Using the fact that R has full row rank, we deduce that R'' also has full row rank. Finally, let $R' = (D':\ N')$, where $R' \in M_{q\times p}(A)$; then we have $D = R''\,D'$ and $N = R''\,N'$, and thus, $\det D' \neq 0$. This proves the result because we have $P = D^{-1}\,N = (R''\,D')^{-1}\,(R''\,N') = D'^{-1}\,N'$.

The result for weak right-coprime factorizations can be proved similarly. □

COROLLARY 2.12. *A transfer matrix $P = D^{-1}\,N = \tilde{N}\,\tilde{D}^{-1}$ admits a weakly doubly coprime factorization iff the A-modules $\overline{A^q\,R}$ and $\overline{A^{p-q}\,\tilde{R}^T}$ are two free A-modules of rank, respectively, q and $p-q$.*

Let us notice that, from Corollary 2.9, Corollary 2.5, and the fact that a coprime factorization is an intrinsic property of the transfer matrix, we deduce that weakly left/right/coprime factorizations are the weakest possible coprime factorizations.

3. Coherent rings and modules.

3.1. Introduction. Any mathematical model of a plant is only an approximation of the real system. Thus, the algebra of SISO stable plants needs to be endowed with a norm in order to take into account some errors in the modelization. For technical reasons, we usually prefer to ask this normed algebra to be complete. Therefore, we generally require an algebra of SISO stable systems to be a *Banach algebra* [16, 49] (e.g., $H_\infty(\mathbb{C}_+)$, $\mathcal{A}$, $\hat{\mathcal{A}}$, $l_1(\mathbb{Z}_+)$ [3, 8, 9, 48, 49]).

However, it is known that all *noetherian k-Banach algebras* ($k = \mathbb{R}, \mathbb{C}$)—namely, Banach algebras such that every ideal is finitely generated—are k-finite-dimensional [41]. Hence, for instance, $H_\infty(\mathbb{C}_+)$, $\hat{\mathcal{A}}$, $L_1(\mathbb{R}_+) + \mathbb{R}\,\delta$, $l_1(\mathbb{Z}_+)$ are not noetherian rings, and thus, an ideal I of these algebras A generally does not have the form $I = \sum_{i=1}^{n} A\,a_i$ for a finite set $\{a_1, \ldots, a_n\}$ of elements of A. A direct consequence is that most of the algebraic objects (kernel, image, quotient, sum, intersection, etc.) are generally not finitely generated. Hence, we cannot study the algebraic properties of systems, defined by matrices whose entries belong to Banach algebras, by means of the concepts and techniques developed for noetherian rings (i.e., the main part of commutative algebra).

The concept of a *coherent ring* was first introduced in 1960 by Chase [5], and the definition of a *coherent module* appeared in [1] in 1964 (see [17] for more details). Coherent rings form a general class of rings including noetherian rings, Boolean rings, Bézout domains, semihereditary rings, etc. [17, 39]. This concept is closely related to the one of a *coherent sheaf* introduced by Cartan [4] and Serre [42] in the study of analytic and algebraic geometries.

In this section, we show that one possible way to cope with the fact that most of the integral domains of SISO stable plants are not noetherian is to require that these domains be coherent rings. In particular, for coherent rings, we give algorithms which compute the A-closure $\overline{A^q\,R}$ of an A-module of the form $A^q\,R$ (see Theorem 2.11) and which check whether or not a finitely generated A-module is torsion-free, reflexive, or projective. Finally, we shall characterize explicitly the class of integral domains A such that every transfer matrix, with entries in $K = Q(A)$, admits a weakly doubly coprime factorization.

3.2. Definitions and results.

DEFINITION 3.1 (see [2, 15, 39]). *We have the following definitions:*

- *An A-module M is* coherent *if M is a finitely generated A-module and every finitely generated submodule of M is finitely presented.*
- *A ring A is* coherent *if it is coherent as an A-module.*

Hence, A is a coherent ring iff every finitely generated ideal $I = \sum_{i=1}^{n} A\,a_i$ of A is finitely presented, i.e., the *module of relations* of I (or *syzygy* of I), defined by

$$S(I) = \left\{ r = (r_1 : \cdots : r_n) \in A^n \;\mid\; \sum_{i=1}^{n} r_i\,a_i = 0 \right\}, \tag{3.1}$$

is finitely generated. In terms of equations, A is a coherent ring iff for every $n \in \mathbb{Z}_+$ and $a = (a_1 : \cdots : a_n)^T \in A^n$ there exist $m \in \mathbb{Z}_+$ and $R \in M_{m\times n}(A)$ such that

$$\forall\, r = (r_1 : \cdots : r_n) \in A^n : \quad r\,a = 0 \Leftrightarrow \exists\, b = (b_1 : \cdots : b_m) \in A^m :\; r = b\,R.$$

Example 3.1. Any *noetherian ring*, namely a ring where any ideal I is finitely generated, i.e., has the form $I = \sum_{i=1}^{n} A\,a_i$ for a finite number of $a_i \in A$, is coherent [1, 39]. In particular, RH_∞ and $k[s]$, with k a field, are coherent domains. An example of a coherent ring which is not noetherian is given by the ring $k[\chi_i, i \in \mathbb{N}]$ of polynomials in infinitely many variables χ_i with coefficients in a field k (see [39]).

We give a few definitions which are related to the extension of (1.4) on the left.

DEFINITION 3.2. *We have the following definitions (see* [1, 14, 24, 39]*):*

- *A* projective *(resp.,* free, flat*)* resolution *of an A-module M is an exact sequence of the form*

$$\cdots \xrightarrow{d_3} F_2 \xrightarrow{d_2} F_1 \xrightarrow{d_1} F_0 \longrightarrow M \longrightarrow 0, \tag{3.2}$$

where F_i is a projective (resp., free, flat) A-module and d_i is an A-morphism.

- *A* finite free resolution *of an A-module M is an exact sequence of the form* (3.2)*, where F_i is a* finite free *A-module, i.e., $F_i \cong A^{r_i}$, $r_i \in \mathbb{Z}_+$, for $i \geq 0$.*
- *The* projective *(resp.,* flat*)* dimension $\mathrm{pd}_A(M)$ *(resp.,* $\mathrm{w.dim}_A(M)$*) of an A-module M is the minimum $n \in \mathbb{Z}_+ \cup \{+\infty\}$ such that there exists a projective (resp., flat) resolution of M of length n, i.e., of the form*

$$0 \longrightarrow F_n \xrightarrow{d_n} F_{n-1} \xrightarrow{d_{n-1}} \cdots \xrightarrow{d_2} F_1 \xrightarrow{d_1} F_0 \longrightarrow M \longrightarrow 0.$$

- *The* global dimension *and* weak global dimension *of A are defined by*

$$\mathrm{gl.dim}(A) = \sup\{\mathrm{pd}_A(M) \mid A - \text{module } M\} \in \mathbb{Z}_+ \cup \{+\infty\},$$
$$\mathrm{w.gl.dim}(A) = \sup\{\mathrm{w.dim}_A(M) \mid A - \text{module } M\} \in \mathbb{Z}_+ \cup \{+\infty\}.$$

For a general ring A, we have the inequality $\mathrm{w.gl.dim}(A) \leq \mathrm{gl.dim}(A)$. If A is a noetherian ring, then the equality holds [2, 39].

Remark 3.1. Using the canonical basis of the free A-module $F_i \cong A^{r_i}$, every finite free resolution of an A-module M has the form

$$\cdots \xrightarrow{.R_2} A^{r_1} \xrightarrow{.R_1} A^{r_0} \longrightarrow M \longrightarrow 0, \tag{3.3}$$

where R_i is an $(r_i \times r_{i-1})$-matrix whose entries belong to A, and $.R_i : A^{r_i} \to A^{r_{i-1}}$ is defined by letting operate a row vector of length r_i on the left of R_i to obtain a row vector of length r_{i-1}. Moreover, M is defined by the system $R_1\, z = 0$, where z_i is the class of e_i in M and $\{e_1, \ldots, e_{r_0}\}$ is the canonical basis of A^{r_0} (see (1.5)).

Definition 3.3. *We have the following definitions (see* [1, 39]*):*

- *A ring A is* semihereditary *if every finitely generated ideal of A is a projective A-module.*
- *A semihereditary integral domain is called a* Prüfer domain.
- *A ring A is a* Bézout domain *if every finitely generated ideal of A is a principal ideal, i.e., generated by a single element of A.*
- *A ring A is* hereditary *if every ideal of A is a projective A-module.*
- *A hereditary integral domain is called a* Dedekind domain.
- *A ring A is a* principal ideal domain *if every ideal of A is generated by a single element of A.*

We shall give in [33] some examples of Prüfer and Dedekind domains, as they will play an important role in internal stabilizability. Coherent rings with small weak global dimensions have been studied and classified largely in algebra [17, 39, 47].

Theorem 3.4 (see [2, 39]). *We have the following results:*

1. *Semihereditary rings and Prüfer domains are coherent rings.*
2. *Hereditary rings and Dedekind domains are noetherian and, thus, coherent rings.*
3. *If A is an integral domain, then* $\mathrm{w.gl.dim}(A) \leq 1 \Leftrightarrow A$ *is a Prüfer domain.*
4. *If A is an integral domain, then* $\mathrm{gl.dim}(A) \leq 1 \Leftrightarrow A$ *is a Dedekind domain.*

We have the following inclusions of rings:

Noetherian rings:	Principal ideal domains	$\subset$	Dedekind domains
$\cap$	$\cap$		$\cap$
Coherent rings:	Bézout domains	$\subset$	Prüfer domains

Example 3.2. The integral domains $E(k)$, $k = \mathbb{R}$, $\mathbb{C}$, and $\mathcal{E} = E(\mathbb{R}) \cap \mathbb{R}(s)[e^{-s}]$, defined in Theorem 1.6, are Bézout domains and, thus, two coherent rings.

PROPOSITION 3.5 (see [15, 47]). *An integral domain with* $\mathrm{gl.dim}(A) \leq 2$ *is coherent.*

General rings with $\mathrm{w.gl.dim}(A) = 2$ are less understood [14, 17, 24, 47].

PROPOSITION 3.6 (see [1, 39]). *If A is a coherent ring, then an A-module M is coherent iff M is a finitely presented A-module.*

DEFINITION 3.7. *We call an* A-system *a system of the form $R\,z = 0$, where z is a set of formal variables and R is a finite matrix whose entries belong to A.*

From Proposition 3.6, we have the following corollary.

COROLLARY 3.8. *If A is a coherent ring, then there is a one-to-one correspondence between coherent A-modules and A-systems.*

Proof. $\Rightarrow$ Let $\sum_{j=1}^{p} R_{ij}\ z_j = 0,\ R_{ij} \in A,\ i = 1, \ldots, q$, be an A-system and $R = (R_{ij}) \in M_{q\times p}(A)$. Let us define the following A-morphism:

$$\begin{array}{rcl} .R = A^q & \longrightarrow & A^p, \\ (a_1 : \cdots : a_q) & \longrightarrow & (a_1 : \cdots : a_q)\, R. \end{array}$$

If we note $M = \mathrm{coker}\, .R = A^p/A^q\, R$, then we have the following exact sequence,

$$A^q \xrightarrow{.R} A^p \xrightarrow{\pi} M \longrightarrow 0, \tag{3.4}$$

and, by Proposition 3.6, M is a coherent A-module because A is a coherent ring.

$\Leftarrow$ Let M be a coherent A-module. Using the fact that A is a coherent ring, by Proposition 3.6, M is a finitely presented A-module, there exists an exact sequence of the form (3.4), and, thus, M is defined by means of a system of equations of the form $R\,z = 0$. □

3.3. Elementary algebraic operations. The next proposition shows that the class (category) of finitely presented A-modules over a coherent ring A, i.e., coherent modules, is invariant under elementary algebraic operations. First, let us notice that any finitely generated submodule of a coherent module is also coherent.

PROPOSITION 3.9 (see [1, 39]). *If two terms in the exact sequence*

$$0 \longrightarrow M' \longrightarrow M \longrightarrow M'' \longrightarrow 0$$

are coherent A-modules, so is the third one.

COROLLARY 3.10 (see [1, 39]). *Let M, N, $M' \subset M$, $M'' \subset M$ be coherent A-modules, $\phi : M \longrightarrow N$ an A-morphism, I a coherent ideal, and S a multiplicative set of A. Then, we have the following:*

1. *$M/M', M \oplus N, M' \cap M'', M' + M''$ are coherent A-modules.*
2. *$\ker\,\phi$, $\mathrm{im}\,\phi$, $\mathrm{coker}\,\phi$, and $\mathrm{coim}\,\phi$ are coherent A-modules.*
3. *$M \otimes_A N$ and $\hom_A(M, N)$ are coherent A-modules.*
4. *$S^{-1}A$ is a coherent A-module.*
5. *$IM = \{\sum_{i=1}^{n} a_i\, m_i \mid a_i \in I,\ m_i \in I\}$ is a coherent A-module.*
6. *$\mathrm{ann}(M) = \{a \in A \mid a\, M = 0\}$ is a coherent ideal of A.*

COROLLARY 3.11. *Let A be a coherent ring and M a finitely presented A-module. Then there exists a finite free resolution of M of the form* (3.3).

Proof. Using Proposition 3.9, we prove by induction that every finite power A^{r_i} of A is a coherent A-module (take $M = A^n, M' = A^{n-1}, M'' = A$). The kernel of a homomorphism d_i between two coherent A-modules is a coherent A-module and, by Proposition 3.6, is a finitely presented A-module. Then, the module of relations of R_i is finitely presented, and thus, M has a finite free resolution. □

DEFINITION 3.12. *Let M be an A-module with a projective resolution of the form* (3.2) *and N another A-module. Then we have the following definitions:*

- *The defects of exactness of*

$$\cdots \xleftarrow{d_3^\star} \hom_A(F_2,N) \xleftarrow{d_2^\star} \hom_A(F_1,N) \xleftarrow{d_1^\star} \hom_A(F_0,N) \longleftarrow 0, \tag{3.5}$$

where $d_i^\star$ *is defined by* $d_i^\star(f) = f \circ d_i$ *for all* $f \in \hom_A(F_{i-1},N)$, *depend only on* M *and* N *and not on* (3.2) *and are called* $\operatorname{ext}_A^i(M,N)$ *(see* [2, 29, 39]*). In particular, we have*

$$\begin{cases} \operatorname{ext}_A^0(M,N) = \ker d_1^\star = \hom_A(M,N), \\ \operatorname{ext}_A^i(M,N) = \ker d_{i+1}^\star / \operatorname{im} d_i^\star, \ i \geq 1. \end{cases}$$

- *The defects of exactness of*

$$\cdots \xrightarrow{\mathrm{id}_N \otimes d_3} N \otimes_A F_2 \xrightarrow{\mathrm{id}_N \otimes d_2} N \otimes_A F_1 \xrightarrow{\mathrm{id}_N \otimes d_1} N \otimes_A F_0 \longrightarrow 0, \tag{3.6}$$

where $\mathrm{id}_N \otimes d_i$ *is defined by* $(\mathrm{id}_N \otimes d_i)(n \otimes m) = n \otimes d_i(m)$ *for all* $n \in N$, *for all* $m \in F_i$, *depend only on* M *and* N *and not on* (3.2) *and are called* $\operatorname{tor}_i^A(M,N)$ *(see* [2, 29, 39]*). In particular, we have*

$$\begin{cases} \operatorname{tor}_0^A(M,N) = \operatorname{coker}(\mathrm{id}_N \otimes d_1) = N \otimes_A M, \\ \operatorname{tor}_i^A(M,N) = \ker(\mathrm{id}_N \otimes d_i)/\operatorname{im}(\mathrm{id}_N \otimes d_{i+1}), \quad i \geq 1. \end{cases}$$

Remark 3.2. If M has a finite free resolution of the form (3.3), then (3.5) is defined by $\cdots \xleftarrow{R_3.} N^{r_2} \xleftarrow{R_2.} N^{r_1} \xleftarrow{R_1.} N^{r_0} \longleftarrow 0$, where $R_i. : N^{r_{i-1}} \to N^{r_i}$ is defined by letting operate a column vector of length r_{i-1}, whose entries belong to N on the right of R_i, to obtain a column vector of length r_i, whose entries belong to N. We have

$$\operatorname{ext}_A^i(M,N) = \ker_N(R_{i+1}.)/\operatorname{im}_N(R_i.) \qquad \forall\, i \geq 1.$$

Similarly, (3.6) becomes the complex $\cdots \xrightarrow{.R_3} N^{r_2} \xrightarrow{.R_2} N^{r_1} \xrightarrow{.R_1} N^{r_0} \longrightarrow 0$, where $.R_i : N^{r_i} \to N^{r_{i-1}}$ is defined by letting operate a row vector of length r_i, whose entries belong to N on the right of R_i, to obtain a row vector of length r_{i-1}, whose entries belong to N and

$$\operatorname{tor}_i^A(M,N) = \ker_N .R_i / \operatorname{im}_N .R_{i+1} \qquad \forall\, i \geq 1.$$

Proposition 3.13 (see [2, 39]). *We have the following results:*

- $\operatorname{ext}_A^i(M,N) = 0\ \forall\, i \geq 1, \forall\, N\ A-$*module* $\Leftrightarrow M$ *is a projective* A*-module.*
- $\operatorname{tor}_A^i(M,N) = 0\ \forall\, i \geq 1, \forall\, N\ A-$*module* $\Leftrightarrow M$ *is a flat* A*-module.*

Corollary 3.14. *If* A *is a coherent ring, and* M *and* N *are two coherent* A*-modules, then* $\operatorname{ext}_A^i(M,N)$ *and* $\operatorname{tor}_i^A(M,N)$ *are coherent* A*-modules for* $i \geq 0$. *Moreover,* $\operatorname{ext}_A^i(M,A)$ *is a torsion* A*-module for* $i \geq 1$.

Proof. Using the fact that $\operatorname{ext}_A^i(M,N)$ (resp., $\operatorname{tor}_i^A(M,N)$) does not depend on the projective resolution of M, by Proposition 3.6 and Corollary 3.11, we choose a finite free resolution (3.3) for M. By Proposition 3.9, $\hom_A(F_i,N)$ (resp., $N \otimes_A F_i$) is a coherent A-module, and thus, $\ker d_i^\star$ and $\operatorname{im} d_i^\star$ (resp., $\ker(\mathrm{id}_N \otimes d_i)$ and $\operatorname{im}(\mathrm{id}_N \otimes d_i)$) are coherent A-modules. Finally, $\operatorname{ext}_A^i(M,N)$ (resp., $\operatorname{tor}_i^A(M,N)$) is also a coherent A-module for $i \geq 0$ as a quotient of two coherent A-modules. The proof of the fact that $\operatorname{ext}_A^i(M,A)$ is a torsion A-module is the same as that of Lemma 1 in [28]. ∎

DEFINITION 3.15. *Let M be an A-module defined by a finite presentation:*

$$F_1 \xrightarrow{d_1} F_0 \longrightarrow M \longrightarrow 0.$$

We call the transposed module[1] *of M, the A-module $T(M) = \operatorname{coker} d_1^\star$ defined by*

$$0 \longleftarrow T(M) \longleftarrow F_1^\star \xleftarrow{d_1^\star} F_0^\star.$$

Hence, if $M = A^p/A^q\,R$, then the transposed module is $T(M) = A^q/R\,A^p$, where the vectors of A^q and A^p are now column ones (duality). Using the fact that A is commutative, we finally have $T(M) = A^q/A^p\,R^T$, where we use only row vectors.

If A is a coherent ring and M a coherent A-module, i.e., finitely presented A-module, then $T(M)$ is also a coherent A-module because it is finitely presented.

Remark 3.3. We commit a little abuse of notation in denoting the transposed A-module of M by $T(M)$: $\operatorname{coker} d_1^\star$ depends on the particular choice of d_1, i.e., on the particular form of the system of equations chosen to represent the module. However, we have (see [29]):

1. If R has full row rank, then $T(M)$ depends only on M and not on R.
2. If R does not have full rank, i.e., $\ker .R \neq 0$, then $\operatorname{coker} d_1^\star$ depends only on M up to *a projective equivalence* [39], a fact which shows that $\operatorname{ext}_A^i(T(M), N)$ depends only on M and N for $i \geq 1$.

The next theorem shows how to characterize the module properties in terms of the extension and torsion functors.

THEOREM 3.16. *Let A be a coherent ring with* $\operatorname{w.gl.dim}(A) \leq n$, *$M$ a finitely presented A-module, and $T(M)$ its transposed A-module. Then, we have*

1. $t(M) \cong \operatorname{ext}_A^1(T(M), A)$,
2. $t(M) \cong \operatorname{tor}_1^A(K/A, M)$,
3. *M is torsion-free iff* $\operatorname{ext}_A^1(T(M), A) = 0$,
4. *M is reflexive iff* $\operatorname{ext}_A^i(T(M), A) = 0,\ i = 1, 2$,
5. *M is projective iff* $\operatorname{ext}_A^i(T(M), A) = 0,\ i = 1, \ldots, n$.

Proof. The proofs of 1, 3, 4, 5 are the same as those given in [27, 28] for noetherian rings: we just need to change finitely generated modules (resp., noetherian rings) into finitely presented (resp., coherent) ones. See also the proof of Proposition 3.4 of [33]. For a proof of 2, see [39]. □

Using Proposition 2.4, Lemma 2.3, and Theorem 3.16, we obtain an algorithm which computes the closure $\overline{A^q\,R}$ of an A-module of the form $A^q\,R$.

ALGORITHM. Input: A coherent ring A and $R \in M_{q\times p}(A)$. Output: $R' \in M_{r\times p}(A)$ such that $\overline{A^q\,R} = A^r\,R'$.

1. Start with $R \in M_{q\times p}(A)$.
2. Transpose R to obtain $R^T \in M_{p\times q}(A)$.
3. Find a family of generators of $\ker .R^T = \{\lambda \in A^p \mid \lambda\,R^T = 0\}$. If $\{\lambda_1, \ldots, \lambda_m\}$ is a family of generators of $\ker .R^T$, then denote by $R_{-1}^T \in M_{m\times p}(A)$ the matrix whose ith row is λ_i.
4. Transpose R_{-1}^T to obtain $R_{-1} \in M_{p\times m}(A)$.
5. Find a family of generators of $\ker .R_{-1} = \{\eta \in A^p \mid \eta\,R_{-1} = 0\}$. If $\{\eta_1, \ldots, \eta_r\}$ is a family of generators of $\ker .R_{-1}$, then denote by $R' \in M_{r\times p}(A)$ the matrix whose ith row is η_i. We have $\overline{A^q\,R} = A^r\,R'$, $A^p/\overline{A^q R} \cong A^p R_{-1}$.

[1] Do not confuse the notation of the transposed module $T(M)$ of an A-module M with the torsion submodule $t(M)$ of M.

If R' has full row rank, then $\overline{A^q R} = A^r R'$ is a free A-module. (See Example 3.4 for the explicit computations of the A-closure $\overline{A^q R}$ of a certain A-module $A^q R$.) To finish, let us note that we can use the previous algorithm to check whether a transfer matrix admits a weakly left/right/doubly coprime factorization (see Theorem 2.11). Indeed, the previous algorithm allows us to have a precise description of the A-closure of an A-module of the form $A^q R$. However, checking whether such an A-closure is free can be a very difficult algebraic problem (see, e.g., the proof of the Quillen–Suslin theorem in [39]).

3.4. Coherent Sylvester domains.

DEFINITION 3.17 (see [6, 10]). *A projective-free coherent domain with*

$$\text{w.gl.dim}(A) \leq 2$$

is called a coherent Sylvester domain.

Example 3.3. A Bézout domain (e.g., $E(k)$, $\mathcal{E}$ by Theorem 1.6) and thus, a principal ideal domain (e.g., RH_∞ by Theorem 1.6, $k[s]$, with k a field) are coherent Sylvester domains. More generally, $A = B[x]$ is a coherent Sylvester domain iff B is a Bézout domain [11] (e.g., $A = \mathbb{Z}[x]$, $A = k[s][z] = k[s, z]$, with k a field, or $A = B[x]$, where B is the ring of all algebraic integers, i.e., the integral closure of $\mathbb{Z}$ in $\mathbb{C}$; see [39]).

DEFINITION 3.18. *A ring A is* regular *if every finitely generated ideal of A has a finite projective dimension.*

THEOREM 3.19 (see [50]). *A coherent regular domain A is a GCDD—every a and b of A have a greatest common divisor $[a, b]$—iff every finitely generated projective ideal of A is principal.*

COROLLARY 3.20.[2] *A coherent Sylvester domain is a GCDD.*

Proof. A coherent Sylvester domain is a projective-free coherent domain with $\text{w.gl.dim}(A) \leq 2$ and, thus, a regular ring which satisfies that every finitely generated projective ideal is free, i.e., is principal, because A is an integral domain. Then, the result follows directly from Theorem 3.19. □

PROPOSITION 3.21. *If A is a coherent Sylvester domain, then, for every A-module M defined by a finite free resolution,*

$$F_1 \xrightarrow{d_1} F_0 \longrightarrow M \longrightarrow 0,$$

there exist a free A-module F_1' and two A-morphisms $d_1' : F_1' \to F_0$ and $d_1'' : F_1 \to F_1'$ such that $d_1 = d_1' \circ d_1''$ and we have the following exact sequences:

$$0 \longrightarrow F_1' \xrightarrow{d_1'} F_0 \longrightarrow M/t(M) \longrightarrow 0, \tag{3.7}$$

$$0 \longrightarrow \ker d_1 \longrightarrow F_1 \xrightarrow{d_1''} F_1' \longrightarrow t(M) \longrightarrow 0. \tag{3.8}$$

[2] We would like to thank Prof. W. Dicks for pointing out to us that this result is already contained in Lemma 4.1 of [11].

Proof. We have the following commutative exact diagram:

$$\begin{array}{ccccccccc}
 & & 0 & & & & 0 & & \\
 & & \downarrow & & & & \downarrow & & \\
 & & \ker d_1'' & & 0 & & t(M) & & \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow \ker d_1 \longrightarrow & F_1 & \xrightarrow{d_1} & F_0 & \xrightarrow{\pi} & M & \longrightarrow 0 & \\
 & & \downarrow d_1'' & & \| & & \downarrow \pi' & & \\
0 & \longrightarrow & \ker\,\phi & \longrightarrow & F_0 & \xrightarrow{\phi} & M/t(M) & \longrightarrow 0, & \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
 & & \operatorname{coker} d_1'' & & 0 & & 0 & & \\
 & & \downarrow & & & & & & \\
 & & 0 & & & & & &
\end{array}$$

where $\phi = \pi' \circ \pi$ and $d_1'' : F_1 \to \ker \phi$ is induced by the identity homomorphism from F_0 to F_0 and $\pi' : M \to M/t(M)$. An easy chase in the diagram shows that $\ker d_1'' \cong \ker d_1$ and $\operatorname{coker} d_1'' \cong t(M)$. Now, let us prove that $\ker \phi$ is a finite free A-module. $M/t(M)$ is a coherent A-module over a coherent ring A [15] and, in particular, $M/t(M)$ is a finitely generated A-module. It is well known that $M/t(M)$ can be imbedded into a finitely generated free A-module F_{-1} (see, e.g., [39] or [15]), and we have the following exact sequence:

$$0 \longrightarrow M/t(M) \longrightarrow F_{-1} \longrightarrow F_{-1}/(M/t(M)) \longrightarrow 0.$$

Hence, we have the following exact sequence:

$$0 \longrightarrow \ker \phi \longrightarrow F_0 \xrightarrow{\phi} F_{-1} \longrightarrow F_1/(M/t(M)) \longrightarrow 0.$$

Using the fact that $\text{w.gl.dim}(A) = 2$, we then have $\operatorname{pd}_A(F_1/(M/t(M))) \leq 2$, and thus, $\operatorname{pd}_A(\ker \phi) = 0$ [2, 39], i.e., $\ker \phi$ is a projective A-module, and thus, a free A-module because A is a projective-free ring. $\ker \phi$ is a finitely generated A-module because $\ker \phi$ is a coherent A-module. ($\ker \phi$ is the kernel of an A-morphism between two finite free A-modules.) Thus, $\ker \phi \cong F_1' \cong A^r$, $r \in \mathbb{Z}_+$, which gives (3.7) and (3.8). □

In the next corollary, we reformulate Proposition 3.21 in terms of weakly doubly coprime factorizations. It generalizes to coherent Sylvester domains a result obtained by Smith in [43] for $H_\infty(\mathbb{C}_+)$ (see section 3.5).

COROLLARY 3.22. *If A is a coherent Sylvester domain, then every matrix P, whose entries belong to $K = Q(A)$, admits a weakly doubly coprime factorization.*

Proof. If $P \in M_{q\times(p-q)}(K)$, then we can define $R = (dI_q : -H) \in M_{q\times p}(A)$, where d is the product of the denominators of the entries of P and $H = dP \in M_{q\times(p-q)}(A)$. By Proposition 3.21, there exist a full rank matrix $R'' \in M_q(A)$ and a weakly left-prime full row rank matrix $R' = (D : -N) \in M_{q\times p}(A)$ such that $R = R''R'$. Thus, we have $(dI_q : -H) = R''(D : -N)$ and $\det R'' \neq 0$. Then, we have

$$\begin{cases} dI_q = R''D \\ H = R''N, \end{cases} \Rightarrow \det D \neq 0, \quad \Rightarrow P = (dI_q)^{-1}H = (R''D)^{-1}(R''N) = D^{-1}N.$$

Dually, we have $P = G(dI_{p-q})^{-1}$, and thus, there exists a weakly right-prime matrix $\tilde{R}' = (\tilde{N}^T : \tilde{D}^T)^T$ such that $\tilde{R} = (G^T : (dI_{p-q})^T)^T = (\tilde{N}^T : \tilde{D}^T)^T\tilde{R}''$.

Therefore, $P = G\,(d\,I_{p-q})^{-1} = (\tilde{N}\,\tilde{R}'')\,(\tilde{D}\,\tilde{R}'')^{-1} = \tilde{N}\,\tilde{D}^{-1}$ is a weakly right-coprime factorization. □

To prove the next theorem, we shall need the next proposition due to Dicks.

PROPOSITION 3.23 (based on [12]). *Let A be an integral domain. If, for every finitely generated free A-module F_0 and every finitely generated free A-submodule F_1 of F_0, the A-closure of F_1 in F_0 is a finitely generated free A-module, then A is a coherent Sylvester domain.*

Proof. Let $K = Q(A)$, p, $q \in \mathbb{Z}_+$, $F_0 = A^p$, and R be any matrix belonging to $M_{p\times q}(A)$. We have the exact sequence $A^q \xleftarrow{.R} A^p \longleftarrow \ker .R \longleftarrow 0$. Applying the tensor product $K\otimes_A$ to the previous exact sequence, we obtain the exact sequence (K is a flat A-module) $K^q \xleftarrow{.R} K^p \longleftarrow K\otimes_A \ker .R \longleftarrow 0$. Therefore, $K\otimes_A \ker .R$ is a K-subvector space of K^p, and thus, there exists a finite basis $\{e_1, \ldots, e_m\}$ of $K\otimes_A \ker .R$, where $m = \dim_K(K\otimes_A \ker .R) \leq p$. Let us note $e_i = f_i/d_i$, with $f_i \in A^p$ and $0 \neq d_i \in A$, and let F_1 be the A-submodule of K^p generated by $\{f_1, \ldots, f_m\}$. Then, F_1 is a free A-submodule of $\ker .R$. Thus, $\overline{F_1} \subseteq \overline{\ker .R} = \ker .R$, because $\ker .R$ is an A-closed submodule of A^q. Moreover, for every $\lambda \in \ker .R$, we have $\lambda = \sum_{i=1}^m a_i\,e_i$, with $a_i \in K$, and clearing the denominators of a_i and $e_i = f_i/d_i$, there exists $0 \neq a \in A$ such that $a\,\lambda \in F_1$, i.e., $\lambda \in \overline{F_1}$, and thus $\overline{F_1} = \ker .R \subseteq A^p = F_0$. Then, by hypothesis, $\ker .R$ is a finitely generated free A-module. Using the implication (v) $\Rightarrow$ (i) of Theorem 10 of [10] (namely, the annihilator of every matrix is free $\Rightarrow$ A is a coherent Sylvester domain), we obtain that A is a coherent Sylvester domain. □

The next theorem characterizes the integral domains over which every transfer matrix admits a weakly doubly coprime factorization.

THEOREM 3.24. *We have the following equivalences:*

1. *Every multi-input multi-output (MIMO) plant admits a weakly doubly coprime factorization.*
2. *A is a coherent Sylvester domain.*

Proof. 1 $\Rightarrow$ 2. Let F_0 be any finitely generated free A-module, and suppose that $F_0 = A^p$ for a certain positive integer p. Let F_1 be any finitely generated free A-submodule of F_0, and suppose that F_1 has rank q. Taking a basis for F_1, then there exists a full row rank matrix $R \in M_{q\times p}(A)$ such that we have $F_1 = A^q\,R$. We can always suppose that R can be written as $R = (D :\ -N)$, where D is a full rank matrix. Then, by hypothesis, $P = D^{-1}\,N$ has a weakly doubly coprime factorization, i.e., there exists a weakly left-prime matrix $R' = (D' :\ -N') \in M_{q\times p}(A)$ such that $\det D' \neq 0$ and $P = {D'}^{-1}\,N'$. Then, by Lemma 2.6 and Theorem 2.11, we have $\overline{A^q\,R} = A^q\,R'$ and, using the fact that $A^q\,R'$ is a free A-module of rank q, we obtain that $\overline{F_1} = \overline{A^q\,R}$ is a finite free A-submodule of F_0. From Proposition 3.23, it follows that A is a coherent Sylvester domain.

2 $\Rightarrow$ 1 was already proved in Corollary 3.22. □

3.5. An example: $H_\infty(\mathbb{C}_+)$.

THEOREM 3.25 (see [23, 37]). *If D is a finitely connected domain of $\mathbb{C}$, then $H_\infty(D)$ is a coherent domain. In particular, if we denote the open right half-plane by $\mathbb{C}_+ = \{s \in \mathbb{C} \mid \mathrm{Re}\,s > 0\}$ and the open unit disc by $\mathbb{D} = \{z \in \mathbb{C} \mid |z| < 1\}$, then $H_\infty(\mathbb{C}_+)$ and $H_\infty(\mathbb{D})$ are two coherent integral domains.*

In the rest of this paper, we shall consider only the case $D = \mathbb{D}$ and, by extension, $D = \mathbb{C}_+$. The proof of the coherence of $H_\infty(D)$ is based on the following theorem, which is a weak-* version of the Beurling–Lax theorem [25]. The condition on m is given by point 2 of the *lemma on the local rank* (p. 44) and Remark (p. 45) of [25].

THEOREM 3.26. *Let $R \in M_{q\times p}(H_\infty(D))$, and let us define the $H_\infty(D)$-morphism:*

$$\begin{array}{rccl} R: & H_\infty(D)^p & \longrightarrow & H_\infty(D)^q, \\ & (a_1 : \cdots : a_p)^T & \longrightarrow & R\,(a_1 : \cdots : a_p)^T. \end{array}$$

Then, there exists $R_{-1} \in M_{p\times m}(H_\infty(D))$ such that

$$\ker R = R_{-1}\, H_\infty(D)^m, \tag{3.9}$$

$$(R_{-1}(e^{i\,\theta}))^\star\, R_{-1}(e^{i\,\theta}) = I_m \qquad \text{for almost every } \theta \in [0,\, 2\pi), \tag{3.10}$$

$$m = p - \operatorname{rank} R, \tag{3.11}$$

where $\operatorname{rank} R$ *is the number of $H_\infty(D)$-linearly independent rows of R.*

COROLLARY 3.27. *Let $A = H_\infty(D)$. If M is a finitely presented A-module, then*

$$\operatorname{pd}_A(M) \leq 2.$$

Proof. Let $A = H_\infty(D)$ and $A^q \xrightarrow{.R} A^p \longrightarrow M \longrightarrow 0$ be a finite presentation of M. Using Theorem 3.26, up to a transposition, there exists an $r \times q$-matrix R_1 whose entries belong to A such that we have the following exact sequence:

$$0 \longrightarrow \ker .R_1 \longrightarrow A^r \xrightarrow{.R_1} A^q \xrightarrow{.R} A^p \longrightarrow M \longrightarrow 0. \tag{3.12}$$

From the exactness of (3.12), we obtain (see Definition 1.9 and Proposition 1.10)

$$\operatorname{rank}(\ker .R_1) + \operatorname{rank} M = r + p - q.$$

From the exact sequence $0 \longrightarrow \operatorname{im}.R \longrightarrow A^p \longrightarrow M \longrightarrow 0$, we obtain $\operatorname{rank} M = p - \operatorname{rank} R$. From (3.11), we have $r = q - \operatorname{rank} R$, and thus, $\operatorname{rank}(\ker .R_1) = 0$, i.e., $\ker .R_1$ is a torsion A-module. However, $\ker .R_1$ is a submodule of the free A-module A^q, and thus, $\ker .R_1 = 0$ because a free module is torsion-free. Hence, every finitely presented A-module M has a finite free resolution of length at most 2, i.e., $\operatorname{pd}_A(M) \leq 2$. □

COROLLARY 3.28. *$H_\infty(D)$ has a weak global dimension 2, i.e.,*

$$\operatorname{w.gl.dim}(H_\infty(D)) = 2. \tag{3.13}$$

Proof. Let $A = H_\infty(D)$. Using Corollary 3.27 and the fact that every finitely presented flat module is projective (see 3 of Proposition 1.4), then every finitely presented A-module M has a finite flat resolution of length at most 2, i.e., $\operatorname{w.dim}_A(M) = \operatorname{pd}_A(M) \leq 2$. Moreover, $\operatorname{w.gl.dim}(A)$ is attained by taking the supremum of the weak dimension of finitely presented modules [14, 24], and thus, $\operatorname{w.gl.dim}(A) \leq 2$. In Example 4.3, we shall give a finitely presented torsion-free $H_\infty(\mathbb{C}_+)$-module which is not projective (similar examples can be exhibited for $D = \mathbb{D}$). Thus, we have $\operatorname{w.gl.dim}(A) = 2$. □

The next corollary follows directly from the fact that $\operatorname{w.gl.dim}(H_\infty(D)) = 2$.

COROLLARY 3.29. *$H_\infty(D)$ is a regular ring.*

The following corollary was first proved in [46] for full row rank matrices.

COROLLARY 3.30. *$H_\infty(D)$ is a projective-free integral domain.*

Proof. Let $A = H_\infty(D)$. Every finitely generated projective module is finitely presented [2]. Hence, let us suppose that M is a finitely presented projective A-module defined by a finite free resolution $F_1 \xrightarrow{d_1} F_0 \longrightarrow M \longrightarrow 0$. Then, $T(M)$ is a coherent

A-module and, by Theorem 3.26, $T(M)$ has a finite free resolution of length 2 of the form $0 \longleftarrow T(M) \longleftarrow F_1^\star \xleftarrow{d_1^\star} F_0^\star \xleftarrow{d_2^\star} F_{-1}^\star \longleftarrow 0$. The fact that M is a projective A-module implies that $\mathrm{ext}_A^i(T(M), A) = 0$, $i \geq 1$ (see Theorem 3.16), and thus, $F_1 \xrightarrow{d_1} F_0 \xrightarrow{d_2} F_{-1} \longrightarrow 0$ is an exact sequence; i.e., $M = \mathrm{coker}\, d_1 \cong \mathrm{im}\, d_2 = F_{-1}$ is a free A-module. □

COROLLARY 3.31. *$H_\infty(D)$ is a coherent Sylvester domain and, thus, a GCDD.*

The fact that $H_\infty(D)$ is a GCDD was first proved in [36] (see also [43]).

Corollary 3.28 shows that, for any finitely presented $A = H_\infty(D)$–module M, we have $\mathrm{ext}_A^i(T(M), A) = 0$ for all $i \geq 3$. Hence, by Theorem 3.16, every finitely presented A-module M satisfies only one of the three following possibilities: M has a nontrivial torsion submodule, M is torsion-free but not free, or M is free.

Example 3.4. In Example 2.1, we proved that the factorization $P = D^{-1}\, N$, defined by (1.3), of the transfer matrix (1.2) was not weakly left-coprime. Let us notice that $R = (D : -N)$ was obtained by clearing the denominators of P once all its entries were written as quotients of (stable) elements of $A = H_\infty(\mathbb{C}_+)$. Hence, clearing the denominators of P does not generally lead to weakly doubly coprime factorizations. In general, we need to use the algorithm developed at the end of section 3.3 to compute a weakly doubly coprime factorization of a transfer matrix. Let us compute a weakly left-coprime factorization of the transfer matrix (1.2).

1. Let us reconsider the A-module $M = A^4/A^2\, R$ defined in Example 1.4. The matrix $R \in M_{2\times 4}(A)$ has full row rank, and thus, we have the finite free presentation

$$0 \longrightarrow A^2 \xrightarrow{.R} A^4 \longrightarrow M \longrightarrow 0.$$

2. The transposed A-module $T(M)$ is defined by the exact sequence

$$0 \longleftarrow T(M) \longleftarrow A^2 \xleftarrow{.R^T} A^4 \longleftarrow \ker .R^T \longleftarrow 0.$$

3. Let $\lambda = (\lambda_1 :\ \lambda_2 :\ \lambda_3 :\ \lambda_4)^T \in \ker .R^T$; then we have

$$\begin{cases} \frac{(s-1)}{(s+1)}\, \lambda_1 - \frac{(s-1)\, e^{-s}}{(s+1)^2}\, \lambda_3 - \left(\frac{s-1}{s+1}\right)^2 \lambda_4 = 0, \\ \frac{(s-1)}{(s+1)}\, \lambda_2 - \frac{1}{(s+1)}\, \lambda_4 = 0. \end{cases} \tag{3.14}$$

By Corollary 3.31, A is a GCDD. The greatest common factor of $\frac{s-1}{s+1}$ and $\frac{1}{s+1}$ is 1; thus, from the second equation of (3.14), we have

$$\begin{cases} \lambda_2 = \frac{1}{(s+1)}\, \mu_1, \\ \lambda_4 = \frac{(s-1)}{(s+1)}\, \mu_1, \end{cases} \qquad \mu_1 \in A.$$

Substituting λ_4 in the first equation of (3.14), we obtain

$$\frac{(s-1)}{(s+1)} \left(\lambda_1 - \frac{e^{-s}}{(s+1)}\, \lambda_3 - \left(\frac{s-1}{s+1}\right)^2 \mu_1 \right) = 0 \Rightarrow \lambda_1 = \frac{e^{-s}}{(s+1)}\, \lambda_3 + \left(\frac{s-1}{s+1}\right)^2 \mu_1$$

because A is an integral domain and $\lambda_i, \mu_1 \in A$. Finally, we have

$$\begin{cases} \lambda_1 = \left(\frac{s-1}{s+1}\right)^2 \mu_1 + \frac{e^{-s}}{(s+1)}\, \mu_2, \\ \lambda_2 = \frac{1}{(s+1)}\, \mu_1, \\ \lambda_3 = \mu_2, \\ \lambda_4 = \frac{(s-1)}{(s+1)}\, \mu_1, \end{cases} \Leftrightarrow (\lambda_1 :\ \lambda_2 :\ \lambda_3 :\ \lambda_4) = (\mu_1 :\ \mu_2) \begin{pmatrix} (\frac{s-1}{s+1})^2 & \frac{1}{s+1} & 0 & \frac{s-1}{s+1} \\ \frac{e^{-s}}{s+1} & 0 & 1 & 0 \end{pmatrix}.$$

If we call the matrix in the second member R_{-1}^T, then we have the following exact sequence:

$$0 \longleftarrow T(M) \longleftarrow A^2 \overset{.R^T}{\longleftarrow} A^4 \overset{.R_{-1}^T}{\longleftarrow} A^2 \longleftarrow 0. \tag{3.15}$$

Moreover, if we note $\mu = (\mu_1 : \mu_2)$, then, from $\lambda = \mu\, R_{-1}^T$, we have

$$\left\{ \begin{array}{l} \mu_1 = 2\,\lambda_2 + \lambda_4, \\ \mu_2 = \lambda_3, \end{array} \right. \Rightarrow S_{-1}^T \triangleq \left(\begin{array}{cc} 0 & 0 \\ 2 & 0 \\ 0 & 1 \\ 1 & 0 \end{array} \right) : \; R_{-1}^T\, S_{-1}^T = I_2 \Rightarrow S_{-1}\, R_{-1} = I_2.$$

4. Dualizing (3.15), we obtain the following complex:

$$0 \longrightarrow A^2 \xrightarrow{.R} A^4 \xrightarrow{.R_{-1}} A^2 \longrightarrow 0.$$

Therefore, we have

$$\left\{ \begin{array}{l} \mathrm{ext}_A^1(T(M), A) = \ker .R_{-1}/A^2\, R, \\ \mathrm{ext}_A^2(T(M), A) = A^2/A^4\, R_{-1}. \end{array} \right.$$

From $S_{-1}\, R_{-1} = I_2$, we deduce that for all $\xi \in A^2$, the element $\eta = \xi\, S_{-1} \in A^4$ is such that $\xi = \eta\, R_{-1}$, i.e., $A^4\, R_{-1} = A^2$, and thus, $\mathrm{ext}_A^2(T(M), A) = A^2/A^4\, R_{-1} = 0$.

5. If $\eta = (\eta_1 : \eta_2 : \eta_3 : \eta_4) \in \ker .R_{-1}$, then we have

$$\left\{ \begin{array}{l} \left(\frac{s-1}{s+1}\right)^2 \eta_1 + \frac{1}{(s+1)}\, \eta_2 + \frac{(s-1)}{(s+1)}\, \eta_4 = 0, \\ \frac{e^{-s}}{(s+1)}\, \eta_1 + \eta_3 = 0, \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \eta_3 = -\frac{e^{-s}}{(s+1)}\, \eta_1, \\ \frac{(s-1)}{(s+1)} \left(\frac{(s-1)}{(s+1)}\, \eta_1 + \eta_4 \right) = -\frac{1}{(s+1)}\, \eta_2. \end{array} \right. \tag{3.16}$$

Using the fact that the greatest common factor of $\frac{s-1}{s+1}$ and $\frac{1}{s+1}$ is 1, we then have:

$$(3.16) \Leftrightarrow \left\{ \begin{array}{l} \eta_3 = -\frac{e^{-s}}{(s+1)}\, \eta_1, \\ \eta_2 = \frac{(s-1)}{(s+1)}\, \zeta_2,\ \zeta_2 \in A, \\ \frac{(s-1)}{(s+1)}\, \eta_1 + \eta_4 = -\frac{1}{(s+1)}\, \zeta_2, \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \eta_1 = \zeta_1,\ \zeta_1 \in A, \\ \eta_2 = \frac{(s-1)}{(s+1)}\, \zeta_2,\ \zeta_2 \in A, \\ \eta_3 = -\frac{e^{-s}}{(s+1)}\, \zeta_1, \\ \eta_4 = -\frac{(s-1)}{(s+1)}\, \zeta_1 - \frac{1}{(s+1)}\, \zeta_2, \end{array} \right. \Leftrightarrow \eta = \zeta\, R',$$

where $\zeta = (\zeta_1 : \zeta_2)$ and the matrix $R' \in M_{2\times 4}(A)$ is defined by

$$R' = \left(\begin{array}{cccc} 1 & 0 & -\frac{e^{-s}}{(s+1)} & -\frac{s-1}{s+1} \\ 0 & \frac{s-1}{s+1} & 0 & -\frac{1}{s+1} \end{array} \right). \tag{3.17}$$

Thus, we have $\overline{A^2\, R} = A^2\, R'$, and R' has full row rank. Hence, $\overline{A^2\, R}$ is a free A-module of rank 2, and, by Theorem 2.11, the transfer matrix P defined by (1.2) admits the following weakly left-coprime factorization:

$$P = \left(\begin{array}{cc} \frac{e^{-s}}{s+1} & \frac{s-1}{s+1} \\ 0 & \frac{1}{s-1} \end{array} \right) = \left(\begin{array}{cc} 1 & 0 \\ 0 & \frac{s-1}{s+1} \end{array} \right)^{-1} \left(\begin{array}{cc} \frac{e^{-s}}{s+1} & \frac{s-1}{s+1} \\ 0 & \frac{1}{s-1} \end{array} \right). \tag{3.18}$$

Moreover, using the fact that R' has full row rank, we have the exact sequence

$$0 \longrightarrow A^2 \xrightarrow{.R'} A^4 \xrightarrow{.R_{-1}} A^2 \longrightarrow 0 \tag{3.19}$$

and

$$\left\{ \begin{array}{l} t(M) \cong \mathrm{ext}^1_A(T(M), A) = A^2\,R'/A^2\,R, \\ M/t(M) = A^4/A^2\,.R' \cong A^4\,R_{-1} = A^2. \end{array} \right.$$

From $t(M) \cong A^2\,R'/A^2\,R$, we obtain that the class of $(1 : \ 0)\,R'$ in $t(M)$ is the torsion element $z_1 = y_1 - \frac{e^{-s}}{(s+1)}\,u_1 - \frac{(s-1)}{(s+1)}\,u_2$, which satisfies $\frac{(s-1)}{(s+1)}\,z_1 = 0$. Similarly, the class z_2 of $(0 : \ 1)\,R'$ in $t(M)$ is the trivial torsion element 0 because we have $z_2 = \frac{(s-1)}{(s+1)}\,y_2 - \frac{1}{(s+1)}\,u_2 = 0$. Thus, M is not a torsion-free A-module and $M/t(M)$ is a free A-module of rank 2. Finally, we have

$$\left(\begin{array}{cccc} \frac{s-1}{s+1} & 0 & -\frac{(s-1)\,e^{-s}}{(s+1)^2} & -\left(\frac{s-1}{s+1}\right)^2 \\ 0 & \frac{s-1}{s+1} & 0 & -\frac{1}{s+1} \end{array} \right) = \left(\begin{array}{cc} \frac{s-1}{s+1} & 0 \\ 0 & 1 \end{array} \right) \left(\begin{array}{cccc} 1 & 0 & -\frac{e^{-s}}{(s+1)} & -\frac{s-1}{s+1} \\ 0 & \frac{s-1}{s+1} & 0 & -\frac{1}{s+1} \end{array} \right).$$

Remark 3.4. For the sake of simplicity, we have treated here just a simple example. Simple computations, which do not require the algorithm developed in section 3.3, can easily give the weakly left-prime matrix (3.17) and, thus, the weakly left-coprime factorization (3.18) of (1.2). However, for more general systems (see, e.g., $P = (\frac{e^{-s}}{(s-1)} \ : \ \frac{e^{-s}}{(s-1)^2})^T$ [32, 35]), it becomes more difficult to guess a weakly left-coprime factorization, and thus, we really need the algorithm to obtain weakly left/right/doubly coprime factorizations.

4. Doubly coprime factorizations.

4.1. Left-coprime factorizations and stably free modules. Let us introduce the concept of a splitting exact sequence.

DEFINITION 4.1 (see [2, 39]). *An exact sequence* $0 \longrightarrow M' \xrightarrow{f} M \xrightarrow{g} M'' \longrightarrow 0$ *is a* splitting exact *sequence if one of the following equivalent assertions is satisfied:*

- *there exists an* A*-morphism* $h : M'' \longrightarrow M$ *such that* $g \circ h = \mathrm{id}_{M''}$,
- *there exists an* A*-morphism* $k : M \longrightarrow M'$ *such that* $k \circ f = \mathrm{id}_{M'}$,
- *there exist* $\phi = \binom{k}{g} : M \longrightarrow M' \oplus M''$ *and* $\psi = (f \ : \ h) : M' \oplus M'' \longrightarrow M$ *such that* $\phi \circ \psi = \mathrm{id}_{M' \oplus M''}$ *and* $\psi \circ \phi = \mathrm{id}_M$, *where* $\mathrm{id}_M(m) = m$, *for all* $m \in M$.

PROPOSITION 4.2. *We have the following results:*

1. *(see* [2, 39]*) Let* $R \in M_{q \times p}(A)$ *be a full row rank matrix. Then, the* A*-module* $M = A^p/A^q\,R$ *is stably free iff the exact sequence*

$$0 \longrightarrow A^q \xrightarrow{.R} A^p \longrightarrow M \longrightarrow 0 \tag{4.1}$$

is a splitting exact sequence, i.e., iff there exists $S \in M_{p \times q}(A)$ *such that*

$$R\,S = I_q. \tag{4.2}$$

2. *(see* [19]*) Let* $R \in M_{q \times p}(A)$ *be a full row rank matrix and* $M = A^p/A^q\,R$ *the corresponding* A*-module. Then,* M *is stably-free iff*

$$T(M) = A^q/A^p\,R^T = 0.$$

Example 4.1. Let us consider the full row rank matrix $R' \in M_{2\times 4}(A)$ defined by (3.17) and $A = H_\infty(\mathbb{C}_+)$. By point 1 of Proposition 4.2, R' admits a right-inverse $S' \in M_{4\times 2}(A)$ iff the A-module $M' = A^4/A^2\,R'$ is stably free, i.e., iff the A-module $T(M') = A^2/A^4\,{R'}^T = 0$ by point 2 of Proposition 4.2. The A-module $T(M') = A^2/A^4\,{R'}^T$ is defined by the following equations:

$$\begin{cases} \lambda_1 = 0, \\ \frac{(s-1)}{(s+1)}\,\lambda_2 = 0, \\ -\frac{e^{-s}}{(s+1)}\,\lambda_1 = 0, \\ -\frac{(s-1)}{(s+1)}\,\lambda_1 - \frac{1}{(s+1)}\,\lambda_2 = 0. \end{cases} \tag{4.3}$$

If we put a second member $\mu = (\mu_1 : \mu_2 : \mu_3 : \mu_4)^T$ in (4.3), then we have

$$\begin{cases} \lambda_1 = \mu_1, \\ \lambda_2 = -2\,\frac{(s-1)}{(s+1)}\,\mu_1 + \mu_2 - 2\,\mu_4, \end{cases}$$

which proves that, from (4.3), we can deduce that $\lambda_1 = \lambda_2 = 0$, i.e., $T(M') = 0$ and M' is a stably free A-module. A right-inverse S' of R', i.e., $R'\,S' = I_2$, is defined by

$$S = \begin{pmatrix} 1 & -2\,\frac{(s-1)}{(s+1)} \\ 0 & 1 \\ 0 & 0 \\ 0 & -2 \end{pmatrix} \in M_{4\times 2}(A). \tag{4.4}$$

Let us give the definition of the *fitting ideals* of a finitely presented A-module M.

Definition 4.3 (see [13]). *Let $d : F_1 \longrightarrow F_0$ be an A-morphism between two finite free A-modules F_0 and F_1. If we choose bases for F_0 and F_1 ($F_0 \cong A^p$, $F_1 \cong A^q$), then d is defined by a matrix $R \in M_{q\times p}(A)$.*

- *We denote by $I_i(R)$ the ideal of A defined by*
  - *all the $i \times i$ minors of R if $1 \leq i \leq \min\{p, q, \}$,*
  - *$I_i(R) = 0$ if $i > \min\{p, q\}$,*
  - *$I_i(R) = A$ if $i \leq 0$.*
- *Let us define the A-module $M = \operatorname{coker} d$, i.e., $M = A^p/A^q\,R$. The ith* fitting ideal $\mathrm{Fitt}_i(M)$ *is the ideal of A defined by $I_{p-i}(R)$. $\mathrm{Fitt}_i(M)$ does not depend on the choice of the finite free presentation of M.*
- *We denote by $I(M)$ the first nonzero fitting ideal $\mathrm{Fitt}_i(M)$ of M.*

Proposition 4.4 (see [13]). *Let M be a finitely presented A-module. Then, we have the following:*

- *M is a projective A-module of rank r iff $\mathrm{Fitt}_r(M) = A$ and $\mathrm{Fitt}_{r-1}(M) = 0$.*
- *M is a projective A-module of a certain rank iff $I(M) = A$.*

Example 4.2. Let us reconsider Example 4.1. We have

$$\mathrm{Fitt}_0(M') = \mathrm{Fitt}_1(M') = 0, \qquad \mathrm{Fitt}_2(M') = \left(\frac{s-1}{s+1}, \frac{1}{s+1}, \frac{e^{-s}}{(s+1)^2}, \frac{(s-1)\,e^{-s}}{(s+1)^2}\right).$$

We can check that $1 = \frac{s-1}{s+1} + \frac{2}{s+1} \in \mathrm{Fitt}_2(M')$, and thus, $\mathrm{Fitt}_2(M') = A$; i.e., M' is a projective A-module of rank 2 by Proposition 4.4.

The next proposition characterizes the projective modules over Banach algebras.

PROPOSITION 4.5. *If A is a Banach algebra which is an integral domain without radical, i.e., $\sqrt{A} = \{a \in A \mid \lim_{n\to+\infty} \|a^n\|_A^{1/n} = 0\} = 0$, then a full row rank matrix $R \in M_{q\times p}(A)$ defines a projective A-module $M = A^p/A^q\,R$ iff $\mathrm{rk}(\hat{R}(\chi))$ is a constant function on the* maximal ideal space $X(A)$ *of A (see* [16, 49]*), where $\hat{R}$ denotes the* Gelfand transform *of R (see* [16, 49]*), or, equivalently, iff*

$$\inf_{\chi\in X(A)} \sum_{i\in I} |\hat{R}_i(\chi)| \geq \delta > 0,$$

where $(\hat{R}_i)_{i\in I}$ is the family of the $q \times q$ minors of R.

Proof. Using the fact that the maximal ideal space $X(A)$ of a Banach algebra is a Hausdorff compact set and A has only two idempotent elements 1 and 0, then, by the *Shilov theorem* [16], $X(A)$ is a connected space. By the *Swan theorem* [45], any *vector bundle* over $X(A)$ is in one-to-one correspondence to a projective module over the ring of continuous functions $C(X(A))$ on $X(A)$. The fact that $X(A)$ is a connected space implies that the rank of any vector bundle over $X(A)$ is globally constant. Finally, using the fact that A is without radical, by the Gelfand transform [16], any matrix whose entries belong to A can be seen as a matrix whose entries belong to $C(X(A))$. Hence, we find that M is a projective A-module iff $\mathrm{rk}(\hat{R}(\chi))$ is a constant function on $X(A)$. □

Example 4.3. $H_\infty(\mathbb{C}_+)$ and $\hat{\mathcal{A}}$ are two integral domains which are Banach algebras without radical. We can use Proposition 4.5 to check whether or not an A-module is projective. For $A = H_\infty(\mathbb{C}_+)$, we can use the fact that $\mathbb{C}_+$ is dense in $X(A)$ (by the *Corona theorem*; see [25]) in order to take χ only in $\mathbb{C}_+$ instead of the whole $X(A)$. Similarly, for $A = \hat{\mathcal{A}}$, we can restrict the evaluation of $\inf_{\chi\in X(A)} \sum_{i\in I} |\hat{R}_i(\chi)|$ to $\chi \in \overline{\mathbb{C}_+}$, where $\overline{\mathbb{C}_+} = \{s \in \mathbb{C} \mid \mathrm{Re}\, s \geq 0\}$ (see [3, 8]).

- Let $A = H_\infty(\mathbb{C}_+)$. Let $R = (\frac{s-1}{s+1} : \frac{e^{-s}}{s+1}) \in M_{1\times 2}(A)$ and the A-module $M = A^2/A\,R$. Then, M is a projective A-module (i.e., free because A is a coherent Sylvester domain) because we have

$$\inf_{s\in\mathbb{C}_+} \left(\left| \frac{s-1}{s+1} \right| + \left| \frac{e^{-s}}{s+1} \right| \right) > 0.$$

  We can check that we have the following Bézout identity:

$$\left(\frac{s-1}{s+1}\right) \left(1 + 2\left(\frac{1-e^{-(s-1)}}{s-1}\right)\right) + 2\,e\left(\frac{e^{-s}}{s+1}\right) = 1.$$

- Let $A = H_\infty(\mathbb{C}_+)$. The matrix $R = (\frac{1}{s+1} : e^{-s}) \in M_{1\times 2}(A)$ does not define a projective A-module $M = A^2/A\,R$ because we have

$$\inf_{s\in\mathbb{C}_+} \left(\left| \frac{1}{s+1} \right| + |\, e^{-s}\, | \right) = 0.$$

  Indeed, if $(x_n)_{n\in\mathbb{Z}_+}$ is a sequence of strictly positive real numbers tending to $+\infty$, we check that $\lim_{n\to+\infty} |\frac{1}{x_n+1}| = 0$ and $\lim_{n\to+\infty} |e^{-x_n}| = 0$. However, the greatest common divisor of $\frac{1}{s+1}$ and e^{-s} is 1, and thus, R is a weakly left-prime matrix by Proposition 2.2 and Corollary 3.31; i.e., M is a torsion-free (see Corollary 2.5) but not free A-module.

DEFINITION 4.6. *Let A be an integral domain and $K = Q(A)$ its field of fractions.*

- *A transfer matrix* $P \in M_{q\times(p-q)}(K)$ *admits a* left-coprime factorization *if there exists a matrix* $R = (D : -N) \in M_{q\times p}(A)$, *with* $\det D \neq 0$, *such that* $P = D^{-1}\,N$ *and* R *has a right-inverse* $S = (X^T : Y^T)^T \in M_{p\times q}(A)$, *i.e.,*

$$R\,S = D\,X - N\,Y = I_q.$$

- *A transfer matrix* $P \in M_{q\times(p-q)}(K)$ *admits a* right-coprime factorization *if there exists a matrix* $\tilde{R} = (\tilde{N}^T : \tilde{D}^T)^T \in M_{p\times(p-q)}(A)$, *with* $\det \tilde{D} \neq 0$, *such that* $P = \tilde{N}\,\tilde{D}^{-1}$ *and* $\tilde{R}$ *has a left-inverse* $\tilde{S} = (-\tilde{Y} : \tilde{X}) \in M_{(p-q)\times p}(A)$, *i.e.,*

$$\tilde{S}\,\tilde{R} = -\tilde{Y}\,\tilde{N} + \tilde{X}\,\tilde{D} = I_{p-q}.$$

PROPOSITION 4.7. *Let* $P = D^{-1}\,N = \tilde{N}\,\tilde{D}^{-1} \in M_{q\times(p-q)}(K)$ *be a transfer matrix, where* $R = (D : -N) \in M_{q\times p}(A)$ *and* $\tilde{R} = (\tilde{N}^T : \tilde{D}^T)^T \in M_{p\times(p-q)}(A)$. *Let us define the* A*-modules* $M = A^p/A^q\,R$ *and* $\tilde{M} = A^p/A^{p-q}\,\tilde{R}^T$. *Then, we have*

1. *P admits a left-coprime factorization iff $\overline{A^q\,R}$ is a free A-module of rank q and $M/t(M) = A^p/\overline{A^q\,R}$ is a stably free A-module.*
2. *P admits a right-coprime factorization iff $\overline{A^{p-q}\,\tilde{R}^T}$ is a free A-module of rank $p-q$ and $\tilde{M}/t(\tilde{M}) = A^p/\overline{A^{p-q}\,\tilde{R}^T}$ is a stably free A-module.*

Proof. 1. $\Rightarrow$ Let us suppose that P admits a left-coprime factorization of the form $P = {D'}^{-1}\,N'$, where the matrix $R' = (D' : -N') \in M_{q\times p}(A)$ has right-inverse $S' = ({X'}^T : {Y'}^T)^T \in M_{p\times q}(A)$. In particular, R' is weakly left-prime and, by Lemma 2.6, we have $\overline{A^q\,R} = A^q\,R'$. Moreover, R' is a full row rank matrix, and thus, $A^q\,R' = \overline{A^q\,R}$ is a free A-module of rank q. We have the exact sequence

$$0 \longrightarrow A^q \xrightarrow{.R'} A^p \longrightarrow M/t(M) = A^p/\overline{A^q\,R} \longrightarrow 0. \tag{4.5}$$

Using the fact that R' has a right-inverse S', we obtain that (4.5) splits, and thus, we have $A^p \cong A^q \oplus M/t(M)$; i.e., $M/t(M)$ is a stably free A-module.

$\Leftarrow$ Let $P = D^{-1}\,N$ be such that $\overline{A^q\,R}$ is a free A-module of rank q and the A-module $M/t(M) = A^p/\overline{A^q\,R}$ is stably free. Using the fact that $\overline{A^q\,R}$ is a free A-module of rank q, there then exists a weakly left-prime matrix $R' = (D' : -N') \in M_{q\times p}(A)$ such that $\overline{A^q\,R} = A^q\,R'$ and $P = {D'}^{-1}\,N'$. Then, we have the exact sequence (4.5), which splits because $M/t(M) = A^p/A^q\,R'$ is a stably free A-module, and thus, there exists $S' = ({X'}^T : {Y'}^T)^T \in M_{p\times q}(A)$ such that $D'\,X' - N'\,Y' = I_q$, i.e., $P = {D'}^{-1}\,N'$ is a left-coprime factorization of P. Part 2 can be proved similarly. □

Example 4.4. In Example 3.4, we proved that $\overline{A^2\,R} = A^2\,R'$, where R (resp., R') is defined by (1.6) (resp., (3.17)), is a free A-module of rank 2. Moreover, in Example 4.1, we proved that $M/t(M) = A^4/A^2\,R'$ is a stably free A-module. Hence, from point 1 of Proposition 4.7, we deduce that (3.18) is a left-coprime factorization of the transfer matrix (1.2). Finally, using S' obtained in Example 4.1, we obtain

$$\begin{cases} P = \begin{pmatrix} \frac{e^{-s}}{s+1} & \frac{s-1}{s+1} \\ 0 & \frac{1}{s-1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{s-1}{s+1} \end{pmatrix}^{-1} \begin{pmatrix} \frac{e^{-s}}{s+1} & \frac{s-1}{s+1} \\ 0 & \frac{1}{s+1} \end{pmatrix}, \\ \begin{pmatrix} 1 & 0 \\ 0 & \frac{s-1}{s+1} \end{pmatrix} \begin{pmatrix} 1 & -2\,\frac{(s-1)}{(s+1)} \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} \frac{e^{-s}}{s+1} & \frac{s-1}{s+1} \\ 0 & \frac{1}{s+1} \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix} = I_2. \end{cases}$$

Example 4.5. Let us consider the example defined in [49, p. 349]. Let us consider the integral domain $A = \mathbb{R}[t_0, t_1, t_2]/(t_0^2+t_1^2+t_2^2-1)$ of polynomials on the unit sphere of $\mathbb{R}^3$. Let x_i be the class of t_i in A. We have $A = \mathbb{R}[x_0, x_1, x_2]$ with the relation $x_0^2 + x_1^2 + x_2^2 = 1$. It is shown in [20, p. 32] that A is a *unique factorization domain* (UFD) [39], and thus, A is a GCDD.

Let us consider $P = -(x_1/x_0 : x_2/x_0) \in M_{1\times 2}(K)$ with $K = \mathbb{R}(x_0, x_1, x_2)$. We have $P = -x_0^{-1}\,(x_1 : x_2)$ and, if we define $R = (x_0 : x_1 : x_2) \in A^3$, then we have $R\,R^T = 1$. Thus, $\overline{A\,R} = A\,R$ is a free A-submodule of rank 1, and $M = A^3/A\,R$ is a stably free A-module, which proves that P admits a (normalized) left-coprime factorization. Moreover, we have $P = -(x_1 : x_2)\,(x_0^{-1}\,I_2)$. Let us define the matrix

$$\tilde{R} = \begin{pmatrix} -x_1 & -x_2 \\ x_0 & 0 \\ 0 & x_0 \end{pmatrix} \in M_{3\times 2}(A)$$

and the corresponding A-module $\tilde{M} = A^3/A^2\,\tilde{R}^T$. We easily check that $\mathrm{Fitt}_0(\tilde{M}) = 0$ and $\mathrm{Fitt}_1(\tilde{M}) = (x_0\,x_1, x_0\,x_2, x_0^2)$. Thus, x_0 is a greatest common factor of all the 2×2 minors, which, by Proposition 2.2, proves that $\tilde{R}^T$ is not weakly left-prime, i.e., the A-module $\tilde{M}$, defined by the equations

$$\begin{cases} -x_1\,y_0 + x_0\,y_1 = 0, \\ -x_2\,y_0 + x_0\,y_2 = 0, \end{cases}$$

has a nonzero torsion submodule. We easily check that $z = -x_2\,y_1 + x_1\,y_2$, satisfying $x_0\,z = 0$, defines the torsion submodule of $\tilde{M}$. Therefore, $A^2\,\tilde{R}^T$ is not A-closed, and we have $\tilde{M}/t(\tilde{M}) = A^3/A^3\,\tilde{R'}^T$, where $\tilde{R'}^T$ is defined by

$$\tilde{R'}^T = \begin{pmatrix} -x_1 & x_0 & 0 \\ -x_2 & 0 & x_0 \\ 0 & -x_2 & x_1 \end{pmatrix} \in M_3(A).$$

We have $\mathrm{Fitt}_0(\tilde{M}/t(\tilde{M})) = 0$ and $x_0^2, x_1^2, x_2^2 \in \mathrm{Fitt}_1(\tilde{M}/t(\tilde{M})) \Rightarrow 1 \in \mathrm{Fitt}_1(\tilde{M}/t(\tilde{M}))$, and thus, by Proposition 4.4, $\tilde{M}/t(\tilde{M})$ is a projective A-module of rank 1. However, a projective module of rank 1 over a UFD is free (see [20, 45]), and thus, $\tilde{M}/t(\tilde{M})$ is a free A-module of rank 1: $u = x_0\,y_0 + x_1\,y_1 + x_2\,y_2$ is a basis of $\tilde{M}/t(\tilde{M})$ because we have $y_i = x_i\,u$ for $i = 1, \ldots, 3$. Thus, we obtain that $\tilde{M}/t(\tilde{M}) \cong A^3\,R^T \cong A$. Moreover, by Proposition 2.8, we know that $\ker.R^T = \overline{A^2\,\tilde{R}^T} = A^3\,\tilde{R'}^T$. However, it is well known that $\ker.R^T$ is a stably free but not a free A-module [20, 49]. By Proposition 4.7, P does not admit a right-coprime factorization.

Corollary 4.8. *Let $P = D^{-1}\,N = \tilde{N}\,\tilde{D}^{-1} \in M_{q\times(p-q)}(K)$ be a transfer matrix, where $R = (D : -N) \in M_{q\times p}(A)$ and $\tilde{R} = (\tilde{N}^T : \tilde{D}^T)^T \in M_{p\times(p-q)}(A)$. Let us define the A-modules $M = A^p/A^q\,R$ and $\tilde{M} = A^p/A^{p-q}\,\tilde{R}^T$. Then, we have*

1. *P admits a left-coprime factorization iff $\tilde{M}/t(\tilde{M}) = A^p/\overline{A^{p-q}\,\tilde{R}^T}$ is a free A-module of rank q.*
2. *P admits a right-coprime factorization iff $M/t(M) = A^p/\overline{A^q\,R}$ is a free A-module of rank $p-q$.*

Proof. 1. $\Rightarrow$ Let us suppose that P admits the left-coprime factorization $P = {D'}^{-1}\,N'$, where $R' = (D' : -N') \in M_{q\times p}(A)$ has a right-inverse S'. Then, $\overline{A^q\,R} = A^q\,R'$ is a free A-module of rank q, and thus, $M/t(M) = A^p/A^q\,R'$, i.e., we have the

following exact sequence:

$$0 \longrightarrow A^q \xrightarrow{.R'} A^p \longrightarrow M/t(M) \longrightarrow 0.$$

By Proposition 4.2, this exact sequence splits, and thus, $A^p\, {R'}^T \cong A^q$. Finally, by Proposition 2.8, we now have $\tilde{M}/t(\tilde{M}) \cong A^p {R'}^T \cong A^q$.

$\Leftarrow$ Let us suppose that the A-module $\tilde{M}/t(\tilde{M})$ is a free A-module of rank q. $t(\tilde{M})$ is a torsion A-module, and thus, we have $(t(\tilde{M}))^\star = 0$. Hence, dualizing the exact sequence $0 \longrightarrow t(\tilde{M}) \longrightarrow \tilde{M} \longrightarrow \tilde{M}/t(\tilde{M}) \longrightarrow 0$, we obtain $\tilde{M}^\star \cong (\tilde{M}/t(\tilde{M}))^\star \cong A^q$. By Proposition 2.8, we know that $\tilde{M}^\star \cong \ker .\tilde{R} = \overline{A^q\, R}$, and thus, $\overline{A^q\, R}$ is a free A-module of rank q. Moreover, by Proposition 1.9 of [33], $A^p\, \tilde{R}$ is a projective A-module because so is $\tilde{M}/t(\tilde{M})$. Thus, the exact sequence $0 \longrightarrow \ker .\tilde{R} \longrightarrow A^p \longrightarrow A^p\, \tilde{R} \longrightarrow 0$ splits, and we obtain that $A^p \cong A^p\, \tilde{R} \oplus \ker .\tilde{R}$. However, by Proposition 2.8, $A^p\, \tilde{R} \cong M/t(M)$. Thus, we have $A^p \cong M/t(M) \oplus A^q$; i.e., $M/t(M)$ is a stably-free A-module. Then, by Proposition 4.7, P admits a left-coprime factorization. Point 2 can be proved similarly. □

Example 4.6. Let us reconsider the system defined in Example 4.5. We proved that the A-module $\tilde{M}/t(\tilde{M})$ is a free A-module of rank 1, and thus, by Corollary 4.8, P admits a left-coprime factorization. Moreover, it is known that $M/t(M) = M$ is a stably free but not a free A-module [20, 45], i.e., P does not admit right-coprime factorizations.

4.2. Doubly coprime factorizations and free modules. The following result characterizes generalized Bézout identities in terms of free A-modules.

PROPOSITION 4.9. *Let $M = A^p/A^q\, R$ be an A-module defined by a full row rank matrix $R \in M_{q\times p}(A)$, i.e., by the following finite free resolution:*

$$0 \longrightarrow A^q \xrightarrow{.R} A^p \longrightarrow M \longrightarrow 0. \tag{4.6}$$

Then, M is a free A-module iff there exist three matrices R_{-1}, S_{-1}, and S such that we have the following splitting exact sequence,

$$0 \longrightarrow A^q \underset{\xleftarrow{.S}}{\xrightarrow{.R}} A^p \underset{\xleftarrow{.S_{-1}}}{\xrightarrow{.R_{-1}}} A^{p-q} \longrightarrow 0, \tag{4.7}$$

or equivalently, iff we have the following generalized Bézout identities:

$$\text{(i)}\ \begin{pmatrix} S & R_{-1} \end{pmatrix} \begin{pmatrix} R \\ S_{-1} \end{pmatrix} = I_p,$$

$$\text{(ii)}\ \begin{pmatrix} R \\ S_{-1} \end{pmatrix} \begin{pmatrix} S & R_{-1} \end{pmatrix} = \begin{pmatrix} I_q & 0 \\ 0 & I_{p-q} \end{pmatrix} = I_p.$$

Proof. $\Rightarrow$ The A-module M is free, and thus, there exists a $p \times (p-q)$ matrix R_{-1} with entries in A such that the exact sequence (4.6) has the form

$$0 \longrightarrow A^q \xrightarrow{.R} A^p \xrightarrow{.R_{-1}} A^{p-q} \longrightarrow 0.$$

This exact sequence finishes by the free A-module A^{p-q}, and thus, by Proposition 4.2, it splits; i.e., there exists a $(p-q)\times p$ matrix S_{-1} such that $R_{-1}\, S_{-1} = I_{p-q}$. By the equivalences of Definition 4.1, we have the Bézout identities (i) and (ii).

$\Leftarrow$ If we have the splitting exact sequence (4.7) or, equivalently, the Bézout identities (i) and (ii), then $M \cong A^p R_{-1} = A^{p-q}$, i.e., M is a free A-module of rank $p-q$. □

DEFINITION 4.10. *A transfer matrix $P \in M_{q\times(p-q)}(K)$ admits a* doubly coprime factorization *if there exist $(D : -N) \in M_{q\times p}(A)$, $(\tilde{N}^T : \tilde{D}^T)^T \in M_{p\times(p-q)}(A)$, $(X^T : Y^T)^T \in M_{p\times q}(A)$, and $(-\tilde{Y} : \tilde{X}) \in M_{(p-q)\times p}(A)$ such that*

- $P = D^{-1} N = \tilde{N} \tilde{D}^{-1}$,
- $\begin{pmatrix} D & -N \\ -\tilde{Y} & \tilde{X} \end{pmatrix} \begin{pmatrix} X & \tilde{N} \\ Y & \tilde{D} \end{pmatrix} = I_p$,
- $\begin{pmatrix} X & \tilde{N} \\ Y & \tilde{D} \end{pmatrix} \begin{pmatrix} D & -N \\ -\tilde{Y} & \tilde{X} \end{pmatrix} = I_p$.

THEOREM 4.11. *Let $P = D^{-1} N = \tilde{N} \tilde{D}^{-1}$, $R = (D : -N)$, $\tilde{R} = (\tilde{N}^T : \tilde{D}^T)^T$, and the A-modules $M = A^p/A^q R$ and $\tilde{M} = A^p/A^{p-q} \tilde{R}^T$. Then, P admits a doubly coprime factorization iff $M/t(M)$ and $\tilde{M}/t(\tilde{M})$ are free A-modules of rank $p-q$ and q.*

Proof. $\Rightarrow$ If P admits a doubly coprime factorization, then P admits left and right-coprime factorizations, and thus, by Proposition 4.8, the A-modules $M/t(M)$ and $\tilde{M}/t(\tilde{M})$ are free A-modules of rank, respectively, $p-q$ and q.

$\Leftarrow$ By Proposition 4.8, there exist a left and a right-coprime factorization of P:

$$P = {D'}^{-1} N' = \tilde{N}' {\tilde{D}'}^{-1}, \qquad \begin{cases} D' X - N' Y = I_q, \\ -\tilde{Y}' \tilde{N}' + \tilde{X}' \tilde{D}' = I_{p-q}. \end{cases}$$

From $P = {D'}^{-1} N' = \tilde{N}' {\tilde{D}'}^{-1}$, we deduce that $(D' : -N') \begin{pmatrix} \tilde{N}' \\ \tilde{D}' \end{pmatrix} = 0$. If we take

$$\begin{cases} X' = X + \tilde{N}' (\tilde{Y}' X - \tilde{X}' Y), \\ Y' = Y + \tilde{D}' (\tilde{Y}' X - \tilde{X}' Y), \end{cases}$$

we can easily check that $P = {D'}^{-1} N' = \tilde{N}' {\tilde{D}'}^{-1}$ is a doubly coprime factorization:

$$\begin{pmatrix} D' & -N' \\ -\tilde{Y}' & \tilde{X}' \end{pmatrix} \begin{pmatrix} X' & \tilde{N}' \\ Y' & \tilde{D}' \end{pmatrix} = I_p, \qquad \begin{pmatrix} X' & \tilde{N}' \\ Y' & \tilde{D}' \end{pmatrix} \begin{pmatrix} D' & -N' \\ -\tilde{Y}' & \tilde{X}' \end{pmatrix} = I_p. \qquad \square$$

Using Proposition 2.8, we obtain the following corollary of Theorem 4.11.

COROLLARY 4.12. *Let $P = D^{-1} N = \tilde{N} \tilde{D}^{-1} \in M_{q\times(p-q)}(A)$ be a transfer matrix, $R = (D : -N) \in M_{q\times p}(A)$, and $\tilde{R} = (\tilde{N}^T : \tilde{D}^T)^T \in M_{p\times(p-q)}(A)$. Then, P admits a doubly coprime factorization iff the A-modules $A^p \tilde{R}$ and $A^p R^T$ are two free A-modules of rank, respectively, $p-q$ and q.*

This corollary was first proved in [44]. We have the following corollary of Proposition 4.12, which was first obtained in [49].

COROLLARY 4.13. *A SISO plant, defined by $p = n/d$ $(0 \neq d, n \in A)$, admits a coprime factorization iff the ideal $I = (n, d)$ of A is principal.*

Proof. By Proposition 4.12, $p = n/d$ has a coprime factorization iff the A-module $I = A^2 R^T = (d, n)$ is free of rank 1, where $R = (d : -n) \in M_{1\times 2}(A)$. Using the fact that A is an integral domain, I is a free A-module iff I is a principal ideal. □

The next corollary of Proposition 4.7 was first proved in [49].

COROLLARY 4.14. *If A is a Hermite ring, then every transfer matrix P with a left-coprime (resp., right-coprime) factorization admits a doubly coprime factorization.*

Proof. Let $P = D^{-1} N$ be a left-coprime factorization of the transfer matrix P, where $R = (D : -N) \in M_{q\times p}(A)$. By Proposition 4.7, the A-module $M = A^p/A^q R$

is stably free. Using the fact that A is a Hermite ring, then M is free, and the result follows directly from Corollary 4.8, and similarly for right-coprime factorizations. □

Example 4.7. In Example 4.4, we proved that the transfer matrix P defined by (1.2) admits a left-coprime factorization. Using the fact that $A = H_\infty(\mathbb{C}_+)$ is a coherent Sylvester domain and, in particular, a Hermite ring, by Corollary 4.14, we know that P admits a doubly coprime factorization. In fact, we have already done all the computations to obtain a right-coprime factorization of P. Indeed, we proved that (3.19) is an exact sequence, and thus it splits. Hence, using the matrices $R_{-1} = (\tilde{N}^T : \tilde{D}^T)^T \in M_{4\times 2}(A)$ and $S_{-1} = (-\tilde{Y} : \tilde{X}) \in M_{2\times 4}(A)$, defined in Example 3.4, we obtain the following right-coprime factorization of P:

$$\begin{cases} P = \tilde{N}\,\tilde{D}^{-1} = \begin{pmatrix} \left(\frac{s-1}{s+1}\right)^2 & \frac{e^{-s}}{s+1} \\ \frac{1}{s+1} & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ \frac{s-1}{s+1} & 0 \end{pmatrix}^{-1}, \\ -\begin{pmatrix} 0 & -2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \left(\frac{s-1}{s+1}\right)^2 & \frac{e^{-s}}{s+1} \\ \frac{1}{s+1} & 0 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ \frac{s-1}{s+1} & 0 \end{pmatrix} = I_2. \end{cases}$$

THEOREM 4.15 (see [49]). *The following assertions are equivalent:*

1. *every MIMO plant admits doubly coprime factorizations,*
2. *every SISO plant admits coprime factorizations,*
3. *A is a Bézout domain.*

Proof. $1 \Rightarrow 2$ is trivial. $2 \Rightarrow 3$ is given by Lemma 4.13.

$3 \Rightarrow 1$. If A is a Bézout domain, then every A-module $M = A^p/A^q\,R$, defined by a full row rank matrix $R = (D : -N) \in M_{q\times p}(A)$, is such that $M/t(M)$ is a free A-module. Moreover, a Bézout domain A is a coherent Sylvester domain, and thus, by Proposition 3.21, there exists a full row rank matrix $R' = (D' : -N') \in M_{q\times p}(A)$ such that $M/t(M) = A^p/A^q\,R'$ and $P = D^{-1}\,N = {D'}^{-1}\,N'$. Finally, using Proposition 4.9, we obtain that P admits a doubly coprime factorization. □

Conclusion. We hope we have convinced the reader that the reformulation of the fractional representation approach to analysis problems within the algebraic analysis framework allows us to obtain some new results. These results will be used in the second part of this work [33] to obtain necessary and sufficient conditions for internal stabilizability and to determine the class of rings A over which every plant is internally stabilizable. For the sake of simplicity, we have treated only the case of integral domains, but all the results are still valid for general rings: we need only to slightly change some definitions (e.g., $K = Q(A) = \{a/b \mid a \in A,\ b \in A \setminus Z(A)\}$, where $Z(A)$ is the set of the nonzero divisors of A, $t(M) = \{m \in M \mid \exists\, a \in A \setminus Z(A) : a\,m = 0\}$, etc.).

Acknowledgements. I would like to thank R. Mortini (University of Metz, France) for pointing me toward the references [23, 37]. I had fruitful discussions with W. Dicks (Universitat Autònoma de Barcelona, Spain) after I read in [6] that coherent Sylvester domains were studied in the papers [10, 11]. In [12], he gave me a proof of Proposition 3.23. I would like to thank him for letting me reproduce his proof in this paper. Finally, I am grateful to J. J. Loiseau (IRCCyN, France) and C. Bonnet (INRIA Rocquencourt, France) for their continuous interest in this work, and to those without whom this paper would never have been written, especially to Céline. I would like to thank the University of Leeds for its hospitality and, especially, J. R. Partington and H. G. Dales.

REFERENCES

[1] N. Bourbaki, *Algèbre Commutative*, Masson, Paris, 1985, Chapters 1–4.
[2] N. Bourbaki, *Algèbre*, Masson, Paris, 1980, Chapter 10.
[3] F. M. Callier and C. A. Desoer, *Stabilization, tracking and disturbance rejection in multivariable convolution systems*, Ann. Soc. Sci. Bruxelles, 94 (1980), pp. 5–51.
[4] H. Cartan, *Idéaux des fonctions analytiques de n variables complexes*, Ann. Ecole Norm. Sup., 61 (1944), pp. 149–197.
[5] S. U. Chase, *Direct product of modules*, Trans. AMS, 97 (1960), pp. 457–473.
[6] P. M. Cohn, *Free Rings and Their Relations*, Academic Press, New York, 1985.
[7] G. Conte and A. M. Perdon, *Systems over a principal ideal domain. A polynomial model approach*, SIAM. J. Control Optim., 20 (1982), pp. 112–124.
[8] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Texts Appl. Math. 21, Springer-Verlag, New York, 1991.
[9] C. A. Desoer, R. -W. Liu, J. Murray, and R. Saeks, *Feedback system design: The fractional representation approach to analysis and synthesis*, IEEE Trans. Automat. Control, 25 (1980), pp. 399–412.
[10] W. Dicks and E. D. Sontag, *Sylvester domains*, J. Pure Appl. Algebra, 13 (1978), pp. 243–275.
[11] W. Dicks, *Free algebras over Bézout domains are Sylvester domains*, J. Pure Appl. Algebra, 27 (1983), pp. 15–28.
[12] W. Dicks, *private communication*, Universitat Autònoma de Barcelona, Barcelona, Spain, 2001.
[13] D. Eisenbud, *Commutative Algebra with a View Toward Algebraic Geometry*, Grad. Texts in Math. 150, Springer-Verlag, New York, 1994.
[14] M. Finkel Jones and M. L. Telpy, *Coherent rings of finite weak global dimension*, Communications in Algebra, 10 (1982), pp. 493–503.
[15] L. Fuchs and L. Salce, *Modules over Non-Noetherian Domains*, Math. Surveys Monogr. 84, American Mathematical Society, Providence, RI, 2000.
[16] I. M. Gelfand, D. A. Raikov, and G. E. Shilov, *Commutative Normed Rings*, Chelsea, New York, 1964.
[17] S. Glaz, *Commutative coherent rings: Historical perspective and current developments*, Nieuw Arch. Wiskd., 10 (1992), pp. 27–56.
[18] O. Helmer, *Divisibility properties of integral functions*, Duke Math. J., 6 (1940), pp. 345–356.
[19] E. Kunz, *Introduction to Commutative Algebra and Algebraic Geometry*, Birkhäuser Boston, Cambridge, MA, 1985.
[20] T. Y. Lam, *Serre's Conjecture*, Lecture Notes in Math., 635, Springer-Verlag, New York, 1978.
[21] J. J. Loiseau, *Algebraic tools for the control and stabilization of time-delay systems*, IFAC Reviews, Annual Reviews in Control, 24 (2000), pp. 135–149.
[22] K. Mori and K. Abe, *Feedback stabilization over commutative rings: Further study of coordinate-free approach*, SIAM J. Control Optim., 39 (2001), pp. 1952–1973.
[23] W. S. McVoy and L. A. Rubel, *Coherence of some rings of functions*, J. Funct. Anal., 21 (1976), pp. 76–87.
[24] D. G. McRae, *Homological dimensions of finitely presented modules*, Math. Scand., 28 (1971), pp. 70–76.
[25] N. K. Nikol'skiĭ, *Treatise on the Shift Operator*, Springer-Verlag, New York, Berlin, 1986.
[26] U. Oberst, *Multidimensional Constant Linear Systems*, Acta Appl. Math., 20 (1990), pp. 1–175.
[27] J. F. Pommaret and A. Quadrat, *Algebraic analysis of linear multidimensional control systems*, IMA J. Math. Control Inform., 16 (1999), pp. 275–297.
[28] J. F. Pommaret and A. Quadrat, *A functorial approach to the behaviour of multidimensional control systems*, in Proceedings of the 2nd International Workshop on Multidimensional Systems (NDS), Czocha Castle, Poland, 2000, pp. 91–96; to appear in Applied Mathematics and Computer Science.
[29] J. F. Pommaret and A. Quadrat, *Equivalences of linear control systems*, in Proceedings of the Conference on Mathematical Theory of Networks and Systems (MTNS), Perpignan, France, CD-ROM, 2000; available online at www.univ-perp.fr/mtns2000.
[30] A. Quadrat, *Coherent $H_\infty(D)$-modules in control theory*, in Proceedings of the First IFAC Symposium on System Structure and Control, Prague, CD-ROM, Pergamon Press, Oxford, UK, 2001.
[31] A. Quadrat, *Internal stabilization of coherent control systems*, in Proceedings of the First IFAC Symposium on System Structure and Control, Prague, CD-ROM, Pergamon Press, Oxford, UK, 2001.

[32] A. Quadrat, *Une approche de la stabilisation par l'analyse algébrique, Partie* I. *Factorisations doublement faiblement copremières, Partie* II. *Stabilisation interne, Partie* III. *Sur une structure générale des contrôleurs stabilisants basée sur le rang stable*, in Proceedings of the Conférence Internationale Francophone d'Automatique (CIFA), Nantes, France, CD-ROM, 2002; available online at www.irccyn.ec-nantes.fr/cifa/.

[33] A. Quadrat, *The fractional representation approach to synthesis problems: An algebraic analysis viewpoint* II. *Internal Stabilization*, SIAM J. Control Optim., 42 (2003), pp. 300–320.

[34] A. Quadrat, *On a generalization of the Youla–Kučera parametrization:* I. *The fractional ideals approach to SISO systems*, Systems Control Lett., to appear.

[35] A. Quadrat, *An introduction to internal stabilization of linear infinite dimensional systems*, "Control of Distributed Parameter Systems: Theory and Applications," Proceedings of the International School in Automatic Control of Lille, Ecole Centrale de Lille, Lille, France, 2002.

[36] M. von Renteln, *Hauptideale und äußere Funktionen im Ring H^∞*, Arch. Math., 28 (1977), pp. 519–524.

[37] J.-P. Rosay, *Sur la cohérence de certains anneaux de fonctions*, Illinois J. Math., 21 (1977), pp. 895–897.

[38] H. H. Rosenbrock, *State Space and Multivariable Theory,* Nelson, London, 1970.

[39] J. J. Rotman, *An Introduction to Homological Algebra*, Academic Press, New York, 1979.

[40] S. Shankar and V. R. Sule, *Algebraic geometric aspects of feedback stabilization*, SIAM J. Control Optim., 30 (1992), pp. 11–30.

[41] A. M. Sinclair and A. W. Tullo, *Noetherian Banach algebras are finite dimensional*, Math. Ann., 211 (1974), pp. 151–153.

[42] J. P. Serre, *Faisceaux algébriques cohérents*, Ann. of Math., 61 (1955), pp. 197–278.

[43] M. C. Smith, *On stabilization and the existence of coprime factorizations*, IEEE Trans. Automat. Control, 34 (1989), pp. 1005–1007.

[44] V. R. Sule, *Feedback stabilization over commutative rings: The matrix case*, SIAM J. Control Optim., 32 (1994), pp. 1675–1695; *Corrigendum*, 36 (1998), pp. 2194–2195.

[45] R. G. Swan, *Vector bundles and projective modules*, Trans. Amer. Math. Soc., 105 (1962), pp. 264–277.

[46] V. Tolokonnikov, *Extension problem to an invertible matrix*, Proc. Amer. Math. Soc., 117 (1993), pp. 1023–1030.

[47] W. V. Vasconcelos, *The Rings of Dimension Two*, Lecture Notes in Pure and Appl. Math. 22, Marcel Dekker, New York, 1976.

[48] M. Vidyasagar, H. Schneider, and B. A. Francis, *Algebraic and topological aspects of feedback stabilization*, IEEE Trans. Automat. Control, 27 (1982), pp. 880–894.

[49] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.

[50] Z. Yicai, *On commutative indecomposable coherent regular rings*, Comm. Algebra, 20 (1992), pp. 1389–1394.

[51] G. Zames, *Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Automat. Control, 26 (1981), pp. 301–320.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 300–320

THE FRACTIONAL REPRESENTATION APPROACH TO SYNTHESIS PROBLEMS: AN ALGEBRAIC ANALYSIS VIEWPOINT PART II: INTERNAL STABILIZATION*

A. QUADRAT†

Abstract. In this second part of the paper [A. Quadrat, *SIAM J. Control Optim.*, 40 (2003), pp. 266–299], we show how to reformulate the fractional representation approach to synthesis problems within an algebraic analysis framework. In terms of modules, we give necessary and sufficient conditions for internal stabilizability. Moreover, we characterize all the integral domains A of SISO stable plants such that every MIMO plant—defined by means of a transfer matrix whose entries belong to the quotient field $K = Q(A)$ of A—is internally stabilizable. Finally, we show that this algebraic analysis approach allows us to recover on the one hand the approach developed in [M. C. Smith, *IEEE Trans. Automat. Control*, 34 (1989), pp. 1005–1007] and on the other hand the ones developed in [K. Mori and K. Abe, *SIAM J. Control Optim.*, 39 (2001), pp. 1952–1973; S. Shankar and V. R. Sule, *SIAM J. Control Optim.*, 30 (1992), pp. 11–30; V. R. Sule, *SIAM J. Control Optim.*, 32 (1994), pp. 1675–1695; M. Vidyasagar, H. Schneider, and B. A. Francis, *IEEE Trans. Automat. Control*, 27 (1982), pp. 880–894; M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985].

Key words. fractional representation approach to synthesis problems, internal stabilization, Prüfer domains, Youla–Kučera parametrization of the stabilizing controllers, (weakly) left/right/doubly coprime factorizations, coherent Sylvester domains, $H_\infty(\mathbb{C}_+)$, algebraic analysis, module theory, homological algebra

AMS subject classifications. 93C05, 93D25, 93B52, 93B25, 93C20, 93C23, 16D40, 16E60

PII. S0363012902417139

Introduction. Using the algebraic analysis viewpoint of the fractional representation approach to analysis and synthesis problems [5, 28, 29], developed in the first part of the paper [17], we give necessary and sufficient conditions for *internal stabilizability*. Moreover, using these results, we prove that every multi-input multi-output (MIMO) plant—defined by means of a transfer matrix $P = D^{-1}N = \tilde{N}\tilde{D}^{-1}$, where $R = (D : -N)$ and $\tilde{R} = (\tilde{N}^T : \tilde{D}^T)^T$ are matrices whose entries belong to an integral domain A of single input single output (SISO) stable plants—is internally stabilizable iff A is a *Prüfer domain* [6, 23]. From the fact that the intersection between coherent Sylvester domains (see [17] for more details) and Prüfer domains are just Bézout domains, we also recover the result of Vidyasagar [29]: every MIMO plant admits doubly coprime factorizations iff A is a *Bézout domain*. Hence, if the algebra A is a Prüfer domain but not a Bézout domain, there exist plants which are internally stabilizable but fail to admit doubly coprime factorizations. Therefore, it is not possible to parametrize all their stabilizing controllers by means of the Youla–Kučera parametrization [4, 28]. These results allow us to explain the counterexamples exhibited in [1, 12]. We prove that, over a *projective-free domain* A (e.g., $H_\infty(\mathbb{C}_+)$, RH_∞), every stabilizable system admits doubly coprime factorizations. Finally, we show that the previous results allow us to recover, on the one hand, the results of [25] and, on

*Received by the editors November 15, 1999; accepted for publication (in revised form) November 4, 2002; published electronically April 17, 2003. This work was supported by grant HPMF-CT-1999-00095 during my stay at the University of Leeds.

http://www.siam.org/journals/sicon/42-1/41713.html

†INRIA Sophia Antipolis, CAFE project, 2004 Route des Lucioles, BP 93, 06902 Sophia Antipolis cedex, France (Alban.Quadrat@sophia.inria.fr).

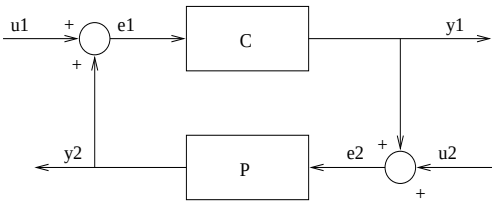


FIG. 1. *Closed-loop.*

the other hand, the ones developed in [12, 13, 24, 26, 27, 28, 29]. We refer to [17] for the development of the algebraic analysis approach used in this second part, as well as for some results that will be continually used in what follows.

Notation. In the course of the text, A denotes a commutative integral domain ($a\,b = 0, a \neq 0 \Rightarrow b = 0$) with a unit, $M_{q\times p}(A)$ (resp., $M_p(A)$), the set of $q \times p$ (resp., $p \times p$) matrices with entries in A and I_p the identity matrix. If $R \in M_{q\times p}(A)$, then R^T is the transposed matrix. *By convention, every vector with entries in A is a row vector.* The positive integers $p, q \in \mathbb{Z}_+$ will always satisfy $p \geq q$. If M and N are two A-modules, then $M \cong N$ means that M and N are isomorphic as A-modules, $\hom_A(M, N)$ is the A-module of the A-morphisms (i.e., A-linear maps) from M to N, and $M^* = \hom_A(M, A)$. Finally, $(a_1, \ldots, a_n)$ denotes the ideal $A\,a_1 + \cdots + A\,a_n$ and $\triangleq$ means "by definition."

1. Closed-loop systems. Let A be an algebra of SISO stable systems which forms an integral domain and let $K = Q(A)$ be its field of fractions. Let us consider the closed-loop formed by a plant $P \in M_{q\times(p-q)}(K)$ and a controller $C \in M_{(p-q)\times q}(K)$ as it is shown in Figure 1. The equations of the closed-loop are

$$\begin{cases} e_1 = u_1 + P\,e_2, \\ e_2 = u_2 + C\,e_1, \\ y_1 = e_2 - u_2, \\ y_2 = e_1 - u_1. \end{cases} \tag{1.1}$$

DEFINITION 1.1 (see [5, 28, 29]). *The plant $P \in M_{q\times(p-q)}(K)$ is* internally stabilizable *if there exists a controller $C \in M_{(p-q)\times q}(K)$ such that all the entries of the following transfer matrix*

$$H(P,C) = \begin{pmatrix} I_q & -P \\ -C & I_{p-q} \end{pmatrix}^{-1} = \begin{pmatrix} (I_q - P\,C)^{-1} & (I_q - P\,C)^{-1}\,P \\ C\,(I_q - P\,C)^{-1} & I_{p-q} + C\,(I_q - P\,C)^{-1}\,P \end{pmatrix} \tag{1.2}$$

are stable, i.e., $H(P,C) \in M_p(A)$.

Let us write P and C in the form $P = {D_p}^{-1}\,N_p$ and $C = {D_c}^{-1}\,N_c$, where $R_p = (D_p : -N_p) \in M_{q\times p}(A)$ and $R_c = (-N_c : D_c) \in M_{(p-q)\times p}(A)$. Thus, we have

$$(1.1) \Leftrightarrow \begin{cases} D_p\,e_1 - N_p\,e_2 - D_p\,u_1 = 0, \\ -N_c\,e_1 + D_c\,e_2 - D_c\,u_2 = 0, \\ y_1 - e_2 + u_2 = 0, \\ y_2 - e_1 + u_1 = 0. \end{cases} \tag{1.3}$$

Let us define the matrices

$$R = \begin{pmatrix} D_p & -N_p & -D_p & 0 \\ -N_c & D_c & 0 & -D_c \end{pmatrix} \in M_{p\times 2p}(A)$$

and

$$R_s = \begin{pmatrix} D_p & -N_p & -D_p & 0 & 0 & 0 \\ -N_c & D_c & 0 & -D_c & 0 & 0 \\ 0 & -I_{p-q} & 0 & I_{p-q} & I_{p-q} & 0 \\ -I_q & 0 & I_q & 0 & 0 & I_q \end{pmatrix} \in M_{2p\times 3p}(A),$$

as well as the following A-modules

$$\begin{cases} M_p = A^p/A^q\,R_p, \\ M_c = A^p/A^{p-q}\,R_c, \\ M = A^{2p}/A^p\,R, \\ M_s = A^{3p}/A^{2p}\,R_s. \end{cases}$$

LEMMA 1.2. *We have $M_s = M \cong M_p \oplus M_c$, and thus*

$$M_s/t(M_s) = M/t(M) \cong M_p/t(M_p) \oplus M_c/t(M_c), \tag{1.4}$$

or equivalently

$$A^{3p}/\overline{A^{2p}\,R_s} = A^{2p}/\overline{A^p\,R} \cong A^p/\overline{A^q\,R_p} \oplus A^p/\overline{A^{p-q}\,R_c},$$

where, for instance, $\overline{A^p\,R}$ is the A-closure of $A^p\,R$ in A^{2p} (see [17] *for more details).*

Proof. We have the following equality:

$$\begin{pmatrix} D_p & -N_p & -D_p & 0 \\ -N_c & D_c & 0 & -D_c \end{pmatrix} \begin{pmatrix} 0 & 0 & I_q & 0 \\ 0 & I_{p-q} & 0 & 0 \\ -I_q & 0 & I_q & 0 \\ 0 & I_{p-q} & 0 & -I_{p-q} \end{pmatrix} = \begin{pmatrix} D_p & -N_p & 0 & 0 \\ 0 & 0 & -N_c & D_c \end{pmatrix}.$$

The second matrix in the left-hand side of the previous equality is unimodular, and thus, invertible. Let us denote this matrix by U. Then, from the previous equality, i.e., $R\,U = R_p \oplus R_c$, we obtain the following commutative exact diagram:

$$\begin{array}{ccccccccc} & & 0 & & & & 0 & & \\ & & \downarrow & & & & \downarrow & & \\ 0 & \longrightarrow & A^p & \xrightarrow{\cdot R} & & & A^{2p} & \xrightarrow{\pi} & M \longrightarrow 0 \\ & & \downarrow \cdot I_p & & & & \downarrow \cdot U & & \\ 0 & \longrightarrow & A^p & \xrightarrow{\cdot (R_p \oplus R_c)} & & & A^{2p} & \xrightarrow{\pi'} & M_p \oplus M_c \longrightarrow 0. \\ & & \downarrow & & & & \downarrow & & \\ & & 0 & & & & 0 & & \end{array}$$

From the previous commutative exact diagram, we deduce that there exists an isomorphism $\phi : M \to M_p \oplus M_c$, defined by $\phi(m) = \pi'(z\,U)$, where $z \in A^{2p}$ is such that $\pi(z) = m$, and thus, $M \cong M_p \oplus M_c$. Moreover, using the equations which define the A-module M_s, we can easily check that $M_s = M$. Finally, using the fact that $M_s = M \cong M_p \oplus M_c$, we obtain $t(M_s) = t(M) \cong t(M_p) \oplus t(M_c)$, and thus, $M/t(M) \cong M_p/t(M_p) \oplus M_c/t(M_c)$. □

2. Internal stabilization: A particular case. We refer the reader to [17] for the definition of a weakly left/right/doubly coprime factorization.

THEOREM 2.1. *Let $P = D_p^{-1} N_p$ and $C = D_c^{-1} N_c$ be two weakly left-coprime factorizations, i.e., $R_p = (D_p : -N_p) \in M_{q\times p}(A)$ and $R_c = (-N_c : D_c) \in M_{(p-q)\times p}(A)$ are weakly left-prime matrices. Then, $P = D_p^{-1} N_p$ is internally stabilized by the controller $C = D_c^{-1} N_c$ iff*

$$\begin{pmatrix} R_p \\ R_c \end{pmatrix}^{-1} \in M_p(A), \text{ i.e., } \begin{pmatrix} R_p \\ R_c \end{pmatrix} \in GL_p(A). \tag{2.1}$$

The same result also holds for weakly right-coprime factorizations.

Proof. $\Rightarrow$ By hypothesis, R_p and R_c are two weakly left-prime matrices, and thus, by Corollary 2.5 of [17], the A-modules $M_p = A^p/A^q R_p$ and $M_c = A^p/A^{p-q} R_c$ are torsion-free. Thus, $t(M) \cong t(M_p \oplus M_c) \cong t(M_p) \oplus t(M_c) = 0$, i.e., M is a torsion-free A-module. Then, by Corollary 2.5 of [17], R is weakly left-prime. Now, the fact that C internally stabilizes P implies (see Definition 1.1)

$$H(P,C) = \begin{pmatrix} I_q & -P \\ -C & I_{p-q} \end{pmatrix}^{-1} = \begin{pmatrix} R_p \\ R_c \end{pmatrix}^{-1} \begin{pmatrix} D_p & 0 \\ 0 & D_c \end{pmatrix} \in M_p(A).$$

Therefore, we have

$$\begin{aligned} \begin{pmatrix} R_p \\ R_c \end{pmatrix}^{-1} R &= \begin{pmatrix} R_p \\ R_c \end{pmatrix}^{-1} \begin{pmatrix} R_p & -D_p & 0 \\ R_c & 0 & -D_c \end{pmatrix} \\ &= \left(I_p \quad - \begin{pmatrix} R_p \\ R_c \end{pmatrix}^{-1} \begin{pmatrix} D_p & 0 \\ 0 & D_c \end{pmatrix} \right) \in M_{p\times 2p}(A). \end{aligned}$$

Finally, using the fact that R is a weakly left-prime full row rank matrix, we obtain (2.1) (see [17] for more details).

$\Leftarrow$ We have

$$(2.1) \Rightarrow \begin{pmatrix} R_p \\ R_c \end{pmatrix}^{-1} \begin{pmatrix} D_p & 0 \\ 0 & D_c \end{pmatrix} = \begin{pmatrix} I_q & -P \\ -C & I_{p-q} \end{pmatrix}^{-1} \in M_p(A),$$

i.e., the controller $C = D_c^{-1} N_c$ internally stabilizes the plant $P = D_p^{-1} N_p$. □

COROLLARY 2.2. *Let $P = D_p^{-1} N_p \in M_{q\times(p-q)}(K)$ be a weakly left-coprime factorization of P. Then, P is internally stabilized by a controller $C \in M_{(p-q)\times q}(K)$ which admits a weakly left-coprime factorization $C = D_c^{-1} N_c$ iff P admits a doubly coprime factorization. The same result also holds for a stabilizable plant P admitting a weakly right-coprime factorization.*

Proof. $\Rightarrow$ Let us suppose that the plant $P = D_p^{-1} N_p$ is internally stabilized by a controller $C = D_c^{-1} N_c$ and R_p and R_c are two weakly left-prime matrices. Then, by Theorem 2.1, we have (2.1). Let us note

$$\begin{pmatrix} R_p \\ R_c \end{pmatrix}^{-1} = \begin{pmatrix} U_1 & V_1 \\ U_2 & V_2 \end{pmatrix} \in M_p(A).$$

Then, we have the following Bézout identities:

$$\begin{pmatrix} D_p & -N_p \\ -N_c & D_c \end{pmatrix} \begin{pmatrix} U_1 & V_1 \\ U_2 & V_2 \end{pmatrix} = I_p, \qquad \begin{pmatrix} U_1 & V_1 \\ U_2 & V_2 \end{pmatrix} \begin{pmatrix} D_p & -N_p \\ -N_c & D_c \end{pmatrix} = I_p. \tag{2.2}$$

In particular, we have

$$\begin{pmatrix} D_p & -N_p \\ 0 & I_{p-q} \end{pmatrix} \begin{pmatrix} U_1 & V_1 \\ U_2 & V_2 \end{pmatrix} = \begin{pmatrix} I_q & 0 \\ U_2 & V_2 \end{pmatrix} \Rightarrow \det D_p \det \begin{pmatrix} U_1 & V_1 \\ U_2 & V_2 \end{pmatrix} = \det V_2,$$

and, using the fact that the second matrix is unimodular and $\det D_p \neq 0$, we obtain that $\det V_2 \neq 0$. Finally, from (2.2), we deduce

$$\begin{cases} D_p V_1 - N_p V_2 = 0, \\ D_p U_1 - N_p U_2 = I_q, \\ -N_c V_1 + D_c V_2 = I_{p-q}, \end{cases}$$

which shows that $P = D_p^{-1} N_p = V_1 V_2^{-1}$ is a doubly coprime factorization of P.

$\Leftarrow$ If $P = D_p^{-1} N_p = \tilde{N}_p \tilde{D}_p^{-1}$ is a doubly coprime factorization of P, then there exist Bézout identities of the form (2.2). Thus, $R_p = (D_p : -N_p) \in M_{q\times p}(A)$ can be complemented into $(R_p^T : R_c^T)^T \in GL_p(A)$, with $R_c \in M_{(p-q)\times p}(A)$. The complement $R_c = (-N_c : D_c)$ to R_p into a unimodular matrix $(R_p^T : R_c^T)^T$ is not uniquely defined (see Corollary 6.1 on the Youla–Kučera parametrization) and we can choose $D_c \in M_{p-q}(A)$ such that $\det D_c \neq 0$. Finally, R_c admits a right-inverse, i.e., $C = D_c^{-1} N_c$ is in particular a weakly left-coprime factorization. Finally, by Theorem 2.1, $C = D_c^{-1} N_c$ internally stabilizes P. □

The next corollary generalizes a result obtained by Smith for $H_\infty(\mathbb{C}_+)$ [25].

COROLLARY 2.3. *If A is a coherent Sylvester domain (e.g., $A = H_\infty(\mathbb{C}_+)$, RH_∞, Bézout domains), then $P \in M_{q\times(p-q)}(K)$ is internally stabilizable iff P admits a doubly coprime factorization.*

Proof. By Theorem 3.24 of [17], every transfer matrix whose entries belong to $K = Q(A)$ admits a weakly doubly coprime factorization. Then, the result follows directly from Corollary 2.2. □

3. Internal stabilization: The general case. In the previous section, we have obtained some results on internal stabilization in the particular case where the transfer matrices admit weakly left- or right-coprime factorizations. In this section, we give some necessary and sufficient conditions for internal stabilizability without any assumption on the transfer matrices.

LEMMA 3.1. *Let $P = D_p^{-1} N_p \in M_{q\times(p-q)}(K)$ (resp., $C = D_c^{-1} N_c \in M_{(p-q)\times q}(K)$) be a plant (resp., a controller). If C internally stabilizes P, then the A-modules $M_p = A^p/A^q R_p$ and $M_c = A^p/A^{p-q} R_c$, where $R_p = (D_p : -N_p) \in M_{q\times p}(A)$, $R_c = (-N_c : D_c) \in M_{(p-q)\times p}(A)$, satisfy*

$$M_p/t(M_p) \oplus M_c/t(M_c) \cong A^p,$$

i.e., $M_p/t(M_p) = A^p/\overline{A^q R_p}$ and $M_c/t(M_c) = A^p/\overline{A^{p-q} R_c}$ are projective A-modules.

Proof. By hypothesis, P is internally stabilized by C, and thus, we have

$$H(P, C) = \begin{pmatrix} I_q & -P \\ -C & I_{p-q} \end{pmatrix}^{-1} = \begin{pmatrix} R_p \\ R_c \end{pmatrix}^{-1} \begin{pmatrix} D_p & 0 \\ 0 & D_c \end{pmatrix} = N \in M_p(A).$$

Let us define the following A-modules $M = A^{2p}/A^p R$ and $M' = A^{2p}/A^p (I_p : -N)$. By Lemma 2.6 of [17], we have $\overline{A^p R} = A^p (I_p : -N)$ because $A^p (I_p : -N)$ is an A-closed submodule of A^{2p}, and thus, we have

$$M/t(M) = A^{2p}/\overline{A^p R} = A^{2p}/A^p (I_p : -N) = M'.$$

Moreover, it is easy to see that the A-module M' is free of rank p and thus, that we have $M/t(M) \cong A^p$. Finally, using (1.4), we obtain

$$M/t(M) \cong M_p/t(M_p) \oplus M_c/t(M_c) \cong A^p,$$

which shows that $M_p/t(M_p) = A^p/\overline{A^q\, R_p}$ and $M_c/t(M_c) = A^p/\overline{A^{p-q}\, R_c}$ are projective A-modules. □

THEOREM 3.2. *A plant $P = D_p^{-1}\, N_p \in M_{q\times(p-q)}(K)$ is internally stabilizable iff $M_p/t(M_p) = A^p/\overline{A^q\, R_p}$ is a projective A-module, with $R_p = (D_p : -N_p) \in M_{q\times p}(A)$ and $M_p = A^p/A^q\, R_p$.*

Proof. $\Rightarrow$ It was proved in Lemma 3.1.

$\Leftarrow$ Let $M_p/t(M_p)$ be a projective A-module. We have the following commutative exact diagram:

$$\begin{array}{ccccccccc}
 & & & & & & & & 0 \\
 & & & & & & & & \downarrow \\
 & & 0 & & 0 & & t(M_p) & & \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & A^q & \xrightarrow{.R_p} & A^p & \xrightarrow{\pi} & M_p & \longrightarrow & 0 \\
 & & \downarrow \kappa & & \| & & \downarrow \pi' & & \\
0 & \longrightarrow & \ker\phi & \longrightarrow & A^p & \xrightarrow{\phi} & M_p/t(M_p) & \longrightarrow & 0, \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
 & & \operatorname{coker}\kappa & & 0 & & 0 & & \\
 & & \downarrow & & & & & & \\
 & & 0 & & & & & &
\end{array} \tag{3.1}$$

where $\phi = \pi' \circ \pi$ and $\kappa : A^q \to \ker\phi$ is induced by $\mathrm{id} : A^p \to A^p$ and $\pi' : M_p \to M_p/t(M_p)$. The fact that $M_p/t(M_p)$ is a projective A-module implies that the exact sequence

$$0 \longrightarrow \ker\phi \longrightarrow A^p \xrightarrow{\phi} M_p/t(M_p) \longrightarrow 0 \tag{3.2}$$

splits (see [17]), and thus, $A^p \cong M_p/t(M_p) \oplus \ker\phi$, i.e., $\ker\phi$ is a projective A-module.

The fact that $\ker\phi$ is a projective A-module is equivalent to the existence of a family $\{a_1, \dots, a_m\}$ of elements of A satisfying [3, 23]:

1. The ideal $(a_1, \dots, a_m)$ is equal to A, i.e., $\exists\, x_i \in A : \sum_{i=1}^m x_i\, a_i = 1$.
2. If $S_{a_i} = \{1, a_i, a_i^2, \dots\}$ is the multiplicative set defined by a_i, then $S_{a_i}^{-1} \ker\phi$ is a free $S_{a_i}^{-1} A$-module (see [17]).

By Proposition 1.10 of [17], we obtain the exact sequence of $S_{a_i}^{-1} A$-modules:

$$0 \longrightarrow S_{a_i}^{-1}(\ker\phi) \longrightarrow (S_{a_i}^{-1} A)^p \xrightarrow{S_{a_i}^{-1}\phi} S_{a_i}^{-1}(M_p/t(M_p)) \longrightarrow 0. \tag{3.3}$$

The fact that $t(M_p)$ is a torsion A-module implies that $K \otimes_A t(M_p) = 0$ (see (1.10) of [17]), and thus, $\mathrm{rank}_A(t(M_p)) = \dim_K(K \otimes_A t(M_p)) = 0$ (see [17] for more details). Applying Proposition 1.10 of [17] to the exact sequence

$$0 \longrightarrow t(M_p) \longrightarrow M_p \longrightarrow M_p/t(M_p) \longrightarrow 0,$$

we obtain $\mathrm{rank}_A(M_p/t(M_p)) = \mathrm{rank}_A(M_p) - \mathrm{rank}_A(t(M_p)) = p - q$. Applying again Proposition 1.10 of [17] to the exact sequence (3.2), we obtain

$$\mathrm{rank}_A(\ker\phi) = p - \mathrm{rank}_A(M_p/t(M_p)) = p - (p-q) = q. \tag{3.4}$$

If we note $S_{a_i}^{-1} A = A_i$, then $S_{a_i}^{-1} \ker \phi$ is a free A_i-module of rank q. Taking a basis of $S_{a_i}^{-1} \ker \phi \cong A_i^q$, there exists a matrix $R_i \in M_{q\times p}(A_i)$ such that (3.3) becomes

$$0 \longrightarrow A_i^q \xrightarrow{.R_i} A_i^p \longrightarrow S_{a_i}^{-1}(M_p/t(M_p)) \longrightarrow 0.$$

By hypothesis, $M_p/t(M_p)$ is a projective A-module, and thus, $S_{a_i}^{-1}(M_p/t(M_p))$ is also a projective A_i-module [3, 23]. Hence, using Proposition 4.2 of [17], the previous exact sequence splits, and thus there exists $S_i \in M_{p\times q}(A_i)$ such that

$$R_i\, S_i = I_q. \tag{3.5}$$

Let us note $R_p = (D_p :\ -N_p) \in M_{q\times p}(A)$ and $R_i = (D_i :\ -N_i) \in M_{q\times p}(A_i)$. First, we prove that $P = D_p^{-1}\, N_p = D_i^{-1}\, N_i$. By localization of (3.1) with respect to S_{a_i}, we obtain the commutative exact diagram ($S_{a_i}^{-1} A$ is a flat A-module [17])

$$\begin{array}{ccccccccc}
 & & & & & & & & 0 \\
 & & & & & & & & \downarrow \\
 & & 0 & & 0 & & & & S_{a_i}^{-1} t(M_p) \\
 & & \downarrow & & \downarrow & & & & \downarrow \\
0 & \longrightarrow & A_i^q & \xrightarrow{.R_p} & A_i^p & \longrightarrow & & & S_{a_i}^{-1} M_p & \longrightarrow & 0 \\
 & & \downarrow .R_i'' & & \| & & & & \downarrow & & \\
0 & \longrightarrow & A_i^q & \xrightarrow{.R_i} & A_i^p & \longrightarrow & & & S_{a_i}^{-1}(M_p/t(M_p)) & \longrightarrow & 0, \\
 & & \downarrow & & \downarrow & & & & \downarrow & & \\
 & & A_i^q/A_i^q\, R_i'' & & 0 & & & & 0 & & \\
 & & \downarrow & & & & & & & & \\
 & & 0 & & & & & & & &
\end{array}$$

where $R_i'' \in M_q(A_i)$ corresponds to $S_{a_i}^{-1}\kappa : A_i^q \to S_{a_i}^{-1} \ker \phi \cong A_i^q$. Hence, we have $R_p = R_i''\, R_i$, i.e.,

$$(D_p :\ -N_p) = R_i''\, (D_i :\ -N_i), \tag{3.6}$$

where $R_i'' \in M_q(A_i)$ has full rank and $S_{a_i}^{-1} t(M_p) \cong A_i^q/A_i^q\, R_i''$. Hence, we have

$$P = D_p^{-1}\, N_p = (R_i''\, D_i)^{-1} (R_i''\, N_i) = D_i^{-1}\, N_i.$$

Cleaning the denominators of each R_i and $S_i = (X_i^T :\ Y_i^T)^T$, there exists $\alpha_i \in \mathbb{Z}_+$ such that all the entries of the matrix $a_i^{\alpha_i}\, S_i\, R_i$ are in A. If $\alpha = \max_{1\leq i\leq m} \alpha_i$, then

$$a_i^\alpha\, S_i\, R_i = a_i^\alpha \begin{pmatrix} X_i\, D_i & -X_i\, N_i \\ Y_i\, D_i & -Y_i\, N_i \end{pmatrix} \in M_p(A), \qquad i = 1, \ldots, m. \tag{3.7}$$

Using the fact that $(a_1, \ldots, a_m) = A$, then there exists a family $\{b_1, \ldots, b_m\}$ of elements of A such that $\sum_{i=1}^m b_i\, a_i^\alpha = 1$. Therefore, we have

$$D_p = R_i''\, D_i \Rightarrow D_p = \sum_{i=1}^m b_i\, a_i^\alpha\, D_p = \sum_{i=1}^m b_i\, a_i^\alpha\, R_i''\, D_i, \tag{3.8}$$

$$N_p = R_i''\, N_i \Rightarrow N_p = \sum_{i=1}^m b_i\, a_i^\alpha\, N_p = \sum_{i=1}^m b_i\, a_i^\alpha\, R_i''\, N_i. \tag{3.9}$$

If we define $S = \sum_{i=1}^{m} b_i\, a_i^{\alpha}\, S_i\, D_i$, then we have

$$S = \left(\left(\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, X_i\, D_i \right)^T : \left(\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, Y_i\, D_i \right)^T \right)^T .$$

We claim that the controller $C \in M_{q\times(p-q)}(K)$, defined by

$$C = \left(\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, Y_i\, D_i \right) \left(\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, X_i\, D_i \right)^{-1},$$

internally stabilizes the plant P; i.e., we have

$$\begin{pmatrix} I_q & -P \\ -C & I_{p-q} \end{pmatrix}^{-1} = \begin{pmatrix} (I_q - PC)^{-1} & (I_q - PC)^{-1}P \\ C(I_q - PC)^{-1} & I_{p-q} + C(I_q - PC)^{-1}P \end{pmatrix} \in M_p(A).$$

We easily check that

$$\begin{aligned}
&I_q - PC \\
&\quad = I_q - D_p^{-1}\, N_p\, \left(\textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, Y_i\, D_i\right) \left(\textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, X_i\, D_i\right)^{-1} \\
&\quad = D_p^{-1}[D_p\, \left(\textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, X_i\, D_i\right) - N_p\, \left(\textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, Y_i\, D_i\right)] \left(\textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, X_i\, D_i\right)^{-1} \\
&\quad = D_p^{-1}[\textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, \left(D_p\, X_i - N_p\, Y_i\right) D_i] \left(\textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, X_i\, D_i\right)^{-1} \\
&\quad = D_p^{-1}[\textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, R_i''\, \left(D_i\, X_i - N_i\, Y_i\right) D_i] \left(\textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, X_i\, D_i\right)^{-1} \quad \text{(by (3.6))} \\
&\quad = D_p^{-1}[\textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, R_i''\, D_i] \left(\textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, X_i\, D_i\right)^{-1} \quad \text{(by (3.5))} \\
&\quad = D_p^{-1}\, D_p\, \left(\textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, X_i\, D_i\right)^{-1} \quad \text{(by (3.8))} \\
&\quad = \left(\textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, X_i\, D_i\right)^{-1}
\end{aligned}$$

$$\Rightarrow (I_q - PC)^{-1} = \textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, X_i\, D_i \in M_q(A),$$

$$\Rightarrow C(I_q - PC)^{-1} = \textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, Y_i\, D_i \in M_{(p-q)\times q}(A),$$

$$\Rightarrow (I_q - PC)^{-1}P = \textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, X_i\, N_i \in M_{q\times(p-q)}(A),$$

$$\Rightarrow I_{p-q} + C(I - PC)^{-1}P = I_{p-q} + \textstyle\sum_{i=1}^{m} b_i\, a_i^{\alpha}\, Y_i\, N_i \in M_{p-q}(A). \qquad \square$$

Remark 3.1. Let us note that the proof of Theorem 3.2 seems to be dual to the one given in [13]. The duality between the approach developed in [26], using the A-modules A^p/R^T and $A^p/\tilde{R}^T$, and the one developed here, using the A-modules $A^p/\overline{A^q\, R}$ and $A^p/\overline{A^{p-q}\, \tilde{R}^T}$, will be explained in Proposition 3.4 (see also Proposition 2.8 of [17]). We refer the reader to [20] for another proof of Theorem 3.2 and where it is shown that

$$C' = \left(\sum_{i=1}^{m} b_i a_i^{\alpha} Y_i R_i^{''-1} \right) \left(\sum_{i=1}^{m} b_i a_i^{\alpha} X_i R_i^{''-1} \right)^{-1}$$

is also a stabilizing controller of P.

Example 3.1. Let $A = H_{\infty}(\mathbb{C}_+)$ and let us consider the following transfer matrix:

$$P = \begin{pmatrix} \frac{e^{-s}}{s-1} & a \\ b & \frac{1}{s-1} \end{pmatrix} \in M_2(K), \tag{3.10}$$

where a, $b \in A$. The A-system (see [17]) which corresponds to P is defined by

$$\begin{cases} \frac{(s-1)}{(s+1)}\, y_1 - \frac{e^{-s}}{(s+1)}\, u_1 - a\, \frac{(s-1)}{(s+1)}\, u_2 = 0, \\ \frac{(s-1)}{(s+1)}\, y_2 - b\, \frac{(s-1)}{(s+1)}\, u_1 - \frac{1}{(s+1)}\, u_2 = 0, \end{cases}$$

i.e., $R\, z = 0$, where $z = (y_1 :\ y_2 :\ u_1 :\ u_2)^T$ and R is the matrix defined by

$$R = \begin{pmatrix} \frac{s-1}{s+1} & 0 & -\frac{e^{-s}}{s+1} & -a\, \frac{(s-1)}{(s+1)} \\ 0 & \frac{s-1}{s+1} & -b\, \frac{(s-1)}{(s+1)} & -\frac{1}{s+1} \end{pmatrix} \in M_{2\times 4}(A). \tag{3.11}$$

Let us check whether or not the A-module $M = A^4/A^2\, R$ is projective. We have $\mathrm{Fitt}_0(M) = 0$, $\mathrm{Fitt}_1(M) = 0$, and

$$\mathrm{Fitt}_2(M) = \left(\left(\frac{s-1}{s+1}\right)^2, \frac{(s-1)}{(s+1)^2}, \frac{e^{-s}}{(s+1)^2}, \frac{(s-1)\, e^{-s}}{(s+1)^2} \right).$$

Then, we have

$$\begin{cases} \left(\frac{s-1}{s+1}\right)^2 + 2\, \frac{(s-1)}{(s+1)^2} = \frac{s-1}{s+1} \in \mathrm{Fitt}_2(M), \\ \frac{(s-1)\, e^{-s}}{(s+1)^2} + 2\, \frac{e^{-s}}{(s+1)^2} = \frac{e^{-s}}{s+1} \in \mathrm{Fitt}_2(M). \end{cases}$$

Moreover,

$$\left(\frac{s-1}{s+1}\right) \left(1 + 2\, \left(\frac{1-e^{-(s-1)}}{s-1}\right)\right) + 2\, e\, \left(\frac{e^{-s}}{s+1}\right) = 1 \in \mathrm{Fitt}_2(M) \Rightarrow \mathrm{Fitt}_2(M) = A, \tag{3.12}$$

and thus, by Proposition 4.4 of [17], M is a projective A-module of rank 2. Thus, by Theorem 3.2, P is internally stabilizable. Let us find a controller C using the construction given in the proof of Theorem 3.2. First, let us notice that the fact that $M = A^4/A^2\, R$ is a projective A-module implies that $M/t(M) = M = A^4/A^2\, R$. Second, from (3.12), with the notations of the proof of Theorem 3.2, we have

$$\begin{cases} a_1 = \frac{s-1}{s+1} \in \mathrm{Fitt}_2(M), \\ a_2 = \frac{e^{-s}}{s+1} \in \mathrm{Fitt}_2(M), \\ b_1 = 1 + 2\frac{(1-e^{-(s-1)})}{(s-1)} \in A, \\ b_2 = 2\, e \in A. \end{cases}$$

In $A_{\frac{s-1}{s+1}}$, we have the following right-inverse $S_{\frac{s-1}{s+1}}$ of $R_{\frac{s-1}{s+1}} = R$:

$$\begin{pmatrix} \frac{s-1}{s+1} & 0 & -\frac{e^{-s}}{s+1} & -a\, \frac{(s-1)}{(s+1)} \\ 0 & \frac{s-1}{s+1} & -b\, \frac{(s-1)}{(s+1)} & -\frac{1}{s+1} \end{pmatrix} \begin{pmatrix} \frac{s+1}{s-1} & 0 \\ 0 & \frac{s+1}{s-1} \\ 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

In $A_{\frac{e^{-s}}{s+1}}$, we have the following right-inverse $S_{\frac{e^{-s}}{s+1}}$ of $R_{\frac{e^{-s}}{s+1}} = R$:

$$\begin{pmatrix} \frac{s-1}{s+1} & 0 & -\frac{e^{-s}}{s+1} & -a\, \frac{(s-1)}{(s+1)} \\ 0 & \frac{s-1}{s+1} & -b\, \frac{(s-1)}{(s+1)} & -\frac{1}{s+1} \end{pmatrix} \begin{pmatrix} 0 & -2\, a \\ -b\, \frac{(s+1)}{e^{-s}} & 1 \\ -\frac{(s+1)}{e^{-s}} & 0 \\ 0 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Hence, S is defined by

$$S = \left(\begin{array}{c} \frac{(s-1)}{(s+1)} \left(1 + 2\,\frac{(1-e^{-(s-1)})}{(s-1)}\right) \begin{pmatrix} \frac{s+1}{s-1} & 0 \\ 0 & \frac{s+1}{s-1} \\ 0 & 0 \\ 0 & 0 \end{pmatrix} + \frac{e^{-s}}{(s+1)}\, 2\, e \begin{pmatrix} 0 & -2\,a \\ -b\,\frac{(s+1)}{e^{-s}} & 1 \\ -\frac{(s+1)}{e^{-s}} & 0 \\ 0 & -2 \end{pmatrix} \end{array} \right) \frac{(s-1)}{(s+1)}\, I_2$$

$$= \frac{(s-1)}{(s+1)} \begin{pmatrix} 1 + 2\frac{(1-e^{-(s-1)})}{(s-1)} & -4\,a\,\frac{e^{-(s-1)}}{s+1} \\ -2\,e\,b & 1 + 2\,\frac{(1-e^{-(s-1)})}{(s-1)} + 2\,\frac{e^{-(s-1)}}{(s+1)} \\ -2\,e & 0 \\ 0 & -4\,\frac{e^{-(s-1)}}{s+1} \end{pmatrix}.$$

Then, a stabilizing controller C of P is defined by

$$C = \begin{pmatrix} -2\,e & 0 \\ 0 & -4\,\frac{e^{-(s-1)}}{s+1} \end{pmatrix} \begin{pmatrix} 1 + 2\,\frac{(1-e^{-(s-1)})}{(s-1)} & -4\,a\,\frac{e^{-(s-1)}}{(s+1)} \\ -2\,e\,b & 1 + 2\,\frac{(1-e^{-(s-1)})}{(s-1)} + 2\,\frac{e^{-(s-1)}}{(s+1)} \end{pmatrix}^{-1}.$$

Remark 3.2. Dually to Theorem 3.2, $P = \tilde{N}_p\,\tilde{D}_p^{-1} \in M_{q\times(p-q)}(K)$ is internally stabilized by $C = \tilde{X}_c^{-1}\,\tilde{Y}_c \in M_{(p-q)\times q}(K)$ iff $\tilde{M}_p = A^p/A^{p-q}\,\tilde{R}_p{}^T$ is such that $\tilde{M}_p/t(\tilde{M}_p)$ is a projective A-module, with $\tilde{R}_p{}^T = (\tilde{N}_p^T :\ \tilde{D}_p^T)^T \in M_{p\times(p-q)}(A)$. In order to shorten the paper, we let the readers check this result themselves. (We can use the fact that C internally stabilizes P iff C^T internally stabilizes P^T.)

COROLLARY 3.3. *If $P = D_p^{-1}\,N_p \in M_{q\times(p-q)}(K)$ is a weakly left-coprime factorization of P, then P is internally stabilizable iff the A-module $M_p = A^p/A^q\,R_p$ is stably free, i.e., iff $P = D_p^1 N_p$ is a left-coprime factorization of P. Moreover, a stabilizing controller C of P has the form*

$$C = Y_c\,X_c^{-1},$$

where $S = (X_c^T :\ Y_c^T)^T \in M_{p\times q}(A)$ is a right inverse of R_p, i.e., $D_p\,X_c - N_p\,Y_c = I_q$.

Proof. $\Rightarrow$ If $P = D_p^{-1}\,N_p$ is internally stabilizable, then, by Theorem 3.2, the A-module $A^p/\overline{A^q\,R_p}$ is a projective A-module, where $R_p = (D_p :\ -N_p) \in M_{q\times p}(A)$. Using the fact that $P = D_p^{-1}\,N_p$ is a weakly left-coprime factorization of P, then, by Lemma 2.6 and Theorem 2.11 of [17], we have $\overline{A^q\,R_p} = A^q\,R_p$. Thus, the A-module $M_p = A^p/A^q\,R_p$ is projective and, using the fact that M_p is a projective A-module and R_p is a full row rank matrix, the exact sequence $0 \longrightarrow A^q \xrightarrow{.R_p} A^p \longrightarrow M_p \longrightarrow 0$ splits [3, 23]. Thus, we have $M_p \oplus A^q \cong A^p$, i.e., M_p is a stably free A-module.

$\Leftarrow$ Let us suppose that M_p is a stably free A-module. In particular, $M_p = M_p/t(M_p)$ is a stably free A-module, and thus, by Theorem 3.2, P is internally stabilizable.

Moreover, we have the exact sequence $0 \longrightarrow A^q \xrightarrow{.R_p} A^p \longrightarrow M_p \longrightarrow 0$. Using the fact that M_p is a stably free A-module, then this exact sequence splits, i.e., there exists $S = (X_c^T :\ Y_c^T)^T \in M_{p\times q}(A)$ such that $R_p\,S = I_q$. We check that $C = Y_c\,X_c^{-1}$ is a stabilizing controller of $P = D_p^{-1}\,N_p$ by computing (1.2) [29]. (We can also use the construction of the stabilizing controller given in the proof of Theorem 3.2: $\ker\phi = A^q$, $C = (Y_c\,D_p)\,(X_c\,D_p)^{-1} = Y_c \cdot X_c^{-1}$.) ∎

Example 3.2. Let us reconsider the transfer matrix P defined by (3.10). In Example 3.1, we proved that the $A = H_\infty(\mathbb{C}_+)$-module $M = A^4/A^2\,R$, where R is defined by (3.11), is projective. Let us check whether or not the A-module M is stably free. The A-module $T(M) = A^2/A^4\,R^T$ is defined by the following equations:

$$\begin{cases} \frac{(s-1)}{(s+1)}\,\lambda_1 = 0, \\ \frac{(s-1)}{(s+1)}\,\lambda_2 = 0, \\ -\frac{e^{-s}}{(s+1)}\,\lambda_1 - b\,\frac{(s-1)}{(s+1)}\,\lambda_2 = 0, \\ -a\,\frac{(s-1)}{(s+1)}\,\lambda_1 - \frac{1}{(s+1)}\,\lambda_2 = 0. \end{cases} \tag{3.13}$$

If we denote by $\mu = (\mu_1 :\ \mu_2 :\ \mu_3 :\ \mu_4)^T$ the second member of (3.13), we have

$$\begin{cases} \lambda_1 = (1 + 2\,\frac{(1-e^{-(s-1)})}{(s-1)})\,\mu_1 - 2\,e\,b\,\mu_2 - 2\,e\,\mu_3, \\ \lambda_2 = -2\,a\,\mu_1 + \mu_2 - 2\,\mu_4, \end{cases}$$

which proves that, from (3.13), we can deduce $\lambda_1 = \lambda_2 = 0$, i.e., $T(M) = 0$, and thus, by 2 of Proposition 4.2 of [17], M is a stably free A-module. Moreover, a right-inverse S of R, i.e., $R\,S = I_2$, is defined by

$$S = \begin{pmatrix} 1 + 2\,\frac{(1-e^{-(s-1)})}{(s-1)} & -2\,a \\ -2\,e\,b & 1 \\ -2\,e & 0 \\ 0 & -2 \end{pmatrix}. \tag{3.14}$$

Thus, a stabilizing controller C of P is defined by

$$C = \begin{pmatrix} -2\,e & 0 \\ 0 & -2 \end{pmatrix} \begin{pmatrix} 1 + 2\,\frac{(1-e^{-(s-1)})}{(s-1)} & -2\,a \\ -2\,e\,b & 1 \end{pmatrix}^{-1}.$$

The next example shows a situation where Corollary 3.3 cannot be used to construct a stabilizing controller for a plant.

Example 3.3. Let us consider the ring $A = \mathbb{R}[t_0, t_1]/(t_0^2 + t_1^2 - 1)$ of polynomials on the unit circle and x_i the class of t_i in A. We have $A = \mathbb{R}[x_0, x_1]$ with the relation $x_0^2 + x_1^2 = 1$. Let $0 \neq a,\, b \in \mathbb{R}$ be such that $a^2 + b^2 = 1$ and let us consider

$$p = (b - x_1)/(x_0 - a) \in K = Q(A). \tag{3.15}$$

It is easy to check that $R = (x_0 - a :\ x_1 - b) \in M_{1\times 2}(A)$ is not weakly left-prime:

$$\left(\tfrac{x_0+a}{x_1-b}\right)\,(x_0 - a :\ x_1 - b) = (-(x_1 + b) :\ x_0 + a) \in A \quad \not\Rightarrow \quad (x_0 + a)/(x_1 - b) \in A.$$

Therefore, by Corollary 2.5 of [17], the A-module $M = A^2/A\,R$ is not torsion-free. We can show that the torsion submodule $t(M)$ of M is generated by

$$z = (x_1 + b)\,y - (x_0 + a)\,u,$$

which satisfies $(x_1 - b)\,z = 0$. In particular, M is not a free A-module, a fact that implies that there do not exist r and s in A such that $(b - x_1)\,s - (x_0 + a)\,r = 1$, i.e.,

p does not admit a coprime factorization. Moreover, we have $M/t(M) = A^2/A^2\,R'$, where R' is defined by

$$R' = \begin{pmatrix} x_0 - a & x_1 - b \\ x_1 + b & -x_0 - a \end{pmatrix} \in M_2(A) \tag{3.16}$$

and we easily check that

$$\begin{cases} \mathrm{Fitt}_0(M/t(M)) = (-x_0^2 + a^2 - x_1^2 + b^2) = 0, \\ \mathrm{Fitt}_1(M/t(M)) = (x_0 - a, x_0 + a, x_1 - b, x_1 + b). \end{cases}$$

Moreover, we have

$$(x_0 + a)/2\,a - (x_0 - a)/2\,a = 1 \in \mathrm{Fitt}_1(M/t(M)) \Rightarrow \mathrm{Fitt}_1(M/t(M)) = A. \tag{3.17}$$

Thus, by Proposition 4.4 of [17], we obtain that $M/t(M)$ is a projective A-module of rank 1 and, then, by Theorem 3.2, p is internally stabilizable. Hence, we are in a situation where Corollary 3.3 cannot be used to determine a stabilizing controller of p because p does not admit any weakly coprime factorization. ($\overline{A\,R} = A^2\,R'$ and $A^2\,R'$ is not a free A-module.)

We show how to construct a stabilizing controller c for p by following the explicit construction given in the proof of Theorem 3.2. Using the fact $M/t(M) = A^2/A^2\,R'$, we obtain that $\ker\phi$ defined by (3.2) satisfies $\ker\phi = A^2\,R'$, where

$$A^2\,R' = \{\lambda_1\,(x_0 - a :\ x_1 - b) + \lambda_2\,(x_1 + b :\ -(x_0 + a)) \mid \lambda_1, \lambda_2 \in A\}.$$

Let $\alpha = (x_0 - a :\ x_1 - b)$ and $\beta = (x_1 + b :\ -(x_0 + a))$. We have the relations

$$\begin{cases} (x_0 + a)\,\alpha + (x_1 - b)\,\beta = 0, \\ (x_1 + b)\,\alpha - (x_0 - a)\,\beta = 0. \end{cases}$$

$A_{x_0+a} \otimes_A \ker\phi$ is a free A_{x_0+a}-module generated by β because we have

$$\alpha = -[(x_1 - b)/(x_0 + a)]\,\beta.$$

Thus, we have $A_{x_0+a} \otimes_A (M/t(M)) = A^2_{x_0+a}/A_{x_0+a}\,(x_1 + b :\ -(x_0 + a))$ and we have

$$\begin{cases} -\frac{(x_1-b)}{(x_0+a)}\ (x_1 + b :\ -(x_0 + a)) = (x_0 - a :\ x_1 - b) \Rightarrow R''_{x_0+a} = -\frac{(x_1-b)}{(x_0+a)}, \\ (x_1 + b :\ -(x_0 + a))\ \begin{pmatrix} 0 \\ \frac{-1}{x_0+a} \end{pmatrix} = 1. \end{cases}$$

$A_{x_0-a} \otimes_A \ker\phi$ is a free A_{x_0-a}-module generated by α because we have

$$\beta = [(x_1 + b)/(x_0 - a)]\,\alpha.$$

Thus, we have $A_{x_0-a} \otimes_A (M/t(M)) = A^2_{x_0-a}/A_{x_0-a}\,(x_0 - a :\ x_1 - b)$, and

$$\begin{cases} (x_0 - a :\ x_1 - b) = (x_0 - a :\ x_1 - b) \Rightarrow R''_{x_0-a} = 1, \\ (x_0 - a :\ x_1 - b)\ \begin{pmatrix} \frac{1}{x_0-a} \\ 0 \end{pmatrix} = 1. \end{cases}$$

Hence, from (3.17), we obtain

$$S = \tfrac{(x_0+a)}{2\,a}\ \begin{pmatrix} 0 \\ \frac{-1}{x_0+a} \end{pmatrix}\ (x_1 + b) - \tfrac{(x_0-a)}{2\,a}\ \begin{pmatrix} \frac{1}{x_0-a} \\ 0 \end{pmatrix}\ (x_0 - a) = -\tfrac{1}{2\,a}\ \begin{pmatrix} x_0 - a \\ x_1 + b \end{pmatrix},$$

and thus, the controller defined by

$$c = \left(-\tfrac{(x_1+b)}{2\,a}\right) / \left(-\tfrac{(x_0-a)}{2\,a}\right) = \tfrac{(x_1+b)}{(x_0-a)}$$

internally stabilizes p. We can easily check that we have

$$\begin{pmatrix} 1 & -p \\ -c & 1 \end{pmatrix}^{-1} = -\frac{1}{2\,a} \begin{pmatrix} x_0-a & -x_1+b \\ x_1+b & x_0-a \end{pmatrix} \in M_2(A).$$

Remark 3.3. Let us notice that Corollary 2.3 also follows from Corollary 3.3: If A satisfies the conditions of Corollary 2.3, then, by Corollary 3.22 of [17], there exists a weakly left-prime matrix $R'_p = (D'_p : -N'_p) \in M_{q\times p}(A)$ such that $P = {D'_p}^{-1} N'_p$. By Corollary 3.3, P is internally stabilizable iff P admits a left-coprime factorization, i.e. the A-module $M'_p = A^p/A^q\,R'_p$ is a stably free A-module (see Proposition 4.7 of [17]). Using the fact that A is a projective-free ring, and thus, a Hermite ring, then M'_p is a free A-module and, by Proposition 4.9 of [17], P is internally stabilizable iff P admits a doubly coprime factorization.

Proposition 3.4. *Let $R \in M_{q\times p}(A)$ and $M = A^p/A^q\,R$ be an A-module. Then, $M/t(M) = A^p/\overline{A^q\,R}$ is a projective A-module iff $A^p\,R^T$ is a projective A-module.*

Proof. $\Rightarrow$ Let $M/t(M)$ be a projective A-module. We have the commutative exact diagram

$$\begin{array}{ccccccccc}
 & & & & 0 & & 0 & & \\
 & & & & \downarrow & & \downarrow & & \\
 & & & & \ker\kappa & & t(M) & & \\
 & & & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & \ker.R & \longrightarrow & A^q & \xrightarrow{.R} A^p & \xrightarrow{\pi} & M & \longrightarrow 0 \\
 & & & & \downarrow\kappa & \| & & \downarrow\pi' & \\
 & & 0 & \longrightarrow & \ker\phi & \longrightarrow A^p & \xrightarrow{\phi} & M/t(M) & \longrightarrow 0, \\
 & & & & \downarrow & \downarrow & & \downarrow & \\
 & & & & \operatorname{coker}\kappa & 0 & & 0 & \\
 & & & & \downarrow & & & & \\
 & & & & 0 & & & &
\end{array} \tag{3.18}$$

where $\phi = \pi' \circ \pi$ and $\kappa : A^q \to \ker\phi$ is induced by $\mathrm{id} : A^p \to A^p$ and $\pi' : M \to M/t(M)$. Thus, by the snake lemma [3, 23], we obtain $\ker\kappa \cong \ker.R$ and $\operatorname{coker}\kappa \cong t(M)$. $M/t(M)$ is a projective A-module, and thus, the last horizontal exact sequence splits and $A^p \cong \ker\phi \oplus M/t(M)$. Then, $\ker\phi$ is a finitely generated projective A-module. Therefore, its dual $(\ker\phi)^\star \triangleq \hom_A(\ker\phi, A)$ is also a projective A-module [3, 23]. Dualizing the previous diagram and using the fact that $t(M)^\star \triangleq \hom_A(t(M), A) = 0$, we obtain the following commutative exact diagram:

$$\begin{array}{ccccccccc}
 & & & & 0 & & 0 & & \\
 & & & & \uparrow & & \uparrow & & \\
0 & \longleftarrow & A^p\,R^T & \xleftarrow{.R^T} & A^p & \longleftarrow & M^\star & \longleftarrow 0 & \\
 & & & & \| & & \uparrow & & \\
0 & \longleftarrow & (\ker\phi)^\star & \longleftarrow & A^p & \longleftarrow & (M/t(M))^\star & \longleftarrow 0. & \\
 & & & & \uparrow & & \uparrow & & \\
 & & & & 0 & & 0 & &
\end{array}$$

Hence, we deduce that $A^p\,R^T \cong (\ker\phi)^\star$, and thus, $A^p\,R^T$ is a projective A-module.

$\Rightarrow$ Let $A^p\,R^T$ be a projective A-module. Then, the exact sequence

$$0 \longleftarrow A^p\,R^T \xleftarrow{.R^T} A^p \longleftarrow M^\star \longleftarrow 0$$

splits, and thus, we have $A^p \cong A^p\,R^T \oplus M^\star$, which implies that $M^\star \triangleq \hom_A(M,A)$ is a finitely generated projective A-module, and thus, $M^{\star\star}$ is also a projective A-module [3, 23]. Moreover, using the fact that $M^\star$ is a finitely generated A-module, then $M^\star$ has a finite free resolution [3], and thus, $T(M) = A^q/A^p\,R^T$ has a finite free resolution:

$$0 \longleftarrow T(M) \longleftarrow A^q \xleftarrow{.R^T} A^p \xleftarrow{.R^T_{-1}} A^n \xleftarrow{.R^T_{-2}} A^m \xleftarrow{.R^T_{-3}} \dots .$$

Dualizing this exact sequence, we obtain the following complex:

$$0 \longrightarrow A^q \xrightarrow{.R} A^p \xrightarrow{.R_{-1}} A^n \xrightarrow{.R_{-2}} A^m \xrightarrow{.R_{-3}} \dots .$$

Therefore, we have the following exact sequence (see [3] for more details):

$$0 \longrightarrow \text{ext}^1_A(T(M),A) \longrightarrow M \longrightarrow \ker .R_{-2} \longrightarrow \text{ext}^2_A(T(M),A) \longrightarrow 0.$$

Moreover, we have the exact sequence $0 \longleftarrow M^\star \longleftarrow A^n \xleftarrow{.R^T_{-2}} A^m$, which gives by duality the exact sequence $0 \longrightarrow M^{\star\star} \longrightarrow A^n \xrightarrow{.R_{-2}} A^m$, from which we deduce that $\ker .R_{-2} = M^{\star\star}$. Hence, we obtain the following exact sequence [14]:

$$0 \longrightarrow \text{ext}^1_A(T(M),A) \longrightarrow M \xrightarrow{\epsilon} M^{\star\star} \longrightarrow \text{ext}^2_A(T(M),A) \longrightarrow 0.$$

We have $\text{ext}^2_A(T(M),A) \cong \text{ext}^1_A(A^p\,R^T,A) = 0$ because $A^p\,R^T$ is a projective A-module [3, 23]. Using the fact that M is a finitely presented A-module, we have the following commutative exact diagram (see [14] for more explanations):

$$\begin{array}{ccccccccc}
 & & & & & & & & 0 \\
 & & & & & & & & \downarrow \\
 & & 0 & & 0 & & 0 & & t(M) \\
 & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\
0 & \longrightarrow & \hom_A(T(M),A) & \longrightarrow & A^q & \xrightarrow{.R} & A^p & \longrightarrow & M & \longrightarrow 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \downarrow & \\
0 & \longrightarrow & \hom_A(T(M),K) & \longrightarrow & K^q & \xrightarrow{.R} & K^p & \longrightarrow & K\otimes_A M & \longrightarrow 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \downarrow & \\
0 & \longrightarrow & \hom_A(T(M),K/A) & \longrightarrow & (K/A)^q & \longrightarrow & (K/A)^p & \longrightarrow & (K/A)\otimes_A M & \longrightarrow 0, \\
 & & \downarrow & & \downarrow & & \downarrow & & \downarrow & \\
 & & \text{ext}^1_A(T(M),A) & & 0 & & 0 & & 0 & \\
 & & \downarrow & & & & & & & \\
 & & \text{ext}^1_A(T(M),K)=0 & & & & & & &
\end{array}$$

where $\text{ext}^1_A(T(M),K) = 0$ because K is an *injective A-module* [3]. Thus, a chase in the diagram shows that $\text{ext}^1_A(T(M),A) \cong t(M)$. Finally, using the fact that $\text{ext}^1_A(T(M),A) \cong t(M)$, we have $M/t(M) \cong M^{\star\star}$. The result follows from the fact that $M^{\star\star}$ is projective, and thus, so is $M/t(M) = A^p/\overline{A^q\,R}$. □

Using Theorem 3.2 and Proposition 3.4, we obtain the following corollary.

COROLLARY 3.5 (see [26]). *The system $P = D^{-1}\,N \in M_{q\times(p-q)}(K)$ is internally stabilizable iff the A-module $A^p\,R^T$ is projective, where $R = (D:\,-N) \in M_{q\times p}(A)$.*

From Corollary 3.5, we deduce the next result. We refer to [20] for more details and a direct proof of this result.

COROLLARY 3.6. *The system $P = D^{-1}\,N \in M_{q\times(p-q)}(K)$ is internally stabilizable iff there exists $S = (X^T : Y^T)^T \in M_{p\times q}(K)$ such that*

1. $S\,R = \begin{pmatrix} X\,D & -X\,N \\ Y\,D & -Y\,N \end{pmatrix} \in M_p(A)$,
2. $R\,S = D\,X - N\,Y = I_q$,

where $R = (D : -N) \in M_{q\times p}(A)$. Then, $C = Y\,X^{-1}$ internally stabilizes P.

PROPOSITION 3.7. *Let $P = (P_1 + P_2) \in M_{q\times(p-q)}(K)$ be a transfer matrix where $P_1 \in M_{q\times(p-q)}(A)$ is the stable part of P and $P_2 \in M_{q\times(p-q)}(K)$ the instable one. Then, we have the following results:*

(1) *P is internally stabilizable iff P_2 is internally stabilizable.*

(2) *If $P_2 = D_2^{-1}\,N_2$ admits a left-coprime factorization and $S_2 = (X_2^T : Y_2^T)^T$ is a right-inverse of $R_2 = (D_2 : -N_2) \in M_{q\times p}(A)$, then a stabilizing controller of P is given by $C = C_2\,(I_q + P_1\,C_2)^{-1}$, where $C_2 = Y_2\,X_2^{-1}$ is a stabilizing controller of P_2. A similar result exists if P_2 admits a right-coprime factorization.*

Proof. (1) Let us suppose that $P_2 = D_2^{-1}\,N_2$ is a fractional representation of P_2 where $R_2 = (D_2 : -N_2) \in M_{q\times p}(A)$. Then, P has the following fractional representation: $P = D_2^{-1}(D_2\,P_1 + N_2)$ with $R = (D_2 : -(D_2\,P_1 + N_2)) \in M_{q\times p}(A)$. Let $M = A^p/A^q\,R$ and $M_2 = A^p/A^q\,R_2$; then we have to prove that the A-module $M/t(M)$ is projective iff $M_2/t(M_2)$ is projective or, equivalently by Proposition 3.4, that the A-module $A^q\,R^T$ is projective iff $A^q\,R_2^T$ is also projective. But, we have trivially $A^q\,R^T = A^q\,R_2^T$.

(2) The A-module $T(M_2) = A^q/A^p\,R_2^T$ is defined by the following equations:

$$\begin{cases} D_2^T\,\lambda = 0, \\ -(N_2^T + P_1^T\,D_2^T)\,\lambda = 0. \end{cases}$$

Putting a second member $\mu = (\mu_1^T : \mu_2^T)^T$ in the previous equations and using the fact that S_2 is a right-inverse of R_2, we obtain $\lambda = (X_2^T + Y_2^T\,P_1^T)\,\mu_1 + Y_2^T\,\mu_2$, i.e., $S = ((X_2 + P_1\,Y_2)^T : Y_2^T)^T$ is a right-inverse of R. Therefore, by Corollary 3.3,

$$\begin{aligned} C &= Y_2\,(X_2 + P_1\,Y_2)^{-1} = Y_2\,((I + P_1\,Y_2\,X_2^{-1})\,X_2)^{-1} \\ &= Y_2\,X_2^{-1}\,(I + P_1\,(Y_2\,X_2^{-1}))^{-1} = C_2\,(I + P_1\,C_2)^{-1} \end{aligned}$$

is a stabilizing controller of P. □

PROPOSITION 3.8. *A system of the form $P \in M_{1\times(p-1)}(K)$ is internally stabilizable iff one of the following assertions is satisfied:*

- *The ideal $I = (a_1, \ldots, a_p)$ is* invertible [22, 23]*, namely we have*

$$I\,(A : I) \triangleq \left\{ \sum_{i=1}^{n} a_i\,b_i \,|\, a_i \in I,\, b_i \in (A : I) \right\} = A, \tag{3.19}$$

  where $(A : I) = \{k \in K = Q(A) \mid (k)\,I \subseteq A\}$ is a fractional ideal of A *and $P = d^{-1}\,N$, $0 \neq d \in A$, $N \in M_{1\times(p-1)}(A)$, $a_1 = d$, and $a_i = N_i$ for $2 \leq i \leq p$.*

- *For $i = 1, \ldots, p$, there exist $x_i \in K = Q(A)$ such that*

$$\begin{cases} \sum_{i=1}^{p} a_i\,x_i = 1, \\ a_i\,x_j \in A, \qquad i, j = 1, \ldots, p. \end{cases} \tag{3.20}$$

Then, the inverse $I^{-1} \triangleq A : I$ *of* I *is defined by* $I^{-1} = (x_1, \ldots, x_p)$ *and*

$$C = -\left(x_2/x_1 : \ldots : x_p/x_1\right)^T \in M_{(p-1)\times 1}(K) \tag{3.21}$$

internally stabilizes P.

Proof. By Theorem 3.2, a plant defined by $P = d^{-1}\,N \in M_{1\times(p-1)}(K)$ is internally stabilizable iff the A-module $M = A^p/A\,R$ is such that the A-module $M/t(M)$ is projective, where $R = (d : -N) = (a_1 : \ldots : a_p) \in M_{1\times p}(A)$. $A^p\,R^T$ is the ideal $I = (a_1, \ldots, a_p)$ of A. Thus, by Proposition 3.4, $M/t(M)$ is a projective A-module iff the ideal $I = (a_1, \ldots, a_p)$ is also a projective A-module. Using the fact that $I \neq 0$, then I is a projective A-module iff I is an invertible ideal, i.e., $I\,(A : I) = A$ [2, 22, 23]. Finally, (3.20) is just (3.19) written in terms of equations (see [23]). □

4. Internal stabilization of SISO plants. The following corollary of Proposition 3.8 gives a characterization of internal stabilization for SISO plants.

COROLLARY 4.1. *A SISO plant, defined by* $p = n/d$ *(*$0 \neq d,\ n \in A$*), is internally stabilizable iff one of the following equivalent assertions is satisfied:*

- *The ideal* $I = (n,\ d)$ *is invertible, i.e., we have*

$$I\,(A : I) = A, \tag{4.1}$$

  where $A : I = \{k \in K = Q(A) \mid k\,n,\ k\,d \in A\}$ *is a fractional ideal of* A.
- *There exist* $x,\ y \in K = Q(A)$ *such that*

$$\begin{cases} d\,x - n\,y = 1, \\ d\,x,\ n\,x,\ d\,y,\ n\,y \in A. \end{cases} \tag{4.2}$$

Then, $I^{-1} = A : I = (x, y)$ *and* $c = y/x$ *internally stabilizes* $p = n/d$.

Remark 4.1. We can also check Corollary 4.1 by computing

$$\begin{pmatrix} 1 & -n/d \\ -y/x & 1 \end{pmatrix}^{-1} = \frac{1}{(d\,x - n\,y)} \begin{pmatrix} d\,x & n\,x \\ d\,y & d\,x \end{pmatrix} \in M_2(A),$$

because $d\,x - n\,y = 1$ and $d\,x,\ n\,x,\ d\,y \in A$. We refer to [16, 19] for more characterizations of stabilization problems of SISO plants in terms of fractional ideals.

Example 4.1. Let us consider the ring $A = \mathbb{R}[t_0, t_1]/(t_0^2 + t_1^2 - 1)$ of polynomials on the unit circle S^1. Let x_i be the class of t_i in $\mathbb{R}_1$ and let us reconsider

$$p = (b - x_1)/(x_0 - a) \in K = Q(A), \quad \text{where } a^2 + b^2 = 1, \quad 0 \neq a, b \in \mathbb{R}.$$

Let us define the ideal $I = (b - x_1,\ x_0 - a)$ of A; then, using the fact that

$$(x_0 - a)\,(x_0 + a) = (b - x_1)\,(b + x_1),$$

we have $A : I = (1, (x_0 + a)/(b - x_1))$ and

$$\left(\tfrac{-1}{2a}\right)\,(x_0 - a) - \left(-\tfrac{x_0+a}{2a(b-x_1)}\right)\,(b - x_1) = 1 \in I\,(A : I) \Rightarrow I\,(A : I) = A.$$

Thus, $c = (x_0+a)/(b-x_1) = (x_1+b)/(x_0-a)$ internally stabilizes $p = (b-x_1)/(x_0-a)$.

Example 4.2. Let us consider $p = (1 + i\sqrt{5})/2$ [1]. Let us define the ideal $I = (2, 1 + i\sqrt{5})$ of $A = \mathbb{Z}[i\sqrt{5}]$. Using the fact that $6 = 2 \times 3 = (1 - i\sqrt{5})(1 + i\sqrt{5})$, we obtain that $A : I = \left(1, (1 - i\sqrt{5})/2\right)$. Moreover, we have

$$(-1)\,2 - \left(-\tfrac{1-i\sqrt{5}}{2}\right)\,(1 + i\sqrt{5}) = 1 \in I\,(A : I) \Rightarrow I\,(A : I) = A,$$

and thus, $c = (1 - i\sqrt{5})/2$ is a stabilizing controller of the plant $p = (1 + i\sqrt{5})/2$.

LEMMA 4.2 (see [15]). *Let $I = (n, d)$ be an ideal of A such that $d \neq 0$; then we have*

$$I\,(A : I) = (d :\ n) + (n :\ d),$$

where $(a :\ b) \triangleq \{c \in A \mid c\,b \in (a)\}$ for all $a,\ b \in A$ [24].

Proof. Let us prove that $(d :\ n) + (n :\ d) \subseteq I\,(A : I)$. Let us choose an element $a \in (d :\ n) = \{b \in A \mid \exists\, k \in A : b\,n = k\,d\}$; then we have

$$\begin{cases} (a/d)\,n = k \in A, \\ (a/d)\,d = a \in A, \end{cases} \Rightarrow (a/d) \in (A : I),\ d \in I \Rightarrow a = d\,(a/d) \in I\,(A : I).$$

Similarly, we prove that $b/n \in (A : I)$, and using the fact that $n \in I$, we obtain that $b = n\,(b/n) \in I\,(A : I)$. Finally, any element $c \in (d :\ n) + (n :\ d)$ can be written as $c = a + b$ with $a \in (d :\ n)$ and $b \in (n :\ d)$, and thus, $c = d\,(a/d) + n\,(b/n) \in I\,(A : I)$, which proves the first inclusion. Second, let us prove that $I\,(A : I) \subseteq (d :\ n) + (n :\ d)$. Any element $c \in I\,(A : I)$ can be written as

$$c = \left(\sum_{i=1}^{l} a_i\,x_i\right) n + \left(\sum_{j=1}^{m} b_j\,x_j\right) d,$$

where $a_i,\ b_j \in A$ and $x_i \in K$ is such that $x_i\,n \in A$ and $x_i\,d \in A$. We have $d\,(\sum_{i=1}^{l} a_i\,x_i\,n) = (\sum_{i=1}^{l} a_i\,x_i\,d)\,n \in (n)$ because $\sum_{i=1}^{l} a_i\,x_i\,d \in A$. In a similar way, we have $n\,(\sum_{j=1}^{m} b_j\,x_j\,d) = (\sum_{j=1}^{m} b_j\,x_j\,n)\,d \in (d)$, and thus, $c \in (d :\ n) + (n :\ d)$, which concludes the proof. □

Using Lemma 4.2, we have the following corollary of Proposition 4.1.

COROLLARY 4.3 (see [24]). *A SISO plant, defined by $p = n/d$ ($0 \neq d,\ n \in A$), is internally stabilizable iff $(d :\ n) + (n :\ d) = A$.*

5. Characterization of the classes of internal stabilizable plants. The following proposition characterizes the integral domains A of SISO stable plants over which every plant is internally stabilizable. We refer to section 3.2 of [17] for the definition of a *Prüfer domain*.

PROPOSITION 5.1 (see [6, 23]). *An integral domain A is a Prüfer domain iff every finitely generated torsion-free A-module M is projective.*

THEOREM 5.2 (see [15]). *We have the equivalences:*

1. *every MIMO plant is internally stabilizable,*
2. *every SISO plant is internally stabilizable,*
3. *A is a Prüfer domain.*

Proof. $1 \Rightarrow 2$ follows from the fact that MIMO plants contain SISO plants.

$2 \Rightarrow 3$ Let us suppose that every SISO system, defined by $p = n/d$, is internally stabilizable. Then, $R = (d :\ -n) \in M_{1\times 2}(A)$ has full row rank. By Theorem 3.2, the A-module $M = A^2/A\,R$ is such that $M/t(M)$ is a projective A-module. But, $A^2\,R^T = (n, d)$ is the ideal of A defined by n and $0 \neq d$. By Proposition 3.4, $M/t(M)$ is a projective A-module iff $I = (n, d)$ is a projective A-module. Hence, every ideal I, generated by two elements n and $0 \neq d$ of A, is a projective A-module, a result which is equivalent to the fact that A is a Prüfer domain (see Lemma 3 of [9]).

$3 \Rightarrow 1$ Let us note $K = Q(A)$ and $P = D^{-1}\,N \in M_{q\times(p-q)}(K)$. Let us define the A-module $M = A^p/A^q\,R$, where $R = (D :\ -N) \in M_{q\times p}(A)$. By hypothesis, A is a

Prüfer domain, and thus, by Proposition 5.1, the torsion-free A-module $M/t(M)$ is projective. Finally, by Theorem 3.2, P is internally stabilizable. □

Example 5.1. We have the following examples of Prüfer domains.

- The domain of entire functions $E(k)$ is a Bézout domain ($k = \mathbb{R}, \mathbb{C}$) [8], and thus, a Prüfer domain [17]. So is $\mathcal{E} = \mathbb{R}(s)[e^{-s}] \cap E(\mathbb{R})$ [11] and RH_∞ [29].
- The integral closure of $\mathbb{Z}$ into a finite extension of $\mathbb{Q}$ is a Dedekind domain, and thus, a Prüfer domain (see section 3.2 of [17] for more details). For instance, the integral closure of $\mathbb{Z}$ in $\mathbb{Q}(i\sqrt{5})$ is the Dedekind domain $\mathbb{Z}[i\sqrt{5}]$.
- If A is a one-dimensional Noetherian domain, K is its field of fractions, and L is a finite algebraic extension field of K, then the integral closure of A in L is a Dedekind domain, and thus, a Prüfer domain. In particular, a nonsingular algebraic surface defines a Dedekind affine domain. For instance, the ring $\mathbb{R}[t_0, t_1]/(t_0^2 + t_1^2 - 1)$ of polynomials on the unit circle is a Dedekind domain.
- If X is an affine irreducible nonsingular real algebraic variety of dimension $m+1$ and Y is any subset of X, then the ring $\mathcal{H}_Y(X)$ of rational functions on X, which are locally bounded on Y (i.e., for all $y \in Y$, there exist a neighborhood $\mathcal{V}(y)$ and a positive real number $M(y)$ such that $|n(x)/d(x)| \leq M(y)$ for all $x \in \mathcal{V}(y) \backslash (d^{-1}(0)y)$, is a Prüfer domain and every finitely generated ideal of $\mathcal{H}_Y(X)$ is generated by $m+1$ elements [10]. More generally, the ring of *meromorphic bounded Nash functions* on a *Nash submanifold* of $\mathbb{R}^m$ is a Prüfer domain [10].
- The integral domain $A = \{P \in \mathbb{Q}[x] \mid P(\mathbb{Z}) \subseteq \mathbb{Z}\}$ of $\mathbb{Z}$-valued polynomials in $\mathbb{Q}[x]$ is a Prüfer domain [6].

6. Youla–Kučera parametrization of the stabilizing controllers. The matrices S and S_{-1} defined in Proposition 4.9 of [17] are defined up to an arbitrary matrix which corresponds to the free parameter in the *Youla–Kučera parametrization* [4, 29].

COROLLARY 6.1. *With the same hypothesis as in Proposition* 4.9 *of* [17]*, we have the following splitting exact sequence:*

$$0 \longrightarrow A^q \underset{\cdot S(Q)}{\overset{\cdot R}{\underset{\longleftarrow}{\longrightarrow}}} A^p \underset{\cdot S_{-1}(Q)}{\overset{\cdot R_{-1}}{\underset{\longleftarrow}{\longrightarrow}}} A^{p-q} \longrightarrow 0, \tag{6.1}$$

with

$$\begin{cases} S_{-1}(Q) = S_{-1} + Q\, R, \\ S(Q) = S - R_{-1}\, Q, \end{cases} \tag{6.2}$$

where R_{-1}, S, *and* S_{-1} *are defined in Proposition* 4.9 *of* [17] *and* $Q \in M_{(p-q)\times q}(A)$. *This is equivalent to the following two Bézout identities:*

(1) $\begin{pmatrix} S(Q) & R_{-1} \end{pmatrix} \begin{pmatrix} R \\ S_{-1}(Q) \end{pmatrix} = I_p$,

(2) $\begin{pmatrix} R \\ S_{-1}(Q) \end{pmatrix} \begin{pmatrix} S(Q) & R_{-1} \end{pmatrix} = \begin{pmatrix} I_q & 0 \\ 0 & I_{p-q} \end{pmatrix} = I_p$.

Proof. We have the following relations which prove the identities (1) and (2):

- $S(Q)\, R + R_{-1}\, S_{-1}(Q) = S\, R + R_{-1}\, S_{-1} = I_p$,
- $R\, S(Q) = R\, S = I_q$,
- $S_{-1}(Q)\, R_{-1} = S_{-1}\, R_{-1} = I_{p-q}$,
- $$\begin{aligned} S_{-1}(Q)\, S(Q) &= S_{-1}\, S - S_{-1}\, R_{-1}\, Q + Q\, R\, S - Q\, R\, R_{-1}\, Q = Q - Q \\ &= 0. \end{aligned}$$ □

COROLLARY 6.2. *Let $P \in M_{q\times(p-q)}(K)$ be a transfer matrix which admits a doubly coprime factorization. Then, all the stabilizing controllers of P are parametrized by means of the Youla–Kučera parametrization*

$$C(Q) = Y(Q)\, X(Q)^{-1} = \tilde{X}(Q)^{-1}\, \tilde{Y}(Q),$$

where $Q \in M_{(p-q)\times q}(A)$ is a free parameter such that $\det X(Q) \neq 0$, $\det \tilde{X}(Q) \neq 0$, and $S_1(Q) = (-\tilde{Y}(Q) : \tilde{X}(Q))$ and $S(Q) = (X(Q)^T : Y(Q)^T)^T$ are defined by (6.2).

Example 6.1. In Example 4.3 of [17], we proved that the $A = H_\infty(\mathbb{C}_+)$-module $M = A^2/A\,R$, with $R = (\frac{s-1}{s+1} : \frac{e^{-s}}{s+1}) \in M_{1\times 2}(A)$, is projective and thus free because A is a coherent Sylvester domain (see Corollary 3.31 of [17]). Few computations lead to the following Bézout identity ($q \in A$):

$$\begin{pmatrix} \frac{s-1}{s+1} & -\frac{e^{-s}}{s+1} \\ 2\,e + \frac{(s-1)}{(s+1)}\,q & 1 + 2\,\frac{(1-e^{-(s-1)})}{(s-1)} - \frac{e^{-s}}{(s+1)}\,q \end{pmatrix} \begin{pmatrix} 1 + 2\,\frac{(1-e^{-(s-1)})}{(s-1)} - \frac{e^{-s}}{(s+1)}\,q & \frac{e^{-s}}{s+1} \\ -2\,e - \frac{(s-1)}{(s+1)}\,q & \frac{s-1}{s+1} \end{pmatrix} = I_2.$$

Thus, all the stabilizing controllers of $p = e^{-s}/(s-1)$ are parametrized by

$$c(q) = \frac{-(2\,e + \frac{(s-1)}{(s+1)}\,q)}{1 + 2\,\frac{(1-e^{-(s-1)})}{(s-1)} - \frac{e^{-s}}{(s+1)}\,q}, \quad q \in A.$$

THEOREM 6.3. *If A is a projective-free domain, then every internally stabilizable plant, defined by a transfer matrix P with entries in $K = Q(A)$, admits doubly coprime factorizations and all the stabilizing controllers of a stabilizable plant can be parametrized by means of the Youla–Kučera parametrization.*

Proof. Using Theorem 3.2 and the exact sequence (3.2), we obtain that $M_p/t(M_p)$ and $\ker\phi$ are two projective A-modules. Using the fact that A is a projective-free ring, we obtain that $M_p/t(M_p)$ and $\ker\phi$ are two free A-modules. From (3.4), we obtain that $\ker\phi \cong A^q$, and thus, we have the following exact sequence:

$$0 \longrightarrow A^q \xrightarrow{.R'} A^p \longrightarrow M_p/t(M_p) \longrightarrow 0,$$

with $R' \in M_{q\times p}(A)$. Using (3.1), we obtain that there exists a full rank matrix $R'' \in M_q(A)$ such that $R = R''\,R'$, i.e., $(D : -N) = R''\,(D' : -N')$, and thus

$$P = D^{-1}\,N = (R''D')^{-1}\,(R''N') = D'^{-1}\,N'.$$

Therefore, by Proposition 4.9 of [17] and Corollary 6.1, the plant P admits doubly coprime factorizations and all the stabilizing controllers of P are parametrized by the Youla–Kučera parametrization. □

COROLLARY 6.4 (see [25]). *If $A = H_\infty(\mathbb{C}_+)$, then a plant is internally stabilizable iff it admits a doubly coprime factorization.*

Example 6.2. The ring $A = \mathbb{R}[t_0, t_1](t_0^2 + t_1^2 - 1)$ (resp., $A = \mathbb{Z}[i\sqrt{5}]$) is a Dedekind domain which is not a principal ideal domain: The ideal $I = (x_0 - a,\, -x_1 + b)$ (resp., $I = (2,\, 1 + i\,\sqrt{5})$) is not a principal ideal [22]. By Corollary 4.13 of [17], it is not possible to parametrize all the stabilizing controllers of $p = (b - x_1)/(x_0 - a)$ (resp., $p = (1 + i\,\sqrt{5})/2$) by means of the Youla–Kučera parametrization.

It is possible to obtain a parametrization of all the stabilizing controllers which generalizes the Youla–Kučera parametrization for a stabilizable plant which does not admit doubly coprime factorizations. We refer the reader to [19, 20] for more details.

PROPOSITION 6.5. *The intersection between the sets of coherent Sylvester domains and Prüfer domains is exactly the set of Bézout domains.*

Proof. $\Rightarrow$ If A is a Prüfer domain, then every ideal $I = (d,\, n)$, generated by two elements $0 \neq d$ and n of A, is invertible [9]. Using the fact that A is also a coherent Sylvester domain, and thus, a greatest common divisor domain (see Corollary 3.20 of [17]), then $I^{-1} = (1,\, 1/[d,\, n])$, where $[d,\, n]$ denotes the greatest common divisor of d and n, and thus, we have

$$I\, I^{-1} = (d/[d,\, n],\, n/[d,\, n]) = A \Rightarrow \exists\, x,\, y \in A :\; d\, x + n\, y = [d,\, n],$$

which proves that I is a principal ideal of A, and thus, A is a Bézout domain.

$\Leftarrow$ By definition, a Bézout domain is a Prüfer and a coherent Sylvester domain. □

Conclusion. We hope we have convinced the reader that the algebraic analysis framework developed in this paper allows us to generalize some results on internal stabilization and to obtain new ones. Due to a lack of space, it was not possible to develop here the strong and the simultaneous stabilization problems [29]. We refer the reader to [16, 18] for a description of a canonical form, based on the concept of *stable range*, that certain stabilizing controllers possess. This canonical form allows us to show that, over a ring A of SISO stable plants of stable range 1 (e.g., $A = H_\infty(\mathbb{C}_+)$), every plant which admits a doubly coprime factorization is strongly stabilizable (i.e., stabilized by means of a stable controller). We also refer the reader to [19, 20] for other results on synthesis problems using fractional ideal and lattice approaches. In particular, a new parametrization of the stabilizing controllers for plants which do not admit doubly coprime factorizations is obtained. Moreover, in this paper the concept of *class group* $C(A)$ and the group $K_0(A)$ of nontrivial isomorphism classes of projective A-modules [22] are introduced. The computations of these groups allow us to check whether or not every internally stabilizable plant admits a doubly coprime factorization (e.g., $\mathcal{C}(\mathbb{R}[t_0, t_1]/(t_0^2 + t_1^2 - 1)) \cong \mathbb{Z}/2\,\mathbb{Z} \neq 0$ and $\mathcal{C}(\mathbb{Z}[i\sqrt{5}]) \cong \mathbb{Z}/2\,\mathbb{Z} \neq 0$ [22] showing that there exist internal stabilizable plants which do not admit a doubly coprime factorization). Finally, in [21], from the algebraic analysis point of view, we show how to recover the operator-theoretic approach developed in [7] (graphs, domains, unbounded operators, etc.) and to obtain new results.

Acknowledgments. I would like to thank H. Lombardi (University of Franche-Comté, France) for discussions on Prüfer domains and for making his paper, "Platitude, localisation et anneaux de Prüfer: une approche constructive," Publications Mathématiques de Besançon, Théorie des nombres, 2002, available to me. I am also grateful to J. J. Loiseau (IRCCyN, France) and C. Bonnet (INRIA Rocquencourt, France) for their continuous interest in this work and to those without whom this paper would never have been written, specially to Céline. I would like to thank the University of Leeds for its hospitality and, specially, J. R. Partington and H. G. Dales.

REFERENCES

[1] V. ANANTHARAM, *On stabilization and existence of coprime factorizations*, IEEE Trans. Automat. Control, 30 (1985), pp. 1030–1031.
[2] N. BOURBAKI, *Algèbre Commutative*, Masson, Paris, 1985, chap. 1–4.
[3] N. BOURBAKI, *Algèbre*, Masson, Paris, 1980, chap. 10.
[4] R. F. CURTAIN AND H. J. ZWART, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Texts Appl. Math. 21, Springer–Verlag, New York, 1991.

[5] C. A. Desoer, R.-W. Liu, J. Murray, and R. Saeks, *Feedback system design: The fractional representation approach to analysis and synthesis*, IEEE Trans. Automat. Control, 25 (1980), pp. 399–412.
[6] L. Fuchs and L. Salce, *Modules over non-Noetherian Domains*, Math. Surveys Monogr. 84, American Mathematical Society, Providence, RI, 2000.
[7] T. T. Georgiou and M. C. Smith, *Graphs, causality, and stabilizability: Linear, shift-invariant systems on* $\mathcal{L}_2[0,\infty)$, Math. Control Signals Systems, 6 (1993), pp. 195–223.
[8] O. Helmer, *Divisibility properties of integral functions*, Duke Math. J., 6 (1940), pp. 345–356.
[9] C. U. Jensen, *On characterizations of Prüfer rings*, Math. Scand., 13 (1963), pp. 90–98.
[10] W. Kucharz and K. Rusek, *On the ring of locally bounded Nash meromorphic functions*, Bull. Austral. Math. Soc., 54 (1996), pp. 503–507.
[11] J. J. Loiseau, *Algebraic tools for the control and stabilization of time-delay systems*, Annu. Rev. Control, 24 (2000), pp. 135–149.
[12] K. Mori, *Feedback stabilization over commutative rings with no right-/left coprime factorizations*, in Proceedings of the 38th IEEE Conference on Decision and Control, Vol. 1, IEEE Press, Piscataway, NJ, 1999, pp. 973–975.
[13] K. Mori and K. Abe, *Feedback stabilization over commutative rings: Further study of the coordinate-free approach*, SIAM J. Control Optim., 39 (2001), pp. 1952–1973.
[14] J. F. Pommaret and A. Quadrat, *A functorial approach to the behaviour of multidimensional control systems*, Appl. Math. Comput. Sci., to appear.
[15] A. Quadrat, *Internal stabilization of coherent control systems*, in System Structure and Control, Pergamon Press, Oxford, CD-ROM, 2001.
[16] A. Quadrat, *Une approche de la stabilisation par l'analyse algébrique Part* I. *Factorisations doublement faiblement copremières, Part* II. *Stabilisation interne, Part* III. *Sur une structure générale des contrôleurs stabilisants basée sur le rang stable*, in the Proceedings of Conference Internationale Francophone d'Automatique CIFA, Nantes, France, CD-ROM, 2002.
[17] A. Quadrat, *The fractional representation approach to synthesis problems: An algebraic analysis viewpoint part* I: *(Weakly) doubly coprime factorizations*, SIAM J. Control Optim., 40 (2003), pp. 266–299.
[18] A. Quadrat, *On a general structure of the stabilizing controllers based on the stable range*, SIAM J. Control Optim., submitted.
[19] A. Quadrat, *On a generalization of the Youla-Kučera parametrization. Part* I: *The fractional ideal approach to SISO systems*, Systems Control Lett., to appear.
[20] A. Quadrat, *On a generalization of the Youla-Kučera parametrization. Part* II*: The lattice approach to MIMO systems*, in preparation.
[21] A. Quadrat, *A behavioural interpretation to the operator-theoretic approach to internal stabilizability*, SIAM J. Control Optim., submitted.
[22] J. Rosenberg, *Algebraic K-Theory and Its Applications*, Grad. Texts Math., 147, Springer-Verlag, 1994.
[23] J. J. Rotman, *An Introduction to Homological Algebra*, Academic Press, New York-London, 1979.
[24] S. Shankar and V. R. Sule, *Algebraic geometric aspects of feedback stabilization*, SIAM J. Control Optim., 30 (1992), pp. 11–30.
[25] M. C. Smith, *On stabilization and the existence of coprime factorizations*, IEEE Trans. Automat. Control, 34 (1989), pp. 1005–1007.
[26] V. R. Sule, *Feedback stabilization over commutative rings: The matrix case*, SIAM J. Control Optim., 32 (1994), pp. 1675–1695.
[27] V. R. Sule, *Corrigendum: Feedback stabilization over commutative rings: The matrix case*, SIAM J. Control Optim., 36 (1998), pp. 2194–2195.
[28] M. Vidyasagar, H. Schneider, and B. A. Francis, *Algebraic and topological aspects of feedback stabilization*, IEEE Trans. Automat. Control, 27 (1982), pp. 880–894.
[29] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 321–341

INCLUSION PRINCIPLE FOR LINEAR TIME-VARYING SYSTEMS*

SRDJAN S. STANKOVIĆ† AND DRAGOSLAV D. ŠILJAK‡

Abstract. The main objective of this paper is to consider time-varying linear systems within the framework of the inclusion principle. Starting from general definitions, conditions are derived for a dynamic system to include a dynamic system of smaller dimension. Particular attention is paid to restriction and aggregation, the most important special cases of inclusion. It is proved, using geometric arguments within the general time-varying context, that any inclusion relationship on a given time-interval can be decomposed into a sequence of restriction-aggregation pairs. Connections between inclusion and zero-state equivalence are discussed. The paper also presents a formulation of the contractibility (expandability) conditions for time-varying state-feedback controllers.

Key words. time-varying linear dynamic systems, inclusion principle, restriction, aggregation, composition, zero-state equivalence, controller contractibility

AMS subject classifications. 93C05, 93C15, 93B17, 93B27

PII. S0363012901390609

1. Introduction.

In modeling of large dynamic systems, it is desirable to end up with a model which is small enough to be manageable and yet one that represents salient features of the original system. A mathematical framework for comparing dynamic systems of different dimensions has been offered by the *inclusion principle*. The interesting part of motions of the original large-scale system is reproduced by the smaller one; that is, the solution space (and, perhaps, performance indices) of the smaller system are *included* in the solution space (and indices) of the original large system.

Alternatively, a system may be expanded into a larger space in order for overlapping systems, which share common parts, to become disjoint. The expansion-contraction process in control engineering is attractive because we can use well-established design methods for decentralized control of disjoint subsystems in the expanded space and, subsequently, contract the design for implementation in the original space.

The expansion-contraction process, which served as a foundation for the inclusion principle, was introduced in [28] within the concept of reliable control involving overlapping multiple controller systems. A formal definition of the principle was presented in [17, 18] using the fundamental geometric arguments [37]. At present, there are generalizations and applications of the inclusion principle in a wide variety of theoretical and practical situations involving model-reduction, time-delay, discrete-time, nonlinear, and large dynamic systems. (For a survey of these areas, see [29, 30].) More recent developments present new results concerning optimal control [13], parallel computations [27, 7], inclusion of dynamic controllers and observers [12, 32, 35], stochastic inclusion [21, 33], expert systems [9], hybrid systems [15], and choice of complementary matrices in dynamic inclusions [3, 4]. Recent applications of the prin-

*Received by the editors June 9, 2001; accepted for publication (in revised form) August 6, 2002; published electronically April 17, 2003. This research was supported by National Science Foundation grant ECS-0099469.

http://www.siam.org/journals/sicon/42-1/39060.html

†Faculty of Electrical Engineering, University of Belgrade, Bulevar Kralja Aleksandra 73, 11000 Belgrade, Yugoslavia (stankovic@etf.bg.ac.yu).

‡Department of Electrical Engineering, Santa Clara University, Santa Clara, CA 95053-0569 (dsiljak@scu.edu).

ciple are in a wide variety of decentralized control problems in mechanical systems [2], electric power systems [33], platoons of vehicles [34], and large segmented telescope [22].

Almost all of the existing results within the inclusion principle paradigm are concerned with the time-invariant dynamic systems. Exceptions are the papers [17] and [10], where the problem of time-varying inclusion was constrained to the time-invariant expansion-contraction framework, without the full generality that can be provided by a natural use of time-varying transformations.

In this paper a general treatment of the inclusion principle applied to linear continuous-time time-varying dynamic systems is presented. Starting from a formulation of the inclusion principle for time-varying systems, necessary and sufficient inclusion conditions are given in terms of the functions characterizing systems under time-varying input/state/output expansion-contraction relationships. A set of conditions, involving time-varying matrices of the basic state models, is derived and compared to the time-invariant case.

We focus our attention on the input/state/output restrictions and aggregations for time-varying systems. Fundamental properties of a variety of the proposed restriction-aggregation types are made evident through a specific state-space realization of the expanded system. A special emphasis is on the *composition* property. It is proved that any input/state/output inclusion on a given time-interval can be decomposed into a sequence of time-varying input/state/output aggregation and input/state/output restriction pairs. The proof of the theorem provides an insight into some fundamental aspects of the inclusion of time-varying systems.

Connections between the zero-state equivalence of time-varying systems and the inclusion principle are also discussed. The results presented in [18] are extended to the time-varying case, demonstrating that the inclusion of dynamic time-varying systems can be regarded as an extension of the equivalent transformation that does not require preservation of dimensionality.

A definition of state-feedback contractibility (expandability) is also presented. Diverse restriction-aggregation types are considered, together with the conditions ensuring inclusion of the resulting closed-loop systems. Examples, related to both expansion and contraction of the state feedback, indicate some possible applications of the inclusion concept for time-varying systems.

2. Inclusion principle for time-varying systems. Consider a pair of linear time-varying dynamic systems

$$\mathbf{S}\colon \quad \dot{x}(t) = A(t)x(t) + B(t)u(t) \quad (x(t_0) = x_0), \qquad y(t) = C(t)x(t), \tag{1}$$

$$\tilde{\mathbf{S}}\colon \quad \dot{\tilde{x}}(t) = \tilde{A}(t)\tilde{x}(t) + \tilde{B}(t)\tilde{u}(t) \quad (\tilde{x}(t_0) = \tilde{x}_0), \qquad \tilde{y}(t) = \tilde{C}(t)\tilde{x}(t), \tag{2}$$

where $x \in \mathcal{X}$, $u \in \mathcal{U}$, and $y \in \mathcal{Y}$ are n-, m-, and l-vectors and $\tilde{x} \in \tilde{\mathcal{X}}$, $\tilde{u} \in \tilde{\mathcal{U}}$, and $\tilde{y} \in \tilde{\mathcal{Y}}$ are $\tilde{n}$-, $\tilde{m}$-, and $\tilde{l}$-vectors, respectively, satisfying $n \le \tilde{n}$, $m \le \tilde{m}$, and $l \le \tilde{l}$. For $t \ge t_0$, the input functions $u_{[t_0,t]}\colon [t_0, t] \to \mathcal{U}$ and $\tilde{u}_{[t_0,t]}\colon [t_0, t] \to \tilde{\mathcal{U}}$ belong to the sets of piecewise continuous functions $\mathcal{U}_f$ and $\tilde{\mathcal{U}}_f$, respectively. The elements of the matrices in $\mathbf{S} = (A(t), B(t), C(t))$ and $\tilde{\mathbf{S}} = (\tilde{A}(t), \tilde{B}(t), \tilde{C}(t))$ are assumed to be piecewise continuous functions of t. Denote by $x(t; t_0, x_0, u_{[t_0,t]})$ and $\tilde{x}(t; t_0, \tilde{x}_0, \tilde{u}_{[t_0,t]})$ the unique solutions of (1) and (2) for the initial time t_0, initial states x_0 and $\tilde{x}_0$, and fixed control functions $u_{[t_0,t]} \in \mathcal{U}_f$ and $\tilde{u}_{[t_0,t]} \in \tilde{\mathcal{U}}_f$, respectively.

It is assumed further that the inputs, states, and outputs of $\mathbf{S}$ and $\tilde{\mathbf{S}}$ are related by linear *time-varying expansion-contraction transformations* defined by the monomor-

phisms $V(t)$: $\mathcal{X} \to \tilde{\mathcal{X}}$, $R(t)$: $\mathcal{U} \to \tilde{\mathcal{U}}$, and $T(t)$: $\mathcal{Y} \to \tilde{\mathcal{Y}}$ and the epimorphisms $U(t)$: $\tilde{\mathcal{X}} \to \mathcal{X}$, $Q(t)$: $\tilde{\mathcal{U}} \to \mathcal{U}$, and $S(t)$: $\tilde{\mathcal{Y}} \to \mathcal{Y}$ so that $\tilde{x}(t) = V(t)x(t)$ or $x(t) = U(t)\tilde{x}(t)$, $\tilde{u}(t) = R(t)u(t)$ or $u(t) = Q(t)\tilde{u}(t)$, and $\tilde{y}(t) = T(t)y(t)$ or $y(t) = S(t)\tilde{y}(t)$. We also assume that $U(t) = V(t)^L$, $Q(t) = R(t)^L$, and $S(t) = T(t)^L$, where the superscript L denotes the left inverse, and that the elements of $U(t), V(t), Q(t), R(t), S(t)$, and $T(t)$ are piecewise continuous functions of t.

DEFINITION 2.1. *The system* $\tilde{\mathbf{S}}$ *includes the system* $\mathbf{S}$ *on* $[t_a, t_b]$*, that is,* $\mathbf{S} \subset \tilde{\mathbf{S}}$ ($\tilde{\mathbf{S}}$ *is an expansion of* $\mathbf{S}$ *and, vice versa,* $\mathbf{S}$ *is a contraction of* $\tilde{\mathbf{S}}$)*, if there exists a quadruplet of full rank matrices* $(U(t), V(t), R(t), S(t))$ *such that, for any* $t_0 \in [t_a, t_b]$ *and any* $x_0 \in \mathcal{X}$ *and* $u_{[t_0,t]} \in \mathcal{U}_f$*, the choice* $\tilde{x}_0 = V(t_0)x_0$ *and* $\tilde{u}(t) = R(t)u(t)$ *implies* $x(t; t_0, x_0, u_{[t_0,t]}) = U(t)\tilde{x}(t; t_0, \tilde{x}_0, \tilde{u}_{[t_0,t]})$ *and* $y[x(t)] = S(t)\tilde{y}[\tilde{x}(t)]$ *for all* $t \in [t_0, t_b]$*,* $t \geq t_0$*.*

In order to obtain compact formulations, we shall introduce two ordered pairs of variables, composed of the independent variables in the above expansion-contraction relationships between the variables in $\mathbf{S}$ and $\tilde{\mathbf{S}}$. The first pair, P_1, corresponds to the pairs of variables $(x_0, \tilde{x}_0)$ and $(u, \tilde{u})$ that can be chosen exogenously, irrespective of the system dynamics, while the second, P_2, corresponds to the pairs $(x, \tilde{x})$ and $(y, \tilde{y})$, reflecting the system behavior. For example, $P_1 = \{x_0, \tilde{u}\}$ stands for $\tilde{x}_0 = V(t_0)x_0$ and $u(t) = Q(t)\tilde{u}(t)$, $P_1 = \{\tilde{x}_0, u\}$ for $x_0 = U(t_0)\tilde{x}_0$ and $\tilde{u}(t) = R(t)u(t)$, $P_2 = \{\tilde{x}, y\}$ for $x(t) = U(t)\tilde{x}(t)$ and $\tilde{y}(t) = T(t)y(t)$, etc. Using the pairs P_1 and P_2, we introduce the form $P_1 \Rightarrow P_2$ as a shorthand notation for different types of inclusion in the sense of Definition 2.1. In such a way, $P_1 \Rightarrow P_2$, where, for example, $P_1 = \{x_0, u\}$ and $P_2 = \{\tilde{x}, y\}$, written also directly as $\{x_0, u\} \Rightarrow \{\tilde{x}, y\}$, will be given the following meaning: conditions $\tilde{x}_0 = V(t_0)x_0$ and $\tilde{u}(t) = R(t)u(t)$ imply that, for any $x_0 \in \mathcal{X}$ and any $u_{[t_0,t]} \in \mathcal{U}_f$, $\tilde{x}(t; t_0, \tilde{x}_0, \tilde{u}_{[t_0,t]}) = V(t)x(t; t_0, x_0, u_{[t_0,t]})$ and $y[x(t)] = S(t)\tilde{y}[\tilde{x}(t)]$, $t \geq t_0$. Definition 2.1 can now be reformulated as follows.

DEFINITION 2.2. $\mathbf{S} \subset \tilde{\mathbf{S}}$ *on* $[t_a, t_b]$ *if* $\exists (U(t), V(t), R(t), S(t))$ $\{x_0, u\} \Rightarrow \{\tilde{x}, \tilde{y}\}$ *for all* $t_0, t \in [t_a, t_b]$*.*

Definitions 2.1 and 2.2 provide a generalization of the approach presented in [17], where only time-invariant state contractions-expansions for time-varying systems have been considered.

Denote by $\Phi(t, \tau)$ and $\tilde{\Phi}(t, \tau)$ the transition matrices of $\mathbf{S}$ and $\tilde{\mathbf{S}}$, respectively, and by $\psi(t) = \Phi(t, 0)$ and $\tilde{\psi}(t) = \tilde{\Phi}(t, 0)$ the corresponding Wronskians [23]. Introduce also $\nu(t) = \tilde{\psi}(t)^{-1}V(t)\psi(t)$, $\mu(t) = \psi(t)^{-1}U(t)\tilde{\psi}(t)$, $b(t) = \psi(t)^{-1}B(t)$, $\tilde{b}(t) = \tilde{\psi}(t)^{-1}\tilde{B}(t)$, $c(t) = C(t)\psi(t)$, and $\tilde{c}(t) = \tilde{C}(t)\tilde{\psi}(t)$; obviously, maps $\nu(t)$ and $\mu(t)$ are monic and epic, respectively [23, 25].

Inclusion conditions expressed in terms of the properties of $\mathbf{S}$ and $\tilde{\mathbf{S}}$ can be formulated as follows.

THEOREM 2.3. $\mathbf{S} \subset \tilde{\mathbf{S}}$ *on* $[t_a, t_b]$ *iff for all* $t_0, t, \tau \in [t_a, t_b], t_0 \leq \tau \leq t$*,*

$$\begin{aligned} &\mu(t)\nu(t_0) = I, \quad b(\tau) = \mu(t)\tilde{b}(\tau)R(\tau), \quad c(t) = S(t)\tilde{c}(t)\nu(t_0), \\ &c(t)b(\tau) = S(t)\tilde{c}(t)\tilde{b}(\tau)R(\tau). \end{aligned} \tag{3}$$

Proof. According to Definition 2.1, condition $x(t) = U(t)\tilde{x}(t)$ is equivalent to

$$\begin{aligned} &\Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau \\ &\qquad = U(t)\tilde{\Phi}(t, t_0)V(t_0)x_0 + U(t)\int_{t_0}^{t} \tilde{\Phi}(t, \tau)\tilde{B}(\tau)R(\tau)u(\tau)d\tau, \end{aligned} \tag{4}$$

whence $\Phi(t,t_0) = U(t)\tilde{\Phi}(t,t_0)V(t_0)$ or $\mu(t)\nu(t_0) = I$, and

$$\Phi(t,\tau)B(\tau) = U(t)\tilde{\Phi}(t,\tau)\tilde{B}(\tau)R(\tau) \quad \Leftrightarrow \quad b(\tau) = \mu(t)\tilde{b}(\tau)R(\tau). \tag{5}$$

The remaining relations in (3) are obtained in a similar way. Hence we have the result. □

Example 2.4. The purpose of this simple example is to illustrate a wide area of diverse relationships between two linear dynamic systems which can be described by using time-varying expansions-contractions.

Let, for all $t, t_0 \in [t_a, t_b]$, $t \geq t_0$,

$$\mathbf{S}: \quad \dot{x} = (\sin t)x + 2u \quad (x(t_0) = x_0), \qquad y = x; \tag{6}$$

that is, in terms of (1), $A(t) = \sin t$, $B(t) = 2$, and $C(t) = 1$. Let $V(t)^T = [1 \ \ \frac{1}{2}\sin t]$. It is easy to verify that

$$\tilde{A}(t)V(t) - V(t)\sin t = \dot{V}(t), \tag{7}$$

where

$$\tilde{A}(t) = \begin{bmatrix} 0 & 2 \\ \frac{1}{2}\cos t & \sin t \end{bmatrix}.$$

Let $\psi(t)$ and $\tilde{\psi}(t)$ denote Wronskians of the homogeneous differential equations $\dot{x} = (\sin t)x$ and $\dot{\tilde{x}} = \tilde{A}(t)\tilde{x}$, $\tilde{x}^T = (\tilde{x}_1, \tilde{x}_2)$. After multiplying (7) by $\tilde{\psi}(t)^{-1}$ from the left and by $\psi(t)$ from the right, one obtains (according to the definition of $\nu(t)$) that $\dot{\nu}(t) = 0$ or $\nu(t) = \nu(t_a)$ for all $t \in [t_a, t_b]$. If $\mu(t)$ is any left inverse of $\nu(t)$, which obviously exists for all $\nu(t)$, one obtains $\mu(t)\nu(t_a) = 1$, that is, the first relation in (3). Notice that the corresponding state map $U(t)$ can always be found from $U(t) = \psi(t)\mu(t)\tilde{\psi}(t)^{-1}$, according to the definition of $\mu(t)$.

Similarly, if for all $\tau \in [t_a, t_b]$

$$\tilde{B}(\tau) = \begin{bmatrix} 2 \\ \sin\tau \end{bmatrix} = V(\tau)B(\tau) = 2V(\tau),$$

after multiplying both sides by $\tilde{\psi}(t)\tilde{\psi}(\tau)^{-1}$ $(t \geq \tau)$ and noticing that the condition $\nu(t) = \nu(\tau)$ is equivalent to $V(t)\psi(t)\psi(\tau)^{-1} = \tilde{\psi}(t)\tilde{\psi}(\tau)^{-1}$, one gets that $\nu(t)b(\tau) = \tilde{b}(\tau)$, which implies $b(\tau) = \mu(t)\tilde{b}(\tau)$, the second relation in (3) (assuming that $R(t) = 1$).

The third and fourth relations in (3) are verifiable analogously if one assumes that $S(t) = 1$ and

$$C(t) = 1 = [1 \ \ 0]V(t) = \tilde{C}(t)V(t).$$

Consequently, the system

$$\tilde{\mathbf{S}}: \quad \dot{\tilde{x}} = \begin{bmatrix} 0 & 2 \\ \frac{1}{2}\cos t & \sin t \end{bmatrix}\tilde{x} + \begin{bmatrix} 2 \\ \sin t \end{bmatrix}u \quad (\tilde{x}(t_0) = \tilde{x}_0), \qquad y = [1 \ \ 0]\tilde{x} \tag{8}$$

includes $\mathbf{S}$ given by (6) in the sense of Theorem 2.3.

Notice that (7) does not have a time-invariant solution for $V(t)$ (satisfying $\dot{V}(t) = 0$).

The definitions of the matrix functions $\nu(t)$, $\mu(t)$, $b(t)$, $\tilde{b}(t)$, $c(t)$, and $\tilde{c}(t)$ involve the transition matrices of both $\mathbf{S}$ and $\tilde{\mathbf{S}}$, implying problems in direct applications of the general inclusion conditions (3) in practice, especially when one has to design one of the systems under the inclusion relation when the other one is given. The need for simpler and more tractable formulations arises within both model-reduction and model-expansion frameworks. One of the principal aims of all further elaborations in this paper is to obtain inclusion conditions in the form suitable for practical applications. However, in the general time-varying case it is hardly possible to express the inclusion conditions through direct and simple relations between the system matrices of $\mathbf{S}$ and $\tilde{\mathbf{S}}$. For example, using the Peano–Baker series (e.g., [25]), one obtains for $\mu(t)\nu(\tau) = I$ the following general formulation:

$$\begin{aligned} &I + \int_\tau^t A(\sigma_1)d\sigma_1 + \int_\tau^t A(\sigma_1)\int_\tau^{\sigma_1} A(\sigma_2)d\sigma_2 d\sigma_1 + \cdots \\ &\quad = U(t)\left(I + \int_\tau^t \tilde{A}(\sigma_1)d\sigma_1 + \int_\tau^t \tilde{A}(\sigma_1)\int_\tau^{\sigma_1} \tilde{A}(\sigma_2)d\sigma_2 d\sigma_1 + \cdots\right)V(\tau). \end{aligned} \tag{9}$$

Assuming that, in addition, the commutativity condition holds for both $A(t)$ and $\tilde{A}(t)$, that is, $A(t)\int_\tau^t A(\sigma)d\sigma = \int_\tau^t A(\sigma)d\sigma A(t)$ and $\tilde{A}(t)\int_\tau^t \tilde{A}(\sigma)d\sigma = \int_\tau^t \tilde{A}(\sigma)d\sigma\tilde{A}(t)$ (for all t, τ), (9) gives only

$$\begin{aligned} e^{\int_\tau^t A(\sigma)d\sigma} &= \sum_{k=0}^{\infty}\frac{1}{k!}\left[\int_\tau^t A(\sigma)d\sigma\right]^k = U(t)e^{\int_\tau^t \tilde{A}(\sigma)d\sigma}V(\tau) \\ &= \sum_{k=0}^{\infty}\frac{1}{k!}U(t)\left[\int_\tau^t \tilde{A}(\sigma)d\sigma\right]^k V(\tau). \end{aligned} \tag{10}$$

More compact formulations can be obtained in two important particular cases.

Time-invariant systems. For time-invariant systems and time-invariant expansion-contraction maps, the known conditions for state inclusion $A^i = U\tilde{A}^i V$, $i = 0, 1, \ldots, \tilde{n} - 1$, follow from either (9) or (10) after replacing $\Phi(t, \tau)$ and $\tilde{\Phi}(t, \tau)$ by $e^{A(t-\tau)}$ and $e^{\tilde{A}(t-\tau)}$, respectively, and applying the Cayley–Hamilton theorem [18, 16, 29, 35].

Periodic systems. When both $A(t)$ and $\tilde{A}(t)$ are T-periodic, i.e., $A(t+T) = A(t)$ and $\tilde{A}(t+T) = \tilde{A}(t)$, the Floquet decomposition gives $\Phi(t, \tau) = \Pi(t)e^{\Gamma(t-\tau)}\Pi(\tau)^{-1}$ and $\tilde{\Phi}(t, \tau) = \tilde{\Pi}(t)e^{\tilde{\Gamma}(t-\tau)}\tilde{\Pi}(\tau)^{-1}$, where $\Pi(t)$ and $\tilde{\Pi}(t)$ are nonsingular T-periodic matrices, while Γ and $\tilde{\Gamma}$ are constant matrices defined by $e^{\Gamma T} = \Phi(T, 0)$ and $e^{\tilde{\Gamma}T} = \tilde{\Phi}(T, 0)$, respectively; see, for example, [24, 25]. Assume that the time-invariant parts of these decompositions satisfy the inclusion conditions in the sense that $\Gamma^i = U_0\tilde{\Gamma}^i V_0$, $i = 0, 1, \ldots$, where U_0 and V_0 are time-invariant full rank matrices satisfying $U_0V_0 = I$. Then

$$\Phi(t, \tau) = \Pi(t)U_0\tilde{\Pi}(t)^{-1}\tilde{\Phi}(t, \tau)\tilde{\Pi}(\tau)V_0\Pi(\tau)^{-1}. \tag{11}$$

Thus $U(t) = \Pi(t)U_0\tilde{\Pi}(t)^{-1}$ and $V(t) = \tilde{\Pi}(t)V_0\Pi(t)^{-1}$ are now T-periodic full rank matrices satisfying $U(t)V(t) = I$, and the inclusion conditions from Theorem 2.3 hold for $\mathbf{S}$ and $\tilde{\mathbf{S}}$.

A set of interesting and useful relations between the matrices in $\mathbf{S}$ and $\tilde{\mathbf{S}}$, enabling comparisons with the time-invariant case, can be derived from the general inclusion

relations (3). Assuming that the indicated derivatives exist and are continuous, define the following sequences of matrix functions for $\mathbf{S}$ and $\tilde{\mathbf{S}}$, respectively:

$$F_i(t;F_0(t)) = -A(t)F_{i-1}(t;F_0(t)) + \dot{F}_{i-1}(t;F_0(t)) \tag{12}$$

and

$$\tilde{F}_i(t;\tilde{F}_0(t)) = -\tilde{A}(t)\tilde{F}_{i-1}(t;\tilde{F}_0(t)) + \dot{\tilde{F}}_{i-1}(t;\tilde{F}_0(t)), \tag{13}$$

$i = 1, 2, \ldots$, where $F_0(t) = F_0(t;F_0(t))$ and $\tilde{F}_0(t) = \tilde{F}_0(t;\tilde{F}_0(t))$.

COROLLARY 2.5. *Assume that* $\mathbf{S} \subset \tilde{\mathbf{S}}$ *and that the matrices in* (1) *and* (2), *together with the expansion-contraction matrices, are continuously differentiable a sufficient number of times. Then, for* $i = 0, 1, 2, \ldots$,

$$\begin{gathered} F_i(t;I) = U(t)\tilde{F}_i(t;V(t)), \qquad F_i(t;B(t)) = U(t)\tilde{F}_i(t;\tilde{B}(t)R(t)), \\ C(t)F_i(t;I) = S(t)\tilde{C}(t)\tilde{F}_i(t;V(t)), \qquad C(t)F_i(t;B(t)) = S(t)\tilde{C}(t)\tilde{F}_i(t;\tilde{B}(t)R(t)). \end{gathered} \tag{14}$$

Proof. The first relation in (14) follows from $\Phi(t,\tau) = U(t)\tilde{\Phi}(t,\tau)V(\tau)$ after differentiating both sides i times with respect to τ and calculating the obtained derivatives at $\tau = t$. Namely, we have $\frac{\partial^i}{\partial\tau^i}\Phi(t,\tau)\,|_{\tau=t} = U(t)\frac{\partial^i}{\partial\tau^i}\tilde{\Phi}(t,\tau)V(\tau)|_{\tau=t}$, whence

$$F_i(t;I) = \frac{\partial^i}{\partial\tau^i}\Phi(t,\tau)\,|_{\tau=t}, \qquad \tilde{F}_i(t;V(t)) = \frac{\partial^i}{\partial\tau^i}\tilde{\Phi}(t,\tau)V(\tau)\,|_{\tau=t},$$

having in mind that $\frac{\partial}{\partial\tau}\Phi(t,\tau) = -\Phi(t,\tau)A(\tau)$ and $\frac{\partial}{\partial\tau}\tilde{\Phi}(t,\tau) = -\tilde{\Phi}(t,\tau)\tilde{A}(\tau)$. The remaining relations in (14) are obtained analogously. □

For $i = 0$, the first relation in (14) gives $U(t)V(t) = I$; for $i = 1$ and $i = 2$, it provides

$$\begin{gathered} A(t) = U(t)[\tilde{A}(t)V(t) - \dot{V}(t)], \\ A(t)^2 - \dot{A}(t) = U(t)[\tilde{A}(t)^2 - \dot{\tilde{A}}(t)]V(t) + U(t)[\ddot{V}(t) - 2\tilde{A}(t)\dot{V}(t)]. \end{gathered} \tag{15}$$

The remaining relations in (14) give for $i = 0$ and $i = 1$

$$\begin{gathered} B(t) = U(t)\tilde{B}(t)R(t), \\ A(t)B(t) - \dot{B}(t) = U(t)[\tilde{A}(t)\tilde{B}(t) - \dot{\tilde{B}}(t)]R(t) - U(t)\tilde{B}(t)\dot{R}(t), \end{gathered} \tag{16}$$

$$\begin{gathered} C(t) = S(t)\tilde{C}(t)V(t), \qquad C(t)A(t) = S(t)\tilde{C}(t)[\tilde{A}(t)V(t) - \dot{V}(t)], \\ C(t)B(t) = S(t)\tilde{C}(t)\tilde{B}(t)R(t), \end{gathered} \tag{17}$$

$$C(t)[A(t)B(t) - \dot{B}(t)] = S(t)\tilde{C}(t)[\tilde{A}(t)\tilde{B}(t) - \dot{\tilde{B}}(t)]R(t) - S(t)\tilde{C}(t)\tilde{B}(t)\dot{R}(t). \tag{18}$$

In the case of time-invariant expansions-contractions, we have

$$\begin{gathered} A(t) = U\tilde{A}(t)V, \qquad A(t)^2 = U\tilde{A}(t)^2V, \qquad B(t) = U\tilde{B}(t)R, \\ A(t)B(t) = U\tilde{A}(t)\tilde{B}(t)R, \quad C(t) = S\tilde{C}(t)V, \quad C(t)A(t) = S\tilde{C}(t)\tilde{A}(t)V, \\ C(t)B(t) = S\tilde{C}(t)\tilde{B}(t)R, \quad C(t)A(t)B(t) = S\tilde{C}(t)\tilde{A}(t)\tilde{B}(t)R. \end{gathered} \tag{19}$$

If, in addition, the systems themselves are time-invariant, one obtains

$$A^i = U\tilde{A}^iV, \quad A^iB = U\tilde{A}^i\tilde{B}R, \quad CA^i = S\tilde{C}\tilde{A}^iV, \\ CA^iB = S\tilde{C}\tilde{A}^i\tilde{B}R, \quad i = 0, 1, \dots, \tilde{n} - 1, \tag{20}$$

the known set of necessary and sufficient inclusion conditions (e.g., [16, 35]).

Notice that (14) gives for $i = 0$ the basic conditions $U(t)V(t) = I$, $B(t) = U(t)\tilde{B}(t)R(t)$, $C(t) = S(t)\tilde{C}(t)V(t)$, and $C(t)B(t) = S(t)\tilde{C}(t)\tilde{B}(t)R(t)$ appearing in all the cases discussed above (compare with (19) and (20)). Notice also the form of these conditions, recalling the form of (3).

3. Restrictions and aggregations. Restriction and aggregation represent the most important special cases of inclusion in the time-invariant case [16, 18, 29]. The situation is the same in the time-varying case. We give here general definitions, covering a variety of input/output expansion-contraction combinations, which represent a generalization to the time-varying case of the definitions proposed in [33, 35].

DEFINITION 3.1. $\mathbf{S}$ *is a restriction of* $\tilde{\mathbf{S}}$ *on* $[t_a, t_b]$ *if one of the following conditions is satisfied for all* $t_0, t \in [t_a, t_b]$, $t_0 \le t$*:*

$R(a)$: $\exists (V(t), R(t), T(t))\ \{x_0, u\} \Rightarrow \{x, y\}$, $R(b)$: $\exists (V(t), R(t), S(t))\ \{x_0, u\} \Rightarrow \{x, \tilde{y}\}$,
$R(c)$: $\exists (V(t), Q(t), T(t))\ \{x_0, \tilde{u}\} \Rightarrow \{x, y\}$, $R(d)$: $\exists (V(t), Q(t), S(t))\ \{x_0, \tilde{u}\} \Rightarrow \{x, \tilde{y}\}$.

Consequently, we shall distinguish four restriction types: restriction (a), denoted as $\mathbf{S} \overset{R(a)}{\subset} \tilde{\mathbf{S}}$ or $R(a)$, restriction (b), denoted as $\mathbf{S} \overset{R(b)}{\subset} \tilde{\mathbf{S}}$ or $R(b)$, etc.

In the time-invariant case, the state restriction (not including input and output contractions-expansions) has been presented in [18], restriction (a) in [16], and restriction (c) in [14] under the name of *extension*. In [33, 32, 35], a general treatment has been presented. The state restriction for time-varying systems is formulated in [17] under the assumption that the expansions-contractions are time-invariant.

DEFINITION 3.2. $\mathbf{S}$ *is an aggregation of* $\tilde{\mathbf{S}}$ *on* $[t_a, t_b]$ *if one of the following conditions is satisfied for all* $t_0, t \in [t_a, t_b], t_0 \le t$*:*

$A(a)$: $\exists (U(t), Q(t), S(t))\ \{\tilde{x}_0, \tilde{u}\} \Rightarrow \{\tilde{x}, \tilde{y}\}$, $A(b)$: $\exists (U(t), R(t), S(t))\ \{\tilde{x}_0, u\} \Rightarrow \{\tilde{x}, \tilde{y}\}$,
$A(c)$: $\exists (U(t), Q(t), T(t))\ \{\tilde{x}_0, \tilde{u}\} \Rightarrow \{\tilde{x}, y\}$, $A(d)$: $\exists (U(t), R(t), T(t))\ \{\tilde{x}_0, u\} \Rightarrow \{\tilde{x}, y\}$.

We have four aggregation types: aggregation (a), denoted as $\mathbf{S} \overset{A(a)}{\subset} \tilde{\mathbf{S}}$ or $A(a)$, etc.

In the time-invariant case, the state aggregation and aggregation (a) have been described in [1, 16, 18, 29]; a general treatment is presented in [33, 32, 35]. In the case of time-varying systems, the state aggregation has been formulated in [17], assuming that the expansions-contractions are time-invariant.

Let us represent, in general, the matrices in $\tilde{\mathbf{S}}$ as

$$\tilde{A}(t) = V(t)A(t)U(t) + M(t), \quad \tilde{B}(t) = V(t)B(t)Q(t) + N(t), \\ \tilde{C}(t) = T(t)C(t)U(t) + L(t),$$

where $M(t)$, $N(t)$, and $L(t)$ are complementary matrices. The following theorems present the restriction and aggregation conditions expressed in terms of the properties of $\mathbf{S}$ and $\tilde{\mathbf{S}}$.

THEOREM 3.3.

$$\begin{aligned}
&\mathbf{S} \overset{R(a)}{\subset} \tilde{\mathbf{S}} \ \textit{iff} \ \ (\nu(t) = \nu(t_0)) \wedge (\nu(t)b(\tau) = \tilde{b}(\tau)R(\tau)) \wedge (T(t)c(t) = \tilde{c}(t)\nu(t_0)),\\
&\mathbf{S} \overset{R(b)}{\subset} \tilde{\mathbf{S}} \ \textit{iff} \ \ (\nu(t) = \nu(t_0)) \wedge (\nu(t)b(\tau) = \tilde{b}(\tau)R(\tau)) \wedge (c(t) = S(t)\tilde{c}(t)\nu(t_0)),\\
&\mathbf{S} \overset{R(c)}{\subset} \tilde{\mathbf{S}} \ \textit{iff} \ \ (\nu(t) = \nu(t_0)) \wedge (\nu(t)b(\tau)Q(\tau) = \tilde{b}(\tau)) \wedge (T(t)c(t) = \tilde{c}(t)\nu(t_0)),\\
&\mathbf{S} \overset{R(d)}{\subset} \tilde{\mathbf{S}} \ \textit{iff} \ \ (\nu(t) = \nu(t_0)) \wedge (\nu(t)b(\tau)Q(\tau) = \tilde{b}(\tau)) \wedge (c(t) = S(t)\tilde{c}(t)\nu(t_0))\\
&(\forall t_0, \tau, t \in [t_a, t_b], t_0 \le \tau \le t).
\end{aligned}$$

Proof. According to Definition 3.1, for all the restriction types, $V(t)\Phi(t,t_0) = \tilde{\Phi}(t,t_0)V(t_0)$ or $\nu(t) = \nu(t_0)$. The remaining conditions follow directly from the corresponding definitions. □

THEOREM 3.4. *Assume that $V(t)$ is continuously differentiable. Then*

$$\begin{aligned}
&\mathbf{S} \overset{R(a)}{\subset} \tilde{\mathbf{S}} \ \textit{if} \ \ (M(t)V(t) = \dot{V}(t)) \wedge (N(t)R(t) = 0) \wedge (L(t)V(t) = 0),\\
&\mathbf{S} \overset{R(b)}{\subset} \tilde{\mathbf{S}} \ \textit{if} \ \ (M(t)V(t) = \dot{V}(t)) \wedge (N(t)R(t) = 0) \wedge (S(t)L(t)V(t) = 0),\\
&\mathbf{S} \overset{R(c)}{\subset} \tilde{\mathbf{S}} \ \textit{if} \ \ (M(t)V(t) = \dot{V}(t)) \wedge (N(t) = 0) \wedge (L(t)V(t) = 0),\\
&\mathbf{S} \overset{R(d)}{\subset} \tilde{\mathbf{S}} \ \textit{if} \ \ (M(t)V(t) = \dot{V}(t)) \wedge (N(t) = 0) \wedge (S(t)L(t)V(t) = 0)\\
&(\forall t \in [t_a, t_b]).
\end{aligned}$$

Proof. As $\nu(t) = \nu(t_0)$, according to Definition 3.1, we have $\dot{\nu}(t) = 0$ for all $t \in [t_a, t_b]$, or

$$-\tilde{\psi}(t)^{-1}\tilde{A}(t)V(t)\psi(t) + \tilde{\psi}(t)^{-1}\dot{V}(t)\psi(t) + \tilde{\psi}(t)^{-1}V(t)A(t)\psi(t) = 0, \tag{21}$$

giving $\tilde{A}(t)V(t) - V(t)A(t) = \dot{V}(t)$ or $M(t)V(t) = \dot{V}(t)$. For $R(a)$ we have $\tilde{\Phi}(t,\tau)\tilde{B}(\tau)$ $R(\tau) = V(t)\Phi(t,\tau)B(\tau)$ so that $\tilde{\Phi}(t,\tau)\ N(\tau)\ R(\tau) = 0 \Leftrightarrow N(t)\ R(t) = 0$ and $T(t)$ $C(t)\ \Phi(t,t_0) = S(t)\ \tilde{C}(t)\ \tilde{\Phi}(t,t_0)\ V(t_0)$, giving $L(t)\ V(t)\ \Phi(t,\tau) = 0 \Leftrightarrow L(t)\ V(t) = 0$. The remaining relations, characterizing $R(b)$, $R(c)$, and $R(d)$, can be derived analogously. □

We remark that systems $\mathbf{S}$ and $\tilde{\mathbf{S}}$ in Example 2.4 satisfy, in fact, restriction conditions with the triplet of expansion-contraction matrices ($V(t) = [1 \ \ \frac{1}{2}\sin t], R(t) = 1, T(t) = 1$).

THEOREM 3.5.

$$\begin{aligned}
&\mathbf{S} \overset{A(a)}{\subset} \tilde{\mathbf{S}} \ \textit{iff} \ \ (\mu(t) = \mu(t_0)) \wedge (b(\tau)Q(\tau) = \mu(t)\tilde{b}(\tau)) \wedge (c(t)\mu(t_0) = S(t)\tilde{c}(t)),\\
&\mathbf{S} \overset{A(b)}{\subset} \tilde{\mathbf{S}} \ \textit{iff} \ \ (\mu(t) = \mu(t_0)) \wedge (b(\tau) = \mu(t)\tilde{b}(\tau)R(\tau)) \wedge (c(t)\mu(t_0) = S(t)\tilde{c}(t)),\\
&\mathbf{S} \overset{A(c)}{\subset} \tilde{\mathbf{S}} \ \textit{iff} \ \ (\mu(t) = \mu(t_0)) \wedge (b(\tau)Q(\tau) = \mu(t)\tilde{b}(\tau)) \wedge (T(t)c(t)\mu(t_0) = \tilde{c}(t)),\\
&\mathbf{S} \overset{A(d)}{\subset} \tilde{\mathbf{S}} \ \textit{iff} \ \ (\mu(t) = \mu(t_0)) \wedge (b(\tau) = \mu(t)\tilde{b}(\tau)R(\tau)) \wedge (T(t)c(t)\mu(t_0) = \tilde{c}(t))\\
&(\forall t_0, \tau, t \in [t_a, t_b], t_0 \le \tau \le t).
\end{aligned}$$

Proof. The condition common for all the aggregation types is $U(t)\tilde{\Phi}(t,t_0) = \Phi(t,t_0)U(t_0) \Leftrightarrow \mu(t) = \mu(t_0)$. The remaining conditions are obtained similarly to those in Theorem 3.3. □

THEOREM 3.6. *Assume that $U(t)$ is continuously differentiable. Then*

$$\begin{aligned}
&\mathbf{S} \overset{A(a)}{\subset} \tilde{\mathbf{S}} \ \textit{iff} \ \ (U(t)M(t) = -\dot{U}(t)) \wedge (U(t)N(t) = 0) \wedge (S(t)L(t) = 0),\\
&\mathbf{S} \overset{A(b)}{\subset} \tilde{\mathbf{S}} \ \textit{iff} \ \ (U(t)M(t) = -\dot{U}(t)) \wedge (U(t)N(t)R(t) = 0) \wedge (S(t)L(t) = 0),\\
&\mathbf{S} \overset{A(c)}{\subset} \tilde{\mathbf{S}} \ \textit{iff} \ \ (U(t)M(t) = -\dot{U}(t)) \wedge (U(t)N(t) = 0) \wedge (L(t) = 0),\\
&\mathbf{S} \overset{A(d)}{\subset} \tilde{\mathbf{S}} \ \textit{iff} \ \ (U(t)M(t) = -\dot{U}(t)) \wedge (U(t)N(t)R(t) = 0) \wedge (L(t) = 0)\\
&(\forall t \in [t_a, t_b]).
\end{aligned}$$

Proof. Condition $\dot{\mu}(t) = 0$, resulting from Definition 3.2, is equivalent to $A(t)U(t) - U(t)\tilde{A}(t) = \dot{U}(t)$ and $U(t)M(t) = -\dot{U}(t)$. The remaining conditions can be derived as in Theorem 3.4. □

It is easy to verify that the conditions for all the types of restriction and aggregation satisfy the general inclusion conditions from Theorem 2.3. For example, conditions $\nu(t) = \nu(t_0)$ and $\mu(t) = \mu(t_0)$ imply that there exist a left inverse $\mu(t) = \nu(t)^L$ and a right inverse $\nu(t_0) = \mu(t_0)^R$, respectively, so that $\mu(t)\nu(t_0) = I$.

The known conditions for the time-invariant case (e.g., [29, 35]) can be easily derived. In [17], the state restriction and aggregation conditions are discussed in the case of time-varying parameters and constant expansion-contraction maps; then we have, for restriction $M(t)V = 0$ and $N(t) = 0$ and for aggregation $UM(t) = 0$ and $UN(t) = 0$, special cases of Theorems 3.4 and 3.6.

Essential properties of input/state/output restrictions and aggregations can be seen from a specific equivalent representation of $\tilde{\mathbf{S}}$, similarly as in [18]. Define a nonsingular transformation $P(t) = [V(t)\vdots W(t)]$, where $W(t)$ is chosen such that $\mathcal{R}(W(t)) = \mathcal{N}(U(t))$. ($\mathcal{R}(.)$ denotes the range space and $\mathcal{N}(.)$ the null space of the indicated map.) Then there exists $W^*(t)$, the unique left inverse of $W(t)$, satisfying $\mathcal{N}(W^*(t)) = \mathcal{R}(V(t))$, and we have $P(t)^{-1} = \begin{bmatrix} U(t) \\ W^*(t) \end{bmatrix}$. Assuming, in addition, that $P(t)$ is continuously differentiable, the change of basis $x^*(t) = P(t)^{-1}\tilde{x}(t)$ leads to the following equivalent representation of $\tilde{\mathbf{S}}$:

$$\begin{aligned}
\mathbf{S}^*: \ \ \dot{x}^*(t) &= [P(t)^{-1}\tilde{A}(t)P(t) - P(t)^{-1}\dot{P}(t)]x^*(t) + P(t)^{-1}\tilde{B}(t)\tilde{u}(t)\\
&= \begin{bmatrix} A(t) & A_a(t) \\ A_r(t) & A^*(t) \end{bmatrix} x^*(t) + \begin{bmatrix} B_1(t) \\ B_2(t) \end{bmatrix} \tilde{u}(t),\\
\tilde{y}(t) &= P(t)\tilde{C}(t)x^*(t) = [C_1(t) \quad C_2(t)]x^*(t).
\end{aligned} \tag{22}$$

In the case in which $\mathbf{S} \overset{R(.)}{\subset} \tilde{\mathbf{S}}$, we have $A_r(t) = W^*(t)\tilde{A}(t)V(t) - W^*(t)\dot{V}(t) = W^*(t)V(t)A(t) = 0$ (having in mind the assumed properties of $W^*(t)$). Similarly, when $\mathbf{S} \overset{A(.)}{\subset} \tilde{\mathbf{S}}$, $A_a(t) = U(t)\tilde{A}(t)W(t) - U(t)\dot{W}(t) = [-\dot{U}(t) + A(t)U(t)]W(t) - U(t)\dot{W}(t) = 0$ because $U(t)W(t) = 0 \Rightarrow \dot{U}(t)W(t) + U(t)\dot{W}(t) = 0$. In both cases, the block $A(t)$ remains invariant since $U(t)\tilde{A}(t)V(t) - U(t)\dot{V}(t) = A(t)$ for restriction, and $U(t)\tilde{A}(t)V(t) + \dot{U}(t)V(t) = A(t)$ for aggregation (having in mind that $U(t)\dot{V}(t) + \dot{U}(t)V(t) = 0$). Also, in both cases, $A^*(t) = W^*(t)\tilde{A}(t)W(t) - W^*(t)\dot{W}(t)$.

The control and output matrices in $\mathbf{S}^*$ depend on the inclusion type. For $R(a)$ and $R(b)$, we have $B_1(t) = B(t)Q(t) + U(t)N(t)$ and $B_2(t) = W^*(t)N(t)$, and, for $R(c)$ and $R(d)$, $B_1(t) = B(t)Q(t)$ and $B_2(t) = 0$. In the case of $A(a)$ and $A(c)$, $B_1(t) = B(t)Q$ and $B_2(t) = W^*(t)N(t)$, and, in the case of $A(b)$ and $A(d)$, $B_1(t) = B(t)Q(t) + U(t)N(t)$ and $B_2(t) = W^*(t)N(t)$. Also, for $R(a)$ and $R(c)$, $C_1(t) = T(t)C(t)$ and

$C_2(t) = L(t)W(t)$, while, for $R(b)$ and $R(d)$, $C_1(t) = T(t)C(t) + L(t)V(t)$ and $C_2(t) = L(t)W(t)$. For $A(a)$ and $A(b)$, $C_1(t) = T(t)C(t) + L(t)V(t)$ and $C_2(t) = L(t)W(t)$, and, for $A(c)$ and $A(d)$, $C_1(t) = T(t)C(t)$ and $C_2(t) = 0$.

Using geometric arguments from [37, 18, 32, 35], it can be concluded that the principal property of all the restriction types is that the two-argument map $\Phi(t, t_0)$ describes, in fact, the action of the restriction of the map $\tilde{\Phi}(t, t_0)$ to $\mathcal{R}(V(t_0))$ with reduced codomain $\mathcal{R}(V(t))$ ($V(t)$ being considered as the insertion map). In the case of time-invariant expansions-contractions, we have, moreover, that $A(t) = \tilde{A}(t)|\mathcal{R}(V)$; that is, $A(t)$ represents the restriction of $\tilde{A}(t)$ to $\mathcal{R}(V)$ with codomain $\mathcal{R}(V)$ (as in the case of time-invariant systems). Therefore, in the case of time-varying expansions-contractions, the "dynamic" restriction of $\tilde{\mathbf{S}}$ does not correspond, in general, to the "static" restriction of the system matrix $\tilde{A}(t)$ itself. Similarly, in the case of aggregation, $\Phi(t, t_0)$ is obtained by aggregating $\tilde{\Phi}(t, t_0)$, according to [1, 29]. $A(t)$ represents an aggregation of $\tilde{A}(t)$ (the map induced in $\mathcal{R}(U)$ by $\tilde{A}(t)$) only in the case of time-invariant expansion-contraction maps.

However, only $R(a)$ can be considered as a "pure" restriction, having in mind that $N(t)R(t) = 0 \Leftrightarrow \tilde{B}(t)R(t) = V(t)B(t)$ and $L(t)V(t) = 0 \Leftrightarrow \tilde{C}(t)V(t) = T(t)C(t)$; that is, $B(t)$ is the restriction of $\tilde{B}(t)$ to $\mathcal{R}(R(t))$ with the reduced codomain $\mathcal{X}$, and $C(t)$ is the restriction of $\tilde{C}(t)$ to $\mathcal{R}(V(t))$ with the reduced codomain $\mathcal{Y}$. In $R(c)$, for example, we have $N(t) = 0 \Leftrightarrow \tilde{B}(t) = V(t)B(t)Q(t)$, that is, $B(t)$ is the map induced in $\mathcal{R}(Q(t))$ (aggregation of $\tilde{B}(t)$ to $\mathcal{R}(Q(t))$), with the reduced codomain $\mathcal{X}$. Analogous statements hold for aggregation. In the case of $A(a)$, we have $U(t)N(t) = 0 \Leftrightarrow U(t)\tilde{B}(t) = B(t)Q(t)$ and $S(t)L(t) = 0 \Leftrightarrow S(t)\tilde{C}(t) = C(t)U(t)$; that is, $B(t)$ and $C(t)$ represent aggregations of $\tilde{B}(t)$ and $\tilde{C}(t)$. The other aggregation types involve restrictions either in the domain or in the codomain.

The controllability and observability of systems including each other represent an important issue (e.g., [36]). It is easy to see from (22) that, in the case of $R(c)$ and $R(d)$, $\mathbf{S}^*$ (and, therefore, $\tilde{\mathbf{S}}$) is not controllable even when $\mathbf{S}$ is controllable (having in mind that $B_2(t) = 0$). However, in the case of $R(a)$ and $R(b)$, controllability of $\mathbf{S}$ could be preserved in the expanded space of $\tilde{\mathbf{S}}$ by a proper choice of the complementary matrix $N(t)$. Analogously, in the case of $A(c)$ and $A(d)$, $\mathbf{S}^*$ is not observable ($C_2(t) = 0$), while, in the case of $A(a)$ and $A(b)$, observability could be achieved by a proper choice of $L(t)$. This is an interesting aspect introduced by input/output expansions-contractions: compare this with the analysis in [18] related to the state restriction-aggregation, and with the general treatment of the time-invariant case in [32, 35]. There is a recent development [4, 5, 6], which expands the freedom in choosing complementary matrices in time-invariant inclusions, that can be applied to the time-varying case as well. It would be also possible to extend to the time-varying case the results presented in [36] related to the implementation of controllability and observability in systems satisfying restriction or aggregation conditions. As will be seen from the next two sections, the notions of *instantaneous controllability* and *instantaneous observability* (see [25]) play important roles in the characterization of time-varying systems under the inclusion relation.

4. Composition. The fundamental importance of restrictions and aggregations for inclusion relationships in the time-varying case will be emphasized by the following theorem, where we shall prove that any input/state/output inclusion can be considered as a specific composition of one input/state/output restriction and one input/state/output aggregation. This property has been shown for the state inclusion of time-invariant linear systems in [18].

Let

$$\bar{\mathbf{S}}\text{:}\quad \dot{\bar{x}}(t) = \bar{A}(t)\bar{x}(t) + \bar{B}(t)\bar{u}(t) \quad (\bar{x}(t_0) = \bar{x}_0), \qquad \bar{y}(t) = \bar{C}(t)\bar{x}(t), \tag{23}$$

where $\bar{x}(t) \in \bar{\mathcal{X}}$, $\bar{u}(t) \in \bar{\mathcal{U}}$, and $\bar{y}(t) \in \bar{\mathcal{Y}}$ with $n \le d(\bar{\mathcal{X}}) \le \tilde{n}$, $m \le d(\bar{\mathcal{U}}) \le \tilde{m}$, and $l \le d(\bar{\mathcal{Y}}) \le \tilde{l}$. ($d(.)$ denotes the dimension of the indicated space.)

THEOREM 4.1. *Assume that the matrices in both* $\mathbf{S}$ *and* $\tilde{\mathbf{S}}$*, as well as the expansion-contraction matrices* $V(t), U(t), R(t), Q(t), T(t)$*, and* $S(t)$ *are continuously differentiable a sufficient number of times* ($\le \tilde{n}$)*. Then* $\mathbf{S} \subset \tilde{\mathbf{S}}$ *on* $[t_a, t_b]$ *iff for each* $t \in [t_a, t_b]$ *there exist both an interval* $[t_s, t_f)$ *containing* t ($[t_s, t_f) \subset [t_a, t_b]$) *and a system* $\bar{\mathbf{S}}$ *defined on the same interval, such that* $\mathbf{S} \overset{A(b)}{\subset} \bar{\mathbf{S}} \overset{R(b)}{\subset} \tilde{\mathbf{S}}$ *on* $[t_s, t_f)$*. (*$t_s = t_f$ *on a set of measure zero.)*

Proof. The sufficiency follows from the basic fact that restrictions and aggregations are special cases of inclusion and that $(\bar{\mathbf{S}} \subset \tilde{\mathbf{S}}) \wedge (\mathbf{S} \subset \bar{\mathbf{S}}) \Rightarrow \mathbf{S} \subset \tilde{\mathbf{S}}$. Namely, it follows from the assumption of the theorem that any $[t_a, t_b]$ can be represented as a union of consecutive nonoverlapping intervals: $\tau_1 = [t_s^1 = t_a, t_f^1)$, $\tau_2 = [t_s^2 = t_f^1, t_f^2), \ldots, \tau_i = [t_s^i, t_f^i), \ldots$, so that $[t_a, t_b] = \bigcup_i \tau_i$. (For finite intervals $[t_a, t_b]$, the last interval reduces to the isolated point t_b.) Take one set $\{\tau_i\}$ covering $[t_a, t_b]$. By assumption, on each τ_i there exists a system $\bar{\mathbf{S}}$ represented by (23) which satisfies $\mathbf{S} \overset{A(b)}{\subset} \bar{\mathbf{S}} \overset{R(b)}{\subset} \tilde{\mathbf{S}}$. Assign quadruplets $(U_i^r(t), V_i^r(t), R_i^r(t), T_i^r(t))$ to $\bar{\mathbf{S}} \subset \tilde{\mathbf{S}}$ and quadruplets $(U_i^a(t), V_i^a(t), R_i^a(t), T_i^a(t))$ to $\mathbf{S} \subset \bar{\mathbf{S}}$ on τ_i, $i = 1, 2, \ldots$. Therefore, the choice $\tilde{x}(t_0^i) = V_i^r(t_0^i)\bar{x}(t_0^i)$ and $\tilde{u}(t) = R_i^r(t)\bar{u}(t)$ implies $\bar{x}(t) = U_i^r(t)\tilde{x}(t; t_0^i, \tilde{x}(t_0^i), \tilde{u}_{[t_0^i, t]})$ and $\bar{y}[\bar{x}(t)] = S_i^r(t)\tilde{y}[\tilde{x}(t)]$, and the choice $\bar{x}(t_0^i) = V_i^a(t_0^i)x(t_0^i)$ and $\bar{u}(t) = R_i^a(t)u(t)$ implies $x(t) = U_i^a(t)\bar{x}(t; t_0^i, \bar{x}(t_0^i), \bar{u}_{[t_0^i, t]})$ and $y[x(t)] = S_i^a(t)\bar{y}[\bar{x}(t)]$ for all $t_0^i, t \in \tau_i = [t_s^i, t_f^i), t_0^i \le t$, $i = 1, 2, \ldots$. Consequently, the choice $\tilde{x}(t_0^i) = V_i^r(t_0^i)V_i^a(t_0^i)x(t_0^i)$ and $\tilde{u}(t) = R_i^r(t)R_i^a(t)u(t)$ implies $x(t; t_0^i, x(t_0^i), u_{[t_0^i, t]}) = U_i^a(t)U_i^r(t)\tilde{x}(t; t_0^i, \tilde{x}(t_0^i), \tilde{u}_{[t_0^i, t]})$ and $y[x(t)] = S_i^a(t)S_i^r(t)\tilde{y}[\tilde{x}(t)]$ for all $t_0^i, t \in \tau_i = [t_s^i, t_f^i), t_0^i \le t$, and, therefore, $\mathbf{S} \subset \tilde{\mathbf{S}}$ on each τ_i, $i = 1, 2, \ldots$. The continuity of both $x(t)$ and $\tilde{x}(t)$, as solutions of (1) and (2), implies that $U_i^a(t_s^i)U_i^r(t_s^i) = U_{i-1}^a(t_f^{i-1})U_{i-1}^r(t_f^{i-1})$, $V_i^r(t_s^i)V_i^a(t_s^i) = V_{i-1}^r(t_f^{i-1})V_{i-1}^a(t_f^{i-1})$, and $S_i^a(t_s^i)S_i^r(t_s^i) = S_{i-1}^a(t_f^{i-1})S_{i-1}^r(t_f^{i-1})$. (The same conclusion holds for isolated points corresponding to $t_s^i = t_f^i$.) Analogously, $R_i^r(t_s^i)R_i^a(t_s^i) = R_{i-1}^r(t_f^{i-1})R_{i-1}^a(t_f^{i-1})$ except for a set of measure zero. Thus $\mathbf{S} \subset \tilde{\mathbf{S}}$ on $[t_a, t_b]$, with the quadruplet $(U(t) = U_i^a(t)U_i^r(t)$, $V(t) = V_i^r(t)V_i^a(t)$, $R(t) = R_i^r(t)R_i^a(t)$, $S(t) = S_i^a(t)S_i^r(t))$.

The necessity will be proven constructively, extending the essential idea presented in [18] to a more complex situation arising in the time-varying case.

Assume that $\mathbf{S} \subset \tilde{\mathbf{S}}$ on $[t_a, t_b]$, with the assigned quadruplet $(U(t), V(t), R(t), S(t))$.

Choose $V^r(t) = [V'(t) \vdots V(t)]$, where $V'(t) = (I - V(t)U(t))\hat{W}(t)$, and $\hat{W}(t)$ is any basis matrix for

$$\hat{\mathcal{W}}(t) = [\mathcal{R}(\langle \tilde{A}(t) | V(t) \rangle) + \mathcal{R}(\langle \tilde{A}(t) | N(t)R(t) \rangle)] \cap \mathcal{R}(V(t))^{\perp},$$

where

$$\langle \tilde{A}(t) | V(t) \rangle = [V(t) \vdots \tilde{F}_1(t; V(t)) \vdots \ldots \vdots \tilde{F}_{\tilde{n}-1}(t; V(t))],$$

$$\langle \tilde{A}(t) | N(t)R(t) \rangle = [N(t)R(t) \vdots \tilde{F}_1(t; N(t)R(t)) \vdots \ldots \vdots \tilde{F}_{\tilde{n}-1}(t; N(t)R(t))]. \tag{24}$$

($(.)^{\perp}$ denotes the annihilator of the indicated map.) Notice that $\tilde{F}_j(t; .)$ for $j \ge \tilde{n}$ represents a linear combination of $\tilde{F}_i(t, .)$, $i = 0, \ldots, \tilde{n} - 1$, so that the matrices

$\langle \tilde{A}(t)|V(t)\rangle$ and $\langle \tilde{A}(t)|N(t)R(t)\rangle$ in (24) have the maximal possible rank. Having in mind that

$$[V^{'}(t)\vdots V(t)] = [\hat{W}(t)\vdots V(t)] \begin{bmatrix} I & 0 \\ -U(t)\hat{W}(t) & I \end{bmatrix}$$

and $U(t)V(t) = I$, we conclude that $V^r(t)$ is monic as well as that $\mathcal{R}(V^r(t)) = \mathcal{R}(\langle \tilde{A}(t)|V(t)\rangle) + \mathcal{R}(\langle \tilde{A}|(N(t)R(t)\rangle)$ and $\mathcal{N}(U(t)) = \mathcal{R}(V^{'}(t))$, implying the existence of the left inverse $V^{'}(t)^L$ such that the choice $U^r(t) = \begin{bmatrix} V^{'}(t)^L \\ U(t) \end{bmatrix}$ implies $U^r(t)V^r(t) = I$. However, the rank of both $\langle \tilde{A}(t)|V(t)\rangle$ and $\langle \tilde{A}(t)|N(t)R(t)\rangle$ is, in general, time-varying, and, consequently, the dimensions of both $V^r(t)$ and $U^r(t)$ are, in general, time-varying. An insight into the properties of these matrices can be obtained by analyzing the symmetric positive semidefinite matrices $\langle \tilde{A}(t)|V(t)\rangle\langle \tilde{A}(t)|V(t)\rangle^T$ and $\langle \tilde{A}(t)|N(t)R(t)\rangle\langle \tilde{A}(t)|N(t)R(t)\rangle^T$. Their elements are, by assumption, continuous functions of time, which implies that their nonnegative real eigenvalues $\lambda_i^v(t)$ and $\lambda_i^{nr}(t)$, $i = 1, \ldots, \tilde{n}$, respectively, are continuous functions of time as well (see, for example, [11]). Consequently, these eigenvalues can be equal to zero either on closed intervals of nonzero length or on a set of time-instants of measure zero. Therefore, $\text{rank}\langle \tilde{A}(t)|V(t)\rangle$ and $\text{rank}\langle \tilde{A}(t)|N(t)R(t)\rangle$ are piecewise continuous functions of time, implying that, for each $t \in [t_a, t_b]$, there exists an interval $[t_s, t_f)$ containing t on which both functions have constant values. (These intervals reduce to single points, that is, $t_s = t_f$, on a set of points of measure zero, corresponding either to simple zeros of the functions $\lambda_i^v(t)$ and $\lambda_i^{nr}(t)$ or to the right end points of the segments of nonzero length, where these functions are equal to zero.) Starting from the differentiability assumption, we shall choose a continuously differentiable $\hat{W}(t)$ for all the continuity points of $\lambda_i^v(t)$ and $\lambda_i^{nr}(t)$ by selecting, for example, the linearly independent columns of $\langle \tilde{A}(t)|V(t)\rangle$ and $\langle \tilde{A}(t)|N(t)R(t)\rangle$.

We shall choose, also, $R^r(t) = [R^{'}(t)\vdots R(t)]$, where $R^{'}(t)$ is any basis matrix for $\bar{\mathcal{U}} \cap \mathcal{U}^{\perp} \subset \mathcal{N}(N(t)) \cap \mathcal{U}^{\perp}$ (if $\mathcal{N}(N(t)) \cap \mathcal{R}(R(t))^{\perp} = \emptyset$, then $R^r(t) = R(t)$), and $T^r(t) = [T^{'}(t)\vdots T(t)]$, where $T^{'}(t)$ is any basis matrix for $\bar{\mathcal{Y}} \subset \mathcal{Y}^{\perp}$. Therefore, there exist left inverses $R^{'}(t)^L$ and $T^{'}(t)^L$ such that the choice $Q^r(t) = \begin{bmatrix} R^{'}(t)^L \\ Q(t) \end{bmatrix}$ and $S^r(t) = \begin{bmatrix} T^{'}(t)^L \\ S(t) \end{bmatrix}$ implies both $Q^r(t)R^r(t) = I$ and $S^r(t)T^r(t) = I$.

Let the matrices in $\bar{\mathbf{S}}$ be

$$\begin{aligned} \bar{A}(t) = U^r(t)\tilde{A}(t)V^r(t) - U^r(t)\dot{V}^r(t), \quad \bar{B}(t) = U^r(t)\tilde{B}(t)R^r(t), \\ \bar{C}(t) = S^r(t)\tilde{C}(t)V^r(t). \end{aligned} \tag{25}$$

From the general expression $\tilde{A}(t) = V^r(t)\bar{A}(t)U^r(t) + M^r(t)$, we obtain

$$\begin{aligned} M^r(t)V^r(t) - \dot{V}^r(t) &= [\tilde{A}(t) - V^r(t)\bar{A}(t)U^r(t)]V^r(t) - \dot{V}^r(t) \\ &= [I - V^r(t)U^r(t)][\tilde{A}(t)V^r(t) - \dot{V}^r(t)] = 0 \end{aligned}$$

after replacing the expression for $\bar{A}(t)$ and using the fact that $\tilde{A}(t)V^r(t) - \dot{V}^r(t) \subset \mathcal{R}(V^r(t))$ by assumption. Note also that

$$\mathcal{R}(\tilde{B}(t)R^r(t)) \subset \mathcal{R}(V(t)) + \mathcal{R}(N(t)R^r(t)) \subset \mathcal{R}(V^r(t)),$$

having in mind that $\tilde{B}(t) = V(t)B(t)Q(t) + N(t)$ and $\mathcal{R}(N(t)R^r(t)) = \mathcal{R}(N(t)R(t))$, according to the definition of $R^r(t)$. From $\tilde{B}(t) = V^r(t)\bar{B}(t)Q^r(t) + N^r(t)$ we conclude, consequently, that

$$N^r(t)R^r(t) = [\tilde{B}(t) - V^r(t)\bar{B}(t)Q^r(t)]R^r(t) = [I - V^r(t)U^r(t)]\tilde{B}(t)R^r(t) = 0.$$

Also, from $\tilde{C}(t) = T^r(t)\bar{C}(t)U^r(t) + L^r(t)$ we obtain

$$S^r(t)L^r(t)V^r(t) = S^r(t)[\tilde{C}(t) - T^r(t)\bar{C}(t)U^r(t)]V^r(t) = 0.$$

Therefore, according to Theorem 3.4, $\bar{\mathbf{S}} \overset{R(b)}{\subset} \tilde{\mathbf{S}}$ on $[t_s, t_f)$ with the quadruplet $(U^r(t), V^r(t), R^r(t), S^r(t))$.

Choose now $V^a(t) = \begin{bmatrix}0\\ I\end{bmatrix}$, $U^a(t) = [0\ I]$, $R^a(t) = \begin{bmatrix}0\\ I\end{bmatrix}$, $Q^a(t) = [0\ I]$, $T^a(t) = \begin{bmatrix}0\\ I\end{bmatrix}$, and $S^a(t) = [0\ I]$. Obviously, $U^a(t)U^r(t) = U(t)$, $V^r(t)V^a(t) = V(t)$, $R^r(t)R^a(t) = R(t)$, and $S^a(t)S^r(t) = S(t)$.

From $\bar{A}(t) = V^a(t)A(t)U^a(t) + M^a(t)$ we obtain

$$\begin{aligned} U^a(t)M^a(t) &= U^a(t)[U^r(t)\tilde{A}(t)V^r(t) - U^r(t)\dot{V}^r(t) - V^a(t)A(t)U^a(t)] \\ &= U(t)\tilde{A}(t)[V^{'}(t)\vdots V(t)] - U(t)[\dot{V}^{'}(t)\vdots\dot{V}(t)] - A(t)[0\vdots I], \end{aligned} \tag{26}$$

having in mind that $\bar{A}(t) = U^r(t)\tilde{A}(t)V^r(t) - U^r(t)\dot{V}^r(t)$ according to (12), since $\bar{\mathbf{S}} \subset \tilde{\mathbf{S}}$.

Consider

$$U(t)F_1(t; V^{'}(t)) = U(t)(-\tilde{A}(t)V^{'}(t) + \dot{V}^{'}(t)), \tag{27}$$

with $V^{'}(t) = (I - V(t)U(t))\hat{W}(t)$; by assumption, $\hat{W}(t) \in \hat{\mathcal{W}}(t)$. Replacing $\hat{W}(t)$ by $\tilde{F}_j(t; V(t))$ in (27), we obtain that

$$\begin{aligned} U(t)\Big\{ &\tilde{A}(t)\tilde{F}_j(t; V(t)) - \dot{\tilde{F}}_j(t; V(t)) - \tilde{A}(t)V(t)U(t)\tilde{F}_j(t; V(t)) \\ &- \frac{d}{dt}[V(t)U(t)\tilde{F}_j(t; V(t))]\Big\} = 0 \end{aligned} \tag{28}$$

for $j = 0, 1, \ldots, \tilde{n} - 1$ since

$$\begin{aligned} &F_{j+1}(t; I) - U(t)[\tilde{A}(t)V(t)F_j(t; I) - \dot{V}(t)F_j(t; I) - V(t)\dot{F}_j(t; I)] \\ &\quad = F_{j+1}(t; I) - A(t)F_j(t; I) - \dot{F}_j(t; I) = 0. \end{aligned}$$

Analogously, we can show that $\mathcal{R}(\tilde{F}_j(t; \tilde{B}(t)R(t))) \subset \mathcal{N}(U(t))$ using the second relation in (3). As $\mathbf{S} \subset \tilde{\mathbf{S}}$ implies $U(t)\tilde{\Phi}(t, \tau)\tilde{B}(\tau) = U(t)\tilde{\Phi}(t, \tau)V(\tau)U(\tau)\tilde{B}(\tau)R(\tau) = \Phi(t, \tau)B(\tau)$, we obtain that

$$\mathcal{R}(\tilde{F}_j(t; V(t)U(t)\tilde{B}(t)R(t))) \subset \mathcal{N}(U(t)),$$

implying $\mathcal{R}(\tilde{F}_j(t; N(t)R(t))) \subset \mathcal{N}(U(t))$. Therefore, according to the definition of $\hat{\mathcal{W}}(t)$, we have $U^a(t)M^a(t) = 0$.

We also have $\bar{B}(t) = V^a(t)B(t)Q^a(t) + N^a(t)$ and

$$U^a(t)N^a(t)R^a(t) = U^a(t)[\bar{B}(t) - V^a(t)B(t)Q^a(t)]R^a(t) = 0,$$

using the relation $U(t)\tilde{B}(t)R(t) = B(t)$ from (16), since $\mathbf{S} \subset \tilde{\mathbf{S}}$ by assumption.

Reasoning similarly, we start from $\tilde{C}(t) = T(t)C(t)U(t) + L(t)$ and the inclusion relations related to $C(t)$ in (14) and obtain $S(t)L(t)\tilde{F}_i(t; V(t)) = 0$ and $S(t)L(t)\tilde{F}_i(t; \tilde{B}(t)R(t)) = 0$. Consequently,

$$\mathcal{N}(S(t)L(t)) \supset \mathcal{R}(\langle\tilde{A}(t)|V(t)\rangle) + \mathcal{R}(\langle\tilde{A}(t)N(t)R(t)\rangle)$$

and $S(t)L(t)V^{'}(t) = 0$, and we get

$$\begin{aligned} S^a(t)L^a(t) &= S^a(t)[\bar{C}(t) - T^a(t)C(t)U^a(t)] \\ &= S(t)(T(t)C(t)U(t) + L(t))[V^{'}(t)\vdots V(t)] - C(t)[0\vdots I] = 0, \end{aligned}$$

having in mind that $\bar{C}(t) = T^a(t)C(t)U^a(t) + L_a(t)$. Therefore, according to Theorem 3.6, $\mathbf{S} \overset{A(b)}{\subset} \bar{\mathbf{S}}$ on $[t_s, t_f)$ with the quadruplet $(U^a(t), V^a(t), R^a(t), S^a(t))$. Hence we have the result. □

The result of the theorem indicates that input/state/output contractions of dynamic time-varying systems may be considered as specific "compositions" of input/state/output restrictions and aggregations, analogously to the time-invariant case [18].

5. Equivalence. The inclusion relation between $\mathbf{S}$ and $\tilde{\mathbf{S}}$ does not imply algebraic equivalence, as in the time-invariant case (see [18]). However, it is possible to compare their external behavior for zero initial states. Let

$$\mathbf{S}_1\colon \dot{x}_1(t) = A_1(t)x_1(t) + B_1(t)u_1(t), \quad (x_1(t_0) = x_{10}), \qquad y_1(t) = C_1(t)x_1(t), \tag{29}$$

$$\mathbf{S}_2\colon \dot{x}_2(t) = A_2(t)x_2(t) + B_2(t)u_2(t), \quad (x_2(t_0) = x_{20}), \qquad y_2(t) = C_2(t)x_2(t), \tag{30}$$

where $x_1(t) \in \mathcal{X}_1$, $u_1(t) \in \mathcal{U}_1$, and $y_1(t) \in \mathcal{Y}_1$ are n_1-, m_1-, and l_1-vectors and $x_2(t) \in \mathcal{X}_2$, $u_2(t) \in \mathcal{U}_2$, and $y_2(t) \in \mathcal{Y}_2$ are n_2-, m_2-, and l_2-vectors, respectively.

DEFINITION 5.1. *We say that* $\mathbf{S}_1$ *and* $\mathbf{S}_2$ *are zero-state RS-equivalent on* $[t_a, t_b]$ *if there exist full rank matrices* $S_1(t)_{l\times l_1}$, $S_2(t)_{l\times l_2}$, $R_1(t)_{m_1\times m}$, *and* $R_2(t)_{m_2\times m}$ $(m \le \min(m_1, m_2)$, $l \le \min(l_1, l_2))$ *such that*

$$S_1(t)H_1(t,\tau)R_1(\tau) = S_2(t)H_2(t,\tau)R_2(\tau) \quad \forall t, \tau \in [t_a, t_b], t \ge \tau, \tag{31}$$

where $H_1(t,\tau)$ *and* $H_2(t,\tau)$ *denote the impulse responses of* $\mathbf{S}_1$ *and* $\mathbf{S}_2$, *respectively.*

We remark that for $S_1(t) = S_2(t) = I$ and $R_1(t) = R_2(t) = I$, systems $\mathbf{S}_1$ and $\mathbf{S}_2$ are zero-state equivalent in the usual sense [20, 25, 23].

THEOREM 5.2. *Systems* $\mathbf{S}$ *and* $\tilde{\mathbf{S}}$ *in* (1), (2) *are zero-state RS-equivalent on* $[t_a, t_b]$ *if* $\mathbf{S} \subset \tilde{\mathbf{S}}$ *on* $[t_a, t_b]$.

Proof. The inclusion relation $c(t)b(\tau) = S(t)\tilde{c}(t)\tilde{b}(\tau)R(\tau)$ from (3) is equivalent to $H(t,\tau) = S(t)\tilde{H}(t,\tau)R(\tau)$, having in mind that $H(t,\tau) = C(t)\Phi(t,\tau)B(\tau)$ and $\tilde{H}(t,\tau) = \tilde{C}(t)\tilde{\Phi}(t,\tau)\tilde{B}(\tau)$, and that $l \le \tilde{l}$ and $m \le \tilde{m}$. □

Thus, for $S_1(t) = S_2(t) = I$ and $R_1(t) = R_2(t) = I$, expansions and contractions are different realizations of a given impulse response. In this sense, as noted in [18] for the time-invariant case, they can be regarded as equivalent transformations not requiring preservation of dimensionality.

THEOREM 5.3. *Two systems* $\mathbf{S}_1$ *and* $\mathbf{S}_2$ *represented by* (29) *and* (30) *are zero-state RS-equivalent on* $[t_a, t_b]$ *if they have a common contraction* $\mathbf{S}$ *on* $[t_a, t_b]$.

Proof. Assume that $\mathbf{S} \subset \mathbf{S}_1$ with quadruplet $(U_1(t), V_1(t), R_1(t), S_1(t))$, and $\mathbf{S} \subset \mathbf{S}_2$ with quadruplet $(U_2(t), V_2(t), R_2(t), S_2(t))$. Then, according to (3),

$$S_1(t)C_1(t)\Phi_1(t,\tau)B_1(\tau)R_1(\tau) = S_2(t)C_2(t)\Phi_2(t,\tau)B_2(\tau)R_2(\tau) = C(t)\Phi(t,\tau)B(\tau), \tag{32}$$

that is, $\mathbf{S}_1$ and $\mathbf{S}_2$ are zero-state *RS*-equivalent, having in mind that $H_1(t,\tau) = C_1(t)\Phi_1(t,\tau)B_1(\tau)$ and $H_2(t,\tau) = C_2(t)\ \Phi_2(t,\tau)\ B_2(\tau)$. □

Conditions of Theorems 5.2 and 5.3 are not sufficient for inclusion. Assume that $\mathbf{S}_1$ with the impulse response $H_1(t,\tau)$ and $\mathbf{S}_2$ with the impulse response $H_2(t,\tau)$ are zero-state RS-equivalent on $[t_a,t_b]$, and that $\mathbf{S}$ represented by (1) is a realization of the impulse response $H(t,\tau) = S_1(t)H_1(t,\tau)R_1(\tau) = S_2(t)H_2(t,\tau)R_2(\tau)$, which is instantaneously controllable, instantaneously observable, and, therefore, minimal (see [19, 20, 25, 23]). Thus rank $\mathcal{I}(t) = n$ and rank $\mathcal{O}(t) = n$ for all $t \in [t_a,t_b]$, where

$$\begin{aligned} \mathcal{I}(t) &= [B(t)\vdots F_1(t;B(t))\vdots \ldots \vdots F_{n-1}(t;B(t))], \\ \mathcal{O}(t) &= [C(t)^T\vdots G_1(t;C(t))^T\vdots \ldots \vdots G_{n-1}(t;C(t))^T]^T, \end{aligned} \tag{33}$$

with $F_i(t;B(t))$ defined by (12) and (13), while $G_0(t;C(t)) = C(t)$ and $G_i(t;C(t)) = G_{i-1}(t;C(t))A(t) + \dot{G}(t;C(t))$, $i = 1,2,\ldots$. Successive differentiation gives

$$\begin{aligned} &\frac{\partial^{i+j}}{\partial t^i \partial \tau^j} C_k(t)\Phi_k(t,\tau)B_k(\tau) = G_i^k(t;C_k(t))\Phi_k(t,\tau)F_j^k(\tau;B_k(\tau)), \\ &\frac{\partial^{i+j}}{\partial t^i \partial \tau^j} C(t)\Phi(t,\tau)B(\tau) = G_i(t;C(t))\Phi(t,\tau)F_j(\tau;B(\tau)) \qquad (k=1,2), \end{aligned} \tag{34}$$

where $F_j^k(t;B_k(t))$ and $G_i^k(t;C_k(t))$ represent the constituent blocks of the controllability and observability matrices $\mathcal{I}_k(t)$ and $\mathcal{O}_k(t)$ of $\mathbf{S}_k$, $k = 1,2$, according to (33). Therefore,

$$\mathcal{O}(t)\Phi(t,\tau)\mathcal{C}(\tau) = S_1(t)\mathcal{O}_1(t)\Phi_1(t,\tau)\mathcal{C}_1(\tau)R_1(\tau) = S_2(t)\mathcal{O}_2(t)\Phi_2(t,\tau)\mathcal{C}_2(\tau)R_2(\tau). \tag{35}$$

Matrices $\mathcal{I}(t)$ and $\mathcal{O}(t)$ have full column and row rank, respectively, for all $t \in [t_a,t_b]$, implying the existence of both the left inverse $\mathcal{O}(t)^L$ and the right inverse $\mathcal{I}(t)^R$. Moreover, it follows from (35) for $\tau = t$ that if $\bar{V}_k(t) = \mathcal{I}_k(t)R_k(t)\mathcal{I}(t)^R$, then $\bar{U}_k(t) = \mathcal{O}(t)^L S_k(t)\mathcal{O}_k(t)$ satisfies $\bar{U}_k(t)\bar{V}_k(t) = I$. Therefore, (35) implies

$$\begin{aligned} \Phi(t,\tau) = \bar{U}_k(t)\Phi_k(t,\tau)\bar{V}_k(\tau), \qquad \Phi(t,\tau)B(\tau) &= \bar{U}_k(t)\Phi_k(t,\tau)B_k(\tau)R_k(\tau), \\ C(t)\Phi(t,\tau) &= S_k(t)C_k(t)\Phi_k(t,\tau)\bar{V}_k(\tau), \\ C(t)\Phi(t,\tau)B(\tau) &= S_k(t)C_k(t)\Phi_k(t,\tau)B_k(\tau)R_k(\tau), \end{aligned} \tag{36}$$

$k = 1,2$. Obviously, according to (3), $\mathbf{S} \subset \mathbf{S}_k$ with the quadruplets $(\bar{U}_k(t), \bar{V}_k(t), R_k(t), S_k(t))$, $k = 1,2$. This conclusion represents an extension to the time-varying case of the result obtained for the time-invariant case in [18].

6. State-feedback contractibility. The problem of inclusion of systems with feedback structure is important because in decentralized control applications it is necessary to carry out feedback control design in the expanded space, contract the obtained control law, and implement it in the original system (see, for example, [14, 29, 31]). Alternatively, a model-reduction scheme can be used to simplify the feedback design; then the obtained control law must be expanded before implementation [26, 8].

Consider the systems $\mathbf{S}$ and $\tilde{\mathbf{S}}$ under the time-varying state feedback defined by

$$\mathbf{C}\text{:}\quad u(t) = K(t)x(t) + r(t), \qquad \tilde{\mathbf{C}}\text{:}\quad \tilde{u}(t) = \tilde{K}(t)\tilde{x}(t) + \tilde{r}(t), \tag{37}$$

where $r(t)$ and $\tilde{r}(t)$ are reference inputs. Define the closed-loop systems by the corresponding triplets of matrices: $\mathbf{S}_f = (A_f(t), B(t), C(t))$ and $\tilde{\mathbf{S}}_f = (\tilde{A}_f(t), \tilde{B}(t), \tilde{C}(t))$, where $A_f(t) = A(t) + B(t)K(t)$ and $\tilde{A}_f(t) = \tilde{A}(t) + \tilde{B}(t)\tilde{K}(t)$. Obviously, $\mathbf{S}_f \subset \tilde{\mathbf{S}}_f$

on $[t_a, t_b]$ iff the conditions of Theorem 2.3 are satisfied, with the transition matrices $\Phi_f(t,\tau)$ and $\tilde{\Phi}_f(t,\tau)$ for $\mathbf{S}_f$ and $\tilde{\mathbf{S}}_f$, respectively. This formulation does not allow, however, getting explicit conditions for $K(t)$ and $\tilde{K}(t)$ to satisfy $\mathbf{S}_f \subset \tilde{\mathbf{S}}_f$, suitable for design purposes. However, sufficient conditions for controller contractibility (expandability) can be formulated starting from the following definition.

DEFINITION 6.1. *Controller $\tilde{\mathbf{C}}$ is contractible to the controller $\mathbf{C}$ (or $\mathbf{C}$ is expandable to $\tilde{\mathbf{C}}$) if the condition $(C1) \vee (C2)$ holds on $[t_a, t_b]$, where*

$$(C1)\colon \exists (U(t), V(t), R(t), S(t))\ \{x_0, u\} \Rightarrow \{\tilde{x}, \tilde{y}\} \wedge (R(t)K(t)x(t) = \tilde{K}(t)\tilde{x}(t)), \tag{38}$$

$$(C2)\colon \exists (U(t), V(t), Q(t), S(t))\ \{x_0, \tilde{u}\} \Rightarrow \{\tilde{x}, \tilde{y}\} \wedge (K(t)x(t) = Q(t)\tilde{K}(t)\tilde{x}(t)). \tag{39}$$

THEOREM 6.2. *$\tilde{\mathbf{C}}$ is contractible to $\mathbf{C}$ on $[t_a, t_b]$ iff the condition $(D0) \wedge [(D1) \vee (D2)]$ holds for all $t_0, \tau, t \in [t_a, t_b]$, $t_0 \le \tau \le t$:*

$$(D0)\colon \quad \mu(t)\nu(t_0) = I, \qquad c(t) = S(t)\tilde{c}(t)\nu(t_0). \tag{40}$$

$$\begin{aligned} (D1)\colon \quad & b(\tau) = \mu(t)\tilde{b}(\tau)R(\tau), \quad c(t)b(\tau) = S(t)\tilde{c}(t)\tilde{b}(\tau)R(\tau), \\ & R(t)K(t)\Phi(t,t_0) = \tilde{K}(t)\tilde{\Phi}(t,t_0)V(t_0), \\ & R(t)K(t)\Phi(t,\tau)B(\tau) = \tilde{K}(t)\tilde{\Phi}(t,\tau)\tilde{B}(\tau)R(\tau). \end{aligned} \tag{41}$$

$$\begin{aligned} (D2)\colon \quad & b(\tau)Q(\tau) = \mu(t)\tilde{b}(\tau), \quad c(t)b(\tau)Q(\tau) = S(t)\tilde{c}(t)\tilde{b}(\tau), \\ & K(t)\Phi(t,t_0) = Q(t)\tilde{K}(t)\tilde{\Phi}(t,t_0)V(t_0), \\ & K(t)\Phi(t,\tau)B(\tau)Q(\tau) = Q(t)\tilde{K}(t)\tilde{\Phi}(t,\tau)\tilde{B}(\tau). \end{aligned} \tag{42}$$

Proof. The proof follows directly from the conditions $(C1)$ and $(C2)$ in Definition 6.1, using the methodology of the proof of Theorem 2.3. In the case of $(C1)$, condition $(D0)$, together with the first two relations in (41), represents the inclusion conditions for $\mathbf{S}$ and $\tilde{\mathbf{S}}$, according to Theorem 2.3, while the remaining conditions in $(D1)$ follow from $R(t)K(t)x(t) = \tilde{K}(t)\tilde{x}(t)$ in (38). Conditions $(D0) \wedge (D2)$ follow from $C(2)$ in a similar way. Hence we have the result. □

It is very important to notice that contractibility in the sense of Definition 6.1 and Theorem 6.2 implies inclusion of the corresponding closed-loop systems. Indeed, condition $(C2)$, for example, ensures that, for any x_0 and $\tilde{u}_{[t_0,t]}$, $x(t; t_0, x_0, u_{[t_0,t]}) = U(t)\tilde{x}(t; t_0, \tilde{x}_0, \tilde{u}_{[t_0,t]})$, where $u_{[t_0,t]}$ is generated by $u(t) = Q(t)\tilde{u}(t)$ for all $t \in [t_0, t]$, and $\tilde{x}_0 = V(t_0)x_0$. Therefore, the input $\tilde{u}(t) = \tilde{K}(t)\tilde{x}(t)$ is just one particular choice satisfying the conditions for $\mathbf{S}_f \subset \tilde{\mathbf{S}}_f$, provided $u(t) = K(t)x(t) = Q(t)\tilde{K}(t)\tilde{x}(t)$. We have a completely analogous situation with $(C1)$. In both cases, $\mathbf{S} \subset \tilde{\mathbf{S}}$, while the additional conditions for $K(t)$ and $\tilde{K}(t)$ can be considered as the regulator output inclusion conditions (if $K(t)$ and $\tilde{K}(t)$ are considered as output matrices and $Q(t)$ and $R(t)$ as the corresponding expansion-contraction maps). Consequently, one of the important implications of contractibility is the preservation of stability. It follows from (3) that, in general,

$$\mathbf{S}_f \subset \tilde{\mathbf{S}}_f \quad \Rightarrow \quad \|\Phi_f(t,t_0)\| \le \|U(t)\|\,\|\tilde{\Phi}_f(t,t_0)\|\,\|V(t_0)\| \tag{43}$$

so that $\|\tilde{\Phi}_f(t,t_0)\| \le \tilde{M} < \infty \Rightarrow \|\Phi_f(t,t_0)\| \le M < \infty$; that is, the stability of $\tilde{\mathbf{S}}_f$ implies the stability of $\mathbf{S}_f$, provided $(\|U(t)\| \le M_u < \infty) \wedge (\|V(t_0)\| \le M_v < \infty)$ for all $t_0, t \in [t_a, t_b], t \ge t_0$.

The methodology of Corollary 2.5 now provides the following corollary.

COROLLARY 6.3. *Assume that* $\tilde{\mathbf{C}}$ *is contractible to* $\mathbf{C}$ *and that all the matrices in* (1) *and* (2), *the expansion-contraction matrices, and matrices* $K(t)$ *and* $\tilde{K}(t)$ *are continuously differentiable a sufficient number of times. Then either*

$$\begin{aligned} R(t)K(t)F_i(t;I) &= \tilde{K}(t)\tilde{F}_i(t;V(t)), \\ R(t)K(t)F_i(t;B(t)) &= \tilde{K}(t)\tilde{F}_i(t;\tilde{B}(t)R(t)) \end{aligned} \tag{44}$$

or

$$\begin{aligned} K(t)F_i(t;I) &= Q(t)\tilde{K}(t)\tilde{F}_i(t;V(t)), \\ K(t)F_i(t;B(t)Q(t)) &= Q(t)\tilde{K}(t)\tilde{F}_i(t;\tilde{B}(t)R(t)). \end{aligned} \tag{45}$$

In the case of restrictions and aggregations the contractibility conditions become the following.

THEOREM 6.4. $\tilde{\mathbf{C}}$ *is contractible to* $\mathbf{C}$ *on* $[t_a, t_b]$ *if one of the following conditions is satisfied for all* $t \in [t_a, t_b]$*:*

$$\begin{aligned} &(R1)\text{:} \quad (\mathbf{S} \overset{R(a)}{\subset} \tilde{\mathbf{S}}) \wedge (R(t)K(t) = \tilde{K}(t)V(t)), \\ &(R2)\text{:} \quad (\mathbf{S} \overset{R(b)}{\subset} \tilde{\mathbf{S}}) \wedge (R(t)K(t) = \tilde{K}(t)V(t)), \\ &(R3)\text{:} \quad (\mathbf{S} \overset{R(c)}{\subset} \tilde{\mathbf{S}}) \wedge (K(t) = Q(t)\tilde{K}(t)V(t)), \\ &(R4)\text{:} \quad (\mathbf{S} \overset{R(d)}{\subset} \tilde{\mathbf{S}}) \wedge (K(t) = Q(t)\tilde{K}(t)V(t)). \end{aligned}$$

THEOREM 6.5. $\tilde{\mathbf{C}}$ *is contractible to* $\mathbf{C}$ *on* $[t_a, t_b]$ *if one of the following conditions is satisfied for all* $t \in [t_a, t_b]$*:*

$$\begin{aligned} &(A1)\text{:} \quad (\mathbf{S} \overset{A(a)}{\subset} \tilde{\mathbf{S}}) \wedge (K(t)U(t) = Q(t)\tilde{K}(t)), \\ &(A2)\text{:} \quad (\mathbf{S} \overset{A(b)}{\subset} \tilde{\mathbf{S}}) \wedge (R(t)K(t)U(t) = \tilde{K}(t)), \\ &(A3)\text{:} \quad (\mathbf{S} \overset{A(c)}{\subset} \tilde{\mathbf{S}}) \wedge (K(t)U(t) = Q(t)\tilde{K}(t)), \\ &(A4)\text{:} \quad (\mathbf{S} \overset{A(d)}{\subset} \tilde{\mathbf{S}}) \wedge (R(t)K(t)U(t) = \tilde{K}(t)). \end{aligned}$$

The inclusion conditions obtained in Theorems 6.4 and 6.5 are completely analogous to the conditions derived in [35] for the time-invariant case and result, essentially, from different combinations of input/output expansions-contractions. It should also be noted that all the conditions from Theorems 6.4 and 6.5 satisfy the contractibility conditions from Theorem 6.2.

Example 6.6. The initial stages of the decentralized overlapping control design in the time-varying case can follow essentially the same lines as in the time-invariant case. Namely, the expansion map, leading to a decomposition of time-varying subsystems, remains, in general, in the same time-invariant form as in [16, 29, 33, 35]. Take, for example, a feedback-linearized time-varying model of a platoon of automotive vehicles, according to [34], in which the ith vehicle is described by

$$\begin{aligned} \dot{d}_i &= v_{i-1} - v_i, \\ \dot{v}_i &= a_i, \\ \dot{a}_i &= -\tau_i(t)^{-1}a_i + \tau_i(t)^{-1}u_i, \end{aligned} \tag{46}$$

where $d_i = x_{i-1} - x_i$ is the distance between two consecutive vehicles, x_{i-1} and x_i being their positions, v_i and a_i are the velocity and acceleration, respectively, u_i is the

input signal chosen to make the closed loop satisfy certain performance criteria, and $\tau_i(t)$ is the time-varying time-constant of the engine (depending, in general, on the velocity) . After forming the state-space model for the platoon with the overall state $X^T = (d_1, v_1, a_1, d_2, v_2, a_2, \ldots, d_N, v_N, a_N)$ and overall input $u^T = (u_1, u_2, \ldots, u_N)$, where N is the number of vehicles, and applying the time-invariant expansions

$$V^T = \begin{bmatrix} I_{3\times3} & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & I_{2\times2} & I_{2\times2} & 0 & & & \\ 0 & 0 & 0 & I_{3\times3} & & & \\ & & & & & \ldots & \\ & & & & & & I_{3\times3} \end{bmatrix}$$

and

$$R^T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 1 & 0 & & & \\ 0 & 0 & 0 & 1 & & & \\ & & & & & \ldots & \\ & & & & & & 1 \end{bmatrix},$$

respectively, one obtains the following model in the expanded space:

$$\tilde{\mathbf{S}}: \quad \begin{bmatrix} \dot{\xi}_1 \\ \dot{\xi}_2 \\ \vdots \\ \dot{\xi}_N \end{bmatrix} = \begin{bmatrix} A_1(t) & 0 & \ldots & 0 \\ 0 & A_2(t) & \ldots & \\ & & \ldots & \\ 0 & & & A_N(t) \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_N \end{bmatrix} + \begin{bmatrix} B_1(t) & 0 & \ldots & 0 \\ 0 & B_2(t) & & \\ & & \ldots & \\ 0 & & & B_N(t) \end{bmatrix} \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_N \end{bmatrix}, \tag{47}$$

in which the *overlapping* subsystems are defined by

$$\mathbf{S}_i: \quad \dot{\xi}_i = A_i(t)\xi_i + B_i(t)\zeta_i = \begin{bmatrix} A_i^l(t) & 0 \\ \bar{A}_d & A_i^v(t) \end{bmatrix} \xi_i + \begin{bmatrix} B_i^l(t) & 0 \\ 0 & B_i^v(t) \end{bmatrix} \zeta_i, \tag{48}$$

where $\xi_i^T = (v_{i-1}, a_{i-1}, d_i, v_i, a_i)$ is the state vector of the ith subsystem, $\zeta_i^T = (u_{i-1}, u_i)$ represents its control vector, and

$$A_i^l(t) = \begin{bmatrix} 0 & 1 \\ 0 & \tau_i(t)^{-1} \end{bmatrix}, \quad \bar{A}_d^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B_i^l(t) = \begin{bmatrix} 0 \\ \tau_i(t)^{-1} \end{bmatrix},$$

$$A_i^v(t) = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -\tau_i(t)^{-1} \end{bmatrix}, \quad B_i^v(t) = \begin{bmatrix} 0 \\ 0 \\ \tau_i(t)^{-1} \end{bmatrix},$$

$i = 1, \ldots, N$. A decentralized time-varying feedback control law can now be designed for the decoupled subsystems. The resulting control law can then be contracted for implementation by using the expansion-contraction matrices as in [16, 18, 29, 34, 35].

Example 6.7. In the case of the reduced order design, time-varying expansions-contractions bring new possibilities; see [26] for the time-invariant case. Consider, for example, the time-varying systems

$$\tilde{\mathbf{S}}: \quad \dot{\tilde{x}} = \begin{bmatrix} 0 & \frac{1}{2}\dot{\alpha}(t) \\ 2 & \alpha(t) \end{bmatrix} \tilde{x} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u \quad (\tilde{x}(t_0) = \tilde{x}_0), \quad y = \begin{bmatrix} 1 & \frac{1}{2}\alpha(t) \end{bmatrix} \tilde{x}, \tag{49}$$

$$\mathbf{S}: \quad \dot{x} = \alpha(t)x + \left(1 + \frac{1}{2}\alpha(t)\right) u \quad (x(t_0) = x_0), \quad y = x, \tag{50}$$

where $\alpha(t)$ is a continuously differentiable function. It is easy to verify that $\mathbf{S} \overset{A(.)}{\subset} \tilde{\mathbf{S}}$ with the triplet $(U(t) = [1 \ \ \frac{1}{2}\alpha(t)], Q(t) = 1, S(t) = 1)$. Assume that the control design is done in the space of $\mathbf{S}$ and that the resulting feedback gain (possibly optimal in some sense) is $K(t) = k(t)$. Then, according to the above considerations, its expansion to the space of $\tilde{\mathbf{S}}$ should be given by $\tilde{K}(t) = k(t)U(t) = [k(t) \ \ \frac{1}{2}k(t)\alpha(t)]$.

7. Conclusion. This paper presents a unified picture of basic aspects of the inclusion concept applied to dynamic time-varying linear continuous-time systems. Starting from the main definitions, involving time-varying expansions-contractions of inputs, states, and outputs of systems under consideration (previous work has been concerned exclusively with time-invariant expansions-contractions), we expressed inclusion conditions in terms of system characteristics. A set of inclusion conditions has been derived involving time-varying system matrices and their derivatives. Restrictions and aggregations of different types have been introduced, covering different combinations of time-varying input/state/output expansions-contractions. Fundamental aspects of restrictions and aggregations are made evident via a specific equivalent representation of the expanded system model. The main theoretical contribution of the paper is contained in the theorem dealing with the composition property. By that any input/state/output inclusion relation on a given time-interval can be decomposed into a specific sequence of pairs of input/state/output restrictions and input/state/output aggregations. The proof uses geometric arguments in the context of time-variability of system characteristics and expansion-contraction maps. The presentation of basic features of the time-varying inclusion covers the problem of zero-state equivalence. It has been proved, extending the approach in [18] to time-varying systems, that two systems are zero-state equivalent on a given time-interval if they have a common contraction on the same interval. The paper also encompasses the practically important problem of contractibility (expandability) of time-varying state-feedback controllers. An indication is given of how to approach the problems of overlapping decentralized and reduced order control designs.

Further research should explore a direct application of the above-developed methodology to the inclusion of time-varying observers and dynamic controllers in the context of the inclusion of performance indices (generalizing the existing approaches related to time-invariant systems). It would be especially important to clarify main practical aspects of decentralized overlapping and reduced order control design of time-varying systems. Moreover, the proposed methodology of analysis, which was presented in the context of the composition property, indicates possible applications to switching and hybrid dynamical systems.

REFERENCES

[1] M. Aoki, *Aggregation, in Optimization Methods for Large-Scale Systems with Applications*, D. A. Wismer, ed., McGraw–Hill, New York, 1971.

[2] L. Bakule and J. Rodellar, *Decentralized control and overlapping decompositions of mechanical systems*, Internat. J. Control, 61 (1995), pp. 559–587.
[3] L. Bakule, J. Rodellar, and J. M. Rossell, *Structure of expansion-contraction matrices in the inclusion principle for dynamic systems*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1136–1155.
[4] L. Bakule, J. Rodellar, and J. M. Rossell, *Generalized selection of complementary matrices in the inclusion principle*, IEEE Trans. Automat. Control, 45 (2000), pp. 1237–1243.
[5] L. Bakule, J. Rodellar, and J. M. Rossell, *Controllability–observability of expanded composite systems*, Linear Algebra Appl., 332–334 (2000), pp. 381–400.
[6] L. Bakule, J. Rodellar, J. M. Rossel, and P. Rubió, *Preservation of controllability–observability in expanded systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 1155–1162.
[7] B. Baran, E. Kaszkurewicz, and A. Bhaya, *Parallel asynchronous team algorithms: Convergence and performance analysis*, IEEE Trans. Parallel Distribributed Systems, 7 (1996), pp. 677–688.
[8] X.-B. Chen, *Some Aspects of Control Systems Design Based on the Inclusion Principle*, Ph.D. Dissertation, University of Belgrade, Belgrade, Yugoslavia, 1994.
[9] M. F. Hassan, R. I. Badr, M. M. Elewa, and H. A. Elnemr, *Expert robust decentralized controller for uncertain large-scale control systems*, IEE Proc. Control Theory Appl., 143 (1996), pp. 519–529.
[10] M. Hodžić and D. D. Šiljak, *Decentralized estimation and control with overlapping information sets*, IEEE Trans. Automat. Control, 31 (1986), pp. 81–86.
[11] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1986.
[12] A. Iftar, *Decentralized estimation and control with overlapping input, state, and output decomposition*, Automatica J. IFAC, 29 (1993), pp. 511–516.
[13] A. Iftar, *Overlapping decentralized dynamic optimal control*, Internat. J. Control, 58 (1993), pp. 187–209.
[14] A. Iftar and U. Özgüner, *Contractible controller design and optimal control with state and output inclusion*, Automatica J. IFAC, 26 (1990), pp. 593–597.
[15] A. Iftar and U. Özgüner, *Overlapping decompositions, expansions, contractions and stability of hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 1040–1055.
[16] M. Ikeda and D. D. Šiljak, *Overlapping decentralized control with input, state and output inclusion*, Control Theory and Advanced Technology, 2 (1986), pp. 155–172.
[17] M. Ikeda, D. D. Šiljak, and D. E. White, *Overlapping decentralized control of linear time-varying systems*, in Advances in Large Scale Systems, J. B. Cruz, ed., JAI Press, Greenwich, CT, 1984, pp. 93–116.
[18] M. Ikeda, D. D. Šiljak, and D. E. White, *An inclusion principle for dynamic systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 244–249.
[19] R. E. Kalman, *Mathematical description of linear dynamical systems*, J. SIAM Control Ser. A, 1 (1963), pp. 152–192.
[20] R. E. Kalman, P. L. Falb, and M. A. Arbib, *Topics in Mathematical System Theory*, McGraw–Hill, New York, 1969.
[21] R. Krtolica, M. Hodžić, and D. D. Šiljak, *A stochastic inclusion principle*, in Differential Equations: Stability and Control, S. Elaydi, ed., Marcel Dekker, New York, 1991, pp. 295–320.
[22] K. Li, E. B. Kosmatopoulos, P. A. Yoannou, and H. Ryciotaki-Boussalis, *Large segmented telescopes: Centralized, decentralized and overlapping control designs*, IEEE Control Systems Magazine, 20 (2000), pp. 59–72.
[23] L. Padulo and M. A. Arbib, *System Theory*, Hemisphere Publishing, Washington, D.C., 1974.
[24] J. A. Richards, *Analysis of Periodically Time-Varying Systems*, Springer-Verlag, New York, 1983.
[25] W. J. Rugh, *Linear System Theory*, Prentice–Hall, Upper Saddle River, NJ, 1996.
[26] M. E. Sezer and D. D. Šiljak, *Validation of reduced order models for control systems design*, J. Guidance Control Dynam., 5 (1982), pp. 430–437.
[27] M. E. Sezer and D. D. Šiljak, *Nested epsilon decompositions of linear systems: Weakly coupled and overlapping blocks*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 521–533.
[28] D. D. Šiljak, *Dynamic reliability using multiple control systems*, in Proceedings of the Second Lawrence Symposium on Systems Decision Science, UC Berkeley, Berkeley, CA, 1978, pp. 173–187.
[29] D. D. Šiljak, *Decentralized Control of Complex Systems*, Academic Press, New York, 1991.
[30] D. D. Šiljak and A. I. Zečević, *Large-scale and decentralized systems*, in Wiley Encyclopedia, J. G. Webster, New York, 1999, pp. 209–224.

[31] S. S. Stanković, X.-B. Chen, and D. D. Šiljak, *Stochastic inclusion principle applied to decentralized overlapping suboptimal LQG control*, in Proceedings of the 13th IFAC World Congress, L, 1996, pp. 12–17.

[32] S. S. Stanković and D. D. Šiljak, *Contractibility of overlapping decentralized control*, in Proceedings of the American Control Conference, Chicago, IL, 2000, pp. 811–818.

[33] S. S. Stanković, X.-B. Chen, M. R. Mataušek, and D. D. Šiljak, *Stochastic inclusion principle applied to decentralized automatic generation control*, Internat. J. Control, 72 (1999), pp. 276–288.

[34] S. S. Stanković, M. J. Stanojević, and D. D. Šiljak, *Decentralized overlapping control of a platoon of vehicles*, IEEE Trans. Control Systems Technology, 8 (2000), pp. 816–832.

[35] S. S. Stanković and D. D. Šiljak, *Contractibility of overlapping decentralized control*, Systems Control Lett., 44 (2001), pp. 189–199.

[36] S. S. Stanković and D. D. Šiljak, *Model abstraction and inclusion principle: A comparison*, IEEE Trans. Automat. Control, 47 (2002), pp. 529–532.

[37] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1979.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 342–362

VARIATIONAL INCLUSIONS UNDER STATE CONSTRAINTS*

MOULKA TAMZALI-LAFOND†

To my son Ander

Abstract. We extend variational equations of ODEs to nonconvex differential inclusions under state constraints. As an application, we provide a direct proof of the maximum principle of the Pontryagin type for an optimal control problem with state constraints.

Key words. differential inclusions, feasible trajectories, maximum principle, state constraints, variational inclusions

AMS subject classifications. 34A60, 49A24, 49J40, 49K24

PII. S036301290139155X

Introduction. Let us consider the differential inclusion

$$x'(t) \in F(t, x(t)), \tag{1}$$

where F is a set-valued map from $[0,T] \times \mathbb{R}^n$ into $\mathbb{R}^n$ and $T > 0$ is given.

We denote by $\mathcal{S}_{[0,T]}(x_0)$ the set of solutions to (1) starting at x_0 and defined on the time interval $[0,T]$.

Investigation of optimization problems is often based on linearization techniques. What should we call linearized inclusion?

For a given trajectory $z \in \mathcal{S}_{[0,T]}(x_0)$, it was suggested in Frankowska [10] to consider the inclusion

$$(**) \begin{cases} w'(t) \in F'(z(t), z'(t)) \cdot w(t) \text{ a.e. in } [0,T], \\ w(0) = 0, \end{cases}$$

where F' denotes a derivative of F at $(z(t), z'(t))$ in a sense we specify later.

It was shown that a solution w to (**) is a limit in $W^{1,1}$ of the difference quotients $\frac{x_h - z}{h}$ when $h \to 0^+$ for some $x_h \in \mathcal{S}_{[0,T]}(x_0)$.

This paper is devoted to an extension of this result to a state constraints case formulated as $x(t) \in K$. Namely, we consider the constrained system

$$\begin{cases} x'(t) \in F(t, x(t)), \\ x(0) = x_0 \ \in C_0, \\ x(t) \in K, \end{cases} \tag{2}$$

where C_0 is a given subset of $\mathbb{R}^n$.

The set of all solutions to (2) defined on $[0,T]$ and starting at some x_0 will be denoted by $\mathcal{S}^K_{[0,T]}(x_0)$.

The main theme of this paper is to replace the above differential inclusion by its "linear approximation" along a trajectory $\overline{x} \in \mathcal{S}^K_{[0,T]}(x_0)$ and to prove an extension of variational equations of ODEs to such a constrained multivalued system.

*Received by the editors June 27, 2001; accepted for publication (in revised form) July 1, 2002; published electronically April 17, 2003.

http://www.siam.org/journals/sicon/42-1/39155.html

†Centre de recherche Viabilité - Jeux - Contrôle, E.R.S. 2064, Université Paris IX Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris cedex 16, Paris, France (mtamzali-lafond@yahoo.fr).

The linearized system along $\overline{x}$ is defined by

$$\begin{cases} w'(t) \in \ dF(t, \overline{x}(t), \overline{x}'(t)) \cdot w(t) \text{ a.e. in } [0,T], \\ w(0) = w_0, \\ w(t) \ \in \mathcal{C}(t) \quad \forall t \text{ in } [0,T], \end{cases}$$

where dF denotes the derivative of the set-valued map F with respect to x, whose definition is recalled in section 1, and $\mathcal{C}(t)$ is a family of closed cones contained in tangent to K at $\overline{x}(t)$.

In the case in which K has sufficiently smooth boundary, $\mathcal{C}(t)$ are usual tangent cones to K at $\overline{x}(t)$. When the boundary of K is nonsmooth, we have to compensate for it by refinement of cones, requiring them to satisfy some additional properties.

We prove, in the first part, that for every solution $w(\cdot)$ of the linearized system, there exist trajectories $\overline{x}_h(\cdot)$ of the original system satisfying

$$\frac{\overline{x}_h(\cdot) - \overline{x}(\cdot)}{h} \longrightarrow w(\cdot) \text{ as } h \to 0^+,$$

using recent results of Frankowska–Rampazzo [12] and Frankowska–Vinter [14]. In other words, the linearized system plays the same role as the variational equation in ODE theory. For this reason, it is called variational inclusion. We provide different instances of constraints for which variational inclusion holds true.

In the second part, we apply this result to obtain a direct proof of the maximum principle under state constraints.

Consider the following optimal control problem:

$$\text{Minimize } \varphi(x(T)) \text{ over solutions to } \begin{cases} x'(t) \in F(t, x(t)) \text{ a.e. in } [0,T], \\ x(0) = x_0, \\ x(t) \in K \quad \forall t \in [0,T]. \end{cases}$$

Let $\overline{x}(\cdot)$ be a minimizer issued from $\overline{x}_0$.

To prove the first order necessary condition for optimality, we apply Fermat's rule: at the point of minimum, the derivative of the function φ is nonnegative in the direction of $w(T)$, where $w(\cdot)$ is a solution of the linearized system. That means we obtain the following inequality:

$$\langle \nabla \varphi(\overline{x}(T)), w(T) \rangle \geq 0.$$

Naturally, to derive this inequality, we use the first part of the article.

Next, using duality theory, we give a direct proof of the maximum principle. By direct proof we mean arguments based on linearization of the original system (the so-called variational inclusions).

A very simple and direct proof of the maximum principle for unconstrained control problems can be found in Pallu de la Barriere [17]. It is based on the variational equation of the ODE.

Linearization of differential inclusion for unconstrained problems and applications to maximum principle was studied in Frankowska [10].

The necessary condition for differential inclusion problems has been known for many years. Since linearization arguments are not applicable in the constrained case, different authors applied tools based on different arguments. For instance, for target control problems, Pontryagin and al. [18] used fixed point arguments. Clarke and

Loewen [6] and Vinter and Pappas [24] derived necessary conditions associated with the optimal control problem involving state constraints in the form $g(x(t)) \leq 0$ when the function g is Lipschitzian and using the nonsmooth analysis tools of Clarke [5] and the ε-variational principle of Ekeland [7]. We notice that their proofs are based on the reduction of optimal control problems to the Bolza type calculus of variations problems without state constraints.

Recently, Arutyunov and Aseev [3] have obtained nondegenerate necessary conditions in Hamiltonian form when F has convex images and is locally Lipschitz in both variables. We do not assume here that F is Lipschitz with respect to the time and do not study additional assumptions to be imposed to get nondegenerate necessary conditions. However, Case 1 of the proof of Theorem 4.1 yields that conditions are nondegenerate, while Case 2 does not allow this conclusion. Also, in [3] the proof uses Ekeland's principle applied to a family of auxiliary problems and a limiting procedure.

Rampazzo and Vinter [20] derived necessary conditions to a nonconvex problem in terms of normal cones and a limiting subdifferential. Their results are proved by replacing the original problem by a family of Bolza problems and applying Ekeland's principle.

In the present work, the maximum principle is derived to a nonconvex F coming from a control system differentiable with respect to x. It is formulated in the usual Jacobian form (adjoint equations and the maximum condition). Recall that all these different formulations are not equivalent.

In this paper, we shall use a direct proof based on our variational inclusion and duality arguments.

In contrast to the work of previous authors (except [20]), we do not suppose any hypotheses about the convexity of F which are basic for the earlier results.

This paper is organized as follows: in section 1, we recall some notions of set-valued analysis we are dealing with. In section 2, we study variational inclusion in the space of absolutely continuous functions, under inequality constraints or constraints given by a closed smooth set. In section 3, we present an analogous result in the space of continuous functions $C[0,T]$, and in the cases of both smooth and nonsmooth constraints. Section 4 provides an application to Mayer's problem of optimal control.

1. Preliminaries and notation. Denote by

- $C([0,T];\mathbb{R}^n)$ the space of continuous maps x from $[0,T]$ into $\mathbb{R}^n$, with the norm $\| x \|_C = \sup_{t\in[0,T]} \| x(t) \|$,
- $W^{1,1}([0,T];\mathbb{R}^n)$ the space of absolutely continuous functions x from $[0,T]$ into $\mathbb{R}^n$ with the norm $\| x \|_{W^{1,1}[0,T]} = \| x(0) \| + \int_0^T \| x'(t) \| \, dt$, and
- $L^1([0,T];\mathbb{R}^n)$ the space of Lebesgue-integrable functions x from $[0,T]$ into $\mathbb{R}^n$ with the norm $\| x \|_{L^1} = \int_0^T \| x(t) \| \, dt$.

Consider the differential inclusion $x'(t) \in F(t,x(t))$, $x(0) = x_0$.

A function $x \in W^{1,1}([0,T];\mathbb{R}^n)$ is called a solution if it satisfies (1) almost everywhere in $[0,T]$.

We recall that $\mathcal{S}_{[0,T]}(x_0)$ denotes the set of solutions starting at x_0 and defined on $[0,T]$, and $\mathcal{S}^K_{[0,T]}(x_0)$ denotes the set of all $x \in \mathcal{S}_{[0,T]}(x_0)$ verifying the state constraints $x(t) \in K$ for all $t \in [0,T]$.

Consider next the convexified (relaxed) differential inclusion

$$\begin{cases} x'(t) \in \overline{co} F(t,x(t)) \text{ a.e. in } [0,T], \\ x(0) = x_0, \end{cases} \tag{3}$$

and denote the set of its solutions by $\mathcal{S}^{rel}_{[0,T]}(x_0)$.

We recall that $\overline{co}F$ denotes the closed convex hull of F.

We shall denote by $\mathcal{S}^{rel,K}_{[0,T]}(x_0)$ the set of all solutions to (3) such that $x(t) \in K$ for all $t \in [0,T]$.

Throughout the paper, $\mathcal{B}$ will denote the closed unit ball.

Let $\mathrm{dist}(\cdot, K) : \mathbb{R}^n \to R$ be the Euclidean distance function from a set $K \subset \mathbb{R}^n$:

$$\mathrm{dist}(x,K) = \inf\{||x-y|| : y \in K\}.$$

We recall the following definitions.

DEFINITION 1.1. *Let $K \subset \mathbb{R}^n$ and $x \in \overline{K}$ (the closure of K).*

- *Contingent cone to K at x:*

$$T_K(x) = \left\{v \in \mathbb{R}^n : \exists h_i \to 0^+ \ \exists v_i \to v \text{ such that } x + h_i v_i \in K\right\}$$

or, equivalently,

$$T_K(x) = \left\{v \in \mathbb{R}^n : \liminf_{h \to 0^+} \mathrm{dist}\left(v, \frac{K-x}{h}\right) = 0\right\}.$$

- *Intermediate (or adjacent) cone to K at x:*

$$I_K(x) = \left\{v \in \mathbb{R}^n : \forall h_i \to 0^+ \ \exists v_i \to v \text{ such that } x + h_i v_i \in K\right\}$$

or, equivalently,

$$I_K(x) = \left\{v \in \mathbb{R}^n : \lim_{h \to 0^+} \mathrm{dist}\left(v, \frac{K-x}{h}\right) = 0\right\}.$$

- *Clarke's tangent cone to K at x:*

$$C_K(x) = \left\{v \in \mathbb{R}^n : \forall h_i \to 0^+ \ \forall y_i \to x \text{ in } K, \exists v_i \to v \text{ such that } y_i + h_i v_i \in K\right\}$$

or, equivalently,

$$C_K(x) = \left\{v \in \mathbb{R}^n : \lim_{\substack{h \to 0^+ \\ y \to x \text{ in } K}} \mathrm{dist}\left(v, \frac{K-y}{h}\right) = 0\right\}.$$

DEFINITION 1.2. *Consider a set-valued map F, from a finite dimensional space X to a finite dimensional space Y, Lipschitzian around x, and let $(x,y) \in GraphF$. The adjacent derivative of F at (x,y) is the set-valued map $dF(x,y)$ from X into subsets of Y defined by*

$$v \in dF(x,y) \cdot u \Longleftrightarrow \lim_{h \to 0^+} \mathrm{dist}\left(v, \frac{F(x+hu)-y}{h}\right) = 0;$$

see Aubin–Frankowska [2] for properties of set-valued derivatives.

In the next sections, we shall impose the following hypotheses on F.

$H.1$

(i) *For all $(t,x) \in [0,T] \times \mathbb{R}^n$, $F(t,x)$ is a nonempty closed set and $F(\cdot,x)$ is measurable for each $x \in \mathbb{R}^n$.*

(ii) *For some $c > 0$, for all $(t,x) \in [0,T] \times \mathbb{R}^n$, $F(t,x) \subset c(||x|| + 1)\mathcal{B}$.*

(iii) *$\exists\, k_F \in L^1([0,T]; \mathbb{R}_+)$ such that for all $x,\ x' \in \mathbb{R}^n$, and $t \in [0,T]$, $F(t,x) \subset F(t,x') + k_F(t)||x - x'||\,\mathcal{B}$.*

We recall that the celebrated Filippov's Theorem (see [8]) provides an estimate of the distance from an absolutely continuous function y to the set $\mathcal{S}_{[t_0,T]}(x_0)$ under assumptions $H.1$. It also allows us to prove the following Filippov–Wazewski relaxation theorem from [12].

THEOREM 1.3 (relaxation theorem). *Consider a set valued map $F : [t_0, T]\times\mathbb{R}^n \rightsquigarrow \mathbb{R}^n$ satisfying $H.1$, and let y be a solution to the relaxed inclusion* (3). *Then for every $\mu > 0$ there exists a solution $x \in \mathcal{S}_{[t_0,T]}(x_0)$ such that $\|x - y\|_C \leq \mu$.*

2. Variational inclusion in $W^{1,1}([0,T];\mathbb{R}^n)$.

2.1. Inequality constraints. Let $g : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be a $C^{1,1}$ function. ($C^{1,1}$ denotes the class of C^1 functions with locally Lipschitz continuous gradients).

Set

$$K = \{x \in \mathbb{R}^n \ : \ g(x) \leq 0\}.$$

Fix $\overline{x}_0 \in C_0$ and $\overline{x} \in \mathcal{S}^K_{[0,T]}(\overline{x}_0)$.

We consider the linearized system along $\overline{x}$, which is formulated as the differential inclusion

$$(P) \quad \begin{cases} w'(t) \in dF(t, \overline{x}(t), \overline{x}'(t)) \cdot w(t) \text{ a.e. in } [0,T], \\ w(0) = w_0 \in I_{C_0}(\overline{x}_0), \\ w(t) \in I_K(\overline{x}(t)), \quad \forall t \text{ in } [0,T], \end{cases}$$

where $dF(t, \overline{x}(t), \overline{x}'(t))$ denotes the partial derivative of $F(t,\cdot)$ at $(\overline{x}(t), \overline{x}'(t))$.

We introduce the constraint qualification (of Mangasarian–Fromowitz type)

$$H.2 \left| \begin{array}{l} \textit{For all } R > 0, \textit{ there exists } \alpha > 0 \textit{ such that} \\ \min_{v\in F(t,x)} \nabla g(x) \cdot v < -\alpha \textit{ for } x \in \mathcal{B}(0,R) \cap \partial K \textit{ and a.e. } t \in [0,T], \end{array} \right.$$

where ∂K denotes the boundary of K.

This assumption implies, in particular, that for all t satisfying $g(\overline{x}(t)) = 0$, we have

$$I_K(\overline{x}(t)) = \{v \ : \ \langle \nabla g(\overline{x}(t)), v\rangle \leq 0\} \tag{4}$$

and that $\mathrm{Int}\{x : g(x) \leq 0\} \neq \emptyset$.

Our main theorem is the following.

THEOREM 2.1. *Fix $\overline{x}_0 \in C_0$ and $\overline{x} \in \mathcal{S}^K_{[0,T]}(\overline{x}_0)$. Assume that $H.1$ and $H.2$ are verified, and let $w(\cdot)$ be a solution to (P). Then, for all $h_i \to 0^+$ and every sequence $\{w_i\}$ such that $\lim_{i\to+\infty} w_i = w_0$ and $\overline{x}_0 + h_i w_i \in K$, there exist solutions $\overline{x}_i \in \mathcal{S}^K_{[0,T]}(\overline{x}_0 + h_i w_i)$ such that*

$$\frac{\overline{x}_i(\cdot) - \overline{x}(\cdot)}{h_i} \longrightarrow w(\cdot) \textit{ in } W^{1,1}([0,T];\mathbb{R}^n) \textit{ when } i \to +\infty.$$

Remark. From the proof provided below, it follows that the assumption on g may be weakened. Namely, it is enough to assume that $g \in C^{1,1}$ on a neighborhood of ∂K and g is continuous on $\mathbb{R}^n$.

Proof of Theorem 2.1.

Step 1. We prove first that, for all solutions $w(\cdot)$ of (P), there exists a sequence x_i such that $x_i \in \mathcal{S}_{[0,T]}(\overline{x}_0 + h_i w_i)$ and $\frac{x_i(\cdot)-\overline{x}(\cdot)}{h_i} \longrightarrow w(\cdot)$ in $W^{1,1}$.

To this aim, we mimic the proof from [2, p. 405], by making a few refinements (namely, inequality (8) below) that would allow us to prove our theorem.

Let $w(\cdot)$ be a solution of (P). Let $w_i \to w_0$ when $i \to +\infty$.

By the very definition of the adjacent derivative, for almost all $t \in [0,T]$,

$$\lim_{i\to\infty} \operatorname{dist}\left(w'(t), \frac{F(t,\overline{x}(t)+h_i w(t)) - \overline{x}'(t)}{h_i}\right) = 0. \tag{5}$$

Moreover, since $\overline{x}'(t) \in F(t,\overline{x}(t))$ almost everywhere in [0,T] and by $H.1$, for all sufficiently large i and for almost every $t \in [0,T]$

$$\operatorname{dist}(\overline{x}'(t)+h_i w'(t), F(t,\overline{x}(t)+h_i w(t))) \le h_i(||w'(t)|| + k_F(t)||w(t)||). \tag{6}$$

Set $y_i(t) = h_i w_i + \overline{x}(t) + h_i(w(t) - w_0)$. Then $y_i'(t) = \overline{x}'(t) + h_i w'(t)$, and $y_i(0) = \overline{x}(0) + h_i w_i$.

We want to estimate $\operatorname{dist}(y_i'(t), F(t,y_i(t)))$. Notice that, because of the Lipschitzianity of F,

$$\begin{aligned}&\operatorname{dist}(y_i'(t), F(t,y_i(t))) \le \operatorname{dist}(y_i'(t), F(t,\overline{x}(t)+h_i w(t)))\\ &+k_F(t)h_i||w_i - w_0||.\end{aligned} \tag{7}$$

By (6), it follows that

$$\operatorname{dist}(y_i'(t), F(t,y_i(t))) \le h_i(||w'(t)|| + k_F(t)||w(t)||) + k_F(t)h_i||w_i - w_0||. \tag{8}$$

Then (5), (7), and the Lebesgue dominated convergence theorem yield

$$\int_0^T \operatorname{dist}(y_i'(t), F(t,y_i(t)))dt = \alpha(h_i),$$

$$\text{where } \lim_{h_i\to 0^+} \frac{\alpha(h_i)}{h_i} = 0.$$

Set

$$\begin{cases} \gamma_i(t) = \operatorname{dist}(y_i'(t), F(t,y_i(t))), \\ \eta_i(t) = \int_0^t \gamma_i(s)\exp(\int_s^t k(\nu)d\nu)ds. \end{cases} \tag{9}$$

By Filippov's existence theorem [8] applied to arcs y_i, there exist solutions $x_i \in \mathcal{S}_{[0,T]}(\overline{x}(0)+h_i w_i)$ such that

$$||x_i(t) - y_i(t)|| \le \eta_i(t),$$

$$||x_i'(t) - y_i'(t)|| \le k_F(t)\eta_i(t) + \gamma_i(t) \text{ a.e. in } [0,T].$$

Hence

$$\frac{x_i(\cdot) - y_i(\cdot)}{h_i} \longrightarrow 0 \text{ in } W^{1,1}.$$

Since $\frac{x_i(0)-\overline{x}(0)}{h_i} = w_i \longrightarrow w_0 = w(0)$, this implies that

$$\lim_{i\to+\infty} \frac{x_i(\cdot) - \overline{x}(\cdot)}{h_i} = w(\cdot) \text{ in } W^{1,1} \text{ on } [0,T]. \tag{10}$$

We can easily see that there exists a constant $M \geq 0$ such that

$$||x_i' - \overline{x}' - h_i w'||_{L^1} \leq M \cdot \alpha(h_i)$$

and

$$||x_i(t) - \overline{x}(t) \ - \ h_i w(t)|| \ \leq \ h_i||w_i - w_0|| + M \cdot \alpha(h_i). \tag{11}$$

The constructed solutions $x_i(\cdot)$ do not necessarily remain in K.

Step 2. Let us introduce the following notation:

$$g^+(x) \ = \ \max\left(g(x) \ , \ 0\right).$$

We shall prove the existence of solutions $\overline{x}_i \in \mathcal{S}^K_{[0,T]}(\overline{x}_0 + h_i w_i)$ satisfying for some $L > 0$ and all $i \geq 1$

$$||\overline{x}_i - x_i||_{W^{1,1}([0,T])} \leq \ L \cdot \max_{t\in[0,T]} \ g^+(x_i(t)). \tag{12}$$

We recall the neighboring feasible trajectories theorem from [14].

THEOREM 2.2 (Frankowska–Vinter). *Assume that H.*1 *and H.*2 *are satisfied. Let* $\widehat{x}_0 \in K$. *Then there exists a constant L, which depends on* $|\widehat{x}_0|$, *with the following property: given any solution* $\widehat{x} \ \in \ \mathcal{S}_{[0,T]}(\widehat{x}_0)$, *a solution* $x \ \in \mathcal{S}^K_{[0,T]}(\widehat{x}_0)$ *can be found such that* $||x - \widehat{x}||_{W^{1,1}([0,T])} \leq \ L \cdot \max_{t\in[0,T]} g^+(\widehat{x}(t))$.

We apply this theorem by taking $\widehat{x} = x_i$ and $\widehat{x}_0 = \overline{x}_0 + h_i w_i$. We get then the existence of $\overline{x}_i(\cdot)$ as claimed.

Step 3. Let us show that the quotients $\frac{\overline{x}_i(.) - \overline{x}(.)}{h_i}$ converge to $w(\cdot)$ in $W^{1,1}$ when $h_i \to 0^+$.

Because of (10), it remains to prove that

$$\frac{\overline{x}_i(\cdot) - x_i(\cdot)}{h_i} \longrightarrow 0 \text{ in } W^{1,1}.$$

For all $t \in [0,T]$, we have to estimate $g^+(x_i(t))$.

Fix any $\varepsilon > 0$, and set

$$Q_\varepsilon = \{s \in [0,T] : \langle \nabla g(\overline{x}(s)), w(s)\rangle \ < \varepsilon\}.$$

Let Q^C_ε denote its complement.

Observe that for all $s \in [0,T]$ such that $g(\overline{x}(s)) = 0$, by (4) we have $s \in Q_\varepsilon$.

- *Case* A. If $t \in Q^C_\varepsilon$, since $g(\overline{x}(\cdot))$ is continuous, we deduce, using (4), that there exists $\delta > 0$ such that

$$\forall \ s \in Q^C_\varepsilon \ , \ g(\overline{x}(s)) \leq -\delta < 0.$$

For i large enough, since Q^C_ε is compact and since $x_i \to \overline{x}$ uniformly, we have

$$g(x_i(s)) \leq -\frac{\delta}{2} < 0.$$

Then we obtain in this case

$$\frac{1}{h_i} \sup_{t\in Q^C_\varepsilon} g^+(x_i(t)) = 0.$$

- *Case* B. If $t \in Q_\varepsilon$, we have

$$g(x_i(t)) \;=\; g(\overline{x}(t)) \;+\; \langle \nabla g(\overline{x}(t)) \,,\; x_i(t) - \overline{x}(t) \rangle \;+\; \theta(|x_i(t) - \overline{x}(t)|),$$

where $\frac{\theta(r)}{r} \to 0$ as $r \to 0$:

$$\begin{aligned} g(x_i(t)) \;=\; g(\overline{x}(t)) \;+\; \langle \nabla g(\overline{x}(t)) \,,\; x_i(t) - \overline{x}(t) \;-\; h_i w(t) \rangle \\ + h_i \langle \nabla g(\overline{x}(t)) \,,\; w(t) \rangle \;+\; \theta(|x_i(t) - \overline{x}(t)|). \end{aligned}$$

Then

$$g(x_i(t)) \leq \; \langle \nabla g(\overline{x}(t)) \,,\; x_i(t) - \overline{x}(t) \;-\; h_i w(t) \rangle \;+\; h_i \varepsilon \;+\; \theta(|x_i(t) - \overline{x}(t)|).$$

We set $N = \sup_{t \in [0,T]} ||\nabla g(\overline{x}(t))||$. Then

$$g(x_i(t)) \leq \; N \cdot [h_i ||w_i - w_0|| \;+\; M \cdot \; \alpha(h_i)] \;+\; h_i \varepsilon \;+\; o(h_i) \;\; \forall t \in Q_\varepsilon.$$

Thus

$$\frac{1}{h_i} g^+(x_i(t)) \;\leq\; N \cdot ||w_i - w_0|| \;+\; N \cdot M \; \frac{\alpha(h_i)}{h_i} \;+\; \varepsilon \;+\; \frac{o(h_i)}{h_i}.$$

Since

$$\lim_{h_i \to 0^+} \frac{\alpha(h_i)}{h_i} \;=\; 0, \;\; \lim_{h_i \to 0^+} \frac{o(h_i)}{h_i} \;=\; 0, \text{ and } w_i \to w_0,$$

we obtain in this case

$$\limsup_{h_i \to 0^+} \left(\sup_{t \in Q_\varepsilon} \frac{1}{h_i} g^+(x_i(t)) \right) \leq \varepsilon.$$

From the above two cases (since ε is arbitrary) and (12), we conclude that

$$\lim_{i \to +\infty} \frac{\overline{x}_i(\cdot) - x_i(\cdot)}{h_i} \;=\; 0 \text{ in } W^{1,1} \text{ on } [0,T],$$

which ends the proof. □

2.2. Constraints given by a closed smooth set. In this section, K is a closed set whose boundary is smooth in the sense explained below.

We consider the same variational inclusion (P), where $\overline{x}$, $\overline{x}_0$ have the same meaning as before.

The constraint set verifies the following "smooth boundary" condition.

Let the signed distance $h : \mathbb{R}^n \mapsto \mathbb{R}$ be defined as

$$h(x) = \begin{cases} -\text{dist}(x, \partial K) & \text{if } x \in K, \\ \text{dist}(x, \partial K) & \text{if } x \notin K \end{cases}$$

and be such that the following hold.

H.3

(i) *There exists $\eta > 0$ such that the function h is of class $C^{1,1}$ on a η-neighborhood of ∂K; i.e., it is differentiable with a locally Lipschitz continuous gradient on $\mathcal{B}(\partial K, \eta)$.*

(ii) *There exists $\varepsilon > 0$ such that, for every $x \in \partial K$ and every t, one has $\min_{v \in F(t,x)} \langle\, v, \mathcal{N}(x) \,\rangle \leq -\varepsilon$, where $\mathcal{N}(x)$ denotes the outer normal to K at x.*

Remarks.

(a) Recall that $\mathcal{N}(x) = \nabla h(x)$.

(b) If F is continuous, it is enough to assume that, for all $(t,x) \in [0,T] \times \partial K$, $\min_{v \in F(t,x)} \langle v, \mathcal{N}(x) \rangle < 0$ instead of $H.3$ (ii) to get the same conclusions as below.

We can state the following theorem.

THEOREM 2.3. *Assume that $H.1$ and $H.3$ are verified, and let $\overline{x} \in \mathcal{S}^K_{[0,T]}(\overline{x}_0)$.*

Fix $w_0 \in I_{C_0}(\overline{x}_0)$ and a solution $w(\cdot)$ to (P). Let $h_i \longrightarrow 0^+$ and $w_i \longrightarrow w_0$ be such that $\overline{x}_0 + h_i w_i \in K$.

Then there exist

$$\overline{x}_i \in \mathcal{S}^K_{[0,T]}(\overline{x}_0 + h_i w_i)$$

$$\text{such that } \frac{\overline{x}_i(\cdot) - \overline{x}(\cdot)}{h_i} \longrightarrow w(\cdot) \text{ in } W^{1,1} \text{ on } [0,T].$$

Proof of Theorem 2.3. Define the constraint set as $K = \{x : h(x) \leq 0\}$, and we have the case of the previous subsection, but we have only $h \in C^{1,1}$ on a neighborhood of ∂K.

Observe next that all estimates of the proof of Theorem 2.1 were used only on a neighborhood of $\{x : g(x) = 0\}$. When $x \in \operatorname{Int} K$, we have the case in which $h(x) < 0$, and so again arguments from the proof of Theorem 2.1 do apply.

3. Variational inclusion in $C([0,T];\mathbb{R}^n)$. We consider the solution map $\mathcal{S}_{[0,T]}$ as the set-valued map from $\mathbb{R}^n$ to $C([0,T];\mathbb{R}^n)$.

Fix $\overline{x}_0 \in C_0$, and let $\overline{x} \in \mathcal{S}^K_{[0,T]}(\overline{x}_0)$, $w_0 \in I_{C_0}(\overline{x}_0)$.

Consider the following linearized inclusion:

$$(L^{rel}) \begin{cases} w'(t) \in d\overline{co}F(t, \overline{x}(t), \overline{x}'(t)) \cdot w(t) \text{ a.e. in } [0,T], \\ w(0) = w_0. \end{cases}$$

In this section, we denote by $\overline{\mathcal{S}}^{lin}_{[0,T]}(w_0)$ the closure in $C([0,T];\mathbb{R}^n)$ of all solutions to (L^{rel}).

We suppose that assumptions $H.1$ are verified.

Observe that, if F satisfies $H.1$, then so does the set-valued map $(t,x) \rightsquigarrow \overline{co}F(t,x)$.

In this section, we establish results which are analogous to those proved in the previous section in cases of inequality and smooth and nonsmooth constraints.

3.1. Inequality constraints. Let $g \in C^{1,1}(\mathbb{R}^n, \mathbb{R}^m)$ and $K = \{x : g(x) \leq 0\}$.

THEOREM 3.1. *Let $w \in \overline{\mathcal{S}}^{lin}_{[0,T]}(w_0)$ be such that $w(t) \in I_K(\overline{x}(t))$ for all $t \in [0,T]$. If $H.2$ holds true, then, for all $h_i \to 0^+$ and $w_i \longrightarrow w_0$ satisfying $\overline{x}_0 + h_i w_i \in K$, there exist $\overline{x}_i \in \mathcal{S}^K_{[0,T]}(\overline{x}_0 + h_i w_i)$ such that*

$$\frac{\overline{x}_i(\cdot) - \overline{x}(\cdot)}{h_i} \longrightarrow w(\cdot) \text{ in } C([0,T];\mathbb{R}^n) \text{ when } i \to +\infty.$$

Proof of Theorem 3.1.

Step 1. Let $w(\cdot)$, h_i, w_i be as above. Consider solutions $\overline{w}_j(\cdot)$ to (L^{rel}) converging to $w(\cdot)$ in $C([0,T];\mathbb{R}^n)$.

Applying the same arguments as in step 1 of the proof of Theorem 2.1, and replacing F by $\overline{co}F$, we deduce the existence of $x_{j,i} \in \mathcal{S}^{rel}_{[0,T]}(\overline{x}_0 + h_i w_i)$ such that

$\frac{x_{j,i}(\cdot)-\overline{x}(\cdot)}{h_i}$ converge uniformly to $\overline{w}_j(\cdot)$ and satisfying the following inequality for all t:

$$||x_{j,i}(t) - \overline{x}(t) - h_i\overline{w}_j(t)|| \leq h_i||w_i - w_0|| + M \cdot \alpha_j(h_i) \tag{13}$$

with M independent from i, j and $\lim_{i\to\infty} \frac{\alpha_j(h_i)}{h_i} = 0$.

By the relaxation theorem, Theorem 1.3, we may also assume that $x_{j,i} \in \mathcal{S}_{[0,T]}(\overline{x}_0 + h_i w_i)$.

We apply the Frankowska–Vinter theorem, Theorem 2.2, by taking $\widehat{x} = x_{j,i}$, $\widehat{x}_0 = \overline{x}_0 + h_i w_i$.

We get then the existence of $y_{j,i}(\cdot) \in \mathcal{S}^K_{[0,T]}(x_0 + h_i w_i)$ satisfying for some $L > 0$ and all $i \geq 1$

$$||y_{j,i} - x_{j,i}||_{W^{1,1}([0,T])} \leq \ L \ \cdot \max_{t\in[0,T]} g^+(x_{j,i}(t)).$$

Step 2. Let us show that $\exists L_1 > 0$, $i(j) \geq 1$, and $\varepsilon_j \to 0$ such that, for all $i \geq i(j)$, $\left\| \frac{y_{j,i}(\cdot)-\overline{x}(\cdot)}{h_i} - \overline{w}_j(\cdot) \right\|_C \leq L_1 \cdot \varepsilon_j$.

We observe that

$$\begin{aligned} &\|y_{j,i}(\cdot) - \overline{x}(\cdot) - h_i\overline{w}_j(\cdot)\|_C \leq L \ \cdot \sup_{t\in[0,T]} g^+(x_{j,i}(t)) \\ &+ \sup_{t\in[0,T]} \|x_{j,i}(t) - \overline{x}(t) - h_i\overline{w}_j(t)\|. \end{aligned} \tag{14}$$

Set for all $r \geq 1$

$$Q_{r,j} = \left\{ s \in [0,T] : \langle \nabla g(\overline{x}(s)) \, , \overline{w}_j(s) \rangle < \frac{1}{r} \right\}.$$

Let $Q^C_{r,j}$ denote its complement.

Since $\overline{w}_j(\cdot) \to w(\cdot)$, $\exists j_0$ such that for all $j \geq j_0$, for all t satisfying $\overline{x}(t) \in \partial K$, we have $t \in Q_{r,j}$.

We mimic step 3 of the proof of Theorem 2.1 by considering two cases.

- *Case* A. If $t \in Q^C_{r,j}$, then $g(\overline{x}(t)) < 0$.
  Then with the same reasoning as step 3 of the proof of Theorem 2.1, for i large enough, we have $g(x_{j,i}(t)) \leq -\frac{\delta}{2} < 0$ and $\frac{1}{h_i} \sup_{t\in Q^C_{r,j}} g^+(x_{j,i}(t)) = 0$.
- *Case* B. If $t \in Q_{r,j}$, we have

  $$\begin{aligned} g(x_{j,i}(t)) \ &= \ g(\overline{x}(t)) \ + \langle \nabla g(\overline{x}(t)) \, , \ x_{j,i}(t) - \overline{x}(t) \ - \ h_i\overline{w}_j(t) \rangle \\ &+ h_i \langle \nabla g(\overline{x}(t)) \, , \ \overline{w}_j(t) \rangle + \theta(|x_{j,i}(t) - \overline{x}(t)|), \end{aligned}$$

  where $\frac{\theta(s)}{s} \to 0$ as $s \to 0$.
  Setting $N = \sup_{t\in[0,T]} ||\nabla g(\overline{x}(t))||$ and using (13), we get, for all $t \in Q_{r,j}$,

  $$g(x_{j,i}(t)) \leq \ N \cdot [h_i||w_i - w_0|| \ + \ M \cdot \ \alpha_j(h_i)] \ + \ h_i \frac{1}{r} \ + \ o_j(h_i).$$

  Thus

  $$\frac{1}{h_i} \sup_{t\in Q_{r,j}} g^+(x_{j,i}(t)) \leq \ N \cdot ||w_i - w_0|| \ + \ N \cdot M \cdot \ \frac{\alpha_j(h_i)}{h_i} \ + \ \frac{1}{r} \ + \ \frac{o_j(h_i)}{h_i}.$$

From the above two cases, and since $\lim_{i\to\infty}\frac{\alpha_j(h_i)}{h_i}=0$, $\lim_{i\to\infty}\frac{o_j(h_i)}{h_i}=0$, and $w_i\to w_0$, and by (14) there exists $L_1\geq 1$ independent from i, j such that for every j we can find $i(j)$ satisfying for all $i\geq i(j)$

$$\left\|\frac{y_{j,i}(\cdot)-\overline{x}(\cdot)}{h_i}-\overline{w}_j(\cdot)\right\|\leq L_1\cdot\frac{3}{r}.$$

Since r is arbitrary, by the diagonalization method we can find sequences $\overline{x}_k\in\{y_{j,i}: i\geq 1,\ j\geq 1\}$ such that $\frac{\overline{x}_k(\cdot)-\overline{x}(\cdot)}{h_k}$ converge to $w(\cdot)$ in $C([0,T];\mathbb{R}^n)$. The proof is complete.

3.2. Smooth constraints. We assume here that the boundary of K is smooth; i.e., assumptions $H.3$ are verified.

THEOREM 3.2. *Let $w\in\overline{\mathcal{S}}^{lin}_{[0,T]}(w_0)$ be such that $w(t)\in I_K(\overline{x}(t))$ for all $t\in[0,T]$. If $H.1$ and $H.3$ hold true, then for all $h_i\to 0^+$ and $w_i\longrightarrow w_0$ satisfying $\overline{x}_0+h_iw_i\in K$, there exist $\overline{x}_i\in\mathcal{S}^K_{[0,T]}(\overline{x}_0+h_iw_i)$ such that $\frac{\overline{x}_i(.)-\overline{x}(.)}{h_i}$ converge to $w(\cdot)$ in $C([0,T];\mathbb{R}^n)$ when $i\to+\infty$.*

Proof of Theorem 3.2. Applying the same arguments as in step 1 of the proof of Theorem 3.1, we deduce the existence of $x_{j,i}(\cdot)\in\mathcal{S}_{[0,T]}(\overline{x}_0+h_iw_i)$ satisfying for all t

$$||x_{j,i}(t)-\overline{x}(t)-h_i\overline{w}_j(t)||\leq h_i||w_i-w_0||+M\cdot\alpha_j(h_i), \tag{15}$$

where $\lim_{i\to\infty}\frac{\alpha_j(h_i)}{h_i}=0$ for all j.

We next estimate $\operatorname{dist}(\overline{x}(t)+h_iw(t),K)$ for all t.

Observe that

$$\operatorname{dist}(\overline{x}(t)+h_iw(t),K)=\begin{cases}0 & \text{if } \overline{x}(t)+h_iw(t)\in K,\\ \operatorname{dist}(\overline{x}(t)+h_iw(t),\partial K) & \text{otherwise.}\end{cases}$$

If i is large enough, $\sup_t(h_i\|w(t)\|)<\eta$.

Let $\varepsilon>0$. Set

$$Q_\varepsilon=\{s\in[0,T]:\langle\nabla h(\overline{x}(s))\,,w(s)\rangle<\varepsilon\},$$

and let Q_ε^C denote its complement.

Observe next that, for all t such that $\overline{x}(t)\in\partial K$, and since $w(t)\in I_K(\overline{x}(t))$, we have $t\in Q_\varepsilon$.

Here again we consider two cases.

- *Case* A. If $t\in Q_\varepsilon^C$, since $\min_{t\in Q_\varepsilon^C}\operatorname{dist}(\overline{x}(t),K^C)>0$, for all i large enough $\min_{t\in Q_\varepsilon^C}\operatorname{dist}(\overline{x}(t)+h_iw(t),K^C)>0$, where K^C denotes the complement of K. So $\sup_{t\in Q_\varepsilon^C}\operatorname{dist}(\overline{x}(t)+h_iw(t),K)=0$.
- *Case* B. If $t\in Q_\varepsilon$, since

  $$h(\overline{x}(t)+h_iw(t))=h(\overline{x}(t))+\langle\nabla h(\overline{x}(t)),h_iw(t)\rangle+o(h_i),$$

  we have for all large i

  $$h(\overline{x}(t)+h_iw(t))\leq 2h_i\varepsilon.$$

From the above two cases, there exists $c>0$ such that for all i large enough

$$\operatorname{dist}(\overline{x}(t)+h_iw(t),K)\leq c\cdot h_i\cdot\varepsilon. \tag{16}$$

Step 1. We prove here that there exist $\overline{y}_{j,i} \in \mathcal{S}^K_{[0,T]}(\overline{x}_0 + h_i w_i)$ satisfying, for some $L \geq 1$, independent from j and i, $i(j) \geq 1$, and $\varepsilon_j \to 0$,

$$\forall i \geq i(j),\ \left\| \frac{\overline{y}_{j,i}(\cdot) - \overline{x}(\cdot)}{h_i} - \overline{w}_j(\cdot) \right\| \leq L \cdot \varepsilon_j.$$

Under assumptions $H.1$ and $H.3$, we have the following result (from [12]).

LEMMA 3.3 (Frankowska–Rampazzo). *There exist $L > 0$ and $\tau > 0$ such that for every $(t_0, x_0) \in [0,T] \times K$, and every trajectory $\overline{x}(\cdot) \in \mathcal{S}_{[t_0,t_0+\tau]}(x_0)$, there is a trajectory $x(\cdot) \in \mathcal{S}^K_{[t_0,t_0+\tau]}(x_0)$ such that*

$$||x - \overline{x}||_{C[t_0,t_0+\tau]} \leq L \sup_{t \in [t_0,t_0+\tau]} \operatorname{dist}(\overline{x}(t), K).$$

Let τ and L_1 be as in the above lemma.

- If $\operatorname{Min}(\tau, T) = T$, then we directly apply Lemma 3.3.
  It says that, for $x_{j,i}(\cdot) \in \mathcal{S}_{[0,T]}(\overline{x}_0 + h_i w_i)$, we have the existence of

  $$\overline{y}_{j,i}(\cdot) \in \mathcal{S}^K_{[0,T]}(\overline{x}_0 + h_i w_i)$$

  such that

  $$||\overline{y}_{j,i} - x_{j,i}||_{C[0,T]} \leq L_1 \cdot \sup_{t \in [0,T]} \operatorname{dist}(x_{j,i}(t), K).$$

  Then, using (15),

  $$\begin{aligned} &\left\| \frac{\overline{y}_{j,i}(\cdot) - \overline{x}(\cdot)}{h_i} - \overline{w}_j(\cdot) \right\| \leq L_1 \cdot \frac{1}{h_i} \sup_{t \in [0,T]} \operatorname{dist}(x_{j,i}(t), K) \\ &+ \|w_i - w_0\| + M \cdot \frac{\alpha_j(h_i)}{h_i}. \end{aligned} \tag{17}$$

  On the other hand, we have

  $$\begin{aligned} &\operatorname{dist}(x_{j,i}(t), K) \leq h_i ||w_i - w_0|| + M \cdot \alpha_j(h_i) \\ &+ \operatorname{dist}(\overline{x}(t) + h_i w(t), K) + h_i ||\overline{w}_j(t) - w(t)||. \end{aligned} \tag{18}$$

  Then, using (16), (17), (18), and the convergence of $\overline{w}_j$ to w, we deduce the existence of L, $i(j) \geq 1$, and $\varepsilon_j \to 0$ such that, for all $i \geq i(j)$,

  $$\left\| \frac{\overline{y}_{j,i}(\cdot) - \overline{x}(\cdot)}{h_i} - \overline{w}_j(\cdot) \right\|_{C([0,\tau])} \leq L \cdot \varepsilon_j. \tag{19}$$

- If $\operatorname{Min}(\tau, T) = \tau < T$, then we apply Lemma 3.3 exactly as above on the time interval $[0, \tau]$ to prove the existence of $y^1_{j,i}(.) \in \mathcal{S}^K_{[0,\tau]}(\overline{x}_0 + h_i w_i)$, L_1, $i(j) \geq 1$, and $\varepsilon_{j,1} \to 0$ as $j \to \infty$ such that, for all $i \geq i(j)$,

  $$\left\| \frac{y^1_{j,i}(\cdot) - \overline{x}(\cdot)}{h_i} - \overline{w}_j(\cdot) \right\|_{C([0,\tau])} \leq L_1 \cdot \varepsilon_{j,1}. \tag{20}$$

  Set $z_{j,i}(\tau) = \frac{y^1_{j,i}(\tau) - \overline{x}(\tau)}{h_i}$. Then $y^1_{j,i}(\tau) = \overline{x}(\tau) + h_i z_{j,i}(\tau) \in K$.

We observe that $\|z_{j,i}(\tau) - \overline{w}_j(\tau)\| \leq L_1 \cdot \varepsilon_{j,1}$.
We set $\overline{z}_{j,i}(t) = \overline{x}(t) + h_i(z_{j,i}(\tau) - \overline{w}_j(\tau)) + h_i\overline{w}_j(t)$ for $t \geq \tau$.
Then $\overline{z}'_{j,i}(t) = \overline{x}'(t) + h_i\overline{w}'_j(t)$ and $\overline{z}_{j,i}(\tau) = \overline{x}(\tau) + h_i z_{j,i}(\tau) = y^1_{j,i}(\tau)$.
We observe that, for all $t \geq \tau$,

$$\begin{aligned} &\frac{1}{h_i}\mathrm{dist}(\overline{z}_{j,i}(t), K) \leq ||z_{j,i}(\tau) - \overline{w}_j(\tau)|| + ||\overline{w}_j(t) - w(t)|| \\ &+\frac{1}{h_i}\mathrm{dist}(\overline{x}(t) + h_i w(t), K). \end{aligned} \tag{21}$$

On the other hand, we do have the same inequality as (8) for all $t \geq \tau$:

$$\mathrm{dist}(\overline{z}'_{j,i}(t), \overline{co}F(t, \overline{z}_{j,i}(t))) \leq h_i(||\overline{w}'_j(t)||+k_F(t)||\overline{w}_j(t)||)+h_i k_F(t)||z_{j,i}(\tau)-\overline{w}_j(\tau)||$$

and

$$\frac{1}{h_i}\mathrm{dist}(\overline{z}'_{j,i}(t), \overline{co}F(t, \overline{z}_{j,i}(t))) \leq ||z_{j,i}(\tau) - \overline{w}_j(\tau)|| + o_t(h_i),$$

where $\frac{o_t(h_i)}{h_i} \to 0$ when $i \to \infty$.
This and the Lebesgue dominated convergence theorem yield

$$\int_\tau^T \mathrm{dist}\left(\overline{z}'_{j,i}(t), \overline{co}F(t, \overline{z}_{j,i}(t))\right) dt = \beta_j(h_i),$$

where $\limsup_{i\to\infty} \frac{\beta_j(h_i)}{h_i} \leq C \cdot \varepsilon_{j,1}$ with C independent from j.
Consider next the time interval $[\tau, 2\tau]$. We apply the Filippov's existence theorem [8] on $[\tau, 2\tau]$ to the arcs $\overline{z}_{j,i}$ and the relaxation theorem, Theorem 1.3.
We have then the existence of $l > 0$ independent from i and j and solutions

$$y_{j,i}(\cdot) \in \mathcal{S}_{[\tau,2\tau]}(\overline{x}(\tau) + h_i z_{j,i}(\tau))$$

such that, for all i large enough,

$$\left\| \frac{y_{j,i}(\cdot) - \overline{z}_{j,i}(\cdot)}{h_i} \right\| \leq l \cdot \varepsilon_{j,1} \text{ on } [\tau, 2\tau]. \tag{22}$$

We now apply Lemma 3.3 on $[\tau, 2\tau]$, considering $y_{j,i}(\cdot) \in \mathcal{S}_{[\tau,2\tau]}(\overline{x}(\tau) + h_i z_{j,i}(\tau))$ to obtain the existence of some L and $y^2_{j,i}(\cdot) \in \mathcal{S}^K_{[\tau,2\tau]}(\overline{x}(\tau) + h_i z_{j,i}(\tau))$ such that

$$||y^2_{j,i} - y_{j,i}||_{C([\tau,2\tau]),\mathbb{R}^n)} \leq L \cdot \sup_{t\in[\tau,2\tau]} \mathrm{dist}(y_{j,i}(t), K).$$

However, since

$$\frac{1}{h_i}\sup_{t\in[\tau,2\tau]} \mathrm{dist}(y_{j,i}(t), K) \leq \frac{1}{h_i}\sup_{t\in[\tau,2\tau]} \|y_{j,i}(t)-\overline{z}_{j,i}(t)\|+\frac{1}{h_i}\sup_{t\in[\tau,2\tau]} \mathrm{dist}(\overline{z}_{j,i}(t), K),$$

using (16), (21), and (22), we obtain the existence of L_2, $i(j)$ and $\varepsilon_{j,2} \to 0^+$ such that

$$\left\| \frac{y^2_{j,i}(\cdot) - \overline{x}(\cdot)}{h_i} - \overline{w}_j(\cdot) \right\|_{C([\tau,2\tau]} \leq L_2 \cdot \varepsilon_{j,2}. \tag{23}$$

– If $\min(2\tau, T) = T$, then concatenate $y^1_{j,i}(\cdot)$ on $[0,\tau]$ and $y^2_{j,i}(\cdot)$ on $[\tau, 2\tau]$ to get the existence of $\overline{y}_{j,i}$ *on* $[0,T]$ such that

$$\overline{y}_{j,i}(\cdot) \in \mathcal{S}^K_{[0,T]}(\overline{x}_0 + h_i w_i).$$

Since $y^1_{j,i}(\tau) = y^2_{j,i}(\tau)$, because of (20) and (23), we get for some L_3 and $\varepsilon_{j,3} \to 0^+$

$$\left\| \frac{\overline{y}_{j,i}(\cdot) - \overline{x}(\cdot)}{h_i} - \overline{w}_j(\cdot) \right\|_{C([0,2\tau])} \leq L_3 \cdot \varepsilon_{j,3} \quad \forall\ i \text{ large enough.}$$

– If $\min(2\tau, T) = 2\tau < T$, we proceed inductively. Assume we have already constructed solutions

$$\overline{y}_{j,i}(.) \in \mathcal{S}^K(\overline{x}_0 + h_i w_i) \text{ on } [0, \min(p\tau, T)] \text{ for some } p \geq 1$$

in such a way that

$$\left\| \frac{\overline{y}_{j,i}(\cdot) - \overline{x}(\cdot)}{h_i} - \overline{w}_j(\cdot) \right\|_{C([0,\min(p\tau,T)];\mathbb{R}^n)} \leq L_p \cdot \varepsilon_{j,p} \quad \forall \text{ large } i,$$

where L_p does not depend on i, j, and $\varepsilon_{j,p} \to 0$ as $j \to \infty$.
If $\min(p\tau, T) = T$, then the proof is complete.
If not, we use the same reasoning as in step 1 on $[p\tau, \min((p+1)\tau, T)]$ to construct $y^p_{j,i}(.) \in \mathcal{S}^K(\overline{y}_{j,i}(p\tau))$ on $[p\tau, \min((p+1)\tau, T)]$ such that $\left\| \frac{\overline{y}^p_{j,i}(\cdot) - \overline{x}(\cdot)}{h_i} - \overline{w}_j(\cdot) \right\| \leq L_{p+1} \cdot \varepsilon_{j,p+1}$, where $\varepsilon_{j,p+1} \to 0^+$ as $j \to \infty$, and we concatenate the two solutions. In this way, we extend the definition of $\overline{y}_{j,i}$ on the whole interval $[0,T]$ in a finite number of steps.

Step 2. We have proved the existence of $\overline{y}_{j,i} \in \mathcal{S}^K_{[0,T]}(\overline{x}_0 + h_i w_i)$, $L > 0$, $i(j)$, and $\varepsilon_j > 0$, satisfying for all $i \geq i(j)$ $\left\| \frac{\overline{y}_{j,i}(\cdot) - \overline{x}(\cdot)}{h_i} - \overline{w}_j(\cdot) \right\| \leq L \cdot \varepsilon_j$ with $\varepsilon_j \to 0^+$. We apply next the diagonalization process to prove the existence of $\overline{x}_i \in \mathcal{S}^K_{[0,T]}(\overline{x}_0 + h_i w_i)$ such that $\frac{\overline{x}_i(\cdot) - \overline{x}(\cdot)}{h_i}$ converge to $w(\cdot)$ in $C([0,T];\mathbb{R}^n)$. The proof is complete. ∎

3.3. Nonsmooth constraints. In this section, we allow the boundary of K to be nonsmooth. We replace hypothesis $H.3$ by the following condition:

$$H.4 \left| \begin{array}{l} \textit{There exists } \eta > 0 \textit{ such that for all } t \in [0,T] \textit{ and } x \in \partial K \\ \exists\ v \in F(t,x) \cap \mathrm{Int}(C_K(x)) \textit{ such that } ||v|| \geq \eta. \end{array} \right.$$

This means that for all $(t,x) \in [0,T] \times \partial K$ there exists at least one velocity $v \in F(t,x)$ pointing strictly in the set K, in the sense that $\exists\ \varepsilon > 0$ such that $y + [0,\varepsilon]\mathcal{B}(v,\varepsilon) \subset K$ for all $y \in \mathcal{B}(x,\varepsilon) \cap K$.

Set for all $t \in [0,T]$

$$\widehat{\mathcal{C}}(t) = \left\{ v \ : \ \lim_{\substack{t' \to t \\ h \to 0^+}} \frac{dist(\overline{x}(t') + hv, K)}{h} = 0 \right\}.$$

Clearly $\widehat{\mathcal{C}}(t) \subset I_K(\overline{x}(t))$.

Set $F(s,x) = F(T,x)$ for all $s \geq T$ and $x \in \mathbb{R}^n$, and assume

$$H_{BV} \left| \begin{array}{l} \textit{There exists } L \geq 0 \textit{ such that for every } d \in [0,T] \\ \textit{and for every Lipschitz map } x(\cdot) \textit{ from } [0,T] \textit{ into } \mathbb{R}^n \\ \int_0^T d_H(F(s,x(s)), F(s+d,x(s)))ds \leq L \cdot d, \end{array} \right.$$

where d_H denotes the Hausdorff distance.

Remark. When $F(\cdot,x)$ is Lipschitz, then the hypothesis H_{BV} is verified.

However, if ∂K is smooth in the sense of subsection 3.2, then, according to [12], this assumption is not needed.

THEOREM 3.4. *Let $w \in \overline{\mathcal{S}}^{lin}_{[0,T]}(w_0)$ be such that $w(t) \in \widehat{\mathcal{C}}(t)$ for all $t \in [0,T]$. If $H.1$, H_{BV}, and $H.4$ hold true, then for all $h_i \to 0^+$ and $w_i \longrightarrow w_0$ satisfying $\overline{x}_0 + h_i w_i \in K$, there exist $\overline{x}_i \in \mathcal{S}^K_{[0,T]}(\overline{x}_0 + h_i w_i)$ such that $\frac{\overline{x}_i(.) - \overline{x}(.)}{h_i}$ converge to $w(\cdot)$ in $C([0,T];\mathbb{R}^n)$ when $i \to +\infty$.*

The proof is very similar to the one of subsection 3.2 above, based on results from [12]. An analogue of Lemma 3.3 can be found in [12] also under assumptions $H.1$, H_{BV}, and $H.4$. The same steps and arguments are applied, and the only difference is the estimation of $\operatorname{dist}(\overline{x}(t) + h_i w(t), K)$.

The assumption $w(t) \in \widehat{\mathcal{C}}(t)$ implies that for all $\varepsilon_1 > 0$ there exist $\delta > 0$ and $i_0 \geq 1$ such that for every $t' \in [0,T] \cap [t-\delta, t+\delta]$ and $i \geq i_0$

$$\operatorname{dist}(\overline{x}(t') + h_i w(t'), K) \leq h_i \, \varepsilon_1. \tag{24}$$

Thus we deduce inequality (16) from (24) by using the compactness of $[0,T]$.

4. A direct proof of the constrained maximum principle. Let K and C_0 be given closed subsets of $\mathbb{R}^n$, and let $\mathcal{Z}$ be a complete separable metric space.

In this section, we study Mayer's problem of optimal control:

Minimize $\varphi(x(T))$ over solutions to

$$\left\{ \begin{array}{l} x'(t) = f(t,x(t),u(t)) \text{ a.e. in } [0,T], \\ u(t) \in U(t), \\ x(0) \in C_0, \\ x(t) \in K \ \forall \, t \in [0,T], \end{array} \right. \tag{25}$$

where the following hold:

- $U : [0,T] \rightsquigarrow \mathcal{Z}$, $U(\cdot)$ is measurable, and $U(t)$ is nonempty and compact.
- $f : [0,T] \times \mathbb{R}^n \times \mathcal{Z} \mapsto \mathbb{R}^n$ is so that $f(t,\cdot,\cdot)$ is continuous, $f(\cdot,x,u)$ is measurable, and $f(t,\cdot,u)$ is differentiable and $k(t)$-Lipschitzian, and for all t, $\sup_{e \in f(t,x,U(t))} ||e|| \leq c(1+||x||)$.
- $\varphi : \mathbb{R}^n \mapsto \mathbb{R}$ is differentiable.

Consider an optimal solution $\overline{x}$, let $\overline{u}$ be a corresponding optimal control, and set $\overline{x}(0) = x_0$. We associate the following linearization along the solution control pair $(\overline{x}, \overline{u})$:

$$(P_a^*) \left\{ \begin{array}{l} w'(t) = \dfrac{\partial f}{\partial x}(t,\overline{x}(t),\overline{u}(t)) \cdot w(t) + v(t), \\ w(0) = w_0 \in K_0, \\ v(t) \in V(t) = T_{\overline{co} f(t,\overline{x}(t),U(t))}(\overline{x}'(t)), \end{array} \right.$$

where K_0 is any closed convex cone contained in $I_{C_0}(x_0)$.

Observe that, for $F(t,x) := f(t,x,U(t))$, we have

$$\frac{\partial f}{\partial x}(t,\overline{x}(t),\overline{u}(t)) \in dF(t,\overline{x}(t),\overline{x}'(t)) \subset d\overline{co}F(t,\overline{x}(t),\overline{x}'(t)),$$

where dF denotes the adjacent derivative with respect to x.

On the other hand, by [2], we have

$$d\overline{co}F(t,\overline{x}(t),\overline{x}'(t)) + T_{\overline{co}f(t,\overline{x}(t),U(t))}(\overline{x}'(t)) \subset d\overline{co}F(t,\overline{x}(t),\overline{x}'(t)).$$

By the last two relations, every solution $w(\cdot)$ to (P_a^*) solves the following differential inclusion:

$$\begin{cases} w'(t) \ \in \ d\overline{co}F(t,\overline{x}(t),\overline{x}'(t)) \cdot w(t), \\ w(0) = w_0 \in K_0. \end{cases}$$

Let $\mathcal{C}(t)$ be closed convex cones satisfying the following assumption:

$H.5$ | *$\mathcal{C}(\cdot) : [0,T] \rightsquigarrow \mathbb{R}^n$ is a lower semicontinuous set-valued map, $\mathcal{C}(t) \subset \widehat{\mathcal{C}}(t)$; there exists a continuous selection $\overline{w}(t) \in \mathcal{C}(t)$ such that for some $\delta > 0$, $B(\overline{w}(t),\delta) \subset \mathcal{C}(t)$ for all t,*

where $\widehat{\mathcal{C}}(t)$ are defined in subsection 3.3.

Remark. If $\mathcal{C}(t) = C_K(\overline{x}(t))$ and Int $C_K(\overline{x}(t)) \neq \emptyset$ for all $t \in [0,T]$, then, using that Int $C_K(\overline{x}(t)) = \{v : \exists \varepsilon > 0$ for all $h \in [0,\varepsilon]$, $y \in \mathcal{B}(\overline{x}(t),\varepsilon) \cap K$ and $z \in \mathcal{B}(v,\varepsilon)$, $y + hz \in K\}$, it is easy to prove that $H.5$ is verified.

We introduce the following notation:

$$Q_1 = \left\{ w \in W^{1,1}([0,T];\mathbb{R}^n) : \ w(\cdot) \text{ solves } (P_a^*) \right\},$$

$$K_2 = \left\{ w \in C([0,T];\mathbb{R}^n) : \ w(t) \in \mathcal{C}(t) \right\}.$$

We notice that K_2 is closed in $C([0,T];\mathbb{R}^n)$ and Int $K_2 \neq \emptyset$.

Let K_1 be the closure in $C([0,T];\mathbb{R}^n)$ of Q_1. We impose the hypothesis $H.2$ in the case of inequality constraints, the hypothesis $H.3$ in the case when the boundary of K is smooth, and $H.4$ together with H_{BV} when the boundary of K is nonsmooth. Now, we can state the maximum principle theorem.

THEOREM 4.1. *If $(\overline{x},\overline{u})$ is optimal, then there exist $\lambda \in \{0,1\}$, a positive Radon measure μ, a measurable function ν, and an absolutely continuous function $p(\cdot)$ not vanishing simultaneously such that*

$$\text{(i)} \quad -p'(t) = \frac{\partial f}{\partial x}(t,\overline{x}(t),\overline{u}(t))^* \cdot \left(p(t) + \int_0^t \nu(s) d\mu(s) \right) \ \text{a.e. on } [0,T],$$

$$\text{(ii)} \quad p(T) + \int_0^T \nu(s) d\mu(s) = \ -\lambda \nabla \varphi(\overline{x}(T)),$$

$$\text{(iii)} \quad p(0) \in K_0^- ,$$

$$\text{(iv)} \quad \nu(t) \in \mathcal{C}(t)^- \ \ \mu \ \textit{a.e. in} \ [0,T],$$

$$\text{(v)} \quad \max_{u \in U(t)} \left\langle p(t) + \int_0^t \nu(s) d\mu(s), f(t, \overline{x}(t), u) \right\rangle$$

$$= \left\langle p(t) + \int_0^t \nu(s) d\mu(s), f(t, \overline{x}(t), \overline{u}(t)) \right\rangle \ \textit{a.e. in} \ [0,T].$$

Remark. If $\mathcal{C}(t) = C_K(\overline{x}(t))$, then $\mathcal{C}(t)^- = N_K(\overline{x}(t))$ (Clarke's normal cone to K at $\overline{x}(t)$).

Proof of Theorem 4.1. We claim the following.

PROPOSITION 4.2. *If $\overline{x}$ is optimal, then, for all $w \in K_1 \cap K_2$,*

$$\langle \nabla \varphi(\overline{x}(T)), w(T) \rangle \geq 0. \tag{26}$$

Indeed, from the previous section, we do have the existence of trajectories $\overline{x}_i$ solving (25) such that $\frac{\overline{x}_i(\cdot) - \overline{x}(\cdot)}{h_i} \longrightarrow w(\cdot)$.

Then

$$\varphi(\overline{x}_i(T)) = \varphi(\overline{x}(T)) + \langle \nabla \varphi(\overline{x}(T)), \overline{x}_i(T) - \overline{x}(T) \rangle + o(h_i).$$

Since $\overline{x}$ is optimal, $\varphi(\overline{x}(T)) \leq \varphi(\overline{x}_i(T))$, and therefore

$$\langle \nabla \varphi(\overline{x}(T)), \overline{x}_i(T) - \overline{x}(T) \rangle + o(h_i) \geq 0.$$

Dividing by h_i and taking the limit when $i \to +\infty$, we obtain the inequality (26).

Define $\gamma : C([0,T]; \mathbb{R}^n) \mapsto \mathbb{R}^n$ by $\gamma(w) = w(T)$. Then the claim (26) is equivalent to, for all $w \in K_1 \cap K_2$,

$$\langle \gamma^* \nabla \varphi(\overline{x}(T)), w \rangle \geq 0, \tag{27}$$

which implies that

$$\gamma^* \nabla \varphi(\overline{x}(T)) \ \in \ (K_1 \cap K_2)^+.$$

We now consider two cases.

Case 1. $0 \in$ Int $(K_2 - K_1)$. Since K_1 and K_2 are nonempty closed convex cones in $C([0,T])$, we infer that

$$(K_1 \cap K_2)^+ = K_1^+ + K_2^+$$

(see, for instance, [4]).

Consequently,

$$\exists \, p_1 \in K_1^+ \ \text{and} \ p_2 \in K_2^+ \ \text{such that} \ \gamma^* \nabla \varphi(\overline{x}(T)) = p_1 + p_2,$$

which yields

$$\langle \gamma^* \nabla \varphi(\overline{x}(T)) - p_2, w \rangle \geq 0 \quad \forall \, w \in K_1,$$

and therefore

$$\langle \nabla\varphi(\overline{x}(T)), w(T)\rangle - \int_0^T w(t)dp_2(t) \geq 0. \tag{28}$$

We need to recall Rockafellar's result; see [21] and [22].

LEMMA 4.3 (Rockafellar). *Let $Q : [0,T] \rightsquigarrow \mathbb{R}^n$ be a lower semicontinuous multifunction such that $Q(t)$ is for every t a nonempty closed convex set. Set $Q = \{w \in C([0,T];\mathbb{R}^n) : w(t) \in Q(t)\}$; then*

$$\begin{aligned} Q^- = \{&\beta \in C([0,T];\mathbb{R}^n)^* \ : \ \exists \text{ a positive measure } \mu \text{ on } [0,T] \text{ and} \\ &\text{a measurable selection } \nu \text{ with } \nu(t) \in Q(t)^- \ \mu \text{ a.e.,} \\ &\text{such that } d\beta(t) = \nu(t)d\mu(t)\}. \end{aligned}$$

Since $-p_2 \in K_2^-$, the above lemma implies that there exist a positive Radon measure μ and a measurable selection ν with $\nu(t) \in \mathcal{C}(t)^-$ such that

$$-dp_2(t) = \nu(t)d\mu(t).$$

Consider next any $w \in Q_1$.

Integrating by parts, we obtain

$$\int_0^T w(t)dp_2(t) = \left\langle w(T), \int_0^T dp_2(t)\right\rangle - \int_0^T w'(t)\left(\int_0^t dp_2(s)\right)dt.$$

Then, replacing this in (28),

$$\left\langle \nabla\varphi(\overline{x}(T)) - \int_0^T dp_2(t), w(T)\right\rangle + \int_0^T w'(t)\left(\int_0^t dp_2(s)\right)dt \geq 0. \tag{29}$$

Let $p(\cdot)$ solve the system

$$\begin{cases} p'(t) = -\dfrac{\partial f}{\partial x}(t,\overline{x}(t),\overline{u}(t))^* \cdot \left(p(t) + \displaystyle\int_0^t \nu(s)d\mu(s)\right), \\ p(T) = -\nabla\varphi(\overline{x}(T)) - \displaystyle\int_0^T \nu(s)d\mu(s). \end{cases} \tag{30}$$

We observe that we have statements (i) and (ii) of Theorem 4.1 with $\lambda = 1$.

Then (29) implies

$$\langle p(T), w(T)\rangle + \int_0^T w'(t)\left(\int_0^t \nu(s)d\mu(s)\right)dt \leq 0. \tag{31}$$

However,

$$\langle p(T), w(T)\rangle = \langle p(0), w(0)\rangle + \int_0^T w'(t)p(t)dt + \int_0^T w(t)p'(t)dt,$$

yielding

$$\begin{aligned} &\langle p(0), w(0)\rangle + \int_0^T w'(t)\left[p(t) + \int_0^t \nu(s)d\mu(s)\right]dt \\ &+ \int_0^T w(t)p'(t)dt \leq 0. \end{aligned} \tag{32}$$

On the other hand, for any integrable selection $v(t) \in V(t)$ and $w(\cdot)$ solving (P_a^*),

$$\begin{aligned}
&\int_0^T w'(t)\left[p(t)+\int_0^t \nu(s)d\mu(s)\right]dt \\
=\ &\int_0^T \left\langle \frac{\partial f}{\partial x}(t,\overline{x}(t),\overline{u}(t))\cdot w(t)+v(t), p(t)+\int_0^t \nu(s)d\mu(s)\right\rangle dt \\
=\ &\int_0^T \left[\left\langle w(t), \frac{\partial f}{\partial x}(t,\overline{x}(t),\overline{u}(t))^*\left(p(t)+\int_0^t \nu(s)d\mu(s)\right)\right\rangle\right. \\
&\left.+\ \left\langle v(t), p(t)+\int_0^t \nu(s)d\mu(s)\right\rangle\right]dt \\
=\ &-\int_0^T w(t)p'(t)dt+\int_0^T \left\langle v(t), p(t)+\int_0^t \nu(s)d\mu(s)\right\rangle dt.
\end{aligned}$$

This and (32) imply

$$\langle w(0), p(0)\rangle + \int_0^T \left\langle v(t), p(t)+\int_0^t \nu(s)d\mu(s)\right\rangle dt \le 0.$$

In particular, taking $v(t) \equiv 0$,

$$\langle w(0), p(0)\rangle \le 0,$$

which means that $p(0) \in K_0^-$.

Taking $w(0) = 0$, we get that, for all integrable selections $v(t) \in V(t)$,

$$\int_0^T \left\langle v(t), p(t)+\int_0^t \nu(s)d\mu(s)\right\rangle dt \le 0,$$

yielding

$$\sup_{v\in V(t)} \left\langle v, p(t)+\int_0^t \nu(s)d\mu(s)\right\rangle \le 0 \text{ a.e.}$$

By [11], for almost every $t \in [0,T]$,

$$\sup_{e\in f(t,\overline{x}(t),U(t))} \left\langle p(t)+\int_0^t \nu(s)d\mu(s), e-\overline{x}'(t)\right\rangle \le 0,$$

implying

$$\begin{aligned}
&\max\nolimits_{u\in U(t)} \left\langle p(t)+\int_0^t \nu(s)d\mu(s), f(t,\overline{x}(t),u)\right\rangle \\
&= \left\langle p(t)+\int_0^t \nu(s)d\mu(s), f(t,\overline{x}(t),\overline{u}(t))\right\rangle \text{ a.e.}
\end{aligned}$$

Case 2. $0 \notin \text{Int}\,(K_2 - K_1)$ or, equivalently, $K_1 \cap (\text{Int}\, K_2) = \emptyset$. Since K_1 and K_2 are closed convex sets in $C([0,T];\mathbb{R}^n)$ and Int $K_2 \neq \emptyset$, they can be separated by a closed hyperplane passing through the origin: i.e., $\exists\, \beta \in (C([0,T];\mathbb{R}^n))^*$, $\beta \neq 0$, such that

$$\langle \beta, a\rangle \le 0 \le \langle \beta, b\rangle \ \forall\, a \in K_2,\ \forall\, b \in K_1. \tag{33}$$

Therefore, $\beta \in K_2^-$. Then, by Lemma 4.3, $\exists$ a positive Radon measure μ and a measurable selection ν with $\nu(t) \in \mathcal{C}(t)^-$ such that $d\beta(t) = \nu(t)d\mu(t)$. On the other hand, (33) implies $\beta \in K_1^+$. Then

$$\forall w \in Q_1\ ,\ \langle \beta, w\rangle \geq 0. \tag{34}$$

However, $\langle \beta, w\rangle = \int_0^T w(t)d\beta(t)$. Integrating by parts, we get

$$\left\langle w(T), \int_0^T d\beta(t)\right\rangle - \int_0^T w'(t)\left(\int_0^t d\beta(s)\right)dt\ \geq 0,$$

and therefore

$$\left\langle w(T), \int_0^T \nu(t)d\mu(t)\right\rangle - \int_0^T w'(t)\left(\int_0^t \nu(s)d\mu(s)\right)dt \geq 0. \tag{35}$$

Let $p(\cdot)$ solve the linear system (30) with the terminal condition $p(T) = -\int_0^T \nu(t)d\mu(t)$.

Thus $p(T) + \int_0^T \nu(t)d\mu(t) = 0$, which is statement (ii) of Theorem 4.1 with $\lambda = 0$. Then, replacing this in (35) and multiplying by -1, we get

$$\langle w(T), p(T)\rangle + \int_0^T w'(t)\left(\int_0^t \nu(s)d\mu(s)\right)dt \leq 0.$$

We have the same inequality as (31). Then statements (iii) and (v) follow in the same way as in Case 1. □

REFERENCES

[1] J.-P. Aubin and A. Cellina, *Differential Inclusions*, Grundlehren Math. Wiss. 264, Springer-Verlag, Berlin, 1984.

[2] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Boston, 1990.

[3] A. V. Arutyunov and S. M. Aseev, *Investigation of the degeneracy phenomenon of the maximum principle for optimal control problems with state constraints*, SIAM J. Control Optim., 35 (1997), pp. 930–952.

[4] J. Borwein, *Weak tangent cones and optimization in a Banach space*, SIAM J. Control Optim., 16 (1978), pp. 512–522.

[5] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[6] F. H. Clarke and P. D. Loewen, *State constraints in optimal control: A case study in proximal normal analysis*, SIAM J. Control Optim., 25 (1987), pp. 1440–1456.

[7] I. Ekeland, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.

[8] A. F. Filippov, *Classical solutions of differential equations with multivalued right-hand side*, SIAM J. Control, 5 (1967), pp. 609–621.

[9] H. Frankowska, *Local controllability and infinitesimal generators of semi-groups of set-valued maps*, SIAM J. Control Optim., 25 (1987), pp. 412–432.

[10] H. Frankowska, *The maximum principle for an optimal solution to a differential inclusion with end points constraints*, SIAM J. Control Optim., 25 (1987), pp. 145–157.

[11] H. Frankowska, *Contingent cones to reachable sets of control systems*, SIAM J. Control Optim., 27 (1989), pp. 170–198.

[12] H. Frankowska and F. Rampazzo, *Filippov's and Filippov-Ważewski's theorems on closed domains*, J. Differential Equations, 161 (2000), pp. 449–478.

[13] H. Frankowska and F. Rampazzo, *Relaxation of control systems under state constraints*, SIAM J. Control Optim., 37 (1999), pp. 1291–1309.

[14] H. Frankowska and R. Vinter, *Existence of neighbouring feasible trajectories: Applications to dynamic programming for state-constrained optimal problems*, J. Optim. Theory Appl., 104 (2000), pp. 21–40.

[15] M. R. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
[16] P. Loewen and R. T. Rockafellar, *The adjoint arc in nonsmooth optimization*, Trans. Amer. Math. Soc., 325 (1991), pp. 39–72.
[17] R. Pallu de la Barriere, *Cours d'automatique théorique*, Dunod, Paris, 1966.
[18] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The mathematical theory of optimal processes*, Translated from Russian by K. N. Trirogoff, L. W. Neustadt, ed., John Wiley, New York, 1962.
[19] F. Rampazzo and R. Vinter, *A theorem on existence of neighbouring trajectories satisfying a state constraint, with applications to optimal control*, IMA J. Math. Control Inform., 16 (1999), pp. 335–351.
[20] F. Rampazzo and R. Vinter, *Degenerate optimal control problems with state constraints*, SIAM, J. Control Optim., 39 (2000), pp. 989–1007.
[21] R. T. Rockafellar, *Integrals which are convex functionals* II, Pacific J. Math., 39 (1971), pp. 439–469.
[22] R. T. Rockafellar, *State constraints in convex control problems of Bolza*, SIAM J. Control, 10 (1972), pp. 691–715.
[23] H. M. Soner, *Optimal control with state-space constraint* I, SIAM J. Control Optim., 24 (1986), pp. 552–561.
[24] R. Vinter and G. Pappas, *A maximum principle for nonsmooth optimal control problems with state constraints*, J. Math. Anal. Appl., 89 (1982), pp. 212–232.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 1, pp. 363–380

GLOBAL UNIQUENESS TESTS AND PERFORMANCE BOUNDS FOR H^∞ OPTIMA*

J. WILLIAM HELTON† AND MARSHALL A. WHITTLESEY‡

Abstract. Optimization of sup norm–type performance functions over the space of H^∞ functions is central to the subject of H^∞ design, that is, design where stability is the key constraint. Problems with large amounts of plant uncertainty are often highly nonconvex and therefore may have many solutions. In this paper, even for highly nonconvex problems, we give a test one can perform, once a local optimum f^* has been computed, to see if it is a global optimum. The uniqueness phenomena we discovered uses H^∞ properties heavily and are considerably stronger than what occurs in other types of general optimization. Also, even when f^* may not be a global optimum we give a way to use it to bound the best performance possible.

Uniqueness results are valuable for assuring an engineer that a local optimum obtained in a computer run is in fact a true global optimum. This can save a practitioner a lot of time and anguish in that it replaces the usual process of initializing an optimization run many times to see if it always goes to the same local optimum; and even after vast numbers of experiments never being sure.

One of the least intuitive properties of SISO (single input, single output) control is that a (local) optimum for a carefully set up H^∞ problem (cf. Theorem 9.4.1 in [J. W. Helton and O. Merino, *Classical Control Using H^∞ Methods: Theory, Optimization, and Design*, SIAM, Philadelphia, 1998], [J. W. Helton and D. E. Marshall, *Indiana Univ. Math. J.*, 39 (1990), pp. 157–184]) even with large amounts of plant uncertainty is unique. Such problems are quite nonconvex, so the fact is surprising. While the result is false in general for MIMO (multiple input, multiple output) control (cf. [J. W. Helton and O. Merino, *Michigan Math. J.*, 41 (1994), pp. 285–287]), in this note we are describing MIMO situations where uniqueness holds.

The setting in this paper is simultaneous (Pareto) optimization of several competing performances $\Gamma_1, \ldots, \Gamma_\ell$ and we obtain uniqueness results for its solutions.

Key words. H^∞ control, frequency response methods, uniqueness, Pareto, optimization, multiple performances, integral quadratic constraint, quantitative feedback theory

AMS subject classifications. 32, 49K

PII. S0363012901389937

1. Introduction.

This paper analyzes a problem in which one optimizes performance functions over the space H_N^∞ of bounded analytic vector-valued functions $f = (f_1, \ldots, f_N)$ defined on the unit circle, $\mathbf{T}$, where each coordinate function f_j belongs to $L^\infty(\mathbf{T})$ and extends to be analytic on the entire unit disk. Let $C(\mathbf{T})$ be the space of continuous complex-valued functions on the circle and let $C^1(\mathbf{T})$ be those elements in $C(\mathbf{T})$ with continuous first derivatives.

1.1. Definition of Pareto optimum.

The performance criteria we optimize are described in terms of nonnegative continuous functions Γ defined on $\mathbf{T} \times \mathbf{C}^N$. We are given positive functions $\Gamma_j(e^{i\theta}, z)$, $j = 1, \ldots, \ell$ for $\ell \leq N$ with $e^{i\theta} \in \mathbf{T}$ and

*Received by the editors May 24, 2001; accepted for publication (in revised form) September 19, 2002; published electronically April 17, 2003. An announcement of some of these results appeared as *Global uniqueness tests for H^∞ optima*, in Proceedings of the 39th IEEE Conference on Decision and Control, Vol. 2, IEEE Press, Piscataway, NJ, pp. 1043–1048.

http://www.siam.org/journals/sicon/42-1/38993.html

†Department of Mathematics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0112 (helton@math.ucsd.edu). This author's work was partially supported by the National Science Foundation, the Office of Naval Research, DARPA, and the Ford Motor Co.

‡Department of Mathematics, California State University, San Marcos, 333 S. Twin Oaks Valley Road, San Marcos, CA 92096 (mwhittle@csusm.edu).

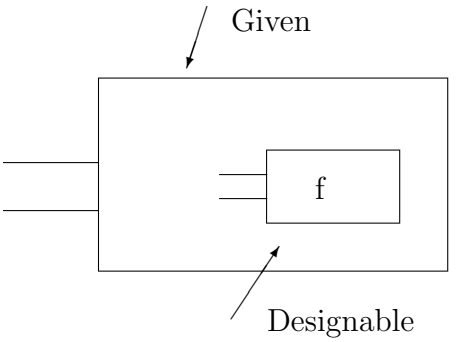


FIG. 1. *For a given plant we want to find the best designable part, represented by* f.

$z \in \mathbf{C}^N$. For function $f \in H_N^\infty$ we define the ℓ performances

$$\gamma_j(f) := \sup_{e^{i\theta} \in \mathbf{T}} \Gamma_j(e^{i\theta}, f(e^{i\theta})), \qquad j = 1, \ldots, \ell.$$

The goals of this paper are best illustrated by the case of two performance functions Γ_1, Γ_2, even though all results hold for ℓ performance functions.

DEFINITION. *A function* $f^* \in H_N^\infty$ *is called a* Pareto optimum *for* Γ_1, Γ_2 *if for each* $f \in H_N^\infty$ *such that* $\gamma_1(f) \le \gamma_1(f^*)$ *and* $\gamma_2(f) \le \gamma_2(f^*)$, *we must have*

$$\gamma_1(f) = \gamma_1(f^*) \quad \textit{and} \quad \gamma_2(f) = \gamma_2(f^*).$$

By the *MultiOPT* problem we shall mean the problem of finding a Pareto optimum. The definition for more Γ_j's is the obvious analogue.

The book of Boyd and Barrat [BB] gives a good discussion of Pareto optimality. The paper [PY95] treats successfully a particular type of frequency domain Pareto optimality, superoptimal H^∞ optimization.

1.2. Engineering motivation. This type of problem is central to frequency domain system design problems where stability is a key constraint. In particular it is important to the area of H^∞-control. The basic physical idea is simple. The following often occurs in a design procedure. We are required to build a system S, but part of the system is given (we are stuck with it) and part of the system is designable (denote its frequency response function by f). Such a system is illustrated in Figure 1. The performance of the system S at frequency ω is a function $\Gamma(\omega, f(i\omega))$ which depends on ω and on our choice of the designable subsystem f. Let us take the convention that large Γ is bad while small Γ is good. Then in a worst case "broadband" design we consider the worst performance over all frequencies

$$\sup_{\omega} \Gamma(\omega, f(i\omega))$$

and try to minimize it over all admissible f. If our main constraint is that the designable subsystem f must be stable, then the design problem becomes the MultiOPT problem with only one Γ, after transforming the right half plane to the unit disk. In this paper we deal with the case where f consists of N designable subsystems $f_1, \ldots, f_N$ and where there are ℓ competing performance criteria $\Gamma_1, \Gamma_2, \ldots, \Gamma_\ell$. Nonconvex performance measures occur in problems with considerable plant uncertainty.

A number of authors, Mayne, Nye, Polak, and Wu [MNPW], Fan, Koninckx, Tits, and Wang [FKTW], Streit [St], Boyd and Barratt [BB], Daleh, Pearson, Balas, Doyle,

Glover, Packard, and Smith [BDGPS], Helton and Merino [HMer98], and Sideris [Si], have theory and computer programs on searching for an optimal f^* with certain kinds of Γ. The main H^∞ optimization problem of quantitative feedback theory (QFT) is essentially the MultiOPT problem. Also integral quadratic constraints (IQCs; see [MR]) address such problems but in a different set of coordinates (behavioral coordinates). Multiple constraints in the frequency domain of a somewhat different flavor are in [KRE96], [PRR97], and [FK93]. There are similar physical problems that they can treat.

1.3. Geometric version of the problem. The MultiOPT problem can be stated geometrically in a way which is physically appealing. The sublevel sets

$$\mathcal{S}^j(\gamma_j) := \{(e^{i\theta}, z) \in \mathbf{T} \times \mathbf{C}^N : \Gamma_j(e^{i\theta}, z) \le \gamma_j\},$$

$$\mathcal{S}^j_\theta(\gamma_j) := \{z \in \mathbf{C}^N : \Gamma_j(e^{i\theta}, z) \le \gamma_j\}$$

of the performance functions Γ_j correspond to values of the frequency response function where the jth performance measure is better (less) than γ_j. For fixed $\vec{\gamma} := (\gamma_1, \ldots, \gamma_\ell)$,

$$\mathcal{S}_\theta(\vec{\gamma}) := \mathcal{S}^1_\theta(\gamma_1) \cap \ldots \cap \mathcal{S}^\ell_\theta(\gamma_\ell) \qquad \forall e^{i\theta} \in \mathbf{T} \tag{1}$$

is the set of values simultaneously yielding peformance level $(\gamma_1, \ldots, \gamma_\ell)$.

Given target sets $\mathcal{S}_\theta(\vec{\gamma})$ in $\mathbf{C}^N$, the suboptimal MultiOPT problem is to find a stable system f whose values $f(e^{i\theta})$ lie in the target sets

$$f(e^{i\theta}) \in \mathcal{S}_\theta(\vec{\gamma}).$$

Standard assumption. Assume that each Γ_j is three times differentiable. Assume that sets $\mathcal{S}_\theta(\vec{\gamma})$ have nonempty interiors for each $\vec{\gamma}$ and are uniformly bounded. Lastly assume that the sets are *uniformly contractible*: there exist mappings $I_t(e^{i\theta}, z)$ from $\mathcal{S}(\vec{\gamma})$ to $\mathcal{S}(\vec{\gamma})$ continuous in t, θ, z such that for each θ, $(e^{i\theta}, z) \mapsto I_t(e^{i\theta}, z)$ is the identity for $t = 0$, the first coordinate of $I_t(e^{i\theta}, z)$ is $e^{i\theta}$, and for each θ, $z \mapsto I_t(e^{i\theta}, z)$ is constant when $t = 1$. Intuitively, uniform contractibility just ensures that each of the domains $\mathcal{S}(\vec{\gamma})$ is arcwise connected (i.e., none have isolated components) and none of them contain holes. Clearly the class of $\mathcal{S}(\vec{\gamma})$ which are uniformly contractible contains the class of $\mathcal{S}(\vec{\gamma})$ whose $\mathcal{S}_\theta(\vec{\gamma})$ are all convex, a class of $\mathcal{S}(\vec{\gamma})$ upon which most uniqueness theory is based. Our assumption is weak and without it most theory and existing algorithms of any existing type appear impossible (unless one is in a situation where the holes do not matter and only one component matters.)[1]

1.4. The gist of the main results. It is common for computer optimization algorithms at the kth step to keep track of both the "primal variables" (in our case f^k) and "dual variables." These are called primal-dual algorithms. We shall see in sections 2 and 4.2 that such an algorithm for MultiOPT which stops in a local

[1]Roughly this means that for all θ, the set $\mathcal{S}_\theta(\vec{\gamma}^*)$ consists only of one piece and has no holes. For such a set to contain holes or to be disconnected, one must be working with a highly nonlinear situation. Under these circumstances, even convergence of one's computer runs to functions f^*, F can be problematic; possibly it would be worth while to reconsider the setup of the original problem.

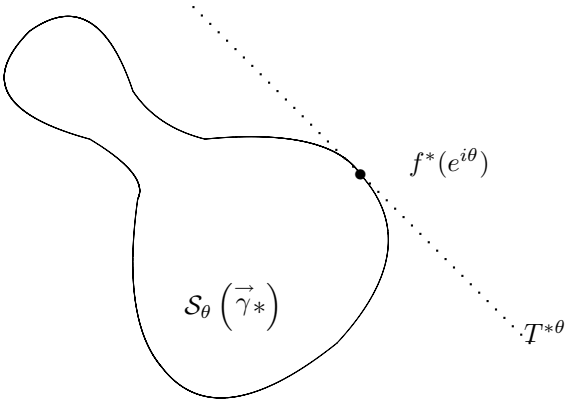


FIG. 2. *Nonconvex but f^* is a unique global optimum.*

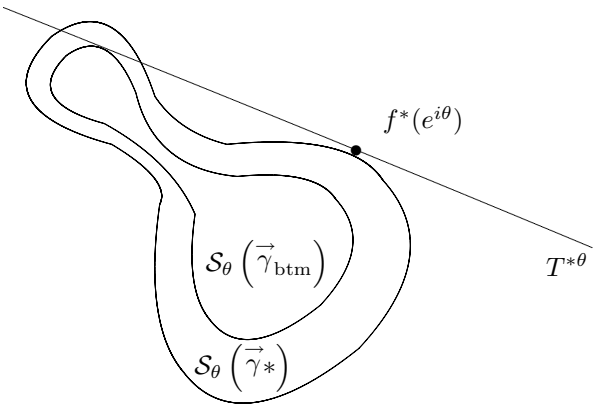


FIG. 3. *Nonconvex but no performance is better than* $\vec{\gamma}^{\rm btm}$.

optimum f^* with performance levels $\vec{\gamma}^* := (\gamma_1(f^*), \ldots, \gamma_\ell(f^*))$ produces information with the simple geometric interpretation that we know the primal local optimum f^* and a[2] tangent plane $T^{*\theta}$ to the boundary $\partial\mathcal{S}_\theta(\vec{\gamma}^*)$ at $f^*(e^{i\theta})$. The tangent plane is a good way to visualize optimal "dual information."

Figure 2 illustrates this situation as well as our uniqueness test when the set $\mathcal{S}_\theta(\vec{\gamma}^*)$ has smooth boundary. Figure 3 illustrates our lower bound for performance. These tests constitute our main results and roughly they say

for each θ, if the tangent plane $T^{*\theta}$ to $\mathcal{S}_\theta(\vec{\gamma}^*)$ at $f^*(e^{i\theta})$

(a) *intersects $\mathcal{S}_\theta(\vec{\gamma}^*)$ at only the point $f^*(e^{i\theta})$, then f^* is the unique global optimum for MultiOPT; see Figure 2;*

(b) *does not intersect $\mathcal{S}_\theta(\vec{\gamma}^{\rm btm})$, then any $f \in H_N^\infty$ has $\vec{\gamma}_j(f) \geq \vec{\gamma}_j^{\rm btm}$ for some $j = 1, 2, \ldots, \ell$. That is, performance can be no better than $\vec{\gamma}^{\rm btm}$; see Figure 3.*

[2]When the set $\mathcal{S}_\theta(\vec{\gamma}^*)$ has smooth boundary (as in problems with a single performance measure Γ^1), the tangent plane $T^{*\theta}$ is unique.

We emphasize that $\mathcal{S}_\theta(\vec{\gamma}^*)$ need not be convex; our conditions are much less stringent. For comparison recall that a closed strictly convex set C with smooth boundary ∂C has the defining property that at each point z_0 on ∂C the tangent plane T_{z_0} to ∂C at z_0 intersects C only at z_0. Thus to check strict convexity one must check this property at *every point* z_0 on ∂C. Our test for uniqueness requires checking this condition at *only one point*, for each θ. This is a surprising property of optimization over spaces of analytic functions.

The tests work in more generality than described here. The sets $\mathcal{S}_\theta(\vec{\gamma}^*)$ can have corners as would be the case with multiperformance problems; see Theorem 4.1. Also, a smaller set than the tangent plane, the *complex tangent plane* suffices in our test; see Theorem 4.2. Section 4.2 gives geometric interpretations of these theorems.

2. Optimality conditions and computation. We begin the detailed description of our global uniqueness test and performance bound by saying precisely what is meant by primal and dual variables.

2.1. Primal-dual optimality conditions. Recall the optimality conditions for MultiOPT. First we introduce the notation

$$\frac{\partial \Gamma}{\partial z} = \begin{pmatrix} \frac{\partial}{\partial z_1}\Gamma_1 & \dots & \frac{\partial}{\partial z_N}\Gamma_1 \\ \dots & \dots & \dots \\ \frac{\partial}{\partial z_1}\Gamma_\ell & \dots & \frac{\partial}{\partial z_N}\Gamma_\ell \end{pmatrix}. \tag{2}$$

THEOREM 2.1 (see [HMer98, Theorem 17.1.1], [HV]). *Assume Γ_j, for $j = 1, \dots, \ell$, satisfies the standard assumption. Suppose that a continuous local optimum f^* does exist with performance values denoted*

$$\gamma_1^*, \dots, \gamma_\ell^*,$$

which makes

$$\frac{\partial \Gamma}{\partial z}(e^{i\theta}, f^*(e^{i\theta}))\frac{\partial \Gamma^T}{\partial z}(e^{i\theta}, f^*(e^{i\theta}))$$

invertible on all $e^{i\theta}$. (This implies $\ell \le N$.) Then there exist functions ψ_j in $L^1(\mathbf{T})$, for $j = 1, \dots, \ell$, which satisfy flatness*,*

$$\psi_j(e^{i\theta})(\gamma_j^* - \Gamma_j(e^{i\theta}, f^*(e^{i\theta}))) = 0 \text{ for } j = 1, \dots, \ell,$$

gradient alignment*,*

$$\psi_1(e^{i\theta})\frac{\partial \Gamma_1}{\partial \bar{z}}(e^{i\theta}, f^*(e^{i\theta})) + \dots + \psi_\ell(e^{i\theta})\frac{\partial \Gamma_\ell}{\partial \bar{z}}(e^{i\theta}, f^*(e^{i\theta})) = e^{-i\theta}\bar{F}(e^{i\theta}), \quad F \in H^2_N, \tag{3}$$

normalization*,*

$$\sum_{j=1}^{\ell} \int_0^{2\pi} \psi_j d\theta = 2\pi,$$

positivity*,*[3]

$$\gamma_j^* - \Gamma_j(e^{i\theta}, f^*(e^{i\theta})) \ge 0 \text{ and } \psi_j \ge 0 \text{ for } j = 1, \dots, \ell.$$

Moreover, it is shown in [HV] *that if f^* is once differentiable, the ψ_j are once differentiable.*

Here A^T denotes the conjugate transpose of A and H^2_N is the set of vector-valued H^2 functions on the circle.

[3]This just reiterates the definition of γ_j^*.

2.2. Computer optimization. It is common for computer optimization algorithms at the kth step to keep track of both the "primal variables" (in our case f^k) and "dual variables" ψ_j^k and consequently F^k. We start with bad guesses $f^0 \in H_N^\infty, \psi_j^0$ and update them in various ways ultimately to approach a solution to the flatness and gradient alignment equations. At optimum a key property is that F is analytic. These are called primal-dual algorithms and are popular. (See [AHO], [VB].) As we shall see in our H^∞ case the dual variable F^k has the interpretation that $e^{-i\theta}\bar{F^k}(e^{i\theta})$ is pointed "normally" to the sets $\mathcal{S}(e^{i\theta})$ at the point $f^k(e^{i\theta})$. Exactly what this means geometrically requires discussion (see section 4.2), but it motivates calling the optimal dual vector function $e^{-i\theta}\bar{F}$ the *conjugate analytic normal* at f^*. The point is that many H^∞ optimization algorithms produce both a primal optimum f^* and a conjugate analytic normal $e^{-i\theta}\bar{F}(e^{i\theta})$ to $\mathcal{S}_\theta(e^{i\theta})$ at $f^*(e^{i\theta})$.

In summary, typical algorithms, c.f. [HMer98], produce a sequence f_k, F_k of approximates to f, F where $f_k \in H_N^\infty \cap C^\infty(\mathbf{T})$. In some algorithms F_k does not belong to H_N^∞, but one can easily compute $\overset{\vee}{F}_k \in H_N^\infty$, which is the best L_N^2 approximate to F_k, and use it as a vague indicator of closeness to optimum f^*, F. In the next section we give a test which removes most of the vagueness from such diagnostics, in that for each $f_k, \overset{\vee}{F}_k$ it gives an absolute lower bound on the best possible performance $\vec{\gamma}^*$. The methods behind proving this lower bound lead to our global uniqueness result.

3. An algorithmic phrasing of our uniqueness test. We now describe our main result in a (high level) algorithmic format. Although a bit redundant it gives a casual reader a description of the method which (except for the most technical hypotheses) is self-contained. The subsequent sections give theorems supplying technical hypotheses and verify that the Algorithm works.

Suppose you have run your favorite numerical algorithm for solving the optimization problem MultiOPT in section 1.1 and that you have obtained a local optimum $f^*, \vec{\gamma}^*$ and the corresponding dual function F.

Before we describe our test, we need a few definitions. For any integer $N > 0$, the N dimensional complex vector space $\mathbf{C}^N$ has the usual inner product $\langle z, w\rangle_C := \sum_j z_j \bar{w}_j =: z \cdot \bar{w}$, but $\mathbf{C}^N$ can be viewed as a $2N$ dimensional real vector space with the inner product $\langle z, w\rangle_R := \operatorname{Re} \sum_j z_j \bar{w}_j =: \operatorname{Re}[z \cdot \bar{w}]$. Define the complex plane which is *complex orthogonal* to the vector $\mathbf{N}$ at location b by

$$\mathbf{N}^{c\perp}(b) := \{z \in \mathbf{C}^N : \bar{\mathbf{N}} \cdot (z - b) = 0\}. \tag{4}$$

This complex orthogonal is a subset of the ordinary *real orthogonal* complement

$$\mathbf{N}^{r\perp}(b) := \{z \in \mathbf{C}^N : \operatorname{Re}[\bar{\mathbf{N}} \cdot (z - b)] = 0\} \tag{5}$$

to $\mathbf{N}$ at b. The b will occasionally be omitted when it is clear from context which point b is intended.

Now we describe our test for global optimality of f^* and, if optimality does not hold, we describe a bound on the best performance.

3.1. The Algorithm. Suppose we are given continuous functions f, F in H_N^∞, where F is never zero on the circle, and the Γ_j satisfy the standard assumption.

1. For each θ, compute[4] a linearly independent set of $N-1$ vectors $\phi^1(e^{i\theta}), \phi^2(e^{i\theta}),$

[4]Given $F(e^{i\theta}) = (F_1(e^{i\theta}), F_2(e^{i\theta}), \dots, F_N(e^{i\theta})) \in \mathbf{C}^N$, there are many ways to compute $\phi^k(e^{i\theta})$ for each θ. One way is to select $\phi^1(e^{i\theta}) = (-F_2(e^{i\theta}), F_1(e^{i\theta}), 0, 0, 0, \dots, 0)$, $\phi^2(e^{i\theta}) = (-F_3(e^{i\theta}), 0, F_1(e^{i\theta}), 0, 0, 0, \dots, 0)$, etc., provided that for that θ, we have $F_1(e^{i\theta}) \neq 0$. The Gram–Schmidt process may be used to obtain an orthogonal set of $\{\phi^k(e^{i\theta})\}_{k=1}^{N-1}$; this may help with numerics.

$\dots, \phi^{N-1}(e^{i\theta})$ satisfying the equation $\phi^k(e^{i\theta}) \cdot F(e^{i\theta}) = 0$ for $k = 1$ to $N-1$. (These $N-1$ vectors form a basis for $\mathbf{N}_\theta^{c\perp}(f(e^{i\theta}))$, where $\mathbf{N}_\theta = e^{-i\theta}\overline{F(e^{i\theta})}$.)

2. For each θ, define $v_\theta^j(w)$ by

$$v_\theta^j(w) := \Gamma_j(e^{i\theta}, f(e^{i\theta}) + w_1\phi^1(e^{i\theta}) + \cdots + w_{N-1}\phi^{N-1}(e^{i\theta})),$$

where $w = (w_1, w_2, \dots, w_{N-1}) \in \mathbf{C}^{N-1}$.

3. For each θ and $j = 1, \dots, \ell$, compute the minimum $\vec{\gamma}_j^{\text{btm}}$ of $v_\theta^j(w)$ for $w \in \mathbf{C}^{N-1}$.

(a) *Global uniqueness of optimum.* If f^* is a local Pareto optimum for MultiOPT with performance $\vec{\gamma}^*$, and F is the corresponding analytic dual, then if for each fixed θ the point $w_0 = 0$ in $\mathbf{C}^{N-1}$ is the unique nondegenerate[5] minimizer for

$$\tilde{\Gamma}(e^{i\theta}, w) := \max\left\{\frac{v_\theta^1(w)}{\gamma_1^*}, \frac{v_\theta^2(w)}{\gamma_2^*}, \dots, \frac{v_\theta^\ell(w)}{\gamma_\ell^*}\right\},$$

then f^* is the only solution to MultiOPT achieving performance level $\vec{\gamma}^*$.

(b) *The multiperformance $\vec{\gamma}^{\text{btm}}$ is a bound on the best performance.* If there exists a $\mathbf{C}^N$-valued continuous[6] function q defined on the circle such that $\operatorname{Re}(e^{i\theta}F(e^{i\theta}) \cdot (q(e^{i\theta}) - f(e^{i\theta}))) < 0$ for all θ and $\Gamma_j(e^{i\theta}, q(e^{i\theta})) \le \vec{\gamma}_j^{\text{btm}}$ for all j and θ,[7] then for any $\tilde{f} \in C(\mathbf{T}) \cap H_N^\infty$, we have $\gamma_j(\tilde{f}) \ge \vec{\gamma}_j^{\text{btm}}$ for at least one $j = 1, \dots, \ell$.

3.2. A tutorial example. Take

$$\Gamma(e^{i\theta}, z_1, z_2) = |(z_1 - e^{-i\theta})(z_1 - ke^{-i\theta})|^2 + |z_2|^2,$$

where $k = e^{\frac{3\pi i}{4}}$. It is easy to see that this is not a convex problem.

One finds that a local optimizer is $f^* := (0, 0)$, giving optimal value equal to $|k|^2 = 1$. We wish to illustrate our global optimality test. Calculate that the partial of Γ with respect to z is

$$\frac{\partial\Gamma}{\partial z}(e^{i\theta}, z) = ((2z_1 - (k+1)e^{-i\theta})\overline{(z_1 - e^{-i\theta})(z_1 - ke^{-i\theta})}, \bar{z_2}),$$

which at $z = (0, 0)$ is

$$\frac{\partial\Gamma}{\partial z}(e^{i\theta}, (0, 0)) = -((k+1)\overline{k}e^{i\theta}, 0).$$

Note this function extends analytically to the disk, which illustrates the gradient alignment condition of local optimality, Theorem 2.1; here $\psi_1 = 1$. There are different ways to implement our global optimality test and we illustrate several of them.

---

[5]Nondegeneracy is a very technical condition which will be defined later in section 4; it involves so fine a distinction that we do not think one would check it in practice.

[6]Note that q does not have to be—in fact, must not be—analytic.

[7]Such a q would exist, for example, if for every j, $\operatorname{Re}(e^{i\theta}F(e^{i\theta}) \cdot (w - f(e^{i\theta}))) < 0$ for all θ and $w \in \mathcal{S}_\theta(\vec{\gamma}^{\text{btm}})$. Here, uniform contractibility of $\mathcal{S}_\theta(\vec{\gamma}^{\text{btm}})$ would guarantee that the continuous function to which those sets may be uniformly contracted would satisfy the conditions that q must satisfy.

The subspace of $\mathbf{C}^N$, which is complex orthogonal to the conjugate of $\frac{\partial \Gamma}{\partial z}(e^{i\theta},(0,0))$, is called the complex tangent plane, and at fixed $e^{i\theta}$ it is

$$\mathbf{N}_\theta^{c\perp}((0,0)) := \{z : z_1 = 0\}.$$

The key issue is whether it intersects

$$\mathcal{S}_\theta(1) = \{z : |(z_1 - e^{-i\theta})(z_1 - ke^{-i\theta})|^2 + |z_2|^2 \le 1\}$$

in more than the one point $(0,0)$. The points in $\mathbf{N}_\theta^{c\perp}((0,0)) \cap \mathcal{S}_\theta(1)$ are

$$\{z : z_1 = 0 \text{ and } |(z_1 - e^{-i\theta})(z_1 - ke^{-i\theta})|^2 + |z_2|^2 \le 1\} = \{z : z_1 = 0 \text{ and } |k|^2 + |z_2|^2 \le 1\}.$$

Since $|k| = 1$ this forces $z_2 = 0$, so $z = (0,0)$. Therefore $(0,0)$ is the unique global optimizer for this MultiOPT problem.

We now give an argument analogous to that just given, but in terms of our Algorithm above. A basis for $\mathbf{N}_\theta^{c\perp}((0,0))$ is $\phi^1(e^{i\theta}) = (0,1)$ for all θ. Then

$$v_\theta^1(w) = \Gamma(e^{i\theta},(0,0) + w(0,1)) = |w|^2.$$

Its minimizer over $w \in \mathbf{C}$ is clearly $(0,0)$ and is unique and nondegenerate.

3.3. Theoretical justification. The lower bound stated in the Algorithm can be rigorously proved under modest assumptions as we now see in Theorem 3.1. See Theorem 3.2 for another bound not requiring the hypothesis concerning q.

THEOREM 3.1. *Suppose the performance functions $\Gamma_1, \ldots, \Gamma_\ell$ satisfy the standard assumption. Suppose that $f, F \in H_N^\infty$ are continuous with F never vanishing. Set $\mathbf{N}_\theta := e^{-i\theta}\bar{F}(e^{i\theta})$. Suppose there exist $\vec{\gamma}_j^{\,\text{btm}}$ such that for every θ and $z \in \mathbf{N}_\theta^{c\perp}(f(e^{i\theta}))$, there exists a $j = 1, 2, \ldots, \ell$ such that*

$$\Gamma_j(e^{i\theta}, z) > \vec{\gamma}_j^{\,\text{btm}} \quad . \tag{6}$$

Suppose that there exists some $\mathbf{C}^N$-valued continuous function $q(e^{i\theta})$ such that $\operatorname{Re}(e^{i\theta}F(e^{i\theta}) \cdot (q(e^{i\theta}) - f(e^{i\theta}))) < 0$ for all θ and $\Gamma_j(e^{i\theta}, q(e^{i\theta})) \le \vec{\gamma}_j^{\,\text{btm}}$ for every $j = 1$ to ℓ. Then there is no $f^{**} \in C(\mathbf{T}) \cap H_N^\infty$ such that $\vec{\gamma}_j\,(f^{**}) \le \vec{\gamma}_j^{\,\text{btm}}$ for $j = 1, 2, \ldots, \ell$.

We soon prove this result which will illustrate the principle behind both 3(a) and 3(b) in the Algorithm.

THEOREM 3.2. *Suppose the performance functions $\Gamma_1, \ldots, \Gamma_\ell$ satisfy the standard assumption. Suppose that $f, F \in H_N^\infty$ are continuous with F never vanishing. Set $\mathbf{N}_\theta := e^{-i\theta}\bar{F}(e^{i\theta})$. Suppose that for every θ and $z \in \mathbf{N}_\theta^{r\perp}(f(e^{i\theta}))$, there exists a $j = 1, 2, \ldots, \ell$ such that*

$$\Gamma_j(e^{i\theta}, z) > \vec{\gamma}_j^{\,\text{btm}} \quad . \tag{7}$$

Then there is no $f^{**} \in C(\mathbf{T}) \cap H_N^\infty$ such that $\vec{\gamma}_j\,(f^{**}) \le \vec{\gamma}_j^{\,\text{btm}}$ for $j = 1, 2, \ldots, \ell$.

COROLLARY 3.3. *If f^* is a local optimum and F is its analytic dual, and if f^*, F are continuous, then f^*, F satisfy the hypotheses of Theorem 3.1 and so its conclusion gives the bound on $\vec{\gamma}_j\,(f^*)$ found in Theorem 3.1.*

Proof of Theorem 3.2. The approach is similar to that which will be used in the proof of the uniqueness theorems in section 5. Consider the transformation of $\mathbf{C}^N$ to $\mathbf{C}$ defined by $\pi_\theta(z) = e^{i\theta}F(e^{i\theta}) \cdot (z - f(e^{i\theta}))$. From assumption (7) we find that $\mathcal{S}_\theta(\vec{\gamma}^{\,\mathrm{btm}})$ does not meet $\mathbf{N}_\theta^{r\perp}$. Thus $\operatorname{Re}\pi_\theta(z)$ is nonzero for all θ and all $z \in \mathcal{S}_\theta(\vec{\gamma}^{\,\mathrm{btm}})$. Thus $\operatorname{Re}\pi_\theta(z)$ has the same sign regardless of the values of θ or $z \in \mathcal{S}_\theta(\vec{\gamma}^{\,\mathrm{btm}})$; by possibly negating F, we may assume that $\pi_\theta(z)$ has strictly negative real part for all $z \in \mathcal{S}_\theta(\vec{\gamma}^{\,\mathrm{btm}})$ and all θ.

Suppose $f^{**} \in H_N^\infty \cap C(\mathbf{T})$ satisfies

$$\Gamma_j(e^{i\theta}, f^{**}(e^{i\theta})) \leq \vec{\gamma}_j^{\,\mathrm{btm}} \tag{8}$$

for every j and θ. The mapping

$$P : \mathbf{T} \to \mathbf{C},$$

$$e^{i\theta} \to \pi_\theta(f^{**}(e^{i\theta}))$$

extends to the analytic function

$$P(s) = sF(s) \cdot (f^{**}(s) - f(s))$$

for s on the closed disk and has a zero at the origin. But since $\pi_\theta(z)$ has strictly negative real part for $z \in \mathcal{S}_\theta(\vec{\gamma}^{\,\mathrm{btm}})$ and all θ, the function P is nonvanishing and has strictly negative real part on the circle. Hence P has strictly negative real part at the origin. This is a contradiction. Thus f^{**} does not exist and the proof is finished. ∎

That P has strictly negative real part on the circle implies that the winding number of P on the circle is zero, while $P(0) = 0$ implies that that winding number is at least 1. This contradiction foreshadows the winding number properties that we shall use in the proof of the Algorithm as stated in Theorem 3.1. The hypothesis involving q could probably be improved.

Proof of Theorem 3.1. Suppose that there exists f^{**} as indicated in the theorem. We construct $\pi_\theta(s)$ and $P(s) = sF(s)(f^{**}(s) - f(s))$ as in the proof of Theorem 3.2. Now assumption (6) and the fact that F is nonzero on the circle imply that for all $\theta, 0 \notin \pi_\theta(\mathcal{S}_\theta(\vec{\gamma}^{\,\mathrm{btm}}))$. By continuity of F and f, this implies that for some $\delta > 0$, $\pi_\theta(\mathcal{S}_\theta(\vec{\gamma}^{\,\mathrm{btm}}))$ excludes the closed disk of radius δ about 0 in $\mathbf{C}$ for all θ. By uniform contractibility of the $\mathcal{S}_\theta(\vec{\gamma}^{\,\mathrm{btm}})$ there exists a homotopy f_t^{**} from f^{**} to q such that $f_t^{**}(e^{i\theta})$ is contained in $\mathcal{S}_\theta(\vec{\gamma}^{\,\mathrm{btm}})$ for every θ, t. Thus $P_t(e^{i\theta}) := \pi_\theta(f_t^{**}(e^{i\theta}))$ does not vanish for any value of θ nor t, and since it is continuous in t, θ, the winding number $\operatorname{wind}_0(P_t)$ of P_t around 0 is independent of t. Indeed, the winding number of $P(e^{i\theta})$ over the circle is the same as the winding number of $e^{i\theta}F(e^{i\theta}) \cdot (q(e^{i\theta}) - f(e^{i\theta}))$ over the circle, which is zero since we assumed that $\operatorname{Re}(e^{i\theta}F(e^{i\theta}) \cdot (q(e^{i\theta}) - f(e^{i\theta}))) < 0$ for all θ. Now we turn to the completely different property of P, namely $P(0) = 0$, to see that $\operatorname{wind}_0(P) > 0$. This contradicts our finding above that $\operatorname{wind}_0(P) = 0$, and thereby shows that f^{**} cannot exist. ∎

To state rigorous results on uniqueness requires technical definitions and we do this in the next section.

4. Uniqueness theorem. In this section we present the precise theorems which establish the validity of the Algorithm in section 3.

4.1. Uniqueness theorem expressed analytically. First we state a consequence of our main uniqueness theorem, which is easier to understand and carries many of the main ideas.

THEOREM 4.1. *Suppose the performance functions* $\Gamma_1, \dots, \Gamma_\ell$ *satisfy the standard assumption. Suppose the following:*

(1) $f^* \in H_N^\infty(\mathbf{T}) \cap C^1(\mathbf{T})$ *and* $\psi_j \geq 0$ *are functions satisfying the flatness, gradient alignment, normalization, and positivity conditions, as well as the condition that* $\frac{\partial \Gamma}{\partial z} \frac{\partial \Gamma^T}{\partial z}$ *be invertible.*

(2) *Set* $\mathbf{N}_\theta := e^{-i\theta}\bar{F}$. *For each* θ

$$\Gamma(e^{i\theta}, z) := \max\left\{ \frac{\Gamma_1(e^{i\theta}, z)}{\gamma_1^*}, \dots, \frac{\Gamma_\ell(e^{i\theta}, z)}{\gamma_\ell^*} \right\}$$

as a function of z *has a unique minimum for* $z \in \mathbf{N}_\theta^{r\perp}(f^*(e^{i\theta}))$ *occurring at* $z_\theta = f^*(e^{i\theta})$. *Here* $\gamma_j^* := \gamma_j(f^*)$ *denotes the performance level produced by* f^* *with respect to* Γ_j, *for* $j = 1, \dots, \ell$.

Then f^* *is a unique Pareto optimum for MultiOPT at performance level* $\vec{\gamma}$. *Namely, there is no other function* $f \in H_N^\infty \cap C^1$, *with the property* $\gamma_j(f) \leq \gamma_j(f^*)$ *for all* $j = 1, \dots, \ell$.

For the remainder of this work, we shall write $\mathbf{N}_\theta^{r\perp}, \mathbf{N}_\theta^{c\perp}$ to mean $\mathbf{N}_\theta^{r\perp}(f^*(e^{i\theta}))$, $\mathbf{N}_\theta^{c\perp}(f^*(e^{i\theta}))$. A stronger theorem (which is considerably harder to prove) requires minimizing Γ on the set $\mathbf{N}_\theta^{c\perp}$ rather than the set $\mathbf{N}_\theta^{r\perp}$, which is one real dimension bigger than $\mathbf{N}_\theta^{c\perp}$. For SISO systems, that is, $N = 1$, it is this stronger theorem and the observation $\mathbf{N}_\theta^{c\perp} = \{0\}$ which leads to the fact, mentioned in the abstract, that for SISO systems H^∞ optima are unique. Before stating the result we introduce a technical condition.

We say that a real-valued function P on an affine subspace $A \subset \mathbf{C}^N$ has a *nondegenerate* (local) *minimum* at $w_0 \in A$ if P grows at least quadratically near w_0;[8] more precisely, there exists a positive C such that for all w in some neighborhood of w_0 in A, we have

$$|P(w) - P(w_0)| \geq C|w - w_0|^2.$$

THEOREM 4.2. *Suppose the performance functions* $\Gamma_1, \dots, \Gamma_\ell$ *and* f^* *and* F *are as in the set up of Theorem* 4.1 *and satisfy the hypotheses* (1). *Replace hypothesis* (2) *of Theorem* 4.1 *by the following weaker hypotheses:*

(2′a) *Set* $\mathbf{N}_\theta := e^{-i\theta}\bar{F}$. *For each* θ

$$\Gamma(e^{i\theta}, z) := \max\left\{ \frac{\Gamma_1(e^{i\theta}, z)}{\gamma_1^*}, \dots, \frac{\Gamma_\ell(e^{i\theta}, z)}{\gamma_\ell^*} \right\}$$

as a function of z *has a unique minimum for* $z \in \mathbf{N}_\theta^{c\perp}$ *occurring at* $z_\theta = f^*(e^{i\theta})$. *Here* γ_j^* *denotes the performance level produced by* f^* *with respect to* Γ_j, *for* $j = 1, \dots, \ell$.

(2′b) *For every* θ, *the real-valued function* $\Gamma(e^{i\theta}, \cdot)$ *on* $\mathbf{N}_\theta^{c\perp}$ *given in* (2′a) *has a nondegenerate minimum on* $\mathbf{N}_\theta^{c\perp}$ *at* $z = f^*(e^{i\theta})$. *(Note*

[8] For our results we could replace "quadratically" with "polynomially" here; some modification would be required in Lemmas 5.1 and 5.2 to prove this.

that F is never zero on the circle. Condition (2′b) *will be replaced later by condition* (2″b).*)*

Then f^* is a unique Pareto optimum for MultiOPT at performance level $\vec{\gamma}=\vec{\gamma}\,(f^*)$. Namely, there is no other function $f \in H^\infty_N \cap C^1$, with the property $\gamma_j(f) \le \gamma_j(f^*)$ for all $j = 1, \dots, \ell$.

These theorems clearly give a test for determining if a local optimum is the unique global optimum for a MultiOPT problem, which is practical to the extent that computing the minimum of the $\Gamma_j(e^{i\theta}, \cdot)$ over subspace $(e^{-i\theta}\bar{F})^{c\perp}$ or, respectively, over $(e^{-i\theta}\bar{F})^{r\perp}$ is practical. At least this is a $2N-2$ real dimensional problem, respectively, $2N-1$ dimensional problem, as opposed to the infinite dimensional MultiOPT problem.

We emphasize that this condition is much less stringent than a global convexity condition that would be required for uniqueness in most optimization problems (ones not involving stability). This will be explained fully in section 4.3, which describes our results geometrically and compares them to conventional convexity.

4.2. Uniqueness theorem expressed geometrically. All of the results of this paper can be stated geometrically. This way of looking at these optimization problems strongly enhances intuition, and also geometry plays a role in our proofs (see section 5). Critical to a geometric understanding are the sublevel sets

$$\mathcal{S}^j_\theta(\gamma_j) := \{z \in \mathbf{C}^N : \Gamma_j(e^{i\theta}, z) \le \gamma_j\}$$

in $\mathbf{C}^N$ of the performance functions Γ_j. Fix $\vec{\gamma}^* := (\gamma^*_1, \dots, \gamma^*_\ell)$. Let $\partial\mathcal{S}_\theta$ denote the topological boundary of $\mathcal{S}_\theta$. Let $d\partial\mathcal{S}_\theta$ denote

$$d\partial\mathcal{S}_\theta(\gamma^*) := \partial\mathcal{S}^1_\theta(\gamma^*_1) \cap \ldots \cap \partial\mathcal{S}^\ell_\theta(\gamma^*_\ell) \qquad \forall e^{i\theta} \in \mathbf{T}. \tag{9}$$

Of course $d\partial\mathcal{S}_\theta \subset \partial\mathcal{S}_\theta$.

The *flatness hypothesis* corresponds to the geometric statement

$$f^*(e^{i\theta}) \in \partial\mathcal{S}^j_\theta \text{ whenever } \psi_j(e^{i\theta}) \neq 0.$$

Hypothesis (2) of Theorem 4.1 corresponds to the geometric statement

$\mathcal{S}_\theta$ intersects $(e^{-i\theta}\bar{F})^{r\perp}$, a "tangent" plane to $\partial\mathcal{S}_\theta$ at $f^*(e^{i\theta})$, only at $f^*(e^{i\theta})$.

Hypothesis (2′) of Theorem 4.2 corresponds to the geometric statement

$\mathcal{S}_\theta$ intersects $(e^{-i\theta}\bar{F})^{c\perp}$,
a complex tangent plane to $\partial\mathcal{S}_\theta$ at $f^*(e^{i\theta})$,
only at $f^*(e^{i\theta})$ and has second order contact there.

To make these last two statements comprehensible we need some definitions and also we do need to prove the statements. Tangent planes to a smooth surface can be defined as the set of points orthogonal to a normal to the surface; there are two notions of orthogonal, real and complex, which lead to two notions of tangent plane, the ordinary tangent plane and the complex tangent. (See below the discussion on tangents.) We are dealing with surfaces which possibly have corners. Then at a corner there are many normal directions and as a consequence many tangent planes. To get formulas for tangent planes to $\partial\mathcal{S}_\theta$ we need some background.

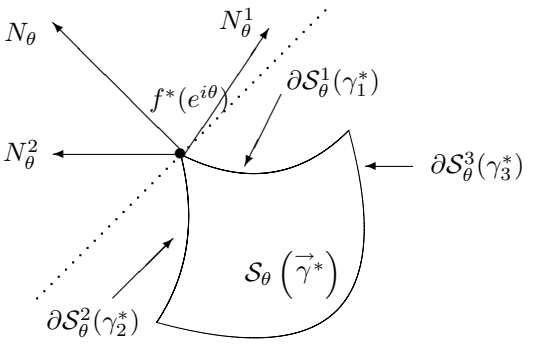


FIG. 4. $\mathbf{N}_\theta^j = \frac{\partial\Gamma_j}{\partial\bar{z}}(e^{i\theta}, f^*(e^{i\theta}))$ *and* $\mathbf{N}_\theta = \psi_1(e^{i\theta})\mathbf{N}_\theta^1 + \psi_2(e^{i\theta})\mathbf{N}_\theta^2 + \psi_3(e^{i\theta})\mathbf{N}_\theta^3$.

Consider a once continuously differentiable function ρ from $\mathbf{C}^N$ to $\mathbf{R}^+$. Let $S = \{z : \rho(z) \le 1\}$ and $\partial S = \{z : \rho(z) = 1\}$ denote its boundary; ρ is called a defining function for ∂S. The surface ∂S is a hypersurface, that is, it has real codimension 1. The gradient $\nabla\rho$ is directed normally to ∂S at z_0, which in complex notation is

$$\nabla\rho(z_0) = \frac{\partial\rho(z_0)}{\partial\bar{z}}.$$

Thus if $z_0 \in \partial\mathcal{S}_\theta^j$, then

$$\frac{\partial\Gamma_j}{\partial\bar{z}}(e^{i\theta}, z_0) \tag{10}$$

is directed normally to $\partial\mathcal{S}_\theta^j$. At a corner of $\mathcal{S}_\theta$ there is a family of normals $\mathcal{N}_\theta$ pointing "out" of $\mathcal{S}_\theta$, which we define to be all vectors of the form

$$\mathcal{N}_\theta := \left\{ \psi_1(e^{i\theta})\frac{\partial\Gamma_1}{\partial\bar{z}}(e^{i\theta}, z_0) + \cdots + \psi_\ell(e^{i\theta})\frac{\partial\Gamma_\ell}{\partial\bar{z}}(e^{i\theta}, z_0) : \text{ for some } \psi_j \ge 0 \right\}. \tag{11}$$

(See Figure 4.) This leads to a formal definition of the *tangent plane* to $\partial\mathcal{S}_\theta$ at $f^*(e^{i\theta})$ as $\mathbf{N}_\theta^{r\perp}$ for some $\mathbf{N}_\theta \in \mathcal{N}_\theta$ and the *complex tangent plane* as $\mathbf{N}_\theta^{c\perp}$ for some $\mathbf{N}_\theta \in \mathcal{N}_\theta$. The *gradient alignment* condition says precisely that

$$e^{-i\theta}\bar{F}(e^{i\theta}) \in \mathcal{N}_\theta.$$

Thus the *gradient alignment* condition amounts to selecting a normal and corresponding tangent planes to $\mathcal{S}_\theta$ at $f^*(e^{i\theta})$. Note the nature of the corner of $\mathcal{S}_\theta$ is determined by which ψ_j are not 0 (called the active ψ_j).

To prove the geometric interpretation of hypotheses (2) and (2′), observe that $\mathcal{S}_\theta(\vec{\gamma}) = \{z : \Gamma(e^{i\theta}, z) \le 1\}$. Thus the hypothesis (2) of Theorem 4.1

$$\Gamma(e^{i\theta}, z) > \min_{\zeta\in T}\Gamma(e^{i\theta}, \zeta) = \Gamma(e^{i\theta}, z_\theta) = 1 \text{ for } z \ne z_\theta \text{ in a set } T$$

says that T touches $\mathcal{S}_\theta(\vec{\gamma})$ only at z_θ where $\Gamma(e^{i\theta}, z_\theta) = 1$. Thus $z_\theta = f^*(e^{i\theta})$ being the location of a unique minimum is equivalent to T touching $\mathcal{S}_\theta(\vec{\gamma})$ only at $f^*(e^{i\theta})$.

The geometrical interpretation of the lower bound presented in section 3 is a variation on what we have just presented which is so straightforward that we will not discuss it in detail.

4.3. Benefits of our uniqueness test and comparisons. The geometric interpretations of this section lead immediately to the statement of our main result given in section 1.4. Recall the striking point is that the test for uniqueness in Theorems 4.2 and 4.1 just requires checking whether a (complex) tangent plane at one point $f^*(e^{i\theta})$ per θ intersects $\mathcal{S}_\theta(\vec{\gamma}^*)$ in other points.

By contrast, convexity requires a test at all points. We now present an analogous "convexity" condition for MultiOPT, a condition which one might try to check a priori; the reader can note how much stronger the hypothesis is than the key condition (2′) presented in Theorem 4.2.

THEOREM 4.3. *Suppose the Γ_j, $j = 1, \dots, \ell$, satisfy the standard assumption. If for every $\theta \in \mathbf{T}$ and every γ_j all complex tangent planes to ∂S_θ touch S_θ in exactly one point and have no more than quadratic order of contact with ∂S_θ,*[9] *then any local optimum is a global optimum.*

Proof. For one smooth Γ this was proved by Vityaev [V] and independently by Whittlesey [W1], [W2]. Multiple Γ_j produce sublevel sets with corners and the slight variation of their proofs required here will be presented in the course of proving Theorem 4.2. Then Theorem 4.3 follows from Theorems 2.1 and 4.2. When "complex tangent plane" is replaced by "tangent plane" thereby producing a stronger assumption in the theorem we have conventional strict convexity; that uniqueness theorem is due to Helton and Howe [HH]. □

4.4. Hypoconvex corners. In this section we are concerned only with uniqueness of optimum and we consider a weaker form of hypothesis (2′) of Theorem 4.2. It is motivated by the fact that often a differentiable local Pareto optimum must be hyperflat; see [HV] for exact hypotheses guaranteeing this. Hyperflatness means that the performance of a particular optimum for every performance Γ_j is flat, i.e., $\Gamma_j(e^{i\theta}, f^*(e^{i\theta}))$ is constant as a function of θ for every $j = 1, 2, \dots, \ell$. The weaker hypothesis is as follows:

(2‴a) Set $\mathbf{N}_\theta := e^{-i\theta}\overline{F}$. The set $\mathbf{N}_\theta^{c\perp}$ intersects $d\partial\mathcal{S}_\theta(\vec{\gamma}^*)$ only at $f^*(e^{i\theta})$ if at all.

(2‴b) There is a homotopy I_t of $d\partial\mathcal{S}_\theta(\vec{\gamma}^*)$ to $f^*(e^{i\theta})$ with $I_t(d\partial\mathcal{S}_\theta(\vec{\gamma}^*))$ lying entirely inside $\mathcal{S}_\theta(\vec{\gamma}^*)$ and missing $\mathbf{N}_\theta^{c\perp}$ for all t, $0 \le t \le 1$.

(2‴c) For every θ, the real-valued function on $\mathbf{N}_\theta^{c\perp}$ given by $z \mapsto \Gamma(e^{i\theta}, z)$ has a nondegenerate minimum on $\mathbf{N}_\theta^{c\perp}$ at $z = f^*(e^{i\theta})$. (Note that this is the same as conditions (2′ b) and (2″ b) of Theorem 4.2.)

THEOREM 4.4. *Assume the hypotheses of Theorem* 4.2 *except replace hypothesis* (2′) *by* (2‴). *Then if f^* and f^{**} are local Pareto optima both with the same performance level $\vec{\gamma}^*$ and if f^{**} is hyperflat, then $f^{**} = f^*$.*

This theorem suggests a class of geometric objects which generalizes the notion of hypoconvexity. We say that $\mathcal{S}_\theta(\vec{\gamma}^*)$ has *hypoconvex corners* provided that at any point p of $d\partial\mathcal{S}_\theta(\vec{\gamma}^*)$ any tangent plane T_p to $d\partial\mathcal{S}_\theta(\vec{\gamma}^*)$ intersects $d\partial\mathcal{S}_\theta(\vec{\gamma}^*)$ only at p and satisfies the order of contact condition (2‴c) at p. Also we require that there be a continuous homotopy I_t satisfying

$$I_t(d\partial\mathcal{S}_\theta(\vec{\gamma}^*)) \subset \mathcal{S}_\theta(\vec{\gamma}^*), \qquad 0 \le t \le 1,$$

[9] This property is called *strict hypoconvexity* and is weaker than strict convexity because complex tangent planes are smaller than tangent planes.

and

$$I_t(d\partial\mathcal{S}_\theta(\vec{\gamma}^{\,*})) \cap T_p = \{p\}.$$

The appeal of having such a geometric condition is as follows. If we have a problem whose sublevel sets $\mathcal{S}_\theta(\vec{\gamma})$ all can be verified to have hypoconvex corners, then clearly no two distinct hyperflat optima f^* and f^{**} can have the same performance levels. Thus if hypoconvexity can be verified in advance, we know (even without any computer runs) that a type of uniqueness must hold.

5. Proofs of Theorem 4.1, Theorem 4.2, and Theorem 4.4. The proof of each theorem begins in the same way and they all follow the pattern laid out in the proofs in section 3. Suppose that $f^* \in H^\infty$, F and performance levels $\gamma_1^*, \ldots, \gamma_\ell^*$ exist meeting hypotheses (1) of Theorem 4.1.

Again consider the transformation

$$\pi_\theta(z) := e^{i\theta}F(e^{i\theta}) \cdot (z - f^*(e^{i\theta})).$$

Transform the sublevel sets $\mathcal{S}_\theta^j(\gamma_j^*)$ with π_θ to obtain sets

$$\tilde{\mathcal{S}}_\theta^j(\gamma_j^*) := \pi_\theta(\mathcal{S}_\theta^j(\gamma_j^*)) \subset \mathbf{C}.$$

This map collapses $\mathcal{S}_\theta^j(\gamma_j^*)$ to $\mathbf{C}$ in a way which makes $\tilde{\mathcal{S}}_\theta^j(\gamma_j^*)$ contain zero.

Suppose f^{**} is an optimizer in $C(\mathbf{T})$ different from f^*. The mapping $P : \mathbf{T} \to \mathbf{C}$ defined by $e^{i\theta} \to \pi_\theta(f^{**}(e^{i\theta}))$ extends to the analytic function

$$P(s) = sF(s) \cdot (f^{**}(s) - f^*(s))$$

for s on the closed disk and has a zero at the origin. P is not identically 0, since the only point $v \in \mathcal{S}_\theta(\vec{\gamma})$ such that $\pi_\theta(v) = 0$ is $f^*(e^{i\theta})$, and for some θ, $f^*(e^{i\theta}) \neq f^{**}(e^{i\theta})$. Thus there is a $\tau > 0$ such that

(P vs. τ) $\quad$ $[0, \tau]$ is in the image of P applied to the unit disk

(by the open mapping theorem applied to P at 0).

5.1. Proof of Theorem 4.1. If $z \in \mathcal{S}_\theta(\vec{\gamma})$, then hypothesis (2) of Theorem 4.1 implies that

$$\operatorname{Re}\pi_\theta(z) \leq 0.$$

To see this, use the geometric interpretation of assumption (2) given in section 4.1, which says

$$\operatorname{Re}\pi_\theta(z) = \operatorname{Re}\left(e^{i\theta}F(e^{i\theta}) \cdot [z - f^*(e^{i\theta})]\right) \neq 0$$

for any $z \in \mathcal{S}_\theta(\vec{\gamma})$ except $z = f^*(e^{i\theta})$. Since $e^{-i\theta}\overline{F}(e^{i\theta})$ is the outward pointing normal we get $\operatorname{Re}\pi_\theta(z) \leq 0$.

Inequality 5.1 implies $\operatorname{Re}P(e^{i\theta}) \leq 0$. Moreover, $P(s)$ is analytic and bounded for s in the unit disk, since f^*, f^{**}, F are. Thus $\operatorname{Re}P(s) \leq 0$ on the unit disk, and in particular $P(s) \neq \tau$ at any $|s| \leq 1$. This contradicts (P vs τ). $\square$

5.2. Proof of Theorem 4.2. In the course of the proof we shall need that fact that condition (2′b.) implies the following condition:

(2″b.) *There exist constants* $C > 0$, $\delta > 0$ *such that for all* θ *and* $z \in \mathcal{S}_\theta(\vec{\gamma})$ *such that* $|z - f^*(e^{i\theta})| < \delta$, *we have* $|F(e^{i\theta}) \cdot (z - f^*(e^{i\theta}))| \geq C|z - f^*(e^{i\theta})|^2$.

We prove this in Lemma 5.1.

Now we must replace the $\operatorname{Re} P \leq 0$ assumption with other weaker structure. This uses the winding number (denoted wind_0) of $(P(e^{i\theta}) - \tau)$ around 0.

The first and greatest difficulty is establishing that it exists. This is accomplished by Lemma 5.2 below, which says we may choose τ small enough that τ is not in $\pi_\theta(\mathcal{S}_\theta(\vec{\gamma}))$ for any θ. Let us assume that this winding number exists and complete the proof.

For perspective note that $\operatorname{Re} P \leq 0$ implies $\operatorname{Re}[P(e^{i\theta}) - \tau] \leq -\tau$, and so $\operatorname{wind}_0[P(e^{i\theta}) - \tau] = 0$. However, even without $\operatorname{Re} P \leq 0$ we can obtain this easily by constructing a homotopy. Recall two functions between which there exists a homotopy not passing through 0 have the same winding number about 0. Begin by constructing a homotopy f_t^{**} of f^{**} to f^* such that every $f_t^{**}(e^{i\theta})$ is contained in $\mathcal{S}_\theta(\vec{\gamma})$. We may do this by using the map I_t from the standard assumption: The maps $I_t(e^{i\theta}, f^*(e^{i\theta}))$ and $I_t(e^{i\theta}, f^{**}(e^{i\theta}))$ construct homotopies of f^* and f^{**} to the same continuous function so the combination is the desired homotopy. Note we do not require the functions f_t^{**} to be analytic, but we merely require them to be continuous. Then the functions P_t

$$e^{i\theta} \mapsto P_t(e^{i\theta}) := \pi_\theta(f_t^{**}(e^{i\theta}))$$

are a homotopy from $P_1 = P$ to $P_0 = 0$. Also,

$$0 = \operatorname{wind}_0(-\tau) = \operatorname{wind}_0(P_0(e^{i\theta}) - \tau) = \operatorname{wind}_0(P(e^{i\theta}) - \tau),$$

since $P_0 = 0$. But from (P vs. τ) the analytic function $P(s) - \tau$ does equal zero for some s in the open disk, so its winding number about 0 is ≥ 1. This is a contradiction, so f^{**} cannot exist. □

LEMMA 5.1. *Condition* (2′b) *implies condition* (2″b).

Proof of Lemma 5.1. Since $z \mapsto \Gamma(e^{i\theta}, z)$ has a nondegenerate minimum on $\mathbf{N}_\theta^{c\perp}$ at $z = f^*(e^{i\theta})$, there exist $\delta(\theta) > 0, C(\theta) > 0$ such that for all $z \in \mathbf{N}_\theta^{c\perp}$ such that $|z - f^*(e^{i\theta})| \leq \delta(\theta)$ we have

$$\Gamma(e^{i\theta}, z) - 1 \geq C(\theta)|z - f^*(e^{i\theta})|^2. \tag{12}$$

By continuity of the functions involved and a compactness argument, we may assume that C and δ are independent of θ. Write $\delta(\theta) = \delta$ and $C(\theta) = C$. Without loss of generality assume that $\delta < 1$. Now suppose that the lemma does not hold; that for each $\epsilon, \gamma > 0$ there exists θ and a $z^1 \in \mathcal{S}_\theta(\vec{\gamma})$ such that $|z^1 - f^*(e^{i\theta})| < \gamma$ but

$$|F(e^{i\theta}) \cdot (z^1 - f^*(e^{i\theta}))| < \epsilon |z^1 - f^*(e^{i\theta})|^2. \tag{13}$$

Let z^2 equal the orthogonal projection of z^1 to $\mathbf{N}_\theta^{c\perp}$. Without loss of generality assume $\gamma < \delta$. Then the distance from z^1 to $\mathbf{N}_\theta^{c\perp}$ is $|z^1 - z^2| = |\frac{F(e^{i\theta})}{|F(e^{i\theta})|} \cdot (z^1 - f^*(e^{i\theta}))| \leq \frac{\epsilon}{|F(e^{i\theta})|}|z^1 - f^*(e^{i\theta})|^2 \leq C_F \epsilon |z^1 - f^*(e^{i\theta})|^2$, where C_F is the reciprocal of the minimum modulus of F on the circle.

Since Γ is uniformly Lipschitz on the set $\{(e^{i\theta}, z) : |z - f^*(e^{i\theta})| \le 1\}$ and $\gamma < \delta < 1$, we find that

$$|z^1 - z^2| \le C_F \epsilon |z^1 - f^*(e^{i\theta})|^2 \tag{14}$$

implies

$$\Gamma(e^{i\theta}, z^2) - 1 \le \Gamma(e^{i\theta}, z^2) - \Gamma(e^{i\theta}, z^1) \le |\Gamma(e^{i\theta}, z^2) - \Gamma(e^{i\theta}, z^1)| \le C_\Gamma C_F \epsilon |z^1 - f^*(e^{i\theta})|^2, \tag{15}$$

where C_Γ depends only on Γ and $|\Gamma(e^{i\theta}, z^2) - \Gamma(e^{i\theta}, z^1)| \le C_\Gamma |z^1 - z^2|$. (Recall $\Gamma(e^{i\theta}, z^1) \le 1$ since $z^1 \in \mathcal{S}_\theta(\vec{\gamma})$.) From (14) and the fact that $|z^1 - f^*(e^{i\theta})| \le \delta < 1$, we obtain $|z^1 - z^2| \le C_F \epsilon |z^1 - f^*(e^{i\theta})|$. Now suppose we choose ϵ so small that $C_F \epsilon < \sqrt{3}/2$. Then the Pythagorean theorem guarantees that $|z^1 - z^2|^2 + |z^2 - f^*(e^{i\theta})|^2 = |z^1 - f^*(e^{i\theta})|^2$, so

$$|z^1 - f^*(e^{i\theta})|^2 \le 4 |z^2 - f^*(e^{i\theta})|^2. \tag{16}$$

Combining (15) and (16),

$$\Gamma(e^{i\theta}, z^2) - 1 \le C_\Gamma C_F 4\epsilon |z^2 - f^*(e^{i\theta})|^2,$$

which contradicts (12) (since $C_\Gamma C_F 4\epsilon$ can be made arbitrarily small), where we recall that $C(\theta) = C$, $z^2 \in \mathbf{N}_\theta^{c\perp}$, and $|z^2 - f^*(e^{i\theta})| \le |z^1 - f^*(e^{i\theta})| < \gamma < \delta$. ∎

Define

$$\tilde{\mathcal{S}}_\theta := \pi_\theta(\mathcal{S}_\theta).$$

We owe the reader the following.

LEMMA 5.2. *The set $\tilde{\mathcal{S}}_\theta$ excludes the set*

$$\mathcal{K} := \left\{ \tilde{z} \in \mathbf{C} : \ \operatorname{Re} \tilde{z} > \frac{1}{2} \sqrt{|\tilde{z}|} \right\} \ \cap \ \{ \tilde{z} \in \mathbf{C} : \ |\tilde{z}| < \varepsilon \},$$

where ε is some positive constant.

The set $\mathcal{K}$ is a "solid cusp" of uniform size (independent of θ) whose interior lies outside of $\tilde{\mathcal{S}}_\theta$ for all θ, and whose singularity touches every $\tilde{\mathcal{S}}_\theta$ at 0.

Proof of Lemma 5.2. The proof splits in two parts. First we show that a $\tilde{z}$ in the image under π_θ of a $z \in \mathcal{S}_\theta$ near $f^*(e^{i\theta})$ lies in the set

$$\left\{ \tilde{z} \in \mathbf{C} : \ \operatorname{Re} \tilde{z} \le \frac{1}{2} \sqrt{|\tilde{z}|} \right\}.$$

In the second part we show that a $\tilde{z}$ in the image of a $z \in \mathcal{S}_\theta$ far from $f^*(e^{i\theta})$ satisfies $|\tilde{z}| \ge \varepsilon$.

CLAIM. *There exists a $\delta > 0$, such that for all*

$$\tilde{z} \in \pi_\theta(\ \mathcal{S}_\theta \cap \{ z \in \mathbf{C}^n \ \ : \ |z - f^*(e^{i\theta})| < \delta \} \ )$$

we have $\operatorname{Re} \tilde{z} \le \frac{1}{2} \sqrt{|\tilde{z}|}$.

Proof of claim. Choose ε so small that for all $e^{i\theta} \in \mathbf{T}$, we have $(\sum_{j=1}^{\ell} \psi_j(e^{i\theta}))\varepsilon / \sqrt{C} < 1/2$, where C comes from assumption (2″b). Choose ε even smaller and then δ so small that assumption (2″b) is satisfied and for $|z - f^*(e^{i\theta})| < \delta$, $z \in \mathcal{S}_\theta$, and $1 \le j \le \ell$,

$$\operatorname{Re} \left(\frac{\partial \Gamma_j}{\partial z}(e^{i\theta}, f^*(e^{i\theta})) \cdot (z - f^*(e^{i\theta})) \right) < \varepsilon |z - f^*(e^{i\theta})|.$$

Then

(17)

$$\mathrm{Re}\,(e^{i\theta}F(e^{i\theta})\cdot(z-f^*(e^{i\theta}))) = \mathrm{Re}\left(\sum_{j=1}^{\ell}\left(\psi_j(e^{i\theta})\frac{\partial\Gamma_j}{\partial z}(e^{i\theta},f^*(e^{i\theta}))\right)\cdot(z-f^*(e^{i\theta}))\right)$$

(18)

$$\leq \left(\sum_{j=1}^{\ell}\psi_j(e^{i\theta})\right)\varepsilon|z-f^*(e^{i\theta})|.$$

By assumption (2″b), $|z-f^*(e^{i\theta})| \leq \frac{1}{\sqrt{C}}\sqrt{|F(e^{i\theta})\cdot(z-f^*(e^{i\theta}))|}$, so

(19) $\mathrm{Re}\,(e^{i\theta}F(e^{i\theta})\cdot(z-f^*(e^{i\theta})))$

(20)

$$\leq \left(\sum_{j=1}^{\ell}\psi_j(e^{i\theta})\right)\frac{\varepsilon}{\sqrt{C}}\sqrt{|e^{i\theta}F(e^{i\theta})\cdot(z-f^*(e^{i\theta}))|}$$

(21)

$$\leq \frac{1}{2}\sqrt{|e^{i\theta}F(e^{i\theta})\cdot(z-f^*(e^{i\theta}))|},$$

i.e.,

$$\mathrm{Re}\,(\pi_\theta(z)) \leq \frac{1}{2}\sqrt{|\pi_\theta(z)|}$$

for all $(s,z)\in\mathcal{S}\cap\{(s,z)\in\mathbf{T}\times\mathbf{C}^n;\ |z-f^*(e^{i\theta})|<\delta\}$. The claim follows.

For the second part of the proof, we consider $\tilde{z}\in\tilde{\mathcal{S}}_\theta$ such that $\tilde{z}=\pi_\theta(z)$ and $|z-f^*(e^{i\theta})|\geq\delta$, where δ is from the first part of the proof. Assumption (2′a) says that $\mathcal{S}_\theta$ misses the complex tangent plane $\mathbf{N}_\theta^{c\perp}$ for all θ (except for $f^*(e^{i\theta})$), so

$$|\pi_\theta(z)|>0$$

for $z\in\mathcal{S}_\theta$ and $z-f^*(e^{i\theta})\neq 0$. By continuity and compactness there exists $\varepsilon>0$ such that if $z\in\mathcal{S}_\theta$ and $|z-f^*(e^{i\theta})|>\delta$ we have

$$|\pi_\theta(z)|>\varepsilon>0$$

uniformly in θ, for some ε and all z,θ, so $|\tilde{z}|>\varepsilon>0$. Combining this with the claim, the proof of the lemma is complete. □

Proof of Theorem 4.4. The proof follows that of Theorem 4.2. The principal difference is the existence of a homotopy f_t^{**} of f^* to f^{**} such that for all θ, $\pi_\theta(f_t^{**}(e^{i\theta}))$ does not belong to the set $\mathcal{K}$. In the proof of Theorem 4.2 the homotopy of the sets $\mathcal{S}$ is the tool that provides us with this fact. Now that we know that the optimum f^* is hyperflat, its graph lies in a smaller set, so the homotopy of the entire $\mathcal{S}$ is not needed; the homotopy of the set where $\Gamma_j=\gamma_j$ for every j (which contains the graph of f^*) will suffice. The only change, then, arises in the second part of the proof of Lemma 5.2, where assumption (2′a) is used. □

REFERENCES

[AHO] F. Alizadeh, J. P. Haeberly, and M. L. Overton, *Primal-dual interior-point methods for semidefinite programming: Convergence rates, stability and numerical results*, SIAM J. Optim., 8 (1998), pp. 746–768.

[BB] S. Boyd and C. Barratt, *Linear Controller Design*, Prentice–Hall, Englewood Cliffs, NJ, 1991.

[BDGPS] G. Balas, J. Doyle, K. Glover, A. Packard, and R. Smith, *The μ-Analysis and Synthesis Toolbox for Use with Matlab*, The Mathworks, Natick, MA, 1990.

[FKTW] M. Fan, L. Wang, J. Koninckx, and A. Tits, *Software package for optimization based design with user-supplied simulators*, IEEE Control Syst. Mag., 9 (1989), pp. 66–71.

[FK93] A. E. Frazho and S. M. Kherat, *On mixed H^2-H^∞ tangential interpolation*, in New Aspects in Interpolation and Completion Theories, Oper. Theory Adv. Appl. 64, Birkhäuser, Basel, 1993.

[HH] J. W. Helton and R. Howe, *A bang-bang principle for the frequency domain*, J. Approx. Theory, 47 (1986), pp. 101–121.

[HMar] J. W. Helton and D. E. Marshall, *Frequency domain design and analytic selections*, Indiana Univ. Math. J., 39 (1990), pp. 157–184.

[HMer94] J. W. Helton and O. Merino, *A fibered polynomial hull without an analytic selection*, Michigan Math. J., 41 (1994), pp. 285–287.

[HV] J. W. Helton and A. Vityaev, *Analytic functions optimizing competing constraints*, SIAM J. Math. Anal., 28 (1997), pp. 749–767.

[HMer98] J. W. Helton and O. Merino, *Classical Control Using H^∞ Methods: Theory, Optimization, and Design*, SIAM, Philadelphia, 1998.

[KRE96] P. P. Khargonekar, M. A. Rotea, and E. Baeyens, *Mixed H_2/H^∞ filtering. H^∞ and robust estimation*, Internat. J. Robust Nonlinear Control, 6 (1996), pp. 313–330.

[MNPW] D. Q. Mayne, W. T. Nye, L. Polak, and T. Wu, *DELIGHT MIMO: An interactive, optimization based multivariable control system design package*, in Computer-Aided Control Systems Engineering, M. Jamshidi and C. J. Herget, eds., North–Holland, Amsterdam, 1985.

[MR] A. Megretski and A. Rantzer, *System analysis via integral quadratic constraints*, IEEE Trans. Automat. Control, 42 (1997), pp. 819–830.

[PY95] V. V. Peller and N. J. Young, *Construction of superoptimal approximants*, Math. Control Signals Systems, 8 (1995), pp. 118–137.

[PRR97] R. K. Prasanth and M. A. Rotea, *Interpolation with multiple norm constraints*, Math. Control Signals Systems, 10 (1997), pp. 165–187.

[Si] T. Sideris, *Robust Feedback Synthesis via Conformal Mappings and H_∞ Optimization*, Ph.D. thesis, University of Southern California, Los Angeles, 1985.

[St] R. Streit, *Solution of systems of complex linear equations in the l_∞ norm with constraints on the unknowns*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 132–149.

[VB] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.

[V] A. E. Vityaev, *Uniqueness of solutions of a H^∞ optimization problem and complex geometric convexity*, J. Geom. Anal., 9 (1999), pp. 161–173.

[W1] M. A. Whittlesey, *Polynomial hulls and H^∞ control for a hypoconvex constraint*, Math. Ann., 317 (2000), pp. 677–701.

[W2] M. A. Whittlesey, *Polynomial Hulls and an Optimization Problem*, preprint, California State University, San Marcos, San Marcos, CA; available online from http://www.csusm.edu/mwhittle/research/publications.html

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 381–396

COMPUTATION FOR MULTIDIMENSIONAL CAUCHY PROBLEM*

T. WEI†, Y. C. HON‡, AND J. CHENG§

Abstract. This paper devises a computational method for solving a Cauchy problem of Laplace's equation in multidimensional space. By using the Green formula, the Cauchy problem is transformed to a moment problem so that numerical computations using a regularization technique can be achieved. A stability estimation and a suitable choice of regularization parameter for the proposed method are also given. For numerical verification, a numerical example in the three-dimensional case is presented.

Key words. numerical computation, Cauchy problem, ill-posedness, multidimension

AMS subject classifications. 65N15, 65M30, 35R25

PII. S0363012901389391

1. Introduction.

Let $\Omega \subset \mathcal{R}^{d+1}, d \in \mathcal{N}$, be a simply connected domain with Lipschitz continuous boundary $\partial\Omega$. Assume that Γ is an open part of the boundary $\partial\Omega$. Without loss of generality, Γ is also assumed to be connected. Consider the following multidimensional Cauchy problem of Laplace's equation:

$$\Delta u(x) = 0, \qquad x \in \Omega, \tag{1.1}$$

$$u(x) = f(x), \qquad x \in \Gamma, \tag{1.2}$$

$$\frac{\partial u(x)}{\partial \nu} = g(x), \qquad x \in \Gamma, \tag{1.3}$$

where Δ is the $(d+1)$-dimensional Laplacian operator and ν is the unit outward normal with respect to $\partial\Omega$.

This is the classical Cauchy problem of Laplace's equation which arises from many real applications such as nondestructive testing techniques [1], [5], [6], [10], [12], geophysics [23], and cardiology [13]. The Cauchy problem is known to be highly ill-posed. That is, any small change in the initial data may result in a dramatic change in the solution (see page 16 in [7] and Chapter 2 in [17]). Under an additional a priori boundedness condition, a continuous dependence of the solution on the initial data can be obtained. This is called conditional stability [20]. Other conditional stability for Laplace's equation can be found in [11].

Due to the highly ill-posedness of the problem, numerical computation is very difficult. To the knowledge of the authors, there is still no numerical solution to the Cauchy problem (1.1)–(1.3) in three- or higher-dimensional cases. To obtain a stable numerical solution for these kinds of ill-posed problems, several regularization

*Received by the editors May 15, 2001; accepted for publication (in revised form) April 5, 2002; published electronically May 12, 2003. This work was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (project CityU 1178/02P). The first and third authors are partially supported by the NSF of China (grants 1027 1050 and 1027 1032).

http://www.siam.org/journals/sicon/42-2/38939.html

†Department of Mathematics, City University of Hong Kong, Hong Kong SAR, China, and Department of Mathematics, Lanzhou University, Lanzhou, Gansu Province 730000, China (tingwei@lzu.edu.cn).

‡Department of Mathematics, City University of Hong Kong, Hong Kong SAR, China (maychon@cityu.edu.hk).

§Department of Mathematics, Fudan University, Shanghai 200433, China (jcheng@fudan.edu.cn).

methods have been proposed: The quasi-reversibility method [18], [19], the Tikhonov regularization method [3], [23], the boundary element method [15], and discretization [14], [21], [8]. Recently, Cheng et al. [9] and Hon and Wei [16] proposed a new computational method for solving the Cauchy problem for Laplace's equation (1.1)–(1.3) in a two-dimensional case through a transformation of the Cauchy problem to a moment problem whose numerical approximation can be achieved.

This paper further extends the method to solve a multidimensional Cauchy problem for Laplace's equation. The proposed method is similar to the existing boundary element method for solving elliptic equations. From the numerical solution of the moment problem, the boundary values of the solution on $\partial\Omega \setminus \Gamma$ are determined. The interior values of the solution can then be obtained after solving a well-posed mixed boundary value problem for Laplace's equation. A stability estimation and a suitable choice of regularization parameter for the proposed method are also given. For numerical verification, a numerical example is presented at the end of section 6.

2. Methodology. Let v satisfy

$$\Delta v(x) = 0, \qquad x \in \Omega, \tag{2.1}$$

$$\frac{\partial v(x)}{\partial \nu} = 0, \qquad x \in \Gamma_1, \tag{2.2}$$

where $\Gamma_1 = \partial\Omega \setminus \Gamma$.

Using the Green formula, we have

$$\int_\Omega (v\Delta u - u\Delta v)d\sigma = \int_{\partial\Omega} \left(v\frac{\partial u}{\partial \nu} - u\frac{\partial v}{\partial \nu} \right) ds, \tag{2.3}$$

where $d\sigma$ and ds represent the volume and area elements, respectively.

Since u is the solution of (1.1)–(1.3) and v is the solution of (2.1)–(2.2), the identity (2.3) can be rewritten as

$$\int_{\Gamma_1} v\frac{\partial u}{\partial \nu} ds = \int_\Gamma \left(f\frac{\partial v}{\partial \nu} - gv \right) ds. \tag{2.4}$$

Denote

$$\mathcal{H} = \{v \mid v \text{ satisfies (2.1)–(2.2)}\}. \tag{2.5}$$

In our recent work [9], [16], we have the following result.

PROPOSITION 2.1. *If the Cauchy problem* (1.1)–(1.3) *has a solution u such that $\frac{\partial u}{\partial \nu}|_{\Gamma_1} \in L^2(\Gamma_1)$, then $\beta = \frac{\partial u}{\partial \nu}|_{\Gamma_1}$ is the unique solution satisfying the following moment problem:*

$$\int_{\Gamma_1} v\beta ds = \int_\Gamma \left(f\frac{\partial v}{\partial \nu} - gv \right) ds \equiv \mu_v(f, g), \tag{2.6}$$

where $v \in \mathcal{H}$.

Conversely if $\beta \in L^2(\Gamma_1)$ is the solution of (2.6)*, then there exists a unique solution u of the Cauchy problem* (1.1)–(1.3) *such that $\frac{\partial u}{\partial \nu}|_{\Gamma_1} = \beta \in L^2(\Gamma_1)$.*

For numerical computational purposes, we choose $\{v_j\}_{j=0}^\infty \subset \mathcal{H}$ such that

$$\overline{Span\{v_j|_{\Gamma_1}\}_{j=0}^\infty} = L^2(\Gamma_1).$$

The moment problem (2.6) becomes

$$\int_{\Gamma_1} v_j \beta ds = \mu_j, \qquad j = 0, 1, 2, \ldots, \tag{2.7}$$

where $\mu_j = \int_\Gamma (f \frac{\partial v_j}{\partial \nu} - g v_j) ds$ and v_j are functions satisfying (2.1)–(2.2). It is noted here that each μ_j is determined uniquely from f, g, and $v_j \in \mathcal{H}$. In the following sections, it will be shown that the basis functions v_j $(j = 0, 1, 2, \ldots)$ can be suitably chosen to give a stable and satisfactory numerical approximation for the solution of the Cauchy problem.

3. Choices of basis functions in the space $\mathcal{H}$. A constructional method for obtaining a set of basis functions from the space $\mathcal{H}$ is given in the following proposition.

PROPOSITION 3.1. *Given is the following special Cauchy problem for Laplace's equation:*

$$\Delta v(x_1, x_2, \ldots, x_d, x_{d+1}) = 0, \qquad x = (x_1, x_2, \ldots, x_{d+1}) \in \mathcal{R}^{d+1}, \tag{3.1}$$

$$\frac{\partial v}{\partial x_{d+1}}(x_1, \ldots, x_d, 0) = 0, \tag{3.2}$$

$$v(x_1, \ldots, x_d, 0) = x_1^{k_1} x_2^{k_2} \cdots x_d^{k_d}, \tag{3.3}$$

where $k_1, k_2, \ldots, k_d$ are fixed numbers in $\mathcal{N} \cup \{0\}$. There exists a homogeneous polynomial in $\mathcal{R}^{d+1}$ that satisfies (3.1)–(3.3) *of the form*

$$v(x) = x_1^{k_1} x_2^{k_2} \ldots x_d^{k_d} + \sum_{n=2}^{s} P_{s-n}(x_1, x_2, \ldots, x_d) x_{d+1}^n, \tag{3.4}$$

where, for $s = k_1 + k_2 + \cdots + k_d$ and $n = 2, 3, \ldots, s$,

$$P_{s-n} = \begin{cases} 0 & \text{when } n \text{ is odd,} \\ (-1)^{\frac{n}{2}} \frac{1}{n!} \triangle^{\frac{n}{2}} (x_1^{k_1} x_2^{k_2} \cdots x_d^{k_d}) & \text{when } n \text{ is even.} \end{cases} \tag{3.5}$$

Proof. Denote $s = k_1 + k_2 + \cdots + k_d$ to be the order of the polynomial in the right-hand side of (3.3). For the cases in which $s = 0, 1$, it is clear that $v(x) = x_1^{k_1} x_2^{k_2} \cdots x_d^{k_d}$ satisfies (3.1)–(3.3). For the cases in which $s \geq 2$, it is easy to obtain from the boundary conditions (3.2) and (3.3) that the solution $v(x)$ is the following homogeneous polynomial of order s:

$$v(x) = x_1^{k_1} x_2^{k_2} \cdots x_d^{k_d} + \sum_{n=2}^{s} P_{s-n}(x_1, x_2, \ldots, x_d) x_{d+1}^n, \tag{3.6}$$

where $P_m(x_1, x_2, \ldots, x_d)$ is a homogeneous polynomial of $x_1, x_2, \ldots, x_d$ with order $m, m = 0, 1, \ldots, s-2$. Denote $w = x_1^{k_1} x_2^{k_2} \cdots x_d^{k_d}$. We then have

$$\begin{aligned} \Delta v(x) = {} & (\Delta w + 2P_{s-2}) + (0 + 3 \cdot 2 P_{s-3}) x_{d+1} \\ & + \sum_{n=2}^{s-2} (\triangle P_{s-n} + (n+2)(n+1) P_{s-n-2}) x_{d+1}^n. \end{aligned}$$

From (3.1), the homogeneous polynomials $P_m, m = 0, 1, \ldots, s-2$, must satisfy the following recursive formula:

$$\triangle w + 2P_{s-2} = 0, \tag{3.7}$$

$$0 + 3 \cdot 2 P_{s-3} = 0, \tag{3.8}$$

$$\triangle P_{s-n} + (n+2)(n+1) P_{s-n-2} = 0, \quad n = 2, 3, \ldots, s-2, \tag{3.9}$$

for $s \geq 4$. In the case when $s = 2$, only (3.7) is valid. When $s = 3$, only (3.7) and (3.8) are valid. From (3.7)–(3.9), we can deduce that, for $n = 2, 3, \ldots, s$,

$$P_{s-n} = \begin{cases} 0 & \text{when } n \text{ is odd,} \\ (-1)^{\frac{n}{2}} \frac{1}{n!} \triangle^{\frac{n}{2}} w & \text{when } n \text{ is even.} \end{cases} \tag{3.10}$$

Substituting (3.10) into (3.6), we have then constructed a homogeneous polynomial of order s that satisfies the Cauchy problem (3.1)–(3.3). ∎

Henceforth we denote the polynomial solution (3.6) for given $k_1, k_2, \ldots, k_d$ as $v_{k_1 k_2 \cdots k_d}(x)$, $\quad x = (x_1, \ldots, x_{d+1})$. For the three-dimensional case ($d = 2$), several basis functions will be given in section 6.

PROPOSITION 3.2. *For $k_1, k_2, \ldots, k_d \in \mathcal{N} \cup \{0\}$, we have*

$$|v_{k_1, k_2, \ldots, k_d}(x)| \leq (2M)^s, \qquad x \in \bar{\Omega}, \tag{3.11}$$

and

$$\left| \frac{\partial v_{k_1, k_2, \ldots, k_d}(x)}{\partial x_i} \right| \leq (4M)^s, \qquad i = 1, 2, \ldots, d+1, \tag{3.12}$$

where $M = \sup_{x \in \bar{\Omega}} |x|$ and $s = k_1 + k_2 + \cdots + k_d$.

Proof. Since the proof will be similar in the general dimensional cases, we shall give the proof of (3.11) only for the case in which $d = 2$. For $d = 2$, we simply denote $(k_1, k_2) = (i, j)$. By using the formula (3.6) and (3.10) in Proposition 3.1, we get

$$v_{ij}(x) = w - \frac{1}{2!}(\triangle w)x_3^2 + \frac{1}{4!}(\triangle^2 w)x_3^4 + \cdots + (-1)^N \frac{1}{(2N)!}(\triangle^N w)x_3^{2N}, \tag{3.13}$$

where $w = x_1^i x_2^j$ and $N = [\frac{i+j}{2}]$ with $[a]$ denoting the largest integer less or equal to a.

Since

$$\begin{aligned} \triangle^n w &= \left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} \right)^n x_1^i x_2^j \\ &= \sum_{r=0}^{n} C_n^r \left(\frac{\partial^2}{\partial x_1^2} \right)^r x_1^i \left(\frac{\partial^2}{\partial x_2^2} \right)^{n-r} x_2^j \\ &= \sum_{r=0}^{n} C_n^r C_i^{2r} C_j^{2n-2r} (2r)!(2n-2r)! x_1^{i-2r} x_2^{j-2(n-r)}, \end{aligned} \tag{3.14}$$

where C_n^r is the usual combinatorial notation, here, $C_n^r = 0$ for $r > n$. Since $C_n^r \leq C_{2n}^{2r}$, we obtain

$$\frac{1}{(2n)!} |\triangle^n w| \leq \sum_{r=0}^{n} C_i^{2r} C_j^{2n-2r} M^{i+j-2n}. \tag{3.15}$$

From (3.13) and (3.15), we then have

$$|v_{ij}| \leq \left(1 + \sum_{n=1}^{N} \sum_{r=0}^{n} C_i^{2r} C_j^{2n-2r} \right) M^{i+j} \tag{3.16}$$

$$\leq 2^i 2^j M^{i+j}. \tag{3.17}$$

A similar proof for (3.12) can be obtained. The proof of the proposition is then complete. □

In the next section, a numerical algorithm with convergence analysis for solving the moment problem is given. The numerical approach will be shown to be both accurate and efficient in section 6.

4. Algorithm for solving the moment problem. Let $\Omega \subset \mathcal{R}_+^{d+1}$ be a bounded and simply connected Lipschitz domain and

$$\Gamma_1 = \partial\Omega \setminus \Gamma = \{\, x \,|\, x_{d+1} = 0,\ 0 \le x_i \le 1,\ \ i = 1, 2, \ldots, d\}, \tag{4.1}$$

where Γ is an open part of the boundary $\partial\Omega$. It is noted here that Γ can be an arbitrary Lipschitz boundary. This makes the proposed method feasible and robust in solving real practical inverse problems arising from the physical world. We choose $v_{k_1 k_2 \cdots k_d}(x)$ as given in Propositions 3.1 and 3.2 for all $k_1, k_2, \ldots, k_d$ in $\mathcal{N} \cup \{0\}$. Since they all satisfy (3.1)–(3.3), we obtain from (2.7) that the Cauchy problem for Laplace's equation (1.1)–(1.3) is equivalent to the following moment problem:

$$\int_{\Gamma_1} x_1^{k_1} \cdots x_d^{k_d} \beta(x_1, \ldots, x_d) dx_1 \cdots dx_d = \mu_{k_1 \cdots k_d} \quad \forall k_1, k_2, \ldots, k_d \in \mathcal{N} \cup \{0\}, \tag{4.2}$$

where $\Gamma_1 = [0, 1]^d$ and

$$\mu_{k_1 \cdots k_d} = \int_\Gamma \left(f \frac{\partial v_{k_1 \cdots k_d}}{\partial \nu} - g v_{k_1 \cdots k_d} \right) ds. \tag{4.3}$$

This moment problem is called the Hausdorff moment problem and has been studied by many researchers (see, e.g., [4], [22], [24], [25]). Particularly, a practical algorithm for the numerical computation of the Hausdorff moment problem in the two-dimensional case can be found in the works of Talenti [22] and Viano [24]. The numerical method for multidimensional Hausdorff moment problems can be found in Ang, Gorenflo, and Trong [2].

Based on the numerical method given by Ang, Gorenflo, and Trong [2], an approximated solution for the moment problem (4.2) can be obtained from the following numerical steps:

Step 1. Calculate the shifted Legendre polynomials and their coefficients:

$$L_m(x) = \sum_{j=0}^{m} C_{mj} x^j, \qquad m = 0, 1, 2, \ldots, \tag{4.4}$$

where

$$C_{mj} = (2m+1)^{\frac{1}{2}} (-1)^j \frac{(m+j)!}{(j!)^2 (m-j)!}, \quad j = 0, 1, \ldots, m. \tag{4.5}$$

Step 2. Calculate the coefficients:

$$\lambda_{k_1 \cdots k_d} = \sum_{p_1=0}^{k_1} \cdots \sum_{p_d=0}^{k_d} C_{k_1 p_1} \cdots C_{k_d p_d} \mu_{p_1 \cdots p_d}. \tag{4.6}$$

Step 3. Calculate the approximated solution:

$$p_N(x_1, \ldots, x_d) = \sum_{k_1, \ldots, k_d = 0}^{k_1 + \cdots + k_d = N} \lambda_{k_1 \cdots k_d} L_{k_1 \cdots k_d}, \tag{4.7}$$

where the orthonormal polynomials $L_{k_1\cdots k_d}$ are defined by

$$L_{k_1\cdots k_d}(x_1,\ldots,x_d) = L_{k_1}(x_1)\cdots L_{k_d}(x_d)$$
$$= \sum_{p_1=0}^{k_1}\cdots\sum_{p_d=0}^{k_d} C_{k_1p_1}\cdots C_{k_dp_d}x_1^{p_1}\cdots x_d^{p_d}.$$

The p_N given in Step 3 then approximates the solution β of the moment problem.

Remark 4.1. The formula given in Step 3 is a little different from the one given in Ang, Gorenflo, and Trong [2].

THEOREM 4.2. *Let $\mu = \{\mu_{k_1\cdots k_d}\}, k_i \in \mathcal{N}\cup\{0\}, i = 1,2,\ldots,d$, be a sequence of real numbers obtained from* (4.3)*, and p_N is given by Step* 3*. Suppose that u is the solution of the Cauchy problem* (1.1)–(1.3)*. Denote $\beta = \frac{\partial u}{\partial \nu}|_{\Gamma_1}$. Then we have*

$$\sum_{k_1,\ldots,k_d=0}^{\infty}\left(\sum_{p_1,\ldots,p_d=0}^{\infty} C_{k_1p_1}\cdots C_{k_dp_d}\mu_{p_1\cdots p_d}\right)^2 \le \infty \tag{4.8}$$

and

$$p_N \to \beta \quad in \quad L^2(\Gamma_1) \quad as \quad n\to\infty, \tag{4.9}$$

where C_{ij} is defined by (4.5) *when $j \le i$ and $C_{ij} = 0$ when $j > i$. Moreover, if β is in $H^1(\Gamma_1)$, then*

$$\|p_N - \beta\| \le \frac{\sqrt{d}}{N+1}(F(\beta))^{\frac{1}{2}}, \qquad N\in\mathcal{N}, \tag{4.10}$$

where $\|\cdot\|$ is the usual $L^2(\Gamma_1)$ norm and

$$F(\beta) = \sum_{i=1}^{d}\int_{\Gamma_1} x_i(1-x_i)\left|\frac{\partial\beta}{\partial x_i}\right|^2 dx_1\cdots dx_d. \tag{4.11}$$

Proof. For simplicity, we give the proof in the case when $d = 2$. For notational convenience, denote $(x_1,x_2) = (x,y)$, $(k_1,k_2) = (i,j)$, $\mu_{k_1k_2} = \mu_{ij}$, $L_{k_1k_2}(x,y) = L_i(x)L_j(y)$, and $\Gamma_1 = [0,1]^2$. From the formula (4.2) and (4.6), we have

$$\int_{\Gamma_1}\beta L_{ij}dxdy = \lambda_{ij}, \qquad i,j = 0,1,\ldots. \tag{4.12}$$

Using the completeness and the orthonormality properties of $\{L_{ij}\}$ in $L^2(\Gamma_1)$, we obtain

$$\beta = \sum_{i,j=0}^{\infty}\lambda_{ij}L_{ij} \tag{4.13}$$

and

$$\|\beta\|^2 = \sum_{i,j=0}^{\infty}|\lambda_{ij}|^2 \le \infty,$$

and hence (4.8) holds. To prove (4.9), we subtract (4.7) from (4.13) to obtain

$$\beta - p_N = \sum_{i+j \geq N+1} \lambda_{ij} L_{ij}. \tag{4.14}$$

Hence

$$\|\beta - p_N\|^2 = \sum_{i+j \geq N+1} |\lambda_{ij}|^2. \tag{4.15}$$

Combining (4.8) and (4.15), we then have (4.9). For (4.10), we rely on the following identity (cf. [2], [22]):

$$\int_0^1 x(1-x)|v'(x)|^2 dx = \sum_{k=0}^{\infty} k(k+1)\alpha_k^2 \qquad \forall v \in H^1(0,1), \tag{4.16}$$

where $\alpha_k = \int_0^1 v(x) L_k(x) dx$. From (4.13) and (4.16), we then have

$$\sum_{i,j=0}^{\infty} i^2 \lambda_{ij}^2 \leq \int_{\Gamma_1} x(1-x) \left| \frac{\partial \beta}{\partial x}(x,y) \right|^2 dxdy \tag{4.17}$$

and

$$\sum_{i,j=0}^{\infty} j^2 \lambda_{ij}^2 \leq \int_{\Gamma_1} y(1-y) \left| \frac{\partial \beta}{\partial y}(x,y) \right|^2 dxdy. \tag{4.18}$$

Adding (4.17) to (4.18), we obtain

$$\sum_{i,j=0}^{\infty} (i^2 + j^2) \lambda_{ij}^2 \leq F(\beta), \tag{4.19}$$

where $F(\beta)$, defined in (4.11), becomes

$$F(\beta) = \int_{\Gamma_1} \left(x(1-x) \left| \frac{\partial \beta}{\partial x}(x,y) \right|^2 + y(1-y) \left| \frac{\partial \beta}{\partial y}(x,y) \right|^2 \right) dxdy. \tag{4.20}$$

Since

$$2(i^2 + j^2) \geq (N+1)^2 \quad \text{for} \quad i + j \geq N+1, \tag{4.21}$$

we finally obtain from (4.15), (4.19), and (4.21) that

$$(N+1)^2 \|p_N - \beta\|^2 \leq 2 \sum_{i+j \geq N+1} (i^2 + j^2) \lambda_{ij}^2 \leq 2F(\beta), \tag{4.22}$$

which then completes the proof. □

5. Stability estimation and choice of regularization parameter. Taking the highly ill-posedness characteristic of the Cauchy problem into consideration, we now assume that the Cauchy data f and g contain some errors.

Let $\widetilde{f}$ and $\widetilde{g}$ be measured data with total error less than ε as

$$\|\widetilde{f} - f\|_{L^2(\Gamma)} + \|\widetilde{g} - g\|_{L^2(\Gamma)} \leq \varepsilon. \tag{5.1}$$

The corresponding moments for $\widetilde{f}$ and $\widetilde{g}$ are

$$\widetilde{\mu}_{k_1\cdots k_d} = \int_\Gamma \left(\widetilde{f} \frac{\partial v_{k_1\cdots k_d}}{\partial \nu} - \widetilde{g} v_{k_1\cdots k_d} \right) ds. \tag{5.2}$$

We have the following error estimate about the moments:

$$\begin{aligned} |\widetilde{\mu}_{k_1\cdots k_d} - \mu_{k_1\cdots k_d}| &\leq \left[\left(\int_\Gamma v_{k_1\cdots k_d}^2 ds \right)^{\frac{1}{2}} + \left(\int_\Gamma \left(\frac{\partial v_{k_1\cdots k_d}}{\partial \nu} \right)^2 ds \right)^{\frac{1}{2}} \right] \varepsilon \\ &= D_{k_1\cdots k_d}\, \varepsilon. \end{aligned} \tag{5.3}$$

Using the choice of $v_{k_1\cdots k_d}$ given in Propositions 3.1 and 3.2, we then have

$$D_{k_1\cdots k_d} \leq 2\sqrt{vol(\Gamma)} A^{k_1+\cdots+k_d} = B A^{k_1+\cdots+k_d}, \tag{5.4}$$

where $A = 4M > 0$ is a constant depending only on Ω.

The approximated solution with noisy data $\widetilde{f}$ and $\widetilde{g}$ is given by

$$\widetilde{p}_N(x_1, \ldots, x_d) = \sum_{k_1,\ldots,k_d=0}^{k_1+\cdots+k_d=N} \widetilde{\lambda}_{k_1\cdots k_d} L_{k_1\cdots k_d}, \tag{5.5}$$

where

$$\widetilde{\lambda}_{k_1\cdots k_d} = \sum_{p_1=0}^{k_1} \cdots \sum_{p_d=0}^{k_d} C_{k_1 p_1} \cdots C_{k_d p_d} \widetilde{\mu}_{p_1\cdots p_d}. \tag{5.6}$$

We can obtain the following stability estimation.

THEOREM 5.1. *Suppose that u is the solution of the Cauchy problem* (1.1)–(1.3). *Denote $\beta = \frac{\partial u}{\partial \nu}|_{\Gamma_1}$. If $\beta \in H^1(\Gamma_1)$, then*

$$\|\beta - \widetilde{p}_N\|_{L^2(\Gamma_1)} \leq \varepsilon B A^N (\sqrt{2\pi})^{-d} (3 + 2\sqrt{2})^{dN+d} + \sqrt{d}(N+1)^{-1} (F(\beta))^{\frac{1}{2}}, \tag{5.7}$$

where $F(\beta)$ is given by (4.11).

Proof. We give a proof only for the case in which $d = 2$. The proof for the general dimension is similar. Since

$$\|\widetilde{p}_N - \beta\| \leq \|p_N - \beta\| + \|p_N - \widetilde{p}_N\|, \tag{5.8}$$

where $\|\cdot\|$ is the usual L^2 norm, from (4.7) and (5.5) we have

$$p_N - \widetilde{p}_N = \sum_{i,j=0}^{i+j=N} \left(\sum_{p=0}^{i} \sum_{q=0}^{j} C_{ip} C_{jq} (\mu_{pq} - \widetilde{\mu}_{pq}) \right) L_i L_j, \tag{5.9}$$

and hence

$$\|p_N - \widetilde{p}_N\|^2 = \sum_{i,j=0}^{i+j=N} \left(\sum_{p=0}^{i} \sum_{q=0}^{j} C_{ip} C_{jq} (\mu_{pq} - \widetilde{\mu}_{pq}) \right)^2. \tag{5.10}$$

From (5.3) and (5.4), we have

$$|\mu_{pq} - \widetilde{\mu}_{pq}| \le \varepsilon B A^{p+q}.$$

Hence we have

$$\begin{aligned} \|p_N - \widetilde{p}_N\|^2 &\le \varepsilon^2 B^2 A^{2N} \sum_{i,j=0}^{i+j=N} \left(\sum_{p=0}^{i} \sum_{q=0}^{j} |C_{ip} C_{jq}| \right)^2 \\ &\le \varepsilon^2 B^2 A^{2N} \left(\sum_{i=0}^{N} \left(\sum_{p=0}^{i} |C_{ip}| \right)^2 \right)^2. \end{aligned} \tag{5.11}$$

We then have (cf. [2, p. 19])

$$\sum_{i=0}^{N} \left(\sum_{p=0}^{i} |C_{ip}| \right)^2 \le (2\pi)^{-1} (3 + 2\sqrt{2})^{2N+2}. \tag{5.12}$$

Substituting this inequality into (5.11), we have

$$\|p_N - \widetilde{p}_N\| \le \varepsilon B (2\pi)^{-1} A^N (3 + 2\sqrt{2})^{2N+2}. \tag{5.13}$$

The proof is then completed by using (4.10), (5.8), and (5.13). □

Theorem 5.1 implies that if $\varepsilon \neq 0$ and N tends to infinity, then the right-hand side of (5.7) will tend to infinity even when ε is very small. This indicates that we have to choose a suitable value for N so that the right-hand side of (5.7) is kept small. This is one kind of regularization technique, with N being the regularization parameter (cf. [2], [3]). The choice of N is given in the following corollary.

COROLLARY 5.2. *Suppose that*

$$f(t) = A^t (3 + 2\sqrt{2})^{td} \tag{5.14}$$

and

$$N(\varepsilon) = [f^{-1}(\varepsilon^{-\frac{1}{2}})] = \left[\frac{\ln(\frac{1}{\varepsilon})}{2 \ln A + 2d \ln(3 + 2\sqrt{2})} \right]. \tag{5.15}$$

If $\beta \in H^1(\Gamma_1)$, we have

$$\|\widetilde{p}_N - \beta\|_{L^2(\Gamma_1)} \le C_1 \varepsilon^{\frac{1}{2}} + C_2 \frac{1}{\ln \frac{1}{\varepsilon}}, \tag{5.16}$$

where

$$C_1 = B \left(\frac{3 + 2\sqrt{2}}{\sqrt{2\pi}} \right)^d \qquad \text{and} \qquad C_2 = 2\sqrt{d} |F(\beta)|^{\frac{1}{2}} (\ln A + d \ln(3 + 2\sqrt{2})).$$

Proof. When $N(\varepsilon)$ is given as in (5.15), the first term of the right-hand side of (5.7) is $C_1\varepsilon^{\frac{1}{2}}$. Since

$$N(\varepsilon)+1 \geq \frac{\ln(\frac{1}{\varepsilon})}{2\ln A + 2d\ln(3+2\sqrt{2})}, \tag{5.17}$$

we have

$$\frac{1}{N(\varepsilon)+1} \leq \frac{2\ln A + 2d\ln(3+2\sqrt{2})}{\ln(\frac{1}{\varepsilon})}. \tag{5.18}$$

Hence (5.16) holds. ∎

Consider the following boundary value problem:

$$\Delta\widetilde{u}_N = 0 \qquad \text{in} \quad \Omega, \tag{5.19}$$

$$\frac{\partial\widetilde{u}_N}{\partial\nu} = \widetilde{p}_N \qquad \text{on} \quad \Gamma_1, \tag{5.20}$$

$$\widetilde{u}_N = \widetilde{f} \qquad \text{on} \quad \Gamma, \tag{5.21}$$

where we suppose that $\widetilde{f} \in H^1(\Omega)$. We also assume that u is the solution of the Cauchy problem (1.1)–(1.3) for which the Cauchy data f in (1.3) is in $H^1(\Omega)$. From the results given in [9], there exists a unique solution for the boundary value problem (5.19)–(5.21) satisfying

$$\|\widetilde{u}_N - u\|_{H^1(\Omega)} \leq C\left(\left\|\widetilde{p}_N - \frac{\partial u}{\partial\nu}\right\|_{L^2(\Gamma_1)} + \|\widetilde{f} - f\|_{H^1(\Omega)}\right), \tag{5.22}$$

where $C>0$ is a constant which depends on Ω and Γ. Therefore, we have the following theorem.

THEOREM 5.3. *Under the assumptions given in Theorem* 5.1 *and the condition*

$$\|\widetilde{f} - f\|_{H^1(\Omega)} + \|\widetilde{g} - g\|_{L^2(\Gamma_1)} \leq \varepsilon, \tag{5.23}$$

there exists a constant $C>0$ which depends on Ω and Γ such that

$$\begin{aligned}\|\widetilde{u}_N - u\|_{L^2(\Omega)} \leq{}& C(\varepsilon + \varepsilon B(\sqrt{2\pi})^{-d}A^N(3+2\sqrt{2})^{dN+d} \\ &+ \sqrt{d}(F(\beta))^{\frac{1}{2}}(N+1)^{-1}).\end{aligned} \tag{5.24}$$

COROLLARY 5.4. *Suppose that*

$$N(\varepsilon) = \left[\frac{\ln(\frac{1}{\varepsilon})}{2\ln A + 2d\ln(3+2\sqrt{2})}\right]. \tag{5.25}$$

There exists a constant $C>0$ which depends on $\|\beta\|_{H^1(\Gamma_1)}$, Ω, and Γ such that

$$\|\widetilde{u}_N(x) - u(x)\|_{L^2(\Omega)} \leq \frac{C}{|\ln\varepsilon|}. \tag{5.26}$$

6. Numerical example. In this section, a numerical example is performed to verify the accuracy and efficiency of the proposed method. Let

$$\Omega = \{(x_1, x_2, x_3) | \quad 0 \le x_i \le 1, \quad i = 1, 2, 3\} \tag{6.1}$$

and

$$\Gamma_1 = \partial\Omega \backslash \Gamma = \{\, x \,|\, 0 \le x_1 \le 1, \quad 0 \le x_2 \le 1, x_3 = 0\}. \tag{6.2}$$

The basis functions are chosen to be

$$\begin{aligned}
v_{00} &= 1, \\
v_{10} &= x_1, \\
v_{01} &= x_2, \\
v_{20} &= x_1^2 - x_3^2, \\
v_{11} &= x_1 x_2, \\
v_{02} &= x_2^2 - x_3^2, \\
v_{30} &= x_1^3 - 3x_1 x_3^2, \\
v_{03} &= x_2^3 - 3x_2 x_3^2, \\
v_{21} &= (x_1^2 - x_3^2) x_2, \\
v_{12} &= (x_2^2 - x_3^2) x_1, \\
v_{22} &= (6x_1^2 x_2^2 - 6x_1^2 x_3^2 - 6x_2^2 x_3^2 + 2x_3^4)/6, \\
v_{40} &= x_1^4 - 6x_1^2 x_3^2 + x_3^4, \\
v_{04} &= x_2^4 - 6x_2^2 x_3^2 + x_3^4, \\
v_{31} &= x_1^3 x_2 - 3x_1 x_2 x_3^2, \\
v_{13} &= x_2^3 x_1 - 3x_1 x_2 x_3^2.
\end{aligned}$$

Example. The exact solution of (1.1)–(1.3) is taken to be

$$u(x_1, x_2, x_3) = \frac{1}{\sqrt{(x_1 - \frac{1}{2})^2 + (x_2 - \frac{1}{2})^2 + (x_3 + 1)^2}}.$$

Following the numerical algorithm given in section 4, we can compute f and g as well as $\mu_{ij} = \int_\Gamma (f \frac{\partial v_{ij}}{\partial \nu} - g v_{ij}) ds$, $i, j = 0, 1, 2, \ldots$. The numerical comparisons between the approximation $p_N = p_N(x_1, x_2)$ as $N = 7$ and the exact $\frac{\partial u}{\partial \nu}$ on Γ_1 is shown in Figures 1a and 1b. As indicated in the surface plots in Figures 1a and 1b, respectively, the p_N approximates the $\frac{\partial u}{\partial \nu}$ very well. The error surface between this approximation and the exact solution is shown in Figure 2.

Finally, we added some noise to the Cauchy data f and g by adding a polynomial of order 4 (error order 10^{-4}) to f and g. The numerical errors surface and line plots for the case when $N = 5$ as shown in Figures 3–4c, respectively, indicate that the proposed method is stable and effective in solving the three-dimensional Cauchy problem for Laplace's equation.

Remark 6.1. For numerical convenience, the given example considers the case in which $\Gamma_1 = \{\, x \,|\, x_{d+1} = 0, 0 \le x_i \le 1, \ i = 1, 2, \ldots, d\}$. With minor modification, the method can readily be applied to more general cases. The only difference is that we have to solve a general moment problem in which the basis functions may not be polynomials.

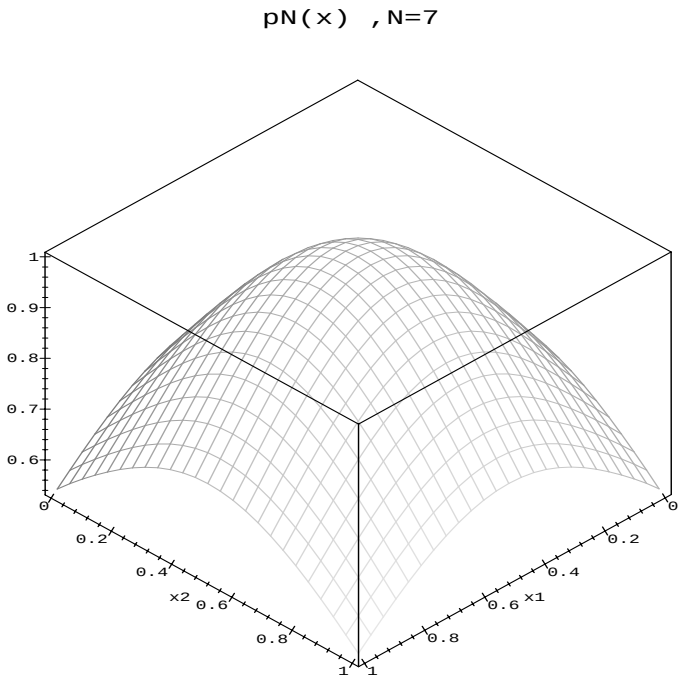


FIG. 1a. *Surface plot of* $p_N(x_1, x_2)$ *for* $N = 7$, $u(x_1, x_2, x_3) = \frac{1}{\sqrt{(x_1-\frac{1}{2})^2+(x_2-\frac{1}{2})^2+(x_3+1)^2}}$.

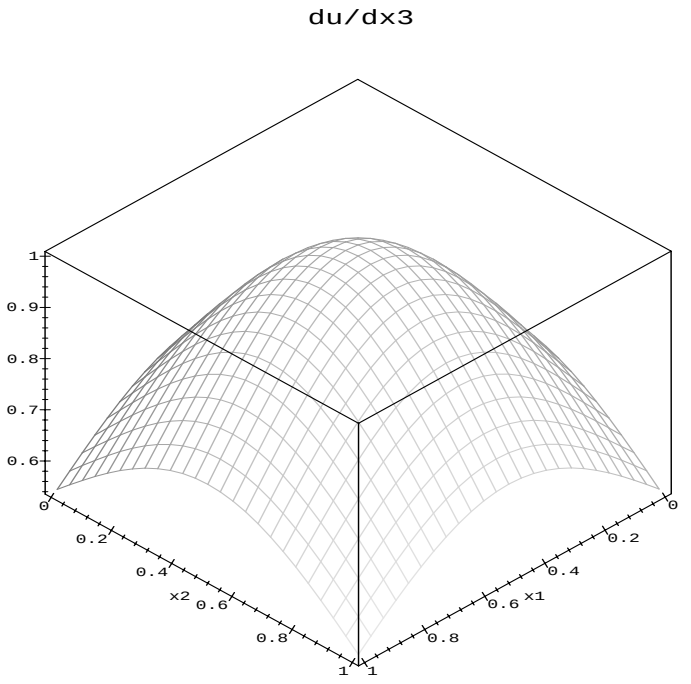


FIG. 1b. *Surface plot of exact* $\beta = \frac{\partial u}{\partial \nu}|_{\Gamma_1}$, $u(x_1, x_2, x_3) = \frac{1}{\sqrt{(x_1-\frac{1}{2})^2+(x_2-\frac{1}{2})^2+(x_3+1)^2}}$.

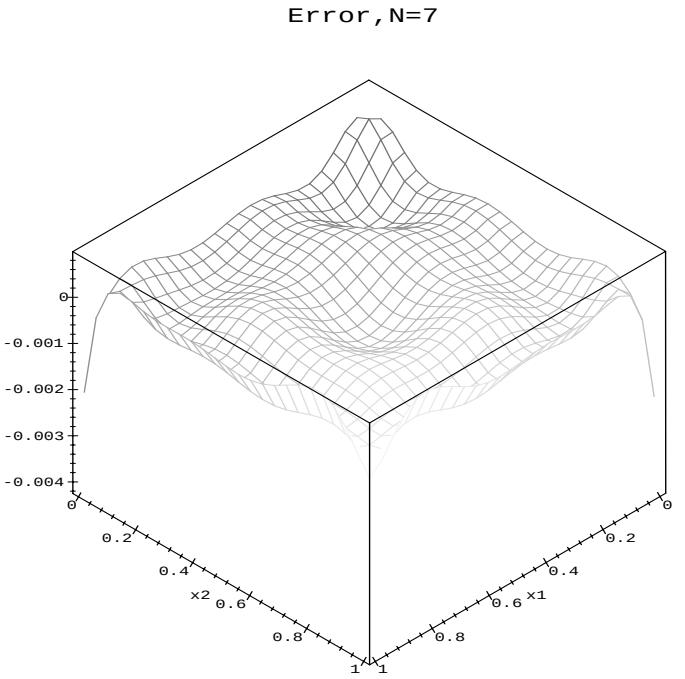


FIG. 2. *Error plot on the difference between* $p_N(x_1, x_2)$ *for* $N = 7$ *and* $\frac{\partial u}{\partial \nu}|_{\Gamma_1}$, $u(x_1, x_2, x_3) = \frac{1}{\sqrt{(x_1-\frac{1}{2})^2+(x_2-\frac{1}{2})^2+(x_3+1)^2}}$.

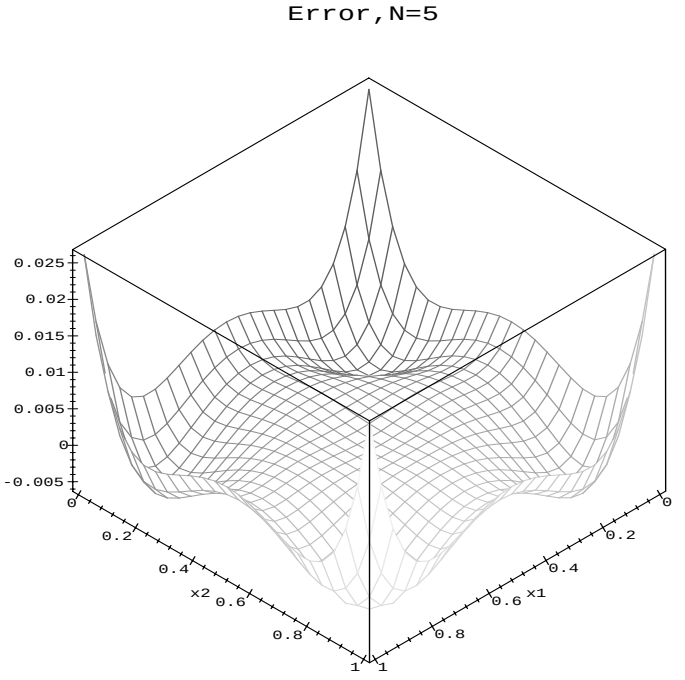


FIG. 3. *Error plot on the difference between* $\widetilde{p}_N(x)$ *for* $N = 5$ *and* $\frac{\partial u}{\partial \nu}|_{\Gamma_1}$ *with noise data* $\widetilde{f}$ *and* $\widetilde{g}$, $\widetilde{f} = f + x_1(1 - x_1)x_2(1 - x_2)10^{-4}$ *on* $x_3 = 1$, $\widetilde{g} = g + x_1(1 - x_1)x_2(1 - x_2)10^{-4}$ *on* $x_3 = 1$.

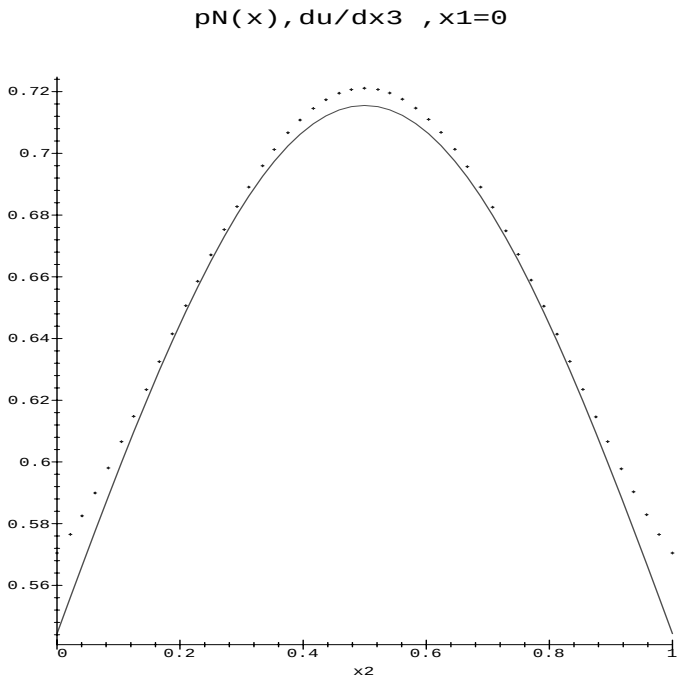


FIG. 4a. *Line plots of* $\frac{\partial u}{\partial \nu}(x_1, x_2)$ *and* $\widetilde{p}_N(x_1, x_2)$ *for* $N = 5$ *and* $x_1 = 0$*; dotted lines: approximated solution; solid lines: exact solution.*

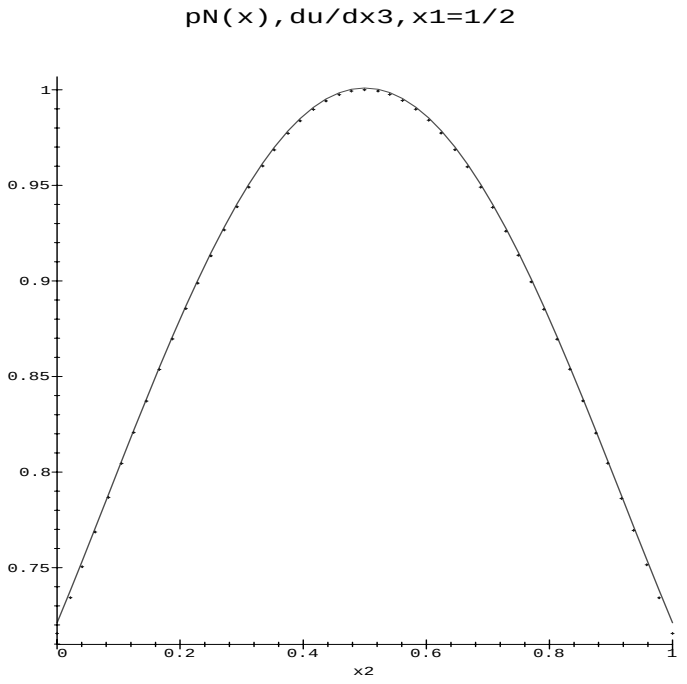


FIG. 4b. *Line plots of* $\frac{\partial u}{\partial \nu}(x_1, x_2)$ *and* $\widetilde{p}_N(x_1, x_2)$ *for* $N = 5$ *and* $x_1 = 1/2$*; dotted lines: approximated solution; solid lines: exact solution.*

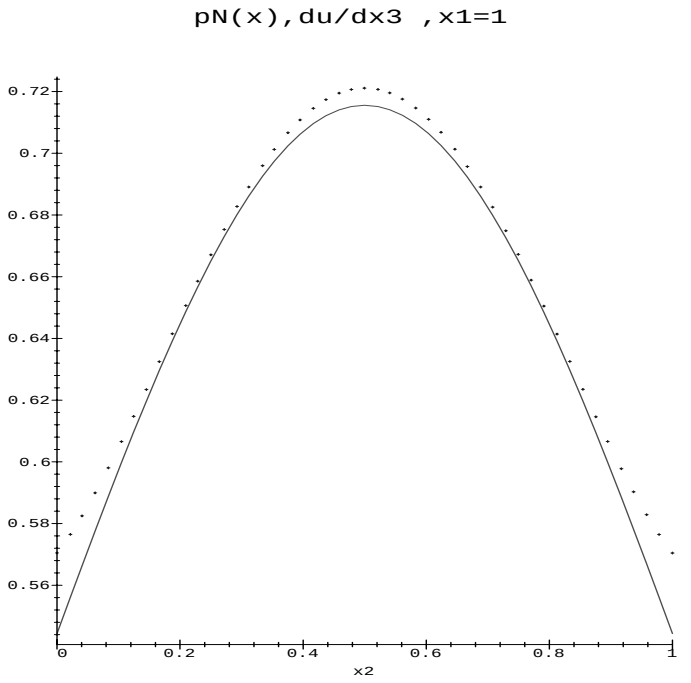


FIG. 4c. *Line plots of* $\frac{\partial u}{\partial \nu}(x_1, x_2)$ *and* $\widetilde{p}_N(x_1, x_2)$ *for* $N = 5$ *and* $x_1 = 1$*; dotted lines: approximated solution; solid lines: exact solution.*

Remark 6.2. The convergence error estimate given in (5.26) is of logarithmical type, which is too weak for efficient numerical simulation. From the theory on ill-posed problems, it is known that, if u is analytic on Γ_1, the estimation can be improved to Hölder type. This can explain the rather good numerical results obtained in section 6.

7. Conclusions. In this paper, a numerical method for solving the multidimensional Cauchy problem for Laplace's equation is proposed. Proofs on the convergence and stability estimation of the method are also given. It is also proven that the regularization parameter for the proposed method can be chosen suitably to give a stable and acceptable approximation to the solution when the Cauchy data have noises. For numerical verification, a numerical example in the three-dimensional case is presented.

Acknowledgment. The authors would like to thank the referees for their careful reading and helpful suggestions on the manuscript.

REFERENCES

[1] G. ALESSANDRINI, *Stable determination of a crack from boundary measurements*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 497–516.

[2] D. D. ANG, R. GORENFLO, AND D. D. TRONG, *A multi-dimensional Hausdorff moment problem: Regularization by finite moments*, Z. Anal. Anwendungen, 18 (1999), pp. 13–25.

[3] D. D. ANG, N. H. NGHIA, AND N. C. TAM, *Regularized solutions of a Cauchy problem for the Laplace equation in an irregular layer: A three-dimensional model*, Acta Math. Vietnam., 23 (1998), pp. 65–74.

[4] J. M. BORWEIN AND A. S. LEWIS, *On the convergence of moment problems*, Trans. Amer. Math. Soc., 325 (1991), pp. 249–271.

[5] A. L. Bukhgeim, J. Cheng, and M. Yamamoto, *Stability for an inverse boundary problem of determining a part of a boundary*, Inverse Problems, 15 (1999), pp. 1021–1032.

[6] A. L. Bukhgeim, J. Cheng, and M. Yamamoto, *On a sharp estimate in a non-destructive testing: Determination of unknown boundaries*, in Applied Electromagnetism and Mechanics, K. Miya, M. Yamamoto, and Nguyen Xuan Hung, eds., Japan Society of Applied Electromagnetism and Mechanics, Tokyo, Japan, 1998, pp. 64–75.

[7] A. Calderón, *Uniqueness in the Cauchy problem for partial differential equations*, Amer. J. Math., 80 (1958), pp. 16–36.

[8] J. R. Cannon and P. DuChateau, *Approximating the solution to the Cauchy problem for Laplace's equation*, SIAM J. Numer. Anal., 14 (1977), pp. 473–483.

[9] J. Cheng, Y. C. Hon, T. Wei, and M. Yamamoto, *Numerical computation of a Cauchy problem for Laplace's equation*, ZAMM Z. Angew. Math. Mech., 81 (2001), pp. 665–674.

[10] J. Cheng, S. Prössdorf, and M. Yamamoto, *Local estimation for an integral equation of first kind with analytic kernel*, J. Inverse Ill-Posed Probl., 6 (1998), pp. 115–126.

[11] J. Cheng and M. Yamamoto, *Unique continuation on a line for harmonic functions*, Inverse Problems, 14 (1998), pp. 869–882.

[12] J. Cheng and M. Yamamoto, *Local stability of a linearized inverse problem in detecting steel reinforcement bars. Proceedings of the International Conference on Inverse Problems and Applications (Quezon City,* 1998*)*, Matimyás Mat., 21 (1998), pp. 18–33.

[13] P. Colli-Franzone, L. Guerri, S. Tentoni, C. Viganotti, S. Baruffi, S. Spaggiari, and B. Taccardi, *A mathematical procedure for solving the inverse potential problem of electrocardiography. Analysis of the time-space accuracy from in vitro experimental data*, Math. Biosci., 77 (1985), pp. 353–396.

[14] R. S. I. Falk and P. B. Monk, *Logarithmic convexity for discrete harmonic functions and the approximation of the Cauchy problem for Poisson's equation*, Math. Comp., 47 (1986), pp. 135–149.

[15] D. N. Hào and D. Lesnic, *The Cauchy problem for Laplace's equation via the conjugate gradient method*, IMA J. Appl. Math., 65 (2000), pp. 199–217.

[16] Y. C. Hon and T. Wei, *Backus-Gilbert algorithm for the Cauchy problem of the Laplace equation*, Inverse Problems, 17 (2001), pp. 261–271.

[17] V. Isakov, *Inverse Problems for Partial Differential Equations*, Springer-Verlag, New York, 1998.

[18] M. V. Klibanov and F. Santosa, *A computational quasi-reversibility method for Cauchy problems for Laplace's equation*, SIAM J. Appl. Math., 51 (1991), pp. 1653–1675.

[19] R. Lattès and J. L. Lions, *The Method of Quasi-Reversibility: Applications to Partial Differential Equations*, Elsevier, New York, 1969.

[20] L. E. Payne, *Bounds in the Cauchy problem for the Laplace's equation*, Arch. Ration. Mech. Anal., 5 (1960), pp. 35–45.

[21] H.-J. Reinhardt, H. Han, and D. N. Hào, *Stability and regularization of a discrete approximation to the Cauchy problem for Laplace's equation*, SIAM J. Numer. Anal., 36 (1999), pp. 890–905.

[22] G. Talenti, *Recovering a function from a finite number of moments*, Inverse Problems, 3 (1987), pp. 501–517.

[23] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, Winston and Sons, Washington, D.C., 1977.

[24] G. A. Viano, *Solution of the Hausdorff moment problem by the use of Pollaczek polynomials*, J. Math. Anal. Appl., 156 (1991), pp. 410–427.

[25] D. V. Widder, *The Laplace Transform*, Princeton University Press, Princeton, NJ, 1941.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 397–421

A PROBLEM OF SEQUENTIAL ENTRY AND EXIT DECISIONS COMBINED WITH DISCRETIONARY STOPPING*

MIHAIL ZERVOS†

Abstract. We consider a stochastic control problem that has emerged in the economics literature as an investment model under uncertainty. This problem combines features of both stochastic impulse control and optimal stopping. The aim is to discover the form of the optimal strategy. It turns out that this has a priori rather unexpected features. The results that we establish are of an explicit nature. We also construct an example whose value function does not possess C^1 regularity.

Key words. stochastic impulse control, optimal switching, optimal stopping, real options

AMS subject classifications. 93E20, 93E03, 62L15, 60G40, 91B70

PII. S036301290038111X

1. Introduction.

Problems that combine features of both stochastic optimal control and optimal stopping have attracted the interest of several researchers. Models of absolutely continuous control of the drift and discretionary stopping have been studied by Krylov [K], Beneš [B], Karatzas and Sudderth [KS], Karatzas and Wang [KW], and Karatzas and Ocone [KO]. Models of combined singular stochastic control where the control effort takes the form of a finite variation process and discretionary stopping have been studied by Davis and Zervos [DZ] and Karatzas, Ocone, Wang, and Zervos [KOWZ]. These two families of problems have been motivated by applications in target tracking, where the controller has to steer a system close to a target and then decide on an engagement time, as well as by applications in finance. The latter ones include the classical consumption/investment problem for a small investor who can decide on the time of their "exit" from the market (see Karatzas and Wang [KW]) as well as the pricing of American contingent claims under constraints or with transaction costs.

In this paper, we consider a problem of stochastic impulse control combined with optimal stopping with a view to discovering the form of the optimal strategy. Note that the impulse control component of the control strategy is not of the standard form because the sizes of the jumps associated with each intervention strategy are not discretionary but are constrained to follow the pattern $\ldots, 1, -1, 1, -1, \ldots$. This simplification makes the problem easier to analyze. However, it is offset by the extra complexity that is introduced by the additional control variable, which is the discretionary stopping.

Problems of this type arise in the context of various applications in which the system dynamics involve discrete actions. For instance, in manufacturing, one needs to choose a machine setup mode over time so as to switch optimally among a finite number of different product types (see Sethi and Zhang [SZ]). The actual motivation of this paper arises from the area of "real options" that has emerged in the economics literature over the past two decades. This area is concerned with the development of new stochastic models that can lead to more accurate pricing of investments in real

*Received by the editors November 14, 2000; accepted for publication (in revised form) September 6, 2002; published electronically May 12, 2003.
http://www.siam.org/journals/sicon/42-2/38111.html
†Department of Mathematics, King's College London, The Strand, London WC2R 2LS, UK (mihail.zervos@kcl.ac.uk).

assets by taking into account the value of managerial flexibility; the interested reader can consult the books by Dixit and Pindyck [DP] and Trigeorgis [T].

To fix ideas, consider an economic activity that is centered on a project that can operate in two modes, an "open" one and a "closed" one. Whenever the project is in its "open" operating mode, it yields a stream of profits or losses that is a functional of the uncertain prices of input and output commodities. Whenever the project is in its "closed" operating mode, it yields neither profits nor losses. The transition of the project from one of its operating modes to the other one forms a sequence of managerial decisions and is associated with certain fixed costs. The problem is to determine the switching strategy that maximizes the expected present value of all profits and losses resulting from the project. Variants of this problem have been developed in the economics literature as models for the valuation of investments in real assets by Brennan and Schwartz [BS], Dixit [D], and Dixit and Pindyck [DP]. Such a problem has the features of stochastic impulse control, and explicit solutions have been obtained in the mathematics literature by Brekke and Øksendal [BØ1, BØ2], Lumley and Zervos [LZ], and Duckworth and Zervos [DuZ1].

Suppose now that the option of totally abandoning the project at a discretionary time and at a certain fixed cost is added in the set of available managerial decisions. The resulting problem then combines stochastic impulse control with discretionary stopping. In fact, such a model is extensively discussed in Dixit and Pindyck [DP, section 7.2], and is a special case of the one developed by Brennan and Schwartz [BS]. However, these authors make very little progress in actually solving the problem. The purpose of this paper is to solve completely the resulting optimization problem under the assumption that the rate at which the project yields profits or losses is a standard Brownian motion. Such an assumption is probably crude as long as real life applications are concerned. However, it leads to explicit, nontrivial results that unveil the qualitative nature of the optimal strategy.

The results of our analysis take qualitatively different forms, depending on parameter values, and can be summarized informally as follows. Suppose that the switching costs are fixed. If the abandonment cost is very large (see Case I in Theorem 6 and Figure 1), then it is optimal to perpetuate the project by switching it to its "closed" mode as soon as its output cash flow falls below a certain level and by switching it to its "open" mode as soon as its potential output cash flow rises above a certain higher level. If the abandonment cost is very small (see Case III of Theorem 6 and Figure 4), then abandonment is optimal sooner or later. If the project is in its "closed" mode at time 0, then it is switched to its "open" mode as soon as its potential output cash flow exceeds a certain level. Once in it, the project should be kept in its "open" operating mode for as long as its output cash flow is above a given level and should be abandoned as soon as its output cash flow falls below this level. For intermediate values of the abandonment cost, we have an a priori rather unexpected combination of the two cases above (see Case II of Theorem 6 and Figure 3). If the project starts from its "closed" mode, then it is never abandoned, and the situation resembles the case in which the abandonment cost is very large. A similar scenario pertains to the case in which the project is originally "open" and its output cash flow assumes sufficiently high levels. However, if the project is originally "open" and its output cash flow assumes very low values, then it is optimal to abandon the project immediately. The most interesting possibility arises when the project is originally "open" and its output cash flow assumes moderately low values. In this case, it is optimal to keep the project alive and keep on accumulating losses until its output cash flow either

falls below a certain level, in which event the project is totally abandoned, or rises above another certain level, in which event its operation enters the perpetual life-cycle pertaining to the case of a large abandonment cost. As a result, the abandonment time of the project is either finite or infinite, and each of the two possibilities has positive probability.

The paper is organized as follows. Section 2 is concerned with the formulation of the stochastic optimization problem that we address. In section 3, we prove a verification theorem that will play a crucial role in our subsequent analysis. The assumptions of the theorem allow for the possibility that the value function is not C^1, and the proof is developed using Itô–Tanaka's formula and relies on the properties of local times. The explicit solution of the nontrivial case discussed above is developed in section 4. Finally, an example whose value function is not C^1 is presented in section 5.

2. Problem formulation. Let $(\Omega, \mathcal{F}, P)$ be a complete probability space equipped with a filtration $(\mathcal{F}_t)$ satisfying the usual conditions of right continuity and augmentation by P-negligible sets and carrying a standard one-dimensional $(\mathcal{F}_t)$-Brownian motion W. We denote by $\mathcal{Z}$ the family of all adapted, finite variation, càglàd processes Z with values in $\{0,1\}$ and by $\mathcal{S}$ the set of all $(\mathcal{F}_t)$-stopping times.

We consider a stochastic system that can operate in two modes, an "open" one and a "closed" one. The system's mode of operation can be changed at a sequence of $(\mathcal{F}_t)$-stopping times. These transition times constitute a decision strategy that we model by a process $Z \in \mathcal{Z}$. Specifically, given any time t, $Z_t = 1$ if the system is "open" at time t, whereas $Z_t = 0$ if the system is "closed" at time t. The stopping times at which the jumps of Z occur are the intervention times at which the system's operating mode is changed. We denote by $z \in \{0,1\}$ the system's mode at time 0. We also assume that the operation of this system can be permanently abandoned at an $(\mathcal{F}_t)$-stopping time T, which is an additional decision variable. We define the set of all admissible strategies to be

$$\Pi_z = \{(Z,T) \mid \ Z \in \mathcal{Z},\ Z_0 = z,\ T \in \mathcal{S}\}.$$

We assume that the rate at which the system yields payoff, the switching costs associated with the transition of the system from its "closed" mode to its "open" one and vice versa, as well as the permanent abandonment cost are all functions of a state process X which satisfies the one-dimensional SDE

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dW_t, \qquad X_0 = x \in \mathcal{I}, \tag{1}$$

where $\mathcal{I}$ is a given interval. We assume that the functions $b, \sigma : \mathcal{I} \to \mathbb{R}$ satisfy assumptions such that this SDE has a unique strong solution with values in $\mathcal{I}$ for all $t \geq 0$, P-a.s. In the problem that we solve in section 4, $\mathcal{I} = \mathbb{R}$. However, if, following several of the references mentioned in the introduction, we use X to model commodity prices, we must have $\mathcal{I} =]0, \infty[$.

With each admissible strategy $(Z,T) \in \Pi_z$ we associate the expected payoff

$$\begin{aligned} J_{z,x}(Z,T) = E\Bigg[& \int_0^T R_s\big[H_1(X_s)Z_s + H_0(X_s)(1-Z_s)\big]\,ds \\ & - \sum_{0 \leq s \leq T} 1_{\{s<\infty\}} R_s[G_1(X_s)\,(\Delta Z_s)^+ + G_0(X_s)\,(\Delta Z_s)^-] \\ & - 1_{\{T<\infty\}} R_T F(X_T)\Bigg], \end{aligned} \tag{2}$$

where $\Delta Z_t = Z_{t+} - Z_t$, $(\Delta Z_t)^{\pm} = \max\{\pm\Delta Z_t, 0\}$, and the discounting process R is given by

$$R_t = \exp\left(-\int_0^t r(X_s)\,ds\right) \tag{3}$$

for some positive function $r : \mathcal{I} \to \mathbb{R}$. Here, $H_1(X_t)$ (resp., $H_0(X_t)$) is the rate at which the system yields payoff assuming that, at time t, it is in its "open" (resp., "closed") operating mode. Also, $G_1(X_t)$, $G_0(X_t)$ are the costs associated with switching the investment from its "closed" to its "open" mode, and vice versa, respectively, at time t, whereas $F(X_t)$ is the cost faced if the system is completely abandoned at time t.

The objective is to maximize $J_{z,x}(Z,T)$ over Π_z. Accordingly, we define the value function

$$v(z,x) = \sup_{(Z,T)\in\Pi_z} J_{z,x}(Z,T).$$

We assume that the problem is well posed in the sense that all of the integrals in (2) are well defined for every admissible strategy, and nontrivial in the sense that $v(z,x) < \infty$ for every initial condition (z,x). For the problem to be well posed, we also need to assume that no strategy associated with a finite payoff involves an infinite number of switchings prior to abandonment on a set of positive probability so that every switching strategy can be modelled by a process in $\mathcal{Z}$. A sufficient condition for this assumption to hold is $G_1(x) + G_0(x) > \epsilon > 0$ for all $x \in \mathcal{I}$. From an economics perspective, this assumption is a natural one because it rules out the unrealistic situation in which arbitrarily high profits can be made by rapidly changing the system's operating mode.

All of the assumptions discussed above are of an implicit nature. Further assumptions will appear in the statement of Theorem 1, again in an implicit way. On the other hand, the results of sections 4 and 5 will assume that the problem's data have specific forms.

At this point, it would be of interest to make a comment on a possible generalization of the model considered here. The dynamics of the state process X can be modified to include an additional, regime switching process so that (1) becomes

$$dX_t = b(\theta_t, X_t)\,dt + \sigma(\theta_t, X_t)\,dW_t, \qquad X_0 = x \in \mathcal{I}.$$

The process θ can be taken to be a finite-state Markov chain representing a number of different economic outlooks (e.g., a state of economic growth and a state of recession). Models involving regime switchings have been considered in the literature and include Guo [G], who solves the problem of pricing a Russian option in such a context. A generalization of the model studied here in this direction would multiply the complexity of the problem by the number of states that the process θ can assume, and we leave it as an interesting open problem.

3. A verification theorem. The problem considered in the previous section combines features of both stochastic impulse control and optimal stopping. Therefore, we expect that the value function v should satisfy the Hamilton–Jacobi–Bellman (HJB) equation

$$\begin{aligned} \max\big\{&\mathcal{L}v(z,x) + zH_1(x) + (1-z)H_0(x), \\ &v(1-z,x) - v(z,x) - zG_0(x) - (1-z)G_1(x), \\ &-v(z,x) - F(x)\big\} = 0, \qquad z = 1,0,\ x \in \mathcal{I}, \end{aligned} \tag{4}$$

where the second order elliptic operator $\mathcal{L}$ is defined by

$$\mathcal{L}v(z,x) = \frac{1}{2}\sigma^2(x)v_{xx}(z,x) + b(x)v_x(z,x) - r(x)v(z,x).$$

The ideas behind the origins of this equation are the following. Suppose that, at time 0, the system is in its "open" operating mode, i.e., $z = 1$. The controller's immediate decision consists of choosing between three actions. The first action is to totally terminate the system's operation at the cost of $-F(x)$. Such a possibility gives rise to the inequality

$$v(1,x) \geq -F(x). \tag{5}$$

The second option is to pay the cost of $G_0(x)$ to switch the system to its "closed" operating mode and then continue optimally. This possibility yields the inequality

$$v(1,x) \geq -G_0(x) + v(0,x). \tag{6}$$

The third action is to leave the system in its "open" operating mode for a short time Δt and then continue optimally. This action is associated with the inequality

$$v(1,x) \geq E\left[\int_0^{\Delta t} R_s H_1(X_s)\,ds + R_{\Delta t} v(1, X_{\Delta t})\right].$$

Under the assumption that $v(1,\cdot)$ is sufficiently smooth, we may apply Itô's formula to the last term and then divide by Δt before letting $\Delta t \downarrow 0$ to obtain

$$\begin{aligned}\mathcal{L}v(1,x) + H_1(x) &\equiv \frac{1}{2}\sigma^2(x)v_{xx}(1,x) + b(x)v_x(1,x) - r(x)v(1,x) + H_1(x)\\ &\leq 0.\end{aligned} \tag{7}$$

Now, each of (5)–(7) can hold with strict inequality because the corresponding action may not be optimal. However, we expect that the three actions considered above form a complete repertoire of optimal tactics. Therefore, given any $x \in \mathcal{I}$, we expect that one of (5)–(7) should hold with equality. Combining all of these relationships, we can conclude that the value function $v(1,\cdot)$ should satisfy

$$\max\bigl\{\mathcal{L}v(1,x) + H_1(x),\ v(0,x) - v(1,x) - G_0(x),\ -v(1,x) - F(x)\bigr\} = 0. \tag{8}$$

Using similar reasoning, we can also conclude that the value function $v(0,\cdot)$ associated with the system in its "closed" operating mode (i.e., when $z = 0$) should satisfy

$$\max\bigl\{\mathcal{L}v(0,x) + H_0(x), v(1,x) - v(0,x) - G_1(x),\ -v(0,x) - F(x)\bigr\} = 0. \tag{9}$$

Now, combining (8) and (9), we conclude that the value function v should satisfy (4). Without any further conditions, this equation has, in general, uncountably many solutions.

Example 1. Suppose that $\mathcal{I} = \mathbb{R}$, and, for all $x \in \mathbb{R}$, $b(x) = 0$, $\sigma(x) = \sqrt{2}$, $r(x) = 4$, $H_1(x) = 3e^x + 4$, $H_0(x) = 0$, $G_1(x) = G_0(x) = 1$ and $F(x) = c$ for some constant $c > 0$. It is straightforward to verify that each of the functions defined by

$$w(z,x) = Ae^{2x} + Be^{-2x} + e^x + z, \qquad A, B \geq 0,$$

satisfies (4).

It turns out that the functions $v(1,\cdot)$ and $v(0,\cdot)$ composing the value function of the special case of the control problem that we explicitly solve in section 4 are both C^1 but not C^2. However, it is clear that, as long as the general problem is concerned, we cannot expect such regularity of the value function unless we impose appropriate assumptions on the problem's data. For instance, we cannot in general expect C^1 regularity unless the abandonment cost function F is C^1. An explicitly solvable example illustrating this issue is presented in section 5.

In the next theorem, we consider candidates for the value functions $v(1,\cdot)$ and $v(0,\cdot)$ which are differences of convex functions; for a survey of the results needed here, see Revuz and Yor [RY, Appendix 3]. In particular, we consider solutions of (4) in the following sense.

Definition 1. A function $w : \{0,1\} \times \mathcal{I} \to \mathbb{R}$ satisfies (4) if each of $w(1,\cdot)$, $w(0,\cdot)$ is a difference of two convex functions and (4) is true Lebesgue-a.e., with $\hat{\mathcal{L}}$ in place of $\mathcal{L}$, where the operator $\hat{\mathcal{L}}$ is defined by

$$\hat{\mathcal{L}}w(z,x) = \frac{1}{2}\sigma^2(x)w^{\mathrm{ac}}_{xx}(z,x) + b(x)w^-_x(z,x) - r(x)w(z,x).$$

Here, $w^-_x(z,\cdot)$ is the left-hand derivative of $w(z,\cdot)$. Also,

$$w_{xx}(z,dx) = w^{\mathrm{ac}}_{xx}(z,x)\,dx + w^{\mathrm{s}}_{xx}(z,dx) \tag{10}$$

is the Lebesgue decomposition of the second distributional derivative $w_{xx}(z,dx)$ of $w(z,\cdot)$ to the measure $w^{\mathrm{ac}}_{xx}(z,x)\,dx$, which is absolutely continuous with respect to the Lebesgue measure, and the measure $w^{\mathrm{s}}_{xx}(z,dx)$, which is mutually singular with the Lebesgue measure.

We can now prove conditions which are sufficient for optimality in our problem.

THEOREM 1. *Consider the control problem described in section* 2. *Suppose that* G_1, G_0, F *are continuous functions,* $\sigma^2(x) > 0$ *for all* $x \in \mathcal{I}$, *and, for every admissible strategy* $(Z,T) \in \Pi_z$, *there exists a sequence of times* $t_m \to \infty$ *such that*

$$\lim_{m\to\infty} J_{z,x}(Z, T \wedge t_m) = J_{z,x}(Z,T). \tag{11}$$

Suppose that there exist functions $w(1,\cdot), w(0,\cdot) : \mathcal{I} \to \mathbb{R}$ *which are differences of convex functions such that*

$$-w^{\mathrm{s}}_{xx}(1,dx) \text{ and } -w^{\mathrm{s}}_{xx}(0,dx) \text{ are positive measures} \tag{12}$$

and which satisfy the HJB equation (4) *in the sense of Definition* 1. *Also, suppose that the process* M *defined by*

$$M_t = \int_0^t R_s\sigma(X_s)w^-_x(Z_s,X_s)\,dW_s \tag{13}$$

is a martingale for every switching strategy $Z \in \mathcal{Z}$. *Then, given any initial condition* $(z,x) \in \{0,1\} \times \mathcal{I}$,

(a) $v(z,x) \le w(z,x)$, *and*

(b) *if*

$$\begin{aligned}\operatorname{supp} w^{\mathrm{s}}_{xx}(z,dx) &\subseteq \mathcal{I} \setminus \operatorname{int}\overline{\{x \in \mathcal{I} \mid \ \hat{\mathcal{L}}w(z,x) + zH_1(x) + (1-z)H_0(x) = 0\}} \\ &=: \mathcal{O}_z\end{aligned} \tag{14}$$

and there exists $Z^* \in \mathcal{Z}$ *such that*

$$\hat{\mathcal{L}} w(Z_t^*, X_t) + Z_t^* H_1(X_t) + (1 - Z_t^*) H_0(X_t) = 0 \tag{15}$$

for Lebesgue almost all $t \le T^*$, *P-a.s., and*

$$[w(1, X_t) - w(0, X_t) - G_1(X_t)]\,(\Delta Z_t^*)^+ = 0, \tag{16}$$

$$[w(0, X_t) - w(1, X_t) - G_0(X_t)]\,(\Delta Z_t^*)^- = 0 \tag{17}$$

for all $t \le T^*$, *P-a.s., where*

$$T^* = \inf\{t \ge 0 \mid \ w(Z_t^*, X_t) = -F(X_t)\}, \tag{18}$$

as well as a sequence of times $t_m \to \infty$ *satisfying* (11) *and*

$$\lim_{m\to\infty} E\big[R_{t_m} \big|w(Z^*_{t_m}, X_{t_m})\big|\big] = 0 \quad \textit{and} \quad \lim_{m\to\infty} E\big[R_{t_m} |F(X_{t_m})|\big] = 0, \tag{19}$$

then $v(z, x) = w(z, x)$, *and the optimal strategy is* (Z^*, T^*).

Proof. Fix any $z = 0, 1$. Using Itô–Tanaka's formula (see Revuz and Yor [RY, Theorem VI.1.5]), we obtain

$$\begin{aligned} w(z, X_t) = w(z, x) &+ \int_0^t b(X_s) w_x^-(z, X_s)\, ds + \int_0^t \sigma(X_s) w_x^-(z, X_s)\, dW_s \\ &+ \frac{1}{2} \int_{\mathcal{I}} L_t^a\, w_{xx}(z, da), \end{aligned} \tag{20}$$

where L^a is the local time of the process X at level a. We assume that

$$\text{the mapping } (t, a) \to L_t^a \text{ is continuous in } t \text{ and càdlàg in } a, \tag{21}$$

P-a.s. (see Revuz and Yor [RY, Theorem VI.1.7]). With reference to (10) and the occupation times formula (see Revuz and Yor [RY, Corollary VI.1.6]),

$$\int_{\mathcal{I}} L_t^a w_{xx}^{\mathrm{ac}}(z, a)\, da = \int_0^t \sigma^2(X_s) w_{xx}^{\mathrm{ac}}(z, X_s)\, ds,$$

and so (20) implies

$$\begin{aligned} w(z, X_t) = w(z, x) &+ \int_0^t \left[\frac{1}{2}\sigma^2(X_s) w_{xx}^{\mathrm{ac}}(z, X_s) + b(X_s) w_x^-(z, X_s)\right] ds \\ &+ \int_0^t \sigma(X_s) w_x^-(z, X_s)\, dW_s + A_t^z, \end{aligned}$$

where

$$A_t^z = \frac{1}{2} \int_{\mathcal{I}} L_t^a\, w_{xx}^{\mathrm{s}}(z, da). \tag{22}$$

For future reference, observe that (12) implies

$$-A^z \text{ is a continuous, increasing process} \tag{23}$$

because such a statement is true for local times. Now, using the integration by parts formula for semimartingales, we obtain

$$R_t w(z,X_t) = w(z,x) + \int_0^t R_s \hat{\mathcal{L}} w(z,X_s)\,ds + \int_0^t R_s \sigma(X_s) w_x^-(z,X_s)\,dW_s + \int_0^t R_s\,dA_s^z. \tag{24}$$

We can now prove the two statements of the theorem.

(a) Fix any admissible strategy $(Z,T) \in \Pi_z$, and suppose that the abandonment time T is bounded by a constant. Define the increasing sequence of $(\mathcal{F}_t)$-stopping times (T_n) by

$$T_1 = \inf\{t \geq 0 \mid \ Z_t \neq z\} \quad \text{and} \quad T_{n+1} = \inf\{t > T_n \mid \ Z_t \neq Z_{T_n+}\}, \tag{25}$$

with the usual convention that $\inf \emptyset = \infty$. Note that the assumption that Z is a finite variation process implies that its discontinuities cannot accumulate within any compact subset of $\mathbb{R}_+$, so $T_n \to \infty$, P-a.s. Therefore,

$$\begin{aligned} &R_T w(Z_T,X_T) \\ &\quad = R_T w(Z_T,X_T) 1_{\{T \leq T_1\}} + \sum_{n=1}^{\infty} \Bigg[R_T w(Z_T,X_T) - R_{T_n} w(Z_{T_n+},X_{T_n}) \\ &\qquad + \sum_{j=1}^{n-1} \big[R_{T_{j+1}} w(Z_{T_{j+1}},X_{T_{j+1}}) - R_{T_j} w(Z_{T_j+},X_{T_j}) \big] \\ &\qquad + R_{T_1} w(Z_{T_1},X_{T_1}) + \sum_{j=1}^{n} R_{T_j} \big[w(Z_{T_j+},X_{T_j}) - w(Z_{T_j},X_{T_j}) \big] \Bigg] 1_{\{T_n < T \leq T_{n+1}\}}. \end{aligned} \tag{26}$$

Now, since Z is constant on the stochastic interval $]T_j, T_{j+1}]$ and T is bounded, (24) implies

$$\begin{aligned} \big[R_{T_{j+1}} w(Z_{T_{j+1}},X_{T_{j+1}}) - R_{T_j} w(Z_{T_j+},X_{T_j}) \big] 1_{\{T_{j+1} < T\}} &= \Bigg[\int_{T_j}^{T_{j+1}} R_s \hat{\mathcal{L}} w(Z_s,X_s)\,ds \\ &\quad + M_{T_{j+1}} - M_{T_j} + \int_{T_j}^{T_{j+1}} R_s Z_s\,dA_s^1 + \int_{T_j}^{T_{j+1}} R_s (1-Z_s)\,dA_s^0 \Bigg] 1_{\{T_{j+1} < T\}}, \end{aligned}$$

where M is defined as in (13). Since the terms

$$\begin{aligned} &\big[R_T w(Z_T,X_T) - w(z,x) \big] 1_{\{T \leq T_1\}}, \\ &\big[R_{T_1} w(Z_{T_1},X_{T_1}) - w(z,x) \big] 1_{\{T_1 \leq T\}}, \\ &\big[R_T w(Z_T,X_T) - R_{T_n} w(Z_{T_n+},X_{T_n}) \big] 1_{\{T_n < T \leq T_{n+1}\}} \end{aligned}$$

admit similar expressions, (26) implies

$$\begin{aligned} R_T w(Z_T,X_T) &= w(z,x) + \int_0^T R_s \hat{\mathcal{L}} w(Z_s,X_s)\,ds + M_T \\ &\quad + \sum_{0 \leq s < T} R_s \big[w(Z_{s+},X_s) - w(Z_s,X_s) \big] + \int_0^T R_s Z_s\,dA_s^1 + \int_0^T R_s (1-Z_s)\,dA_s^0. \end{aligned}$$

It follows that

$$\begin{aligned}
&\int_0^T R_s\big[H_1(X_s)Z_s + H_0(X_s)(1-Z_s)\big]\,ds \\
&\qquad - \sum_{0\le s\le T} R_s[G_1(X_s)\,(\Delta Z_s)^+ + G_0(X_s)\,(\Delta Z_s)^-] - R_T F(X_T) \\
&\quad = w(z,x) - R_T\big[w(Z_{T+},X_T) + F(X_T)\big] + M_T \\
&\quad + \int_0^T R_s\big[\hat{\mathcal{L}}w(Z_s,X_s) + H_1(X_s)Z_s + H_0(X_s)(1-Z_s)\big]\,ds \\
&\quad + \sum_{0\le s\le T} R_s\big[w(1,X_s) - w(0,X_s) - G_1(X_s)\big](\Delta Z_s)^+ + \int_0^T R_s Z_s\,dA_s^1 \\
&\quad + \sum_{0\le s\le T} R_s\big[w(0,X_s) - w(1,X_s) - G_0(X_s)\big](\Delta Z_s)^- + \int_0^T R_s(1-Z_s)\,dA_s^0.
\end{aligned} \tag{27}$$

In view of (23) and the fact that w satisfies (4) in the sense of Definition 1, this implies

$$\begin{aligned}
&\int_0^T R_s\big[H_1(X_s)Z_s + H_0(X_s)(1-Z_s)\big]\,ds \\
&\qquad - \sum_{0\le s\le T} R_s[G_1(X_s)\,(\Delta Z_s)^+ + G_0(X_s)\,(\Delta Z_s)^-] - R_T F(X_T) \\
&\qquad \le w(z,x) + M_T.
\end{aligned}$$

Taking expectations and noting that the stochastic integral has expectation 0, we obtain $J_{z,x}(Z,T) \le w(z,x)$.

Now, consider the general case in which the abandonment time T is not necessarily bounded by a constant, and let (t_m) be a sequence satisfying (11). From our analysis above, it follows that $J_{z,x}(Z, T\wedge t_m) \le w(z,x)$ for all m. However, this and (11) imply $J_{z,x}(Z,T) \le w(z,x)$, which establishes this part of the theorem.

(b) Suppose that there exists a strategy (Z^*,T^*) satisfying (15)–(18), let (T_n^*) be the associated sequence of stopping times defined as in (25), and let (t_m) be a sequence satisfying (11) as well as (19). Fix any of the stochastic intervals $]T_n^*\wedge T^*\wedge t_m,\ T_{n+1}^*\wedge T^*\wedge t_m]$, and observe that Z^* is constant on this interval, i.e., $Z_t^* = z$ for some $z\in\{0,1\}$, for all $t\in\,]T_n^*\wedge T^*\wedge t_m,\ T_{n+1}^*\wedge T^*\wedge t_m]$, P-a.s. Since the measure dL_t^a is carried by the set $\{t\ge 0 \mid\ X_t = a\}$, P-a.s., (14) and (15) imply

$$L^a_{T_n^*\wedge T^*\wedge t_m} = L^a_{T_{n+1}^*\wedge T^*\wedge t_m}, \quad P\text{-a.s.}, \quad \forall a\in\mathcal{O}_z.$$

Therefore,

$$L^a_{T_n^*\wedge T^*\wedge t_m} = L^a_{T_{n+1}^*\wedge T^*\wedge t_m} \quad \forall a\in\mathcal{O}_z^d, \tag{28}$$

P-a.s., where $\mathcal{O}_z^d$ is any countable subset of $\mathcal{O}_z$ which is dense in $\mathcal{O}_z$. Now, let $A\in\mathcal{F}$ be such that $P(A)=1$ and (21), (28) are true for all $\omega\in A$. Given any $a\in\mathcal{O}_z\setminus\mathcal{O}_z^d$ and any sequence (a_m) in $\mathcal{O}_z^d$ such that $a_m\downarrow a$, (21) implies

$$L_t^a(\omega) = \lim_{k\to\infty} L_t^{a_k}(\omega) \quad \forall t\ge 0, \quad \forall\omega\in A.$$

However, this and (28) imply

$$L^a_{T_n^*\wedge T^*\wedge t_m}(\omega) = L^a_{T_{n+1}^*\wedge T^*\wedge t_m}(\omega) \quad \forall a\in\mathcal{O}_z, \quad \forall\omega\in A.$$

Combining this with (22), we can see that

$$A^z_{T^*_n\wedge T^*\wedge t_m} = A^z_{T^*_{n+1}\wedge T^*\wedge t_m} \qquad \text{if } Z^*_t = z \text{ for } t \in]T^*_n \wedge T^* \wedge t_m,\, T^*_{n+1} \wedge T^* \wedge t_m].$$

It follows that

$$\int_0^{T^*\wedge t_m} R_s Z^*_s\, dA^1_s = \int_0^{T^*\wedge t_m} R_s(1 - Z^*_s)\, dA^0_s = 0.$$

Therefore, in view of (15)–(18), (27) implies

$$\begin{aligned}
&\int_0^{T^*\wedge t_m} R_s\big[H_1(X_s)Z^*_s + H_0(X_s)(1 - Z^*_s)\big]\, ds \\
&\qquad - \sum_{0\le s\le T^*\wedge t_m} R_s[G_1(X_s)\,(\Delta Z^*_s)^+ + G_0(X_s)\,(\Delta Z^*_s)^-] - R_{T^*\wedge t_m} F(X_{T^*\wedge t_m}) \\
&= w(z,x) - 1_{\{T^*>t_m\}} R_{t_m}\left[w(Z^*_{t_m+}, X_{t_m}) + F(X_{t_m})\right] + M_{T^*\wedge t_m}.
\end{aligned}$$

Taking expectations and letting $m \to \infty$, we obtain $J_{z,x}(Z^*, T^*) = w(z,x)$, by virtue of (11) and (19), and the proof is complete. □

Remark 1. To obtain some further insight into the assumptions of the theorem above, suppose that, given a finite number of points $a^1_1 < a^1_2 < \cdots < a^1_{N^1}$ (resp., $a^0_1 < a^0_2 < \cdots < a^0_{N^0}$), $w(1,\cdot)$ (resp., $w(0,\cdot)$) is twice continuously differentiable at every point $x \in \mathcal{I} \setminus \{a^1_1, \ldots, a^1_{N^1}\}$ (resp., $x \in \mathcal{I} \setminus \{a^0_1, \ldots, a^0_{N^0}\}$). Also, suppose that each of the functions $w^-_x(z,\cdot)$ is locally bounded. In this case, assumptions (12) and (14) are equivalent to

$$w^-_x(z, a^z_{i^z}) \ge w^-_x(z, a^z_{i^z}+) \equiv \lim_{x\downarrow a^z_{i^z}} w^-_x(z,x) \quad \forall i^z = 1, 2, \ldots, N^z \text{ and} \tag{29}$$

$$a^z_1, \ldots, a^z_{N^z} \subseteq \mathcal{I} \setminus \operatorname{int}\overline{\{x \in \mathcal{I} \mid \hat{\mathcal{L}}w(z,x) + zH_1(x) + (1-z)H_0(x) = 0\}}, \tag{30}$$

$z = 0, 1$, respectively. For future reference, we should stress that we cannot dispense with either of these two assumptions. Also, it is worth observing the *asymmetry* presented by (12) or (29): had the optimization problem been a minimization one, we would have to replace (12) by the assumption that $w^{\mathrm{s}}_{xx}(1, dx)$ and $w^{\mathrm{s}}_{xx}(0, dx)$ are positive measures, and we would have to consider the reverse inequalities in (29). With regard to (30), we can conclude that the points where C^1 regularity fails should not belong to the interior of the "continuation" region but can be allowed in the closure of the "switching" or "stopping" regions.

Remark 2. The result proved above can be trivially extended to the case in which the system's operating modes are not just two, namely, "open" and "closed," but are any finite positive integer. On the other hand, the proof cannot be trivially modified to account for the case in which the process X assumes values in a higher dimensional state space because it relies heavily on the use of local times and Itô–Tanaka's formula.

To analyze the problem arising if the state process X is an n-dimensional diffusion under similarly general assumptions, one would have to resort to the use of viscosity solutions of the associated HJB equation (see Fleming and Soner [FS] and Yong and Zhou [YZ]). This project would aim at proving that the value function identifies with the unique viscosity solution of the HJB equation. Furthermore, characterizing the optimal strategy would require a viscosity solution version of the verification Theorem 1 in the spirit of Theorem 5.5.3 in Yong and Zhou [YZ]. Such an analysis lies beyond the scope of this article, and we leave it as an interesting open problem.

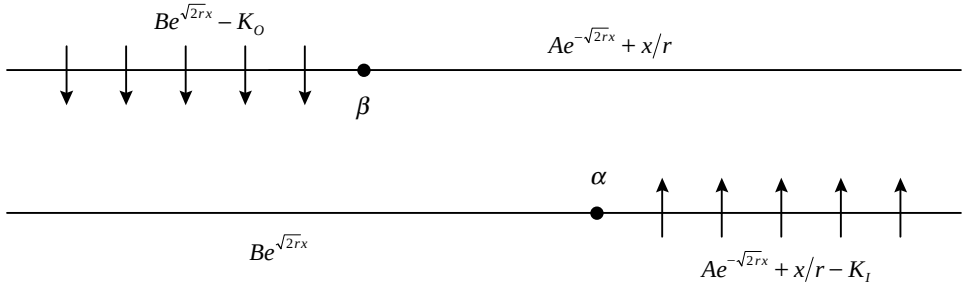


FIG. 1. *The "no-abandonment case."*

4. The explicit solution of a special case. We now solve completely the special case of the general control problem formulated in section 2 that arises if we impose the following assumption.

Assumption 1. $\mathcal{I} = \mathbb{R}$, and $b(x) = 1$, $\sigma(x) = 1$, $r(x) = r$, $H_1(x) = x$, $H_0(x) = 0$, $G_1(x) = K_1$, $G_0(x) = K_0$, and $F(x) = K$ for some constants $r, K_1, K_0, K > 0$, for all $x \in \mathbb{R}$.

In this case, the HJB equation (4) reduces to the following pair of coupled quasi-variational inequalities:

$$\max\left\{\frac{1}{2}v_1''(x) - rv_1(x) + x,\ v_0(x) - v_1(x) - K_0,\ -v_1(x) - K\right\} = 0, \tag{31}$$

$$\max\left\{\frac{1}{2}v_0''(x) - rv_0(x),\ v_1(x) - v_0(x) - K_1,\ -v_0(x) - K\right\} = 0. \tag{32}$$

Here, we write v_1 and v_0 in place of $v(1,\cdot)$ and $v(0,\cdot)$, respectively, to simplify the notation.

To make some headway, we first make some qualitative observations. Since the system yields 0 payoff whenever it operates in its "closed" mode and the abandonment cost K is positive, it follows that abandonment cannot be optimal when the system is in its "closed" mode. As a consequence, abandonment can be part of the optimal strategy only if the system is in its "open" operating mode. Moreover, the system should be in its "open" operating mode if the state process X assumes sufficiently large values and should be in its "closed" operating mode or should be abandoned if the state process X takes sufficiently low values.

Now, a first possibility arises if abandonment is not part of the optimal scenario. In such a case, we should switch the system from its "closed" to its "open" mode whenever the state process X exceeds a level specified by a constant α, and we should switch the system from its "open" to its "closed" mode whenever the state process X falls below a level given by a constant β. Clearly, such a strategy is well defined only if $\beta < \alpha$. It is depicted by Figure 1.

If such a strategy is indeed optimal, the value function should be given by a solution w_1, w_0 of the HJB equations (31)–(32) described as follows. For $x > \beta$, w_1 should satisfy $\frac{1}{2}w_1''(x) - rw_1(x) + x = 0$, namely, $w_1(x) = Ae^{-\sqrt{2r}x} + C_1e^{\sqrt{2r}x} + x/r$ for some constants $A, C_1 \in \mathbb{R}$, whereas, for $x \le \beta$, w_1 should be given by $w_1(x) = w_0(x) - K_0$. On the other hand, if $x < \alpha$, w_0 should satisfy $\frac{1}{2}w_0''(x) - rw_0(x) = 0$, namely $w_0(x) = C_2e^{-\sqrt{2r}x} + Be^{\sqrt{2r}x}$ for some constants $C_2, B \in \mathbb{R}$, whereas, for $x \ge \alpha$, w_0 should be given by $w_0(x) = w_1(x) - K_1$. Now, we must have $C_1 = C_2 = 0$

because, otherwise, the assumptions of Theorem 1 cannot be satisfied. In view of these conditions, w_1, w_0 should be given by

$$w_1(x) = \begin{cases} Be^{\sqrt{2r}x} - K_0 & \text{if } x \leq \beta, \\ Ae^{-\sqrt{2r}x} + x/r & \text{if } x > \beta, \end{cases} \tag{33}$$

$$w_0(x) = \begin{cases} Be^{\sqrt{2r}x} & \text{if } x < \alpha, \\ Ae^{-\sqrt{2r}x} + x/r - K_1 & \text{if } x \geq \alpha, \end{cases} \tag{34}$$

respectively. To specify the parameters A, B, α, β, we postulate that w_1, w_0 are C^1 at the free boundary points β, α, respectively. This requirement gives rise to the system of equations

$$Be^{\sqrt{2r}\alpha} - Ae^{-\sqrt{2r}\alpha} = \frac{\alpha - rK_1}{r}, \tag{35}$$

$$Be^{\sqrt{2r}\alpha} + Ae^{-\sqrt{2r}\alpha} = \frac{1}{r\sqrt{2r}}, \tag{36}$$

$$Be^{\sqrt{2r}\beta} - Ae^{-\sqrt{2r}\beta} = \frac{\beta + rK_0}{r}, \tag{37}$$

$$Be^{\sqrt{2r}\beta} + Ae^{-\sqrt{2r}\beta} = \frac{1}{r\sqrt{2r}.} \tag{38}$$

It is straightforward to verify that these are equivalent to

$$A = -\frac{\beta + rK_0 - 1/\sqrt{2r}}{2r} e^{\sqrt{2r}\beta}, \tag{39}$$

$$B = \frac{\beta + rK_0 + 1/\sqrt{2r}}{2r} e^{-\sqrt{2r}\beta}, \tag{40}$$

$$(\alpha - rK_1 - 1/\sqrt{2r})e^{\sqrt{2r}\alpha} = (\beta + rK_0 - 1/\sqrt{2r})e^{\sqrt{2r}\beta}, \tag{41}$$

$$(\alpha - rK_1 + 1/\sqrt{2r})e^{-\sqrt{2r}\alpha} = (\beta + rK_0 + 1/\sqrt{2r})e^{-\sqrt{2r}\beta}. \tag{42}$$

The next lemma is concerned with the solvability of (41)–(42) and with necessary and sufficient conditions under which the functions w_1, w_0 given above satisfy the HJB equations (31)–(32). To derive the results of Lemma 4 below, we assume here that the constants K_1, K_0 can take negative as well as positive values subject to the condition that $K_1 + K_0 > 0$.

LEMMA 2. *Suppose that $r > 0$ and $K_1, K_0 \in \mathbb{R}$ satisfy $K_1 + K_0 > 0$. There exists a unique pair of points $\alpha = \alpha(r, K_1, K_0)$ and $\beta = \beta(r, K_1, K_0)$ which satisfies* (41)–(42). *Point β is the unique solution of*

$$H(\beta) := \frac{\beta + rK_0 + 1/\sqrt{2r}}{\beta + rK_0 - 1/\sqrt{2r}} \exp(-\sqrt{2r}(2\beta + rK_0 - rK_1)) = -1 \tag{43}$$

and satisfies

$$-rK_0 - \frac{1}{\sqrt{2r}} < \beta < -rK_0, \tag{44}$$

$$-2e^{-2} > \sqrt{2r}(\beta + rK_0 - 1/\sqrt{2r}) \exp(\sqrt{2r}(\beta + rK_0 - 1/\sqrt{2r})) > -e^{-1}, \tag{45}$$

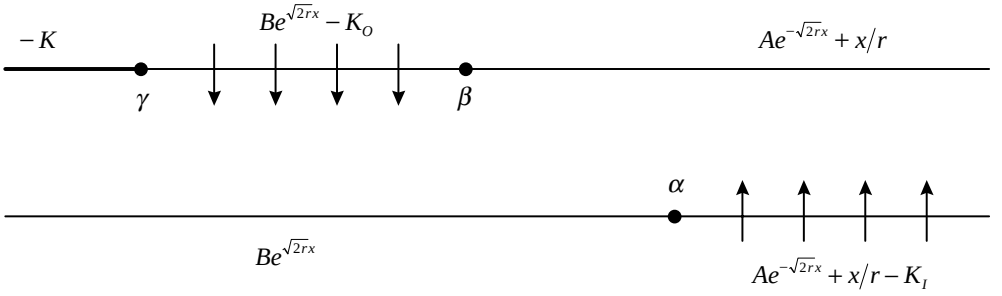


FIG. 2. *An obvious modification of the "no-abandonment case."*

whereas

$$\alpha = -\beta - rK_0 + rK_1 > \beta. \tag{46}$$

The functions w_1, w_0 *defined by* (33)–(34), *where* α *and* β *are as above and* $A, B > 0$ *are given by* (39), (40), *respectively, are convex, nondecreasing,* C^1 *for all* $x \in \mathbb{R}$ *and* C^2 *for all* $x \in \mathbb{R} \setminus \{\beta\}$, $x \in \mathbb{R} \setminus \{\alpha\}$, *respectively, and satisfy*

$$\max\left\{\frac{1}{2}w_1''(x) - rw_1(x) + x,\ w_0(x) - K_0 - w_1(x)\right\} = 0 \tag{47}$$

for all $x \in \mathbb{R} \setminus \{\beta\}$ *and*

$$\max\left\{\frac{1}{2}w_0''(x) - rw_0(x),\ w_1(x) - K_1 - w_0(x),\ -K - w_0(x)\right\} = 0 \tag{48}$$

for all $x \in \mathbb{R} \setminus \{\alpha\}$. *Moreover,* $w_1(x) \geq -K$ *if and only if* $K \geq K_0$.

We collect in the appendix the proofs of those results that are not developed in the text.

If the condition $K \geq K_0$ is not satisfied, we expect that abandonment becomes part of the optimal scenario. Now, assuming that the optimal strategy has a continuous qualitative character, we should expect that, as K_0 rises above K, abandonment should become optimal if the system is "open" and the state process X assumes sufficiently small values. The obvious modification of the strategy studied above is depicted by Figure 2. Such a possibility involves five parameters and three free boundary points, so we cannot impose a C^1 fit at all of the free boundary points.

By an obvious symmetry argument, we can conclude that the value function is C^1 at the points α, β, and C^0 at the point γ. However, by elementary considerations, we can see that the value function is nondecreasing in x. Therefore, if the optimal strategy identifies with the one depicted by Figure 2, we must have $w_1(\gamma-) = 0 < w_1(\gamma+)$, which is unacceptable in light of Remark 1. Alternatively, we can postulate that the value function is C^1 at γ and β (resp., α), and C^0 at α (resp., β). However, such a possibility would impose a discontinuity of the first derivative of the candidate value functions inside the interior of the "continuation" region, which is again contradicting the conclusions of Remark 1. It turns out that a strategy having the form depicted by Figure 2 cannot be optimal. However, the idea that the optimal strategy should possess a character which depends continuously on the problem's data leads us to the conclusion that we should look for a further modification of this strategy. Such

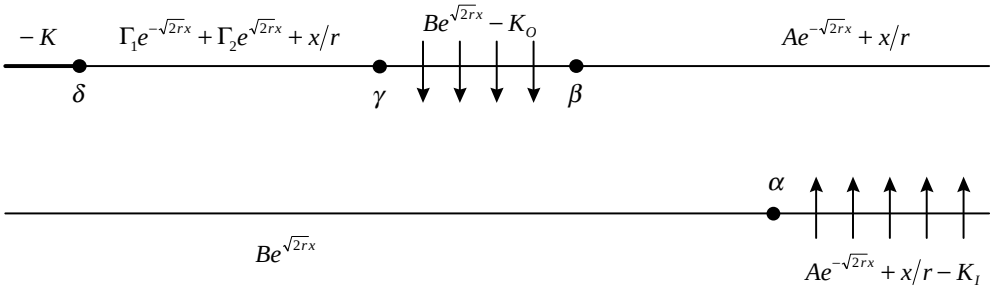


FIG. 3. *The case in which abandonment becomes part of the optimal tactics.*

a modification can be obtained by inserting a "do-not-abandon-or-switch-off" region around γ so that the interface of the "abandonment" and the "switch-off" regions is not just a point but an interval. This strategy is depicted by Figure 3.

If this case is indeed optimal, the value function of the control problem should identify with a solution w_1, w_0 of the HJB equations (31)–(32) described by

$$w_1(x) = \begin{cases} -K & \text{if } x \le \delta, \\ \Gamma_1 e^{-\sqrt{2r}x} + \Gamma_2 e^{\sqrt{2r}x} + x/r & \text{if } \delta < x < \gamma, \\ Be^{\sqrt{2r}x} - K_0 & \text{if } \gamma \le x \le \beta, \\ Ae^{-\sqrt{2r}x} + x/r & \text{if } x > \beta, \end{cases} \tag{49}$$

$$w_0(x) = \begin{cases} Be^{\sqrt{2r}x} & \text{if } x < \alpha, \\ Ae^{-\sqrt{2r}x} + x/r - K_1 & \text{if } x \ge \alpha. \end{cases} \tag{50}$$

The parameters A, B, Γ_1, Γ_2, α, β, γ, δ can then be specified by the requirement that w_1, w_0 are C^1 at the free boundary points α, β, γ, δ. Now, it is a straightforward calculation to verify that this requirement implies that α, β, A, B should satisfy (39)–(42),

$$\Gamma_1 = -\frac{\gamma + rK_0 - 1/\sqrt{2r}}{2r} e^{\sqrt{2r}\gamma}, \tag{51}$$

$$\Gamma_2 = B - \frac{\gamma + rK_0 + 1/\sqrt{2r}}{2r} e^{-\sqrt{2r}\gamma}, \tag{52}$$

and γ, δ should satisfy the system of equations

$$\begin{aligned} F_1(\gamma, \delta) &:= (\delta + rK - 1/\sqrt{2r}) e^{\sqrt{2r}\delta} - (\gamma + rK_0 - 1/\sqrt{2r}) e^{\sqrt{2r}\gamma} \\ &= 0, \end{aligned} \tag{53}$$

$$\begin{aligned} F_2(\gamma, \delta) &:= (\delta + rK + 1/\sqrt{2r}) e^{-\sqrt{2r}\delta} - (\gamma + rK_0 + 1/\sqrt{2r}) e^{-\sqrt{2r}\gamma} + 2rB \\ &= 0. \end{aligned} \tag{54}$$

The next lemma is concerned with the solvability of (53)–(54) as well as with necessary and sufficient conditions under which the functions w_1, w_0 considered above satisfy the HJB equations (31)–(32).

LEMMA 3. *Let $\alpha = \alpha(r, K_1, K_0)$, $\beta = \beta(r, K_1, K_0)$, A, B be as in Lemma 2. The system of equations* (53)–(54) *has a unique solution $\gamma = \gamma(r, K_1, K_0, K)$, $\delta =$*

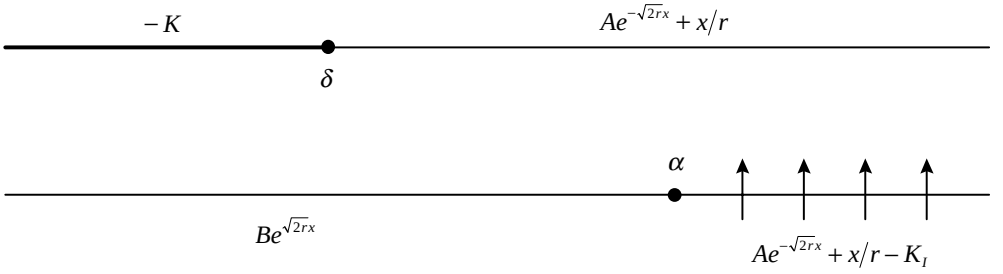


FIG. 4. *The case in which switching the system to its "closed" mode is never optimal.*

$\delta(r, K_1, K_0, K)$ *such that* $\delta < \gamma < \beta$ *if and only if*

$$K_* \vee 0 < K < K_0, \tag{55}$$

where $K_* = K_*(r, K_1, K_0) < K_0$ *is defined by*

$$K_* = -\frac{1}{r\sqrt{2r}} \ln\left(-\frac{\sqrt{2r}}{2}(\beta + rK_0 - 1/\sqrt{2r}) \exp(\sqrt{2r}(\beta + 1/\sqrt{2r}))\right). \tag{56}$$

If $K_* > 0$ *and* $K = K_*$, *then* $\gamma = \beta$, $\delta = -rK - 1/\sqrt{2r}$, $\Gamma_1 = A$, *and* $\Gamma_2 = 0$. *If* (55) *is true, then the functions* w_1, w_0 *defined by* (49), (50), *respectively, where* $\Gamma_1, \Gamma_2 > 0$ *are given by* (51)–(52), *are convex, nondecreasing,* C^1 *for all* $x \in \mathbb{R}$ *and* C^2 *for all* $x \in \mathbb{R} \setminus \{\delta, \gamma, \beta\}$, $x \in \mathbb{R} \setminus \{\alpha\}$, *respectively, and satisfy the HJB equations* (31)–(32).

The optimality of the case considered in the previous lemma depends crucially on the parameter K_*. If $K_* \leq 0$ for every admissible choice of the problem's data, then our solution is complete. However, it turns out that this is not the case in general.

LEMMA 4. *Given any values of the parameters* $r, K_1 > 0$, *the function* $K_*(r, K_1, \cdot)$ *is well defined on* $]-K_1, \infty[$, *is strictly increasing and at least* C^1 *on this interval, and satisfies* $\lim_{K_0 \to \infty} K_*(r, K, K_0) = \infty$ *and* $K_*(r, K, 0) < 0$.

In Lemma 3, we proved that if $K_* > 0$ and $K = K_*$, then $\gamma = \beta$, so the "switch-from-open-to-closed" region disappears, and the optimal strategy is depicted by Figure 4. For $K < K_*$, we can expect that it is not optimal to switch the system from its "open" to its "closed" operating mode at any time, so that the optimal strategy can again be depicted by Figure 4.

If this strategy is indeed optimal, the value function should be given in terms of the functions

$$w_1(x) = \begin{cases} -K & \text{if } x \leq \delta, \\ Ae^{-\sqrt{2r}x} + x/r & \text{if } x > \delta, \end{cases} \tag{57}$$

$$w_0(x) = \begin{cases} Be^{\sqrt{2r}x} & \text{if } x < \alpha, \\ Ae^{-\sqrt{2r}x} + x/r - K_1 & \text{if } x \geq \alpha. \end{cases} \tag{58}$$

Again, we require that w_1, w_0 are C^1 at the free boundary points δ, α, respectively. Straightforward calculations show that C^1 fit at δ yields

$$A = \frac{1}{r\sqrt{2r}} \exp(-\sqrt{2r}(rK + 1/\sqrt{2r})), \tag{59}$$

$$\delta = -rK - \frac{1}{\sqrt{2r}}, \tag{60}$$

whereas C^1 fit at α yields the system of equations

$$Be^{\sqrt{2r}\alpha} - Ae^{-\sqrt{2r}\alpha} = \frac{\alpha - rK_1}{r}, \tag{61}$$

$$Be^{\sqrt{2r}\alpha} + Ae^{-\sqrt{2r}\alpha} = \frac{1}{r\sqrt{2r}}, \tag{62}$$

which is equivalent to

$$B = \frac{\alpha - rK_1 + 1/\sqrt{2r}}{2r} e^{-\sqrt{2r}\alpha}, \tag{63}$$

$$\begin{aligned} G(\alpha) &:= \frac{\sqrt{2r}}{2}(\alpha - rK_1 - 1/\sqrt{2r}) \exp(\sqrt{2r}(\alpha + rK + 1/\sqrt{2r})) \\ &= -1. \end{aligned} \tag{64}$$

The next lemma is concerned with the solvability of (64) and with necessary and sufficient conditions under which this case is optimal.

LEMMA 5. *Equation* (64) *has a unique solution* $\alpha = \alpha(r, K_1, K)$ *such that* $\alpha > -rK - 1/\sqrt{2r}$. *For this value of* α, *and for* $A, B > 0$, *and* $\delta = \delta(r, K)$ *given by* (59), (63), *and* (60), *respectively, the functions* w_1, w_0 *defined by* (57)–(58) *are convex, nondecreasing and* C^1 *for all* $x \in \mathbb{R}$ *and* C^2 *for all* $x \in \mathbb{R} \setminus \{\delta\}$, $\mathbb{R} \setminus \{\alpha\}$, *respectively. Moreover, assuming that* $K_* > 0$, *they satisfy the HJB equations* (31)–(32) *if and only if* $0 < K \leq K_*$.

We can now prove the main result of the section.

THEOREM 6. *Consider the stochastic optimization problem defined in section* 2, *and suppose that its data are as in Assumption* 1. *The value function* v *is* C^1, *convex and nondecreasing in* x, *and is given by* $v(1, \cdot) = w_1$ *and* $v(0, \cdot) = w_0$, *where the following hold:*

(I) *If* $K_0 \leq K$, w_1, w_0 *are given by Lemma* 2 *(see Figure* 1).

(II) *If* $K_* < K < K_0$, *where* $K_* < K_0$ *is given by* (56), w_1, w_0 *are given by Lemma* 3 *(see Figure* 3).

(III) *If* $K_* > 0$ *and* $K \leq K_*$, w_1, w_0 *are given by Lemma* 5 *(see Figure* 4).

In each of the three cases, the optimal strategy can be constructed as in the proof below.

Proof. First, observe that, given any sequence $t_m \to \infty$, (11) and the second limit in (19) are true for all $(Z, T) \in \Pi_z$. Also, in each of the three cases, the functions w_1 and w_0 are convex and nondecreasing.

Now, consider any of the three cases. Since $w_1 \equiv w(1, \cdot)$ and $w_0 \equiv w(0, \cdot)$ are C^1 for all x and C^2 for all x outside a finite set, their second distributional derivatives are measures that are absolutely continuous with respect to the Lebesgue measure. With regard to the notation of Definition 1, this implies that $w^{\mathrm{s}}_{xx}(1, dx) \equiv 0$ and $w^{\mathrm{s}}_{xx}(0, dx) \equiv 0$, and, therefore, (12) as well as (14) are true. Moreover, $w_1 \equiv w(1, \cdot)$ and $w_0 \equiv w(0, \cdot)$ satisfy the HJB equations (31)–(32) in the classical sense, by construction, and, therefore, in the sense of Definition 1.

Since $w(1, \cdot)$ and $w(0, \cdot)$ have bounded first derivatives, the process M defined as in (13) is a square integrable martingale for all $Z \in \mathcal{Z}$. Furthermore, there exist constants C_1 and C_2 such that

$$w(z, x) \leq C_1 + C_2|x| \quad \forall z = 1, 0 \text{ and } x \in \mathbb{R}.$$

It follows that, given any $Z \in \mathcal{Z}$,

$$\lim_{t\to\infty} E\left[e^{-rt}\,|w(Z_t, X_t)|\right] \le \lim_{t\to\infty} E\left[e^{-rt}\,(C_1 + C_2\,|x + W_t|)\right] = 0.$$

However, this shows that (18) is satisfied for all $Z \in \mathcal{Z}$ and, therefore, for the optimal switching process.

The above arguments prove that, in any of the three cases, $w_1 \equiv w(1,\cdot)$ and $w_0 \equiv w(0,\cdot)$ satisfy all of the assumptions related to part (a) of Theorem 1 as well as (14) and (19). As a consequence, to complete the proof, we have to construct a strategy (Z^*, T^*) satisfying (15)–(18).

Now, in Case I, if $z = 1$, then we can see that the strategy (Z^*, T^*), where $T^* = \infty$ and the process $Z^* \in \mathcal{Z}$ is defined by

$$Z_t^* = 1_{\{t=0\}} + \sum_{j=0}^{\infty} 1_{\{T_{2j}^* < t \le T_{2j+1}\}^*}, \tag{65}$$

where $T_0^* = 0$ and the stopping times T_n^*, $n \in \mathbb{N}^*$, are defined recursively by

$$T_{2n+1}^* = \inf\{t \ge T_{2n}^* \mid\ X_t \le \beta\}, \quad n = 0, 1, 2, \ldots, \tag{66}$$

$$T_{2n}^* = \inf\{t \ge T_{2n-1}^* \mid\ X_t \ge \alpha\}, \quad n = 1, 2, \ldots, \tag{67}$$

satisfies (15)–(18). If $z = 0$, we again have $T^* = \infty$, and the optimal switching process Z^* can be constructed in a similar fashion.

In Case II, if $z = 1$ and $x \ge \gamma$ or if $z = 0$, the optimal strategy is the same as in Case I. If $z = 1$ and $x \le \delta$, then the optimal strategy is characterized by $T^* = 0$. If $z = 1$ and $\delta < x < \gamma$, then define

$$T_\delta = \inf\{t \ge 0 \mid\ X_t \le \delta\} \quad \text{and} \quad T_\gamma = \inf\{t \ge 0 \mid\ X_t \ge \gamma\}, \tag{68}$$

and let

$$T^* = T_\delta 1_{\{T_\delta < T_\gamma\}} + \infty 1_{\{T_\delta > T_\gamma\}} \in \mathcal{S}.$$

Also let $T_0^* = 0$, $T_1^* = T_\gamma$, and define T_n^*, $n \ge 2$, as in (67). Then we can see that the strategy (Z^*, T^*) where Z^* is defined as in (65), satisfies (15)–(18).

In Case III, if $z = 1$, then $T^* = T^\delta$, where T_δ is defined as in (68), and Z^*, defined by $Z_t^* = 1$ for all $t \ge 0$, provide the optimal strategy. Finally, if $z = 0$, then $T^* = \inf\{t \ge T_\alpha \mid\ X_t \le \delta\}$, where $T_\alpha = \inf\{t \ge 0 \mid\ X_t \ge \alpha\}$, and Z^* defined by $Z_t^* = 1_{\{T_\alpha < t\}}$, $t \ge 0$, are the optimal strategy, and the proof is complete. □

Remark 3. The rather unexpected qualitative nature of Case II is intimately related to optimal stopping. To understand this claim, consider the case as a perturbation of Case I, where stopping is not part of the optimal strategy. With regard to the heuristic discussion at the beginning of the section, abandonment can be optimal only if the system is "open" and the state process X assumes sufficiently low values. As a result, the optimal strategy should possess the same qualitative nature as in Case I if the system is "closed" or if the system is "open" and the state process X assumes sufficiently large values. Now, as the abandonment cost K falls marginally below the critical value K_0 and abandonment comes into the picture, a continuity argument dictates that the switching boundary points α and β as given by Lemma 2 should be "close" to the optimal ones. However, these points determine completely the function w_0. As a consequence, for sufficiently low values of X, the function w_1

should be "close" to the value function of the purely optimal stopping problem which seeks to maximize

$$E\left[\int_0^\tau e^{-rs}X_s\,ds - e^{-r\tau}\left[K \vee (K_0 - w_0(X_\tau))\right]\right]$$

over all stopping times $\tau \in \mathcal{S}$. In fact, we have proved that w_1 identifies with the value function of this purely optimal stopping problem for appropriate parameter values. Note that the terminal payoff function $-K \vee (K_0 - w_0(\cdot))$ of this problem is not C^1. From these observations, we can conclude that the existence of a "continuation" region such as the interval $]\delta, \gamma[$ in Case II should characterize the optimal strategy in purely optimal stopping problems where the first derivative of the terminal payoff function has appropriate discontinuities.

5. An example in which C^1 regularity of the value function fails. Based on the results established in the previous sections, we can easily construct an example whose value function is not composed by C^1 functions. To this end, consider the problem formulated in section 2, and assume that $\mathcal{I} = \mathbb{R}$, and

$$b(x) = 0, \quad \sigma(x) = 1, \quad r(x) = \frac{1}{2}, \quad H_1(x) = x, \quad H_0(x) = 0,$$
$$G_1(x) = 2, \quad G_0(x) = 10, \quad F(x) = \begin{cases} 5 - e^{x+10} & \text{if } x < -10, \\ 4 & \text{if } x \geq -10 \end{cases}$$

for all $x \in \mathbb{R}$.

With regard to (43) and (56), we calculate

$$\beta\left(\frac{1}{2}, 2, 10\right) = -5.999328399 \quad \text{and} \quad K_*\left(\frac{1}{2}, 2, 10\right) = 9.999328511. \tag{69}$$

Since $K_* > F(x)$, for all $x \in \mathbb{R}$, this example is akin to Case III of Theorem 6. The values of the associated parameters are

$$\delta\left(\frac{1}{2}, 4\right) = -3, \quad \alpha\left(\frac{1}{2}, 2, 4\right) = 1.986338745,$$
$$A = 0.099574137, \quad \text{and} \quad B = 0.272519358.$$

The value function of the example under consideration is given by

$$v(1,x) = \begin{cases} e^{x+10} - 5 & \text{if } x < -10, \\ -4 & \text{if } -10 \leq x < -3, \\ 0.099574137e^{-x} + 2x & \text{if } -3 \leq x, \end{cases}$$
$$v(0,x) = \begin{cases} 0.272519358e^{x} & \text{if } x < 1.9886338745, \\ 0.099574137e^{-x} + 2x - 2 & \text{if } x \geq 1.9886338745. \end{cases}$$

To see this, observe first that the only point where C^1 regularity fails is given by $z = 1$ and $x = -10$. Clearly, (12) and (14) are satisfied (see also Remark 1). Now, with reference to the proofs of Lemma 5 and Theorem 6, all of the assumptions of Theorem 1 will follow if we verify that

$$\frac{1}{2}w_1''(x) - \frac{1}{2}w_1(x) + x \leq 0 \quad \text{for } x < -3,$$
$$w_0(x) - 10 - w_1(x) \leq 0 \quad \text{for } x < -3,$$
$$w_1(x) - 2 - w_0(x) \leq 0 \quad \text{for } x < -3.$$

However, this is a trivial exercise.

Appendix: Proofs of results in section 4.

Proof of Lemma 2. Multiplying (41) and (42) side by side and solving for α, we obtain

$$\alpha = \beta + r(K_1 + K_0) \quad \text{or} \quad \alpha = -\beta - rK_0 + rK_1.$$

Substituting $\beta + r(K_1 + K_0)$ for α in (41) and (42), we obtain

$$(\beta + rK_0 - 1/\sqrt{2r})e^{r\sqrt{2r}(K_1+K_0)} = \beta + rK_0 - 1/\sqrt{2r},$$
$$(\beta + rK_0 + 1/\sqrt{2r})e^{-r\sqrt{2r}(K_1+K_0)} = \beta + rK_0 + 1/\sqrt{2r},$$

respectively. Since $K_1 + K_0 > 0$, there is no β satisfying both of these equations. Therefore, α must be as in (46). Now, (46) and either (41) or (42) yield (43).

Since $H(\beta) > 0$, for all $\beta < -rK_0 - 1/\sqrt{2r}$ and all $\beta > -rK_0 + 1/\sqrt{2r}$, if (43) has a solution, then this has to satisfy $-rK_0 - 1/\sqrt{2r} \le \beta \le -rK_0 + 1/\sqrt{2r}$. Now,

$$H'(\beta) = -2\sqrt{2r}\frac{(\beta + rK_0)^2}{(\beta + rK_0 - 1/\sqrt{2r})^2}\exp(-\sqrt{2r}(2\beta + rK_0 - rK_1)),$$

which implies that H is strictly decreasing in $\mathbb{R} \setminus \{-rK_0, -rK_0 + 1/\sqrt{2r}\}$. Combining this with the facts that $H(-rK_0 - 1/\sqrt{2r}) = 0$ and $H(-rK_0) = -\exp(r\sqrt{2r}(K_1 + K_0)) < -1$, we conclude that (43) has a unique solution which satisfies (44). Observe that the inequality $\beta < -rK_0$ and the expression $\alpha = -\beta - rK_0 + rK_1$ imply trivially that $\beta < \alpha$. Also, (44) implies (45) because the function $x \to xe^x$ is strictly decreasing in $]-\infty, -1[$. Furthermore, (44) along with (39) and (40) imply that $A, B > 0$.

Since $A, B > 0$, the functions w_1, w_0 are convex and nondecreasing. As a consequence, $-K \le w_1$ if and only if $K \ge K_0$, and $-K \le w_0$. Now, to verify that w_1, w_0 satisfy (47) and (48), we have to prove that

$$\frac{1}{2}w_1''(x) - rw_1(x) + x \le 0 \quad \text{for } x < \beta, \tag{70}$$
$$w_1(x) - K_1 - w_0(x) \le 0 \quad \text{for } x \le \beta, \tag{71}$$
$$w_0(x) - K_0 - w_1(x) \le 0 \quad \text{for } \beta \le x \le \alpha, \tag{72}$$
$$w_1(x) - K_1 - w_0(x) \le 0 \quad \text{for } \beta \le x \le \alpha, \tag{73}$$
$$w_0(x) - K_0 - w_1(x) \le 0 \quad \text{for } x \ge \alpha, \tag{74}$$
$$\frac{1}{2}w_0''(x) - rw_0(x) \le 0 \quad \text{for } x > \alpha. \tag{75}$$

Each of (71) and (74) is equivalent to $-K_1 - K_0 \le 0$, which is true. Inequality (70) is trivially implied by $rK_0 + \beta < 0$ (see (44)), whereas (75) is trivially implied by $-\alpha + rK_1 = \beta + rK_0 < 0$.

Now, consider the function g defined by

$$g(x) := Be^{\sqrt{2r}x} - Ae^{-\sqrt{2r}x} - \frac{x}{r} - K_0.$$

Since $g(x) = w_0(x) - K_0 - w_1(x)$ for $x \in [\beta, \alpha]$, (72) and (73) will follow if we prove that

$$-K_1 - K_0 \le g(x) \le 0 \quad \forall x \in [\beta, \alpha]. \tag{76}$$

The function g' is strictly convex because

$$g'''(x) = 2r\sqrt{2r}(Be^{\sqrt{2r}x} + Ae^{-\sqrt{2r}x}) > 0,$$

the inequality being true because $A, B > 0$. As a consequence, $g'(x) < 0$ for all $x \in]\beta, \alpha[$ because $g'(\beta) = g'(\alpha) = 0$ by construction. However, combining this observation with the fact that $g(\beta) = 0$ and $g(\alpha) = -K_0 - K_1$, we conclude that (76) is true. □

Proof of Lemma 3. Fix any $\gamma < -rK_0$, and consider the equation

$$f(\delta) := F_1(\gamma, \delta) = 0. \tag{77}$$

From the calculations

$$\lim_{\delta \to -\infty} f(\delta) = -(\gamma + rK_0 - 1/\sqrt{2r})e^{\sqrt{2r}\gamma} > 0, \tag{78}$$

$$f'(\delta) = \sqrt{2r}(\delta + rK)e^{\sqrt{2r}\delta}, \tag{79}$$

$$f(\gamma) = r(K - K_0)e^{\sqrt{2r}\gamma}, \tag{80}$$

we can see that (77) has a unique solution $\delta < \gamma$ if $K < K_0$. These calculations also imply that (77) does not have a solution $\delta < \gamma$ if $K > K_0$ and $\gamma < -rK$. If $K > K_0$ and $-rK < \gamma$, (77) will have a solution only if f is negative at $\delta = -rK$, where its minimum over $\delta \in]-\infty, \gamma]$ occurs. However,

$$f(-rK) = -(\gamma + rK_0)e^{\sqrt{2r}\gamma} + \frac{1}{\sqrt{2r}}[e^{\sqrt{2r}\gamma} - e^{\sqrt{2r}(-rK)}] > 0,$$

the inequality following because $-rK < \gamma < -rK_0$. From these considerations, we conclude that, given any $\gamma < -rK_0$, (77) has a unique solution $\delta < \gamma$ if and only if $K < K_0$. For the rest of this proof, we assume that this condition is satisfied.

From the above, we can see that, as γ varies, (77) uniquely defines a function $\delta = \delta(\gamma)$ on $]-\infty, -rK_0[$ such that $\delta(\gamma) < \gamma$ for all $\gamma < -rK_0$. Also, by implicit differentiation of (77), we obtain

$$\delta'(\gamma) = \frac{\gamma + rK_0}{\delta(\gamma) + rK} e^{\sqrt{2r}(\gamma - \delta(\gamma))}. \tag{81}$$

Now consider the equation

$$g(\gamma) := F_2(\gamma, \delta(\gamma)) = 0 \quad \text{for } \gamma < \beta. \tag{82}$$

Since

$$g(\gamma) = \sqrt{2r} \int_{\delta(\gamma)}^{\gamma} e^{-\sqrt{2r}s}(s + rK)\, ds + r(K - K_0)e^{-\sqrt{2r}\gamma} + 2rB$$

and $K < K_0$, it follows that

$$\lim_{\gamma \to -\infty} g(\gamma) = -\infty. \tag{83}$$

Also, using (81), we can calculate

$$g'(\gamma) = -2\sqrt{2r}(\gamma + rK_0)e^{-\sqrt{2r}\delta(\gamma)} \sinh[\sqrt{2r}(\gamma - \delta(\gamma))] > 0, \tag{84}$$

whereas, in view of (40),

$$g(\beta) = (\delta(\beta) + rK + 1/\sqrt{2r})e^{-\sqrt{2r}\delta(\beta)}. \tag{85}$$

From (83)–(85), we can see that (82) has a unique solution $\gamma < \beta$ if and only if $\delta(\beta) > -rK - \frac{1}{\sqrt{2r}}$. With regard to the analysis relating to (77), this will be true if and only if $F_1(\beta, -rK - 1/\sqrt{2r}) > 0$, i.e., if and only if

$$-2e^{-2} > \sqrt{2r}(\beta + rK_0 - 1/\sqrt{2r}) \exp(\sqrt{2r}(\beta + rK - 1/\sqrt{2r})). \tag{86}$$

With reference to (45), this is true for $K = K_0$. Furthermore, the right-hand side of this inequality is increasing as K decreases. As a consequence, (86) is true for all $K \in]K_* \vee 0, K_0[$, where $K_* < K_0$ is given by (56).

With regard to the arguments above, if $K_* > 0$ and $K = K_*$, then (86) holds with equality, $\gamma = \beta$, and $\delta = -rK - 1/\sqrt{2r}$. From (39), (51) and (40), (52), it then follows that $\Gamma_1 = A$ and $\Gamma_2 = 0$, respectively.

Since $\gamma < \beta < -rK_0$, (51) implies that $\Gamma_1 > 0$. Furthermore, since

$$\frac{d}{dy}\left[\frac{y + rK_0 + 1/\sqrt{2r}}{2r} e^{-\sqrt{2r}y}\right] = -\frac{1}{\sqrt{2r}}(y + rK_0)e^{-\sqrt{2r}y} > 0 \quad \forall y < -rK_0,$$

it follows that

$$\frac{\beta + rK_0 + 1/\sqrt{2r}}{2r} e^{-\sqrt{2r}\beta} > \frac{\gamma + rK_0 + 1/\sqrt{2r}}{2r} e^{-\sqrt{2r}\gamma}.$$

Therefore, (40) and (52) imply $\Gamma_2 > 0$.

Since $A, B, \Gamma_1, \Gamma_2 > 0$, the functions w_1, w_0 are convex and nondecreasing, so $w_1, w_0 \geq -K$. Also, all of the inequalities associated with the HJB equations (31)–(32) for $x \geq \gamma$ follow from Lemma 2. Therefore, to verify that w_1, w_0 satisfy the HJB equations (31)–(32), it remains to show that

$$\frac{1}{2}w_1''(x) - rw_1(x) + x \leq 0 \quad \text{for } x < \delta, \tag{87}$$

$$w_0(x) - K_0 - w_1(x) \leq 0 \quad \text{for } x \leq \delta, \tag{88}$$

$$w_1(x) - K_1 - w_0(x) \leq 0 \quad \text{for } x \leq \delta, \tag{89}$$

$$w_0(x) - K_0 - w_1(x) \leq 0 \quad \text{for } \delta \leq x \leq \gamma, \tag{90}$$

$$w_1(x) - K_1 - w_0(x) \leq 0 \quad \text{for } \delta \leq x \leq \gamma. \tag{91}$$

The inequalities $\delta + rK < \beta + rK_0 < 0$ trivially imply (87). Also, since $B > 0$, (89) is straightforward, whereas (88) is implied by (90) and the continuity of w_1, w_0.

Now, consider the function g defined by

$$g(x) := -\Gamma_1 e^{-\sqrt{2r}x} + (B - \Gamma_2)e^{\sqrt{2r}x} - \frac{x + rK_0}{r}.$$

Since $g(x) = w_0(x) - K_0 - w_1(x)$, if $x \in [\delta, \gamma]$, (90) and (91) will follow if we show that

$$-K_1 - K_0 \leq g(x) \leq 0 \quad \forall x \in [\delta, \gamma]. \tag{92}$$

By construction,

$$g(\gamma) = g'(\gamma) = 0 \quad \text{and} \quad g''(\gamma) = 2(\gamma + rK_0) < 0. \tag{93}$$

If $B - \Gamma_2 < 0$, then

$$g''(x) = 2r[-\Gamma_1 e^{-\sqrt{2r}x} + (B - \Gamma_2)e^{\sqrt{2r}x}] < 0,$$

so g' is strictly decreasing, which, combined with $g'(\gamma) = 0$, implies $g'(x) > 0$ for all $x < \gamma[$. On the other hand, if $B - \Gamma_2 > 0$, then

$$g'''(x) = 2r\sqrt{2r}[\Gamma_1 e^{-\sqrt{2r}x} + (B - \Gamma_2)e^{\sqrt{2r}x}] > 0,$$

which proves that g' is strictly convex. However, this observation and (93) imply that $g'(x) > 0$ for all $x < \gamma$. Finally, since g is increasing in $[\delta, \gamma]$ and $g(\gamma) = 0$, (92) follows from the observation that

$$g(\delta) = Be^{\sqrt{2r}\delta} - K_0 + K > -K_1 - K_0. \qquad \square$$

Proof of Lemma 4. Suppose that the values of $r, K_1 > 0$ are fixed, and consider the unique solution $\beta = \beta(r, K_1, K_0)$ of (43) as a function of K_0 on $]-K_1, \infty[$. By implicit differentiation of (43), we obtain

$$\frac{\partial \beta}{\partial K_0} + r = \frac{r(\beta + rK_0 + 1/\sqrt{2r})(\beta + rK_0 - 1/\sqrt{2r})}{2(\beta + rK_0)^2}. \tag{94}$$

Now, differentiating (56) with respect to K_0, we obtain

$$\frac{\partial K_*}{\partial K_0} = -\frac{1}{r(\beta + rK_0 - 1/\sqrt{2r})}\left[(\beta + rK_0)\left(\frac{\partial \beta}{\partial K_0} + r\right) - r(\beta + rK_0 - 1/\sqrt{2r})\right].$$

Substituting for $\partial\beta/\partial K_0 + r$ from (94), we obtain

$$\frac{\partial K_*}{\partial K_0} = \frac{\beta + rK_0 - 1/\sqrt{2r}}{2(\beta + rK_0)} > 0,$$

the inequality following because $\beta < -rK_0$. As a consequence, $K_*(r, K_1, \cdot)$ is strictly increasing in $]-K_1, \infty[$.

Now, (45) and (56) imply

$$K_0 > K_* > -\frac{1 - \ln 2}{r\sqrt{2r}} + K_0.$$

However, these inequalities imply $\lim_{K_0 \to \infty} K_*(r, K_1, K_0) = \infty$ and $K_*(r, K_1, 0) < 0$, and the proof is complete. $\square$

Proof of Lemma 5. The fact that (64) has a unique solution $\alpha > -rK - 1/\sqrt{2r}$ follows from the calculations

$$\begin{aligned} G(-rK - 1/\sqrt{2r}) &= -\frac{\sqrt{2r}}{2} r(K + K_1) - 1 < -1, \\ G'(\alpha) &= r(\alpha - rK_1)\exp(\sqrt{2r}(\alpha + rK + 1/\sqrt{2r})), \\ \lim_{\alpha \to \infty} G(\alpha) &= \infty. \end{aligned}$$

Also, this solution satisfies

$$rK_1 < \alpha < rK_1 + 1/\sqrt{2r}, \tag{95}$$

the second inequality holding because $G(rK_1 + 1/\sqrt{2r}) = 0 > -1$.

Now, (63) and (95) imply $B > 0$, whereas $A > 0$ is obvious from (59). Since $A, B > 0$, w_1, w_0 are convex and nondecreasing, $w_1, w_0 \geq -K$. To verify that they satisfy the HJB equations (31)–(32), we have to establish conditions under which

$$\frac{1}{2}w_1''(x) - rw_1(x) + x \leq 0 \quad \text{for } x \leq \delta, \tag{96}$$

$$w_0(x) - K_0 - w_1(x) \leq 0 \quad \text{for } x \leq \delta, \tag{97}$$

$$w_1(x) - K_1 - w_0(x) \leq 0 \quad \text{for } x \leq \delta, \tag{98}$$

$$w_0(x) - K_0 - w_1(x) \leq 0 \quad \text{for } \delta \leq x \leq \alpha, \tag{99}$$

$$w_1(x) - K_1 - w_0(x) \leq 0 \quad \text{for } \delta \leq x \leq \alpha, \tag{100}$$

$$w_0(x) - K_0 - w_1(x) \leq 0 \quad \text{for } x \geq \alpha, \tag{101}$$

$$\frac{1}{2}w_0''(x) - rw_0(x) \leq 0 \quad \text{for } x \geq \alpha. \tag{102}$$

Inequalities (96) and (102) are implied trivially by the fact that $\delta = -rK - 1/\sqrt{2r}$ and the first inequality in (95), respectively. Also, (101) is equivalent to $-K_1 - K_0$, which is true, whereas (98) follows immediately because $B > 0$. In view of the continuity of w_1, w_0 and the fact that $B > 0$, we can also see that (97) is implied by (99).

To study (99), (100), define the function g by

$$g(x) := Be^{\sqrt{2r}x} - Ae^{-\sqrt{2r}x} - \frac{x + rK_0}{r}$$

so that $g(x) = w_0(x) - K_0 - w_1(x)$ if $x \in [\delta, \alpha]$. By construction,

$$g(\alpha) = -K_1 - K_0, \quad g'(\alpha) = 0, \quad g''(\alpha) = 2(\alpha - rK_1) > 0, \tag{103}$$

the inequality following by virtue of (95). Now, since $A, B > 0$,

$$g'''(x) = 2r\left[g'(x) + \frac{1}{r}\right] = 2r\sqrt{2r}[Be^{\sqrt{2r}x} + Ae^{-\sqrt{2r}x}] > 0 \tag{104}$$

imply that g' is strictly convex and $\lim_{x\to-\infty} g'(x) = \infty$. However, these observations and (103) imply that there exists a unique $\hat{x} < \alpha$ such that $g'(\hat{x}) = 0$. Furthermore, $\delta > \hat{x}$ because $g'(\delta) = \sqrt{2r}Be^{\sqrt{2r}\delta} > 0$. From these considerations, we conclude that

$$g'(x) > 0 \quad \forall x \in [\delta, \hat{x}[\qquad \text{and} \qquad g'(x) < 0 \quad \forall x \in]\hat{x}, \alpha[. \tag{105}$$

Now, (100) follows from the fact that

$$-K_1 - K_0 \leq g(x) \quad \forall x \in [\delta, \alpha],$$

which is true in view of (98) and the continuity of w_1, w_0, (103), and (105). On the other hand, (99) will follow if we show that $g(x) \leq 0$ for all $x \in [\delta, \alpha]$. In view of (105), this will be true if and only if $g(\hat{x}) \leq 0$, i.e., if and only if

$$Be^{\sqrt{2r}\hat{x}} - Ae^{-\sqrt{2r}\hat{x}} \leq \frac{\hat{x} + rK_0}{r}. \tag{106}$$

All of the results proved above are true for any positive values of the problem's data r, K_1, K_0, K. Therefore, given any positive value of these parameters, there exists a unique $\Delta \in \mathbb{R}$ such that

$$Be^{\sqrt{2r}\hat{x}} - Ae^{-\sqrt{2r}\hat{x}} = \frac{\hat{x} + r(K_0 + \Delta)}{r}. \tag{107}$$

Clearly, (106) will be true if and only if $\Delta \leq 0$. Now, recall that $\hat{x} \in]\delta, \alpha[$ satisfies $g'(\hat{x}) = 0$, i.e.,

$$Be^{\sqrt{2r}\hat{x}} + Ae^{-\sqrt{2r}\hat{x}} = \frac{1}{r\sqrt{2r}}. \tag{108}$$

With regard to these two equations, we can eliminate B, substitute for A from (59), and solve for K to obtain

$$K = -\frac{1}{r\sqrt{2r}} \ln\left(-\frac{\sqrt{2r}}{2}(\hat{x} + r(K_0 + \Delta) - 1/\sqrt{2r}) \exp(\sqrt{2r}(\hat{x} + 1/\sqrt{2r})) \right). \tag{109}$$

Furthermore, by comparing (61), (62), (107), (108) with (35), (36), (37), (38), respectively, we can see that $\hat{x} = \beta(r, K_1, K_0 + \Delta)$, where β is given by Lemma 2. Therefore, (56) and (109) imply $K = K_*(r, K_1, K_0 + \Delta)$. Since $K_*(r, K_1, \cdot)$ is strictly increasing (see Lemma 4), $\Delta \leq 0$ if and only if $K \leq K_*(r, K_1, K_0)$. However, these arguments establish that (106) is true if and only if $K \leq K_*(r, K_1, K_0)$, and the proof is complete.

Acknowledgments. I am grateful to Kate Duckworth for several discussions on the problem solved in this paper. A preliminary exposition of part of the results obtained here has been presented in Duckworth and Zervos [DuZ2].

REFERENCES

[B] V. E. Beneš, *Some combined control and stopping problems*, paper presented at the CRM Workshop on Stochastic Systems, Montréal, Canada, 1992.

[BØ1] K. A. Brekke and B. Øksendal, *The high contact principle as a sufficiency condition for optimal stopping*, in Stochastic Models and Option Values, D. Lund and B. Øksendal, eds., North–Holland, Amsterdam, 1991, pp. 187–208.

[BØ2] K. A. Brekke and B. Øksendal, *Optimal switching in an economic activity under uncertainty*, SIAM J. Control Optim., 32 (1994) pp. 1021–1036.

[BS] M. J. Brennan and E. S. Schwartz, *Evaluating natural resource investments*, J. Business, 58 (1985), pp. 135–157.

[D] A. Dixit, *Entry and exit decisions under uncertainty*, J. Political Economy, 97 (1989), pp. 620–638.

[DP] A. K. Dixit and R. S. Pindyck, *Investment under Uncertainty*, Princeton University Press, Princeton, NJ, 1994.

[DZ] M. H. A. Davis and M. Zervos, *A problem of singular stochastic control with discretionary stopping*, Ann. Appl. Probab., 4 (1994), pp. 226–240.

[DuZ1] K. Duckworth and M. Zervos, *A model for investment decisions with switching costs*, Ann. Appl. Probab., 11 (2001), pp. 239–260.

[DuZ2] K. Duckworth and M. Zervos, *A problem of stochastic impulse control with discretionary stopping*, in Proceedings of the 39th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 2000, pp. 222–227.

[FS] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[G] X. Guo, *An explicit solution to an optimal stopping problem with regime switching*, J. Appl. Probab., 38 (2001), pp. 464–481.

[KO] I. Karatzas and D. Ocone, *A leavable bounded-velocity stochastic control problem*, Stochastic Process. Appl., 99 (2002), pp. 31–51.

[KOWZ] I. Karatzas, D. Ocone, H. Wang, and M. Zervos, *Finite-fuel singular control with discretionary stopping*, Stochastics Stochastics Rep., 71 (2000), pp. 1–50.

[KS] I. Karatzas and W. D. Sudderth, *Control and stopping of a diffusion process on an interval*, Ann. Appl. Probab., 9 (1999), pp. 188–196.

[KW] I. Karatzas and H. Wang, *Utility maximization with discretionary stopping*, SIAM J. Control Optim., 39 (2000), pp. 306–329.

[K] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.

[LZ] R. R. Lumley and M. Zervos, *A model for investments in the natural resource industry with switching costs*, Math. Oper. Res., 26 (2001), pp. 637–653.

[RY] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*, Springer-Verlag, New York, 1991.

[SZ] S. P. Sethi and Q. Zhang, *Hierarchical Decision Making in Stochastic Manufacturing Systems*, Birkhäuser Boston, Boston, 1994.

[T] L. Trigeorgis, *Real Options: Managerial Flexibility and Strategy in Resource Allocation*, MIT Press, Cambridge, MA, 1996.

[YZ] J. Yong and X. Y. Zhou, *Stochastic Controls*, Springer-Verlag, New York, 1999.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 422–441

STATE CONSTRAINED FEEDBACK STABILIZATION*

F. H. CLARKE† AND R. J. STERN‡

Abstract. A standard finite dimensional nonlinear control system is considered, along with a state constraint set S and a target set Σ. It is proven that open loop S-constrained controllability to Σ implies closed loop S-constrained controllability to the closed δ-neighborhood of Σ, for any specified $\delta > 0$. When the S-constrained minimum time function to Σ satisfies a local continuity condition, conclusions on closed loop S-constrained stabilizability ensue. The (necessarily discontinuous) feedback laws in question are implemented in the sample-and-hold sense and possess a robustness property with respect to state measurement errors. The feedback constructions involve the quadratic infimal convolution of a control Lyapunov function with respect to a certain modification of the original dynamics. The modified dynamics in effect provide for constraint removal, while the convolution operation provides a useful semiconcavity property.

Key words. asymptotic controllability, state constraint, semiconcave control Lyapunov function, constraint removal, feedback, robustness

AMS subject classifications. 93D15, 93D20

PII. S036301290240453X

1. Introduction.

We shall consider a control system of the form

$$\dot{x}(t) = f(x(t), u(t)) \quad \text{a.e.}, \quad u(t) \in U. \tag{1}$$

The state trajectory $x(\cdot)$ evolves in $\mathbb{R}^n$ and control functions $u(\cdot)$ are Lebesgue measurable functions $u : \mathbb{R} \to U$, where $U \subset \mathbb{R}^m$ is a compact control constraint set. We shall assume throughout that the above dynamics satisfy the following standard hypotheses:

(F1) The function $f : \mathbb{R}^n \times U \to \mathbb{R}^n$ is continuous and is locally Lipschitz in the state variable x, uniformly for $u \in U$; that is, for each bounded set $\Gamma \subset \mathbb{R}^n$, there exists $K_\Gamma > 0$ such that

$$\|f(x,u) - f(y,u)\| \le K_\Gamma \|x - y\|,$$

whenever (x,u) and (y,u) are in $\Gamma \times U$.

(F2) The function f possesses linear growth; that is, there exist positive numbers c_1, c_2 such that

$$\|f(x,u)\| \le c_1\|x\| + c_2 \qquad \forall\, (x,u) \in \mathbb{R}^n \times U.$$

(F3) The velocity set

$$f(x,U) := \{f(x,u) : u \in U\}$$

is convex for every $x \in \mathbb{R}^n$.

*Received by the editors March 25, 2002; accepted for publication (in revised form) October 31, 2002; published electronically May 12, 2003.

http://www.siam.org/journals/sicon/42-2/40453.html

†Institut Desargues (Bât 101), Université Claude Bernard Lyon I, 69622 Villeurbanne, France (clarke@desargues.univ-lyon1.fr).

‡Department of Mathematics and Statistics, Concordia University, Montreal, Quebec H4B 1R6, Canada (stern@vax2.concordia.ca). The research of this author was supported by the Natural Sciences Engineering Research Council of Canada and Le Fonds pour la Formation de Chercheurs et l'Aide à la Recherche du Québec.

Under (F1)–(F2), for every initial phase $(\tau, \alpha) \in \mathbb{R} \times \mathbb{R}^n$ and every control function $u(\cdot)$, there exists a unique trajectory $x(t) = x(t; \tau, \alpha, u(\cdot))$ defined for $t \geq \tau$ and satisfying $x(\tau) = \alpha$.

Remark 1.1. Actually, for our purposes, (F2) could be replaced by the somewhat weaker hypothesis that $f(\Gamma, U)$ is bounded for any bounded set $\Gamma \subseteq \mathbb{R}^n$. (See also section 5.3 below with regard to this issue.) Assumption (F3) will be needed below in order to have available a required sequential compactness property of trajectories. On the other hand, in the absence of (F3), the results of this article could be framed in the context of relaxed controls.

A general problem of considerable theoretical as well as applied interest, and one which has received much attention in recent years, is whether *open loop asymptotic controllability* of the origin implies *closed loop stabilization*. We need not give precise definitions of these properties here, but roughly speaking, open loop asymptotic controllability means that for every initial state in $\mathbb{R}^n$, there exists a control function so that the resulting trajectory of (1) is driven asymptotically to the origin and that this property holds in a certain uniform and Lyapunov stable manner. Closed loop stabilizability of the origin involves the existence of a feedback law $k : \mathbb{R}^n \to U$ such that all solutions of the ordinary differential equation

$$\dot{x}(t) = f(x(t), k(x(t))) \tag{2}$$

asymptotically approach the origin, again, in a uniform and Lyapunov stable manner.

A minimal condition for the existence of classical solutions to the ordinary differential equation (2) is that the feedback law $k(\cdot)$ be continuous on $\mathbb{R}^n \backslash \{0\}$. However, as was shown by Sontag and Sussman [36], even when $m = n = 1$, such a feedback law $k(\cdot)$ need not exist. A further negative result in this regard was provided by Brockett [4], who derived a topological condition on the dynamics, which is necessary for the existence of a stabilizing feedback law which is continuous on $\mathbb{R}^n$, and exhibited an example violating this condition, in spite of its global open loop controllability to the origin. In addition, Ryan [32] showed that Brockett's necessary condition persists even when Filippov solutions are considered. The upshot is that in addressing the above problem, due to the fact that continuity of feedback laws cannot be expected, it is advantageous to work with an alternative solution concept for (2), rather than the classical or Filippov ones. On the other hand, if nonautonomous feedbacks of the form $k(t, x)$ are allowed (and they are not, in our problem, which calls for a purely positional feedback law $k(x)$), then continuity is not precluded; see Coron [15] and Coron and Rosier [16].

Clarke et al. [8] obtained an affirmative answer to the above problem in terms of the following "sample-and-hold" solution concept for (2), where $k(\cdot)$ is in general discontinuous. Let an initial state $\alpha \in \mathbb{R}^n$ be specified. Then given a partition

$$\pi = \{t_0, t_1, t_2, \ldots\} \tag{3}$$

of $[0, \infty)$ (where $t_0 = 0$), the associated π-*trajectory* $x(\cdot)$ on $[0, \infty)$ with $x(0) = x(t_0) = \alpha$ is the curve satisfying interval-by-interval dynamics as follows: Set $x_0 = \alpha$. Then on the interval $[t_0, t_1]$, x is the classical solution of the differential equation

$$\dot{x}(t) = f(x(t), k(x_0)), \quad x(t_0) = x_0, \quad t \in (t_0, t_1). \tag{4}$$

We then set $x_1 := x(t_1)$ and restart the system on the next interval as follows:

$$\dot{x}(t) = f(x(t), k(x_1)), \quad x(t_1) = x_1, \quad t \in (t_1, t_2). \tag{5}$$

The process is continued in this manner through each interval. Note that x is the unique solution on $[0,\infty)$ of the differential equation $\dot{x}(t) = f(x(t),u(t))$ satisfying $x(\tau) = \alpha$, with a certain piecewise constant control function u determined by the control feedback $k(x)$. The sample-and-hold solution procedure is sometimes referred to as "closed loop system sampling," and is the same as the "step-by-step" solution concept employed by Krasovskiĭ and Subbotin [23] in differential game theory. We refer the reader to the introduction in Clarke et al. [7] for more detail on the history of the problem, nonsmooth Lyapunov functions, Filippov solutions, and related topics. Other references relevant to the present work are Sontag [35], Clarke et al. [9], [11], Rifford [27], [28], [29], Hermes [19], [20], Kokotovic and Sussman [22], Bacciotti [2], Ancona and Bressan [1], Teel and Praly [37], Kellett and Teel [21], and Prieur [25], [26].

In the present article, we shall address a variant of the problem discussed above, in which a state constraint is imposed; to the best of our knowledge, this is the first such endeavor. Specifically, for a given constraint set $S \subset \mathbb{R}^n$ and target set Σ such that $S \cap \Sigma \neq \emptyset$, we introduce the following definitions.

DEFINITION 1.2. Open loop S-controllability *to* Σ *holds provided that for any initial state* $\alpha \in S$, *there exists a control function* $u(\cdot)$ *and a time* $t(\alpha) \geq 0$ *such that*

$$x(t) = x(t;0,\alpha,u(\cdot)) \in S \qquad \forall\, t \in [0,t(\alpha)] \tag{6}$$

and

$$x(t(\alpha)) \in \Sigma. \tag{7}$$

Note that the controllability property in the preceding definition is not an asymptotic one. On the other hand, if open loop *asymptotic* S-controllability holds for a given target, then obviously open loop S-controllability to the closed γ-neighborhood of Σ holds for every $\gamma > 0$.

DEFINITION 1.3. Closed loop S-controllability *to* Σ *holds provided that there exists a feedback law* $k : \mathbb{R}^n \to U$ *along with reals* $T_1 > 0$ *and* $\beta > 0$ *such that the following holds: If*

$$\operatorname{diam}(\pi) := \max\{t_{i+1} - t_i \,:\, i = 0,1,\ldots\} \leq \beta,$$

then for every $\alpha \in S$, *there exists* $t_1(\alpha) \in [0,T_1]$ *such that the* π*-trajectory associated with the ordinary differential equation*

$$\dot{x}(t) = f(x(t),k(x(t)))$$

and initial condition $x(0) = \alpha$ *satisfies*

$$x(t) \in S \qquad \forall\, t \in [0,t_1(\alpha)] \tag{8}$$

and

$$x(t_1(\alpha)) \in \Sigma. \tag{9}$$

Our first main result (Theorem 4.1) asserts that when certain geometric conditions are imposed upon S, then open loop S-controllability to Σ implies closed loop S-controllability to the closed δ-neighborhood of Σ, for any specified $\delta > 0$. No geometric assumptions are imposed upon the target set Σ, beyond nonemptiness of $S \cap \Sigma$.

DEFINITION 1.4. Closed loop S-stabilizability to Σ *holds provided that closed loop S-controllability to Σ holds, with* (9) *in Definition* 1.3 *fortified to*

$$x(t) \in S \cap \Sigma \qquad \forall\, t \geq t_1(\alpha). \tag{10}$$

In the second main result (Theorem 4.8), we impose a local continuity condition on the S-constrained minimum time function to the target Σ and prove that in the presence of that hypothesis, open loop S-controllability to Σ implies closed loop S-stabilizability to the closed δ-neighborhood of Σ, for any specified $\delta > 0$. Our feedback constructions in these results involve the quadratic infimal convolution of a control Lyapunov function with respect to a certain modification of the original dynamics. The modified dynamics in effect provide for constraint removal, while the convolution operation provides a useful semiconcavity property.

The layout of this article is as follows. In the next section, we will present preliminaries from nonsmooth analysis. Then in section 3, certain required geometric results pertaining to the constraint removal method in Clarke, Rifford, and Stern [12] are recalled. The main results are provided in section 4, while section 5 contains concluding comments, including a robustness property with respect to state measurement error of the feedback laws constructed in section 4.

2. Nonsmooth analysis background. Our general reference on nonsmooth analysis employed in this article is [11]. Other useful references are [9], Clarke [5], [6], Loewen [24], and Vinter [38].

2.1. Notation and definitions. The Euclidean norm is denoted $\|\cdot\|$, and $\langle\, ,\, \rangle$ is the usual inner product. The open unit ball in $\mathbb{R}^n$ is denoted B. For a nonempty set $Z \subset \mathbb{R}^n$, we denote by $\mathrm{co}(Z)$, $\mathrm{cl}(Z)$, $\mathrm{bdry}(Z)$, and $\mathrm{int}(Z)$ the convex hull, closure, boundary, and interior of Z, respectively. We denote the closure of the complement of Z by $\hat{Z} := \mathrm{cl}\{\mathbb{R}^n \backslash Z\}$. Given $\delta > 0$, we denote $Z^\delta := Z + \delta \overline{B}$. The distance of a point u to Z is denoted

$$d_Z(u) := \inf\{\|u - x\| : x \in Z\}.$$

For closed S, the "inf" is replaced with a "min". The set of closest points in Z to a point $x \in \mathbb{R}^n$ is given by

$$\mathrm{proj}_Z(x) := \{z \in Z : d_Z(x) = \|x - z\|\}.$$

This set is nonempty for every x when (the nonempty set) Z is closed.

Let $g : \mathbb{R}^n \to (-\infty, \infty]$ be an extended real valued function which is lower semicontinuous; that is, for each $x \in \mathbb{R}^n$, $g(x) \leq \liminf_{y \to x} g(y)$. A vector $\zeta \in \mathbb{R}^n$ is said to be a *proximal subgradient* (or P-subgradient) of g at a point x such that $g(x) < \infty$ provided that there exists $\sigma > 0$ such that

$$g(y) - g(x) + \sigma \|y - x\|^2 \geq \langle \zeta, y - x \rangle \tag{11}$$

for all y near x; this is known as the *proximal subgradient inequality*. The set of all such vectors ζ is called the P-subdifferential of g at x, denoted $\partial_P g(x)$. One can show that $\partial_P g(x) \neq \emptyset$ for a dense subset of $\mathrm{dom}(g)$, the set of points where g is finite. The limiting, or L-subdifferential, is defined via limits as

$$\partial_L g(x) := \{\lim \zeta_i : \zeta_i \in \partial_P g(x_i), x_i \to x, g(x_i) \to g(x)\},$$

and for g locally Lipschitz, the C-subdifferential is defined as $\partial_C g(x) := \mathrm{co}\{\partial_L g(x)\}$. For such g, one has the inclusions

$$\partial_P g(x) \subset \partial_L g(x) \subset \partial_C g(x) \qquad \forall\, x \in \mathbb{R}^n. \tag{12}$$

2.2. Semiconcavity.

DEFINITION 2.1. *Let $U \subset \mathbb{R}^n$ be open. Then a continuous function $\varphi : U \to \mathbb{R}$ is* semiconcave *on U provided that there exists $c \geq 0$ such that for any $x \in U$*

$$\varphi(x+h) + \varphi(x-h) - 2\varphi(x) \leq c\|h\|^2$$

whenever $\|h\|$ is sufficiently small (depending on x).

Semiconcavity is an important regularity property in the theory of nonlinear partial differential equations and, as first demonstrated by Rifford [28], [29], in Lyapunov theory as well. The following proposition summarizes some useful equivalences involving semiconcavity. It is essentially known, following as it does from facts in Bardi and Capuzzo-Dolcetta [3] and [11].

PROPOSITION 2.2. *For $U \subset \mathbb{R}^n$ open and $\varphi : U \to \mathbb{R}$ locally Lipschitz, the following three properties are equivalent:*

(i) *φ is semiconcave on U.*

(ii) *There exists $c \geq 0$ such that*

$$g(y) := \varphi(y) - c\|y\|^2$$

is locally concave on U; that is, for every $x \in U$, there exists $r_x > 0$ such that $x + r_x B \subset U$ and $g(\cdot)$ is concave on $x + r_x B$.

(iii) *There exists $c \geq 0$ such that given $x \in U$, there exists $r_x > 0$ for which*

$$-\varphi(y) + \varphi(x) + c\|y-x\|^2 \geq \langle \zeta, y - x\rangle \qquad \forall\, \zeta \in \partial_P(-\varphi)(x) \quad \forall\, y \in x + r_x B. \tag{13}$$

Furthermore, if φ is semiconcave on U, then at every $x \in U$ one has

$$\partial_P(-\varphi)(x) = \partial_L(-\varphi)(x) = \partial_C(-\varphi)(x) = -\partial_C\varphi(x). \tag{14}$$

Of particular use to us below will be the following.

COROLLARY 2.3. *Suppose that $U \subset \mathbb{R}^n$ is open and that φ is semiconcave on U. Then for any open convex subset U' of U and any $x, y \in U'$, one has*

$$-\varphi(y) + \varphi(x) + c\|y-x\|^2 \geq \langle \zeta, y - x\rangle \qquad \forall\, \zeta \in \partial_P(-\varphi)(x) \tag{15}$$

and

$$\varphi(y) - \varphi(x) \leq \langle \zeta, y - x\rangle + c\|y-x\|^2 \qquad \forall\, \zeta \in \partial_L\varphi(x), \tag{16}$$

where c is as in (13).

3. State constrained tracking and constraint removal. Our methods will utilize recent results in Clarke, Rifford, and Stern [12] (see also Clarke and Stern [13]), which dealt with the construction of feedback control laws for a general class of state constrained optimal control problems, via a constraint removal method. In that work, extra hypotheses were imposed upon S so as to have available certain geometric properties of *inner approximations* of S, given by

$$S_r := \{x \in S : d_{\hat{S}}(x) \geq r\},$$

for $r \geq 0$; note that $S_0 = S$. Inner approximations were studied in earlier work by Clarke, Ledyaev, and Stern [10], as well as in [12]. Several important properties

verified in [12] regarding inner approximations will also be required in the present article and will be summarized in this section.

The augmented geometric hypotheses on S are now posited. We denote the *lower Hamiltonian* $h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ by

$$h(x,p) := \min_{u\in U}\langle f(x,u), p\rangle. \tag{17}$$

(S1) S is a compact subset of $\mathbb{R}^n$ which is *wedged* at each $x \in S$; that is, $N_S^C(x)$, the *Clarke normal cone* to S at x, is pointed. This means that $N_S^C(x) \cap \{-N_S^C(x)\} = \{0\}$, or equivalently $\mathrm{int}[T_S^C(x)] \neq \emptyset$ for each $x \in S$, where $T_S^C(x)$ denotes the *Clarke tangent cone* to S at x. (We again refer the reader to [11] for the definitions and properties of these geometric constructs.)

(S2) The following "strict inwardness" condition holds:

$$h(x,\zeta) < 0 \quad \forall\, \zeta \in N_S^C(x)\backslash\{0\} \quad \forall\, x \in \mathrm{bdry}(S). \tag{18}$$

Hypotheses (S1)–(S2) *will be assumed to hold in all that follows.*

Remark 3.1.

(a) Any convex body (i.e., a convex set with nonempty interior) is wedged, but convexity is not required; for example, the closed complement of a convex body is necessarily wedged. Wedgedness of S at $x \in \mathrm{bdry}(S)$ is also referred to in the literature as *epi-Lipschitzness* at x, since the property amounts to S being locally linearly homeomorphic to the epigraph of a Lipschitz function; see Rockafellar [30] and Clarke [5].

(b) The set S is said to be *weakly invariant* provided that for any initial state $\alpha \in S$, there exists a control function $u(\cdot)$ such that

$$x(t) = x(t;0,\alpha,u(\cdot)) \in S \qquad \forall\, t \geq 0.$$

This is equivalent to the proximal condition

$$h(x,\zeta) \leq 0 \quad \forall\, \zeta \in N_S^P(x) \quad \forall\, x \in S; \tag{19}$$

see [11]. Hence conditions (S1)–(S2) are sufficient for weak invariance.

3.1. State constrained trajectory tracking. Required properties of inner approximations are summarized in the following lemma. Part (a) asserts that for small positive r, inner approximations S_r of S are weakly invariant, while parts (b) and (c) provide trajectory tracking properties relative to inner approximations in a uniform manner with respect to r. Among references involving state constrained tracking are the seminal results of Soner [33]; see also Forcellini and Rampazzo [17] and Frankowska and Rampazzo [18]. Part (a) of the lemma is included in Corollary 3.4 of [12], while parts (b) and (c) are provided by the proofs of Theorem 3.10 and Proposition 3.13 of [12], respectively.

LEMMA 3.2. *There exists a constant $r_0 > 0$ satisfying the following three properties:*

(a) *For every $r \in [0, r_0]$, the set S_r is nonempty and weakly invariant.*

(b) *Given $T > 0$, there exists a constant $M(T) > 0$ such that for every $r \in [0, r_0]$, the following holds: Let α_0 and α_1 be initial states in S_r, and let $u_0(\cdot)$ be a control function producing a trajectory which satisfies*

$$x(t;0,\alpha_0,u_0(\cdot)) \in S_r \qquad \forall\, t \in [0,T]. \tag{20}$$

Then there exists a control function $u_1(\cdot)$ which produces a trajectory which satisfies

$$\|x(t;0,\alpha_1,u_1(\cdot)) - x(t;0,\alpha_0,u_0(\cdot))\| \le M(T)\|\alpha_1-\alpha_0\| \qquad \forall\, t\in[0,T] \tag{21}$$

and

$$x(t;0,\alpha_1,u_1(\cdot)) \in S_r \qquad \forall\, t\in[0,T]. \tag{22}$$

(c) *Given $T>0$, there exists a constant $W(T)>0$ such that for any initial state $\alpha\in\operatorname{int}(S)$, if $r\in[0,r_0]$ is such that $\alpha\in S_r$ and $u(\cdot)$ is a control function such that*

$$x(t;0,\alpha,u(\cdot)) \in S \qquad \forall\, t\in[0,T], \tag{23}$$

then there exists a control function $\bar u(\cdot)$ such that

$$\|x(t;0,\alpha,\bar u(\cdot)) - x(t;0,\alpha,u(\cdot))\| \le rW(T) \qquad \forall\, t\in[0,T] \tag{24}$$

and

$$x(t;0,\alpha,\bar u(\cdot)) \in S_r \qquad \forall\, t\in[0,T]. \tag{25}$$

3.2. Modified dynamics and constraint removal. It will be convenient to denote

$$F(x) := f(x,U) = \{f(x,u) : u\in U\}.$$

Let us recall that in view of Filippov's lemma, an absolutely continuous arc $x(\cdot)$ is a trajectory of the differential inclusion

$$\dot x(t) \in F(x(t)) \quad \text{a.e.} \tag{26}$$

on a given time interval if and only if for some control function $u(\cdot)$, $x(\cdot)$ is a trajectory of the original control system (1).

We will require the following lemma, which summarizes certain technical facts from [12].

LEMMA 3.3. *If $r_0>0$ is sufficiently small, then for each $r\in[0,r_0]$, there exists a multifunction F_r with the following properties.*

(a) *$F_r(x)$ is a compact convex subset of $\mathbb{R}^n$ for every $x\in\mathbb{R}^n$. Also,*

$$F_r(x) = F(x) \qquad \forall\, x\in S_r \tag{27}$$

and

$$F_r(x) \subset F(x) \qquad \forall\, x\in S. \tag{28}$$

(b) *There exists $K>0$ (independent of r) such that*

$$F_r(x) \subset F_r(y) + K\|y-x\|B \qquad \forall\, x,y\in\mathbb{R}^n;$$

that is, $F_r(\cdot)$ is globally Lipschitz of rank K.

(c) *For every initial phase* $(\tau, \alpha) \in \mathbb{R} \times \mathbb{R}^n$, *there exists a trajectory* $x(\cdot)$ *satisfying the differential inclusion*

$$\dot{x}(t) \in F_r(x(t)) \quad \text{a.e.} \tag{29}$$

on $[\tau, \infty)$ *such that* $x(\tau) = \alpha$.

(d) *The set* S *is strongly invariant with respect to* F_r; *that is, for every initial state* $\alpha \in S$, *every trajectory* $x(\cdot)$ *of the differential inclusion* (29) *with* $x(0) = \alpha$ *satisfies* $x(t) \in S$ *for all* $t \geq 0$.

(e) *There exists* $\varepsilon_0 > 0$ *such that for any* $\alpha \in S + \varepsilon_0 B$, *there exists a trajectory* $x(\cdot)$ *of* (29) *with* $x(0) = \alpha$ *such that* $x(1) \in S$.

(f) *There exists* $C > 0$ *(independent of* r*) such that the following holds: For any* $\alpha \in S \backslash S_r$, *there exists a trajectory of* $x(\cdot)$ *of* (29) *such that* $x(0) = \alpha$ *and*

$$t(\alpha) := \sup\{t : d_{\hat{S}}(x(t)) \leq r\} \leq Cr. \tag{30}$$

(g) *There exists* $T_2(r) > 0$ *such that if* $\alpha \in S$ *and* $u_\alpha \in U$ *are such that* $f(\alpha, u_\alpha) \in F_r(\alpha)$, *then the (unique) solution* $x_\alpha(\cdot)$ *of the differential equation*

$$\dot{x}(t) = f(x(t), u_\alpha), \tag{31}$$

with $x(0) = \alpha$, *satisfies* $x_\alpha(t) \in S$ *for all* $t \in [0, T_2(r)]$.

According to (a) and (b), the multifunction F_r in the statement is globally Lipschitz and agrees with F on the inner approximation S_r and is contained in $F(x)$ for other points $x \in S$. Part (c) of the lemma follows from the standard existence theory for differential inclusions; observe that the usual "linear growth" condition is implied by the *global* Lipschitz nature of F_r. Note that (d) provides for what we refer to as "constraint removal," in the sense that for any initial state $x(0) \in S$, the state constraint

$$x(t) \in S \qquad \forall\, t \geq 0$$

is *implicit* for the differential inclusion (29). The strong invariance asserted in (d) is a consequence (see [11]) of the fact that the multifunction F_r satisfies

$$F_r(x) \subset \text{int} T_S^C(x) \qquad \forall\, x \in S.$$

(The latter condition is nontrivial just for $x \in \text{bdry}(S)$, since $T_S^C(x) = \mathbb{R}^n$ when x is in the interior of S.) Full details of the construction of F_r satisfying (a)–(d) and (f)–(g) are provided in [12]. We mention that part (e) follows independently from that construction and a result on set attainability in [11].

4. Main results. From this point on, we take $\Sigma \subset \mathbb{R}^n$ to be closed set such that $S \cap \Sigma \neq \emptyset$. Our first main result is the following.

THEOREM 4.1. *Let open loop* S*-controllability to* Σ *hold and let* $\delta > 0$ *be given. Then closed loop* S*-controllability to* Σ^δ *holds.*

We shall require the following lemma.

LEMMA 4.2. *Assume that open loop* S*-controllability to* Σ *holds. Then for any* $\gamma > 0$, *there exists* $T_3 = T_3(\gamma) > 0$ *such that the following hold:*

(a) *For any initial state* $\alpha \in S$, *there exists a control function* $u(\cdot)$ *such that for some* $\bar{t} = \bar{t}(\alpha, \gamma) \in [0, T_3]$ *one has*

$$x(t) = x(t; 0, \alpha, u(\cdot)) \in S \qquad \forall\, t \in [0, \bar{t}\,] \tag{32}$$

and

$$x(\bar{t}) \in \Sigma^{\gamma}. \tag{33}$$

(b) *There exists* $r(\gamma) \in (0, r_0]$ *such that if* $0 \le r \le r(\gamma)$, *then the following holds: For any initial state* $\alpha \in S$ *there exist* $\tilde{t} = \tilde{t}(\alpha, \gamma) \in [0, T_3 + Cr_0]$ *and a trajectory* $x(\cdot)$ *of the differential inclusion* (29) *satisfying* $x(0) = \alpha$ *such that*

$$x(t) \in S \qquad \forall\, t \in [0, \tilde{t}\,] \tag{34}$$

and

$$x(\tilde{t}) \in \Sigma^{2\gamma}. \tag{35}$$

Proof. In order to prove part (a), note that open loop S-controllability to Σ implies that for given $\alpha \in S$, there exists $t(\alpha) \ge 0$ such that for some trajectory $x(\cdot)$ of the control system (1) with $x(0) = \alpha$, one has $x(t) \in S$ for all $t \in [0, t(\alpha)]$ and $x(t(\alpha)) \in \Sigma$. By the S-constrained tracking property (b) of Lemma 3.2, there exists $Q(\alpha) > 0$ such that the following holds: For each

$$\alpha_1 \in N(\alpha) := \{\alpha + Q(\alpha)B\} \cap S,$$

there exists a trajectory $x_1(\cdot)$ of (1) with $x_1(0) = \alpha_1$ such that $x_1(t) \in S$ for all $t \in [0, t(\alpha)]$ and $x_1(t(\alpha)) \in \Sigma^{\gamma}$. In particular, using the notation of Lemma 3.2(b), we can take

$$Q(\alpha) = \frac{\gamma}{M(t(\alpha))}.$$

The family of sets $N(\alpha)$ forms a relatively open cover of S, and since S is compact, we have a finite subcover $\{N(\alpha_i)\}_{i=1}^{k}$. It is readily noted that

$$T_3 = \max\{t(\alpha_i) : 1 \le i \le k\}$$

has the required properties.

As for part (b) of the assertion, consider any initial state $\alpha \in S$, and note that by part (f) of Lemma 3.3, for $r \in [0, r_0]$, there exists a trajectory $x_1(\cdot)$ of differential inclusion (29) emanating from α such that $\alpha_1 := x_1(t_1) \in S_r$ for some $t_1 \in [0, Cr_0]$. Furthermore, by part (d) of that lemma (strong invariance), $x_1(t) \in S$ on the interval $[0, t_1]$. By part (a) of the present lemma, there exists a trajectory $x_2(\cdot)$ of the control system (1) such that $x_2(0) = \alpha_1$ and such that for some $t_2 \in [0, T_3]$ one has $x_2(t) \in S$ for all $t \in [0, t_2]$, and $x_2(t_2) \in \Sigma^{\gamma}$. According to tracking property (c) of Lemma 3.2, if $r \in [0, r(\gamma)]$, where

$$r(\gamma) := \min\left\{r_0, \frac{\gamma}{W(T_3)}\right\},$$

then there also exists a trajectory $x_3(\cdot)$ of the control system (1) such that $x_3(0) = \alpha_1$, $x_3(t) \in S_r$ for all $t \in [0, t_2]$, and $\|x_3(t_2) - x_2(t_2)\| \le \gamma$, implying $x_3(t_2) \in \Sigma^{2\gamma}$. Now note that in view of (27), on the interval $[0, t_2]$, the trajectory $x_3(\cdot)$ is also a trajectory of the differential inclusion (29). The required trajectory $x(\cdot)$ of the assertion is the concatenation of $x_1(\cdot)$ and $x_3(\cdot)$. ∎

Assume that the hypotheses and notations of the preceding lemma are still in effect. For a given $r \in [0, r(\gamma)]$, we introduce the following modification of the multifunction F_r:

$$F_{r,\gamma}(x) := \begin{cases} \operatorname{co}\{F_r(x) \cup \overline{B}\} & \text{if } \|x\| \in \Sigma^{2\gamma}, \\ \operatorname{co}\{F_r(x) \cup \frac{[3\gamma - d_\Sigma(x)]}{\gamma}\overline{B}\} & \text{if } x \notin \Sigma^{2\gamma}, x \in \Sigma^{3\gamma}, \\ F_r(x) & \text{if } x \notin \Sigma^{3\gamma}. \end{cases} \tag{36}$$

By part (b) of the preceding lemma and Lemma 3.3(e), any $\alpha \in S + \varepsilon_0\overline{B}$ is the startpoint of some trajectory of the differential inclusion (29) which reaches the target $\Sigma^{2\gamma}$ at a time not exceeding $T_3 + Cr_0 + 1$. Hence the same is true for the differential inclusion

$$\dot{x}(t) \in F_{r,\gamma}(x(t)) \quad \text{a.e.} \tag{37}$$

This is due to the fact that one has the obvious inclusion

$$F_r(x) \subset F_{r,\gamma}(x) \qquad \forall\, x \in \mathbb{R}^n. \tag{38}$$

Furthermore, since the values of the multifunction $F_{r,\gamma}$ are compact convex subsets of $\mathbb{R}^n$ and since this multifunction is globally Lipschitz, it follows from the standard theory of differential inclusions that the set of trajectories of (37) on any compact time interval, emanating from a given startpoint, is nonempty and sequentially compact in the uniform topology. Hence the minimum time $\tau_{r,\gamma}(\alpha)$ to the target $\Sigma^{2\gamma}$ from any startpoint $\alpha \in S + \varepsilon_0\overline{B}$ is attained; here

$$\tau_{r,\gamma}(\alpha) := \min\{\tilde{t} \geq 0 : x(\tilde{t}) \in \Sigma^{2\gamma}, \dot{x}(t) \in F_{r,\gamma}(x(t)) \text{ a.e.}, x(0) = \alpha\}. \tag{39}$$

Note that in this minimum time problem, there is no state constraint imposed.

We go on to define, for r, γ as above, an extended real valued function $V_{r,\gamma} : \mathbb{R}^n \to (-\infty, \infty]$ as

$$V_{r,\gamma}(\alpha) := \begin{cases} \tau_{r,\gamma}(\alpha) & \text{if } \alpha \in S + \varepsilon_0\overline{B}, \\ \infty & \text{if } \alpha \in \mathbb{R}^n \backslash \{S + \varepsilon_0\overline{B}\}. \end{cases} \tag{40}$$

Important properties of the function $V_{r,\gamma}$ are provided by the following lemma, which follows directly from the more general result—Theorem 3 in [29].

LEMMA 4.3. *Let $\gamma > 0$, assume that the origin is open loop S-controllable to Σ, and assume that $0 < r \leq r(\gamma)$. Then the following properties hold:*

(a) *$V_{r,\gamma}$ is Lipschitz on $S + \varepsilon_0\overline{B}$, where ε_0 is as in Lemma* 3.3(e).

(b) *One has*

$$\min_{v \in F_{r,\gamma}(x)} \langle v, \zeta \rangle \leq -1 \quad \forall\, \zeta \in \partial_P V_{r,\gamma}(x) \quad \forall\, x \in \{S + \varepsilon_0\overline{B}\} \backslash \{\Sigma^{2\gamma}\}. \tag{41}$$

Remark 4.4. Actually, (41) holds in equality form, but it is the stated inequality, which encapsulates *weak decrease* of the "control Lyapunov function" $V_{r,\gamma}$ that is of interest to us; see [11] for a discussion of this property.

For $\lambda > 0$, the *quadratic infimal convolution* of (the lower semicontinuous extended real valued function) $V_{r,\gamma}$ is the function $V^\lambda_{r,\gamma} : \mathbb{R}^n \to \mathbb{R}$ given by

$$V^\lambda_{r,\gamma}(x) := \inf_{y \in \mathbb{R}^n} \{V_{r,\gamma}(y) + \lambda \|y - x\|^2\}. \tag{42}$$

This function, which is clearly dominated by $V_{r,\gamma}$, is locally Lipschitz on $\mathbb{R}^n$. Furthermore, if $x \in \mathbb{R}^n$ is such that $\partial_P V^\lambda_{r,\gamma}(x) \neq \emptyset$, then there exists a point $\bar{y} \in \mathbb{R}^n$ such that the infimum in (42) is uniquely attained at $\bar{y}$ and

$$\partial_P V^\lambda_{r,\gamma}(x) \subset \partial_P V_{r,\gamma}(\bar{y}). \tag{43}$$

It can also be shown that in fact, $\partial_P V^\lambda_{r,\gamma}(x)$ reduces to a singleton, the Fréchet derivative of $V^\lambda_{r,\gamma}$ at x, a fact we will not require. These properties of quadratic infimal convolutions are all verified in [11]. We shall in addition require the following lemma concerning the function $V^\lambda_{r,\gamma}$; the hypotheses of Lemma 4.3 are assumed to still be in effect.

LEMMA 4.5. *There exists $\lambda(\gamma) > 0$ such that if $\lambda > \lambda(\gamma)$, then $V^\lambda_{r,\gamma}$ is semiconcave on the set $S + \frac{\varepsilon_0}{2}B$, and*

$$\min_{v \in F_{r,\gamma}(x)} \langle v, \zeta \rangle \leq -\frac{1}{2} \quad \forall\, \zeta \in \partial_P V^\lambda_{r,\gamma}(x) \quad \forall\, x \in \left\{ S + \frac{\varepsilon_0}{2}B \right\} \setminus \left\{ \Sigma^{3\gamma} \right\}. \tag{44}$$

Proof. One has

$$0 \leq V^\lambda_{r,\gamma}(x) \leq V_{r,\gamma}(x) \leq T_3 + Cr_0 + 1 \quad \forall\, x \in S + \varepsilon_0 \overline{B} \quad \forall\, \lambda > 0.$$

Hence for any $x \in S + \varepsilon_0 \overline{B}$ and any given $\lambda > 0$, there exist points $y \in \mathbb{R}^n$ such that

$$V_{r,\gamma}(y) + \lambda \|y - x\|^2 \leq T_3 + Cr_0 + 2. \tag{45}$$

Then

$$\|y - x\| \leq \sqrt{\frac{T_3 + Cr_0 + 2}{\lambda}} =: w(\gamma, \lambda), \tag{46}$$

and therefore

$$V^\lambda_{r,\gamma}(x) = \min_{y \in x + w(\gamma,\lambda)\overline{B}} \{ V_{r,\gamma}(y) + \lambda \|y - x\|^2 \}, \tag{47}$$

or equivalently

$$-V^\lambda_{r,\gamma}(x) = \max_{y \in x + w(\gamma,\lambda)\overline{B}} \{ -V_{r,\gamma}(y) - \lambda \|y - x\|^2 \}. \tag{48}$$

Now consider $x \in S + \frac{\varepsilon_0}{2}B$, and note that the condition

$$\lambda > \hat{\lambda}(\gamma) := 4\left(\frac{T_3 + Cr_0 + 2}{(\varepsilon_0)^2} \right) \tag{49}$$

implies

$$x + w(\gamma, \lambda)\overline{B} \subset S + \varepsilon_0 B. \tag{50}$$

By Lemma 4.3(a), $V_{r,\gamma}$ is Lipschitz on the ball $x + w(\gamma, \lambda)\overline{B}$, and it follows from (48) that the function $-V^\lambda_{r,\gamma}$ is λ-*lower* C^2 on that ball, in the terminology of Rockafellar, who studied this class of functions in [31]. Furthermore, as was shown in Clarke,

Stern, and Wolenski [14], this property implies that one has the uniform proximal subgradient inequality given by

$$-V^{\lambda}_{r,\gamma}(y) + V^{\lambda}_{r,\gamma}(x) + \lambda\|y-x\|^2 \geq \langle \zeta, y-s\rangle \quad \forall\, \zeta \in \partial_P(-V^{\lambda}_{r,\gamma})(x) \qquad \forall\, y \in x + w(\gamma,\lambda)B. \tag{51}$$

According to Proposition 2.2(iii) (with $c=\lambda$), since (51) holds for every $x \in S + \frac{\varepsilon_0}{2}B$, it follows that the function $V^{\lambda}_{r,\gamma}$ is semiconcave on that set, for every $\lambda > \hat{\lambda}(\gamma)$.

Now let $x \in \{S + \frac{\varepsilon_0}{2}B\}\backslash\{\Sigma^{3\gamma}\}$, and assume that

$$\lambda > \tilde{\lambda}(\gamma) := \max\left\{\hat{\lambda}(\gamma)\ ,\ \frac{T_3 + Cr_0 + 2}{\gamma^2}\right\}. \tag{52}$$

Then

$$x + w(\gamma,\lambda)\overline{B} \subset \{S + \varepsilon_0 B\}\backslash\{\Sigma^{2\gamma}\}, \tag{53}$$

and therefore by (41), for any $y \in x + w(\gamma,\lambda)\overline{B}$ one has

$$\min_{v \in F_{r,\lambda}(y)} \langle v, \zeta\rangle \leq -1 \qquad \forall\, \zeta \in \partial_P V_{r,\lambda}(y). \tag{54}$$

Suppose that $\zeta \in \partial_P V^{\lambda}_{r,\lambda}(x)$. Then by property (43) (in the present context), one has that $\zeta \in \partial_P V_{r,\lambda}(\bar{y})$ for some $\bar{y} \in x + w(\gamma,\lambda)B$. In view of (54), there exists $\bar{v} \in F_{r,\lambda}(\bar{y})$ such that $\langle \bar{v}, \zeta\rangle \leq -1$. Let us denote by $K_{r,\lambda}$ a (global) Lipschitz constant for the multifunction $F_{r,\lambda}$. Then there exists $\hat{v} \in F_{r,\lambda}(x)$ such that $\|\bar{v} - \hat{v}\| \leq K_{r,\lambda} w(\gamma,\lambda)$. Now denote by $K'_{r,\lambda}$ a Lipschitz constant for $V_{r,\lambda}$ on $S + \varepsilon_0\overline{B}$. Then $\|\zeta\| \leq K'_{r,\lambda}$, by a standard fact concerning norm bounds on proximal subgradients of Lipschitz functions. We then obtain

$$\langle \hat{v}, \zeta\rangle \leq \langle \bar{v}, \zeta\rangle + K_{r,\lambda}K'_{r,\lambda} w(\gamma,\lambda). \tag{55}$$

It follows that $\langle \hat{v}, \zeta\rangle \leq -\frac{1}{2}$ provided that

$$\lambda > \bar{\lambda}(\gamma) := \frac{T_3 + Cr_0 + 2}{4(K_{r,\lambda}K'_{r,\lambda})^2}.$$

Upon setting

$$\lambda(\gamma) := \max\{\tilde{\lambda}(\gamma)\ ,\ \bar{\lambda}(\gamma)\},$$

(44) holds and the proof is completed. □

Proof of Theorem 4.1. As in the preceding lemma, let us fix $0 < r \leq r(\gamma)$ and $\lambda > \lambda(\gamma)$, where $\gamma > 0$ has been chosen a priori so that $4\gamma < \delta$.

For ease of notation, from this point on we will denote $V^{\lambda}_{r,\gamma} = V$.

Since

$$F_{r,\gamma}(x) = F_r(x) \qquad \forall\, x \in \mathbb{R}^n\backslash\{\Sigma^{3\gamma}\}, \tag{56}$$

the inequality (44) can be written as

$$\min_{v \in F_r(x)} \langle v, \zeta\rangle \leq -\frac{1}{2} \quad \forall\, \zeta \in \partial_P V(x) \qquad \forall\, x \in \left\{S + \frac{\varepsilon_0}{2}B\right\}\backslash\{\Sigma^{3\gamma}\}. \tag{57}$$

This proximal Hamilton–Jacobi inequality is in turn readily seen to be equivalent to the limiting version

$$\min_{v\in F_r(x)} \langle v,\zeta\rangle \le -\frac{1}{2} \quad \forall\, \zeta \in \partial_L V(x) \qquad \forall\, x \in \left\{S + \frac{\varepsilon_0}{2}B\right\}\backslash\{\Sigma^{3\gamma}\}. \tag{58}$$

A feedback law $k : \mathbb{R}\times\mathbb{R}^n \to U$ is now defined as follows:

- Let $x \in \mathbb{R}^n$.
  - If $x \in S\backslash\{\Sigma^{4\gamma}\}$, arbitrarily choose $\zeta \in \partial_L V(x)$, and then set $k(x) = u \in U$ such that $f(x,u) \in F_r(x)$ and

$$\min_{v\in F_r(x)} \langle v,\zeta\rangle = \langle f(x,u),\zeta\rangle.$$

  - Otherwise take $k(x)$ to be any element of U.

Remark 4.6. It is the L-subdifferential of V that features in the definition of the feedback, and not the P-subdifferential. The advantage of this choice is that $\partial_L V(x) \neq \emptyset$ for every x, whereas the possible emptiness of $\partial_P V(x)$ would be problematic in our ensuing construction. Also observe that in the construction of a π-trajectory associated with the control feedback $k(x)$, the choice of $\zeta \in \partial_L V(x_i)$ does not need to be "remembered" at the next node x_{i+1}, so in the case of an "on-line" procedure, it suffices to calculate an *arbitrary* L-subgradient when a given state is reached.

Given an initial state

$$\alpha \in S\backslash\{\Sigma^{\delta}\} \subset S\backslash\{\Sigma^{4\gamma}\},$$

and a partition

$$\pi = \{t_0 = 0, t_1, t_2, \ldots\}$$

of $[0,\infty)$, let us consider the π-trajectory $x(\cdot)$ associated with the ordinary differential equation $\dot{x}(t) = f(x(t), k(x(t)))$, where the initial node is $x(t_0) = x(0) = x_0 = \alpha$. Our goal is to produce positive numbers β and T_1 such that if $\operatorname{diam}(\pi) \le \beta$, then for every such α, there exists $t_1(\alpha) \in [0, T_1]$ such that

$$x(t) \in S \qquad \forall\, t \in [0, t_1(\alpha)] \tag{59}$$

and

$$x(t_1(\alpha)) \in \Sigma^{4\gamma} \subset \Sigma^{\delta}. \tag{60}$$

Of course, if $\alpha \in S \cap \{\Sigma^{4\gamma}\}$, then we can take $t_1(\alpha) = 0$.

We shall assume that

$$\operatorname{diam}(\pi) \le T_2(r), \tag{61}$$

where $T_2(r)$ is as in Lemma 3.3(g). Then the π-trajectory satisfies

$$x_i \in S\backslash\{\Sigma^{4\gamma}\} \Longrightarrow x(t) \in S \qquad \forall\, t \in [t_i, t_{i+1}], \tag{62}$$

since $F_{r,\gamma}(x_i) = F_r(x_i)$.

By Lemma 4.5, V is semiconcave on $S + \frac{\varepsilon_0}{2}B$, and therefore Corollary 2.3 yields

$$V(y) - V(x) \le \langle \zeta, y - x\rangle + \lambda\|y - x\|^2 \quad \forall\, \zeta \in \partial_L V(x) \quad \forall\, x, y \in U', \tag{63}$$

for every open convex set $U' \subset S + \frac{\varepsilon_0}{2}B$. Let

$$M := \max\{\|f(x,u)\| : x \in S + \varepsilon_0 \overline{B}\ ,\ u \in U\}. \tag{64}$$

In order to be able to apply (63) to the evolving π-trajectory, we will assume (further to (61)) that

$$\operatorname{diam}(\pi) \leq \frac{\min\{\gamma, \frac{\varepsilon_0}{2}\}}{M} =: \rho. \tag{65}$$

This together with (62) yields the implication

$$x_i \in S\backslash\{\Sigma^{4\gamma}\} \Longrightarrow x(t) \in x_i + \rho M B \subset S\backslash\{\Sigma^{3\gamma}\} \qquad \forall\, t \in [t_i, t_{i+1}]. \tag{66}$$

Let us further consider the evolution of the π-trajectory. Pick $\zeta_0 \in \partial_L V(x_0)$. Then by (63) and (66), for $t \in [t_0, t_1]$ one has

$$\begin{aligned} V(x(t)) - V(x_0) &\leq \langle \zeta_0, x(t) - x_0\rangle + \lambda\|x(t) - x_0\|^2 \\ &= \left\langle \zeta_0, \int_{t_0}^{t_1} f(x(s), k(x_0)ds \right\rangle + \lambda\|x(t) - x_0\|^2. \end{aligned}$$

For ease of notation, let us abbreviate $K = K_{r,\lambda}$, which we recall denotes a Lipschitz constant for $V = V_{r,\gamma}$ on $S + \varepsilon_0\overline{B}$. We also denote by $\hat{K} = K_\Gamma$ a Lipschitz constant for f as in (F1), with $\Gamma = S + \varepsilon_0\overline{B}$. Then, bearing (58), (66) in mind, one has

$$V(x_1) - V(x_0) \leq -\frac{1}{2}(t_1 - t_0) + (KM\hat{K} + \lambda M^2)(t_1 - t_0)^2. \tag{67}$$

Now pick $\zeta_1 \in \partial_L V(x_1)$. Upon repeating the above steps on the interval $[t_1, t_2]$ and combining this with (67), we obtain

$$V(x_2) - V(x_0) \leq -\frac{1}{2}(t_2 - t_0) + (KM\hat{K} + \lambda M^2)[(t_1 - t_0)^2 + (t_2 - t_1)^2], \tag{68}$$

which readily yields

$$V(x_2) - V(x_0) \leq \left[-\frac{1}{2} + (KM\hat{K} + \lambda M^2)\operatorname{diam}(\pi)\right](t_2 - t_0). \tag{69}$$

Continuing this process, we arrive at

$$V(x_i) - V(x_0) \leq \left[-\frac{1}{2} + (KM\hat{K} + \lambda M^2)\operatorname{diam}(\pi)\right](t_i - t_0). \tag{70}$$

Let us assume that

$$\operatorname{diam}(\pi) \leq \frac{1}{4(KM\hat{K} + \lambda M^2)} =: \beta_1. \tag{71}$$

Then due to (70),

$$V(x_i) - V(x_0) \leq -\frac{1}{4}(t_i - t_0). \tag{72}$$

It is readily noted that $V = V^{\lambda}_{r,\gamma} \equiv 0$ on $\Sigma^{2\gamma}$. Then since V is continuous on $S + \varepsilon_0 \overline{B}$ and $V \equiv 0$ on S, it follows from (72) that $x(\cdot)$ must enter $\Sigma^{4\gamma}$ at a time not exceeding

$$T_1 := 4 \max\{V(x) : x \in S + \varepsilon_0 \overline{B}\} \leq 4(T_3 + Cr_0 + 1). \tag{73}$$

Upon taking

$$\beta := \min\{T_2(r), \rho, \beta_1\}, \tag{74}$$

the proof of the theorem is completed. □

Remark 4.7. Suppose that in Theorem 4.1, the hypothesis of open loop S-controllability to Σ is strengthened to open loop S-controllability to Σ prior to time T, for some $T > 0$, where this means that in Definition 1.2, one has $0 \leq t(\alpha) \leq T$ for every $\alpha \in S$. Then in the proof of Lemma 4.2, one can take $T_3 = T_3(\gamma) = T$ for every $\gamma > 0$. It follows that the conclusion of Theorem 4.1 can be correspondingly strengthened, to "closed loop S-controllability to Σ^{δ} prior to time $\hat{T} := 4(T + Cr_0 + 1)$, for every $\delta > 0$," meaning that for every $\delta > 0$, there exists a feedback which, in the sample-and-hold sense, steers any $\alpha \in S$ to Σ^{δ} along an S-constrained trajectory, prior to time $\hat{T}$. (An analogous remark could be made regarding the next main result, Theorem 4.8.)

Theorem 4.8, below, provides a strengthening of the conclusion of Theorem 4.1 from S-constrained controllability to S-constrained stabilizability, when an extra assumption is posited. The need for strengthened hypotheses is illustrated by the following example. Consider

$$S = \{x \in \mathbb{R}^2 : 1 \leq \|x\| \leq 2\}, \qquad \Sigma = \{(2, 0)\},$$

where the dynamics are given by the bilinear system (a perturbed harmonic oscillator)

$$\dot{x}(t) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} x(t) + u(t)x(t), \qquad U = [-1, 1].$$

It is easy to check that all the hypotheses of Theorem 4.1 hold, including open loop S-controllability, but closed loop S-stabilization does not hold.

THEOREM 4.8. *Let open loop S-controllability to Σ hold, and let $\delta > 0$ be given. Further assume that there exists $\varepsilon > 0$ such that the S-constrained minimum time function $\tau : S \to \mathbb{R}$ to the target Σ,*

$$\tau(\alpha) := \min\{\tilde{t} \geq 0 : x(\tilde{t}) \in \Sigma,\ \dot{x}(t) \in F(x(t)) \text{ a.e., } x(t) \in S\ \forall\, t \in [0, \tilde{t}],\ x(0) = \alpha\}, \tag{75}$$

is continuous on $S \cap \{\Sigma^{2\varepsilon}\}$. Then closed loop S-stabilizability to Σ^{δ} holds.

In the absence of state constraints, in the case of a point target, the above continuity condition is equivalent to "small time controllability"; see Bardi and Capuzzo-Dolcetta [3]. We also remark that in view of sequential compactness of trajectories of the differential inclusion (26) on any compact time interval, the minimum in (75) is indeed attained, for any $\alpha \in S$.

The proof of Theorem 4.8 is actually a continuation of the proof of Theorem 4.1. There we showed that if $\operatorname{diam}(\pi) \leq \beta$, then for any startpoint $\alpha \in S$, the π-trajectory associated with our feedback $k(\cdot)$ enters $\Sigma^{4\gamma}$ not later than time T_1 and is contained in S until its entry into $\Sigma^{4\gamma}$.

In the present proof, we will specialize (but not violate the definition of) the feedback $k(\cdot)$ from the proof of Theorem 4.1, as follows:

- Let $x \in \mathbb{R}^n$.
  - If $x \in S \backslash \{\Sigma^{4\gamma}\}$, arbitrarily choose $\zeta \in \partial_L V(x)$, and then set $k(x) = u \in U$ such that $f(x,u) \in F_r(x)$ and

$$\min_{v \in F_r(x)} \langle v, \zeta \rangle = \langle f(x,u), \zeta \rangle.$$

  - If $x \in S \cap \{\Sigma^{4\gamma}\}$, choose $k(x) = u \in U$ such that $f(x,u) \in F_r(x)$.
  - Otherwise (i.e., when $x \notin S$), take $k(x)$ to be any element of U.

We shall require the following consequence of S-constrained trajectory tracking.

LEMMA 4.9. *There exists $\hat{r}(\gamma) > 0$ such that if $0 < r < \hat{r}(\gamma)$, the following holds: For any $\alpha \in S \cap \{\Sigma^{\varepsilon}\}$, there exists a trajectory of the differential inclusion* (29) *($\dot{x}(t) \in F_r(x(t))$) such that $x(0) = \alpha$ and $x(\tau(\alpha)) \in \Sigma^{2\gamma}$.*

Proof. Let $\alpha \in S \cap \{\Sigma^{\varepsilon}\}$. In view of Lemma 3.3(f), there exists a trajectory of (29) (necessarily S-constrained by part (d) of that lemma) such that $x(0) = \alpha$ and $x(t') \in S_r$ for some time $t' = t'(\alpha) \in [0, Cr]$. We now denote $\alpha' = x(t')$. If r is small enough, then $\alpha' \in S \cap \{\Sigma^{2\varepsilon}\}$, for any α as above.

Let $u(\cdot)$ be a control function such that

$$x(t; 0, \alpha', u(\cdot)) \in S \qquad \forall\, t \in [0, \tau(\alpha')]$$

and

$$x(\tau(\alpha'); 0, \alpha', u(\cdot)) \in \Sigma.$$

That is, $u(\cdot)$ is optimal in the S-constrained minimum time problem with target Σ and startpoint α'; recall that we are presently assuming open loop S-controllability to Σ. In view of the tracking result given by Lemma 3.2(c), there exists a trajectory $\bar{x}(\cdot)$ of (29) such that $\bar{x}(0) = \alpha'$, $\bar{x}(t) \in S_r$ for all $t \in [0, \tau(\alpha')]$, and

$$\bar{x}(\tau(\alpha')) \in \Sigma + rW(\tilde{T})\overline{B}, \tag{76}$$

where $\tilde{T}$ is an upper bound on (the continuous function) $\tau(\cdot)$ on $S \cap \{\Sigma^{2\varepsilon}\}$. It follows that

$$\tilde{x}(t' + \tau(\alpha')) \in \Sigma + rW(\tilde{T})\overline{B}, \tag{77}$$

where $\tilde{x}(\cdot)$ denotes the concatenation of the trajectories $x(\cdot)$ and $\bar{x}(\cdot)$ of differential inclusion (29); it is therefore a trajectory of (29), and as such remains in S. One has

$$\|\tilde{x}(t' + \tau(\alpha')) - \tilde{x}(\tau(\alpha))\| \leq Mt' + M\|\tau(\alpha') - \tau(\alpha)\|, \tag{78}$$

where M is as in (64), although any norm bound on $F(x) = f(x,U)$ over S will do. Now, $0 < t' < Cr$, $\tau(\cdot)$ is continuous on $S \cap \{\Sigma^{2\varepsilon}\}$ (where both α and α' lie when r is sufficiently small), and $\|\alpha - \alpha'\| \leq MCr$. It then follows from (77) and (78) that $\tilde{x}(\tau(\alpha)) \in \Sigma^{2\gamma}$ if r is sufficiently small, independently of $\alpha \in S$. □

Completing the proof of Theorem 4.8. In view of the definitions of the functions $V_{r,\gamma}$, $V = V^{\lambda}_{r,\gamma}$, and the preceding lemma, one has

$$V(\alpha) \leq V_{r,\gamma}(\alpha) \leq \tau(\alpha) \qquad \forall\, \alpha \in S \cap \{\Sigma^{\varepsilon}\}. \tag{79}$$

Let $\omega : [0, \infty) \to [0, \infty]$ be a modulus of continuity for $\tau(\cdot)$ on $S \cap \{\Sigma^{2\varepsilon}\}$. That is, $\omega(\cdot)$ is continuous, strictly increasing, $\omega(0) = 0$, and

$$|\tau(x) - \tau(y)\| \leq \omega(\|x - y\|) \qquad \forall\, x, y \in S \cap \{\Sigma^{2\varepsilon}\}. \tag{80}$$

It is not hard to show that

$$\tau(x) \le \omega(d_\Sigma(x)) \qquad \forall\, x \in S \cap \{\Sigma^{2\varepsilon}\}. \tag{81}$$

We shall choose

$$0 < \gamma < \min\left\{\frac{2\varepsilon}{5}, \frac{\delta}{5 + 4M\omega(5\gamma)}\right\},$$

$$0 < r < \min\{r(\gamma), \hat{r}(\gamma)\},$$

and we take $\lambda > \lambda(\gamma)$ as well as $\operatorname{diam}(\pi) \le \beta$, where β was defined in (74).

Let us now reconsider the π-trajectory $x(\cdot)$ generated by the feedback $k(\cdot)$, emanating from an arbitrary $\alpha \in S$. We know that $x(\cdot)$ enters $\Sigma^{4\gamma}$ not later than time T_1. Denote by t_{i^*} the first node after the π-trajectory enters $\Sigma^{4\gamma}$. (If $\alpha \in \Sigma^{4\gamma}$, then $i^* = 1$.) Since $M\operatorname{diam}(\pi) \le \gamma$,

$$x(t) \in \Sigma^{5\gamma} \subset S \cap \{\Sigma^{2\varepsilon}\} \qquad \forall\, t \in [t_{i^*}, t_{i^*+1}]. \tag{82}$$

Then (79) and (81) imply

$$V(x(t)) \le \omega(5\gamma) \qquad \forall\, t \in [t_{i^*}, t_{i^*+1}], \tag{83}$$

and in particular one has

$$V(x(t_{i^*+1})) \le \omega(5\gamma). \tag{84}$$

Similarly to the proof of Theorem 4.1, $x(\cdot)$ re-enters $\Sigma^{4\gamma}$ not later than time $t_{i^*+1} + 4\omega(5\gamma)$, and consequently, from time t_{i^*+1} until this re-entry, one has

$$\|x(t) - x(t_{i^*+1}\| \le 4M\omega(5\gamma).$$

Hence, in view of our choice of γ, from time t_{i^*} until re-entry to $\Sigma^{4\gamma}$, the π-trajectory remains in Σ^{δ}. The above arguments show that for any $\alpha \in S$, after the π-trajectory enters $\Sigma^{4\gamma} \subset \Sigma^{\delta}$, it thereafter remains in Σ^{δ}. During its evolution, the π-trajectory never leaves S, since it is a trajectory of (29); recall Lemma 3.3(d). □

5. Concluding remarks.

5.1. Robustness. It transpires that the feedback law in Theorem 4.8 possesses a robustness property with respect to state measurement errors which are small in an appropriate sense and when the partition in the discretization scheme has the additional requirement of being "reasonably uniform," an insight first brought to light in [7].

The perturbed system under study is modeled by

$$\dot{x}(t) = f(x(t), \tilde{k}(x(t) + p(t))), \tag{85}$$

where the function $p(\cdot)$ represents the observational error present in applying the feedback law.

Given a partition π of $[0, \infty]$, the π-trajectory x_π obtained in the model (85) is the curve satisfying the following interval-by-interval dynamics: Upon setting $x_0 = \alpha$, on the interval $[t_0, t_1]$, x_π is the classical solution of

$$\dot{x}_\pi(t) = f(x_\pi(t), \tilde{k}(x_0 + p_0)), \quad x_\pi(t_0) = x_0, \quad t \in (t_0, t_1). \tag{86}$$

We then set $x_1 := x_\pi(t_1)$ and restart the process on the next interval:

$$\dot{x}_\pi(t) = f(x_\pi(t), \tilde{k}(x_1 + p_1)), \quad x_\pi(t_1) = x_1, \quad t \in (t_1, t_2). \tag{87}$$

Here the continuous function $x_\pi(t)$ is the *actual* state of the system at time t, and the values $x_i + p_i$ correspond to the inexact measurements used to generate the piecewise constant control function in the scheme.

We have the following robust version of Theorems 4.1 and 4.8. The result allows for erroneous measurements of the state giving values exterior to S, while the π-trajectory that is generated remains in S.

THEOREM 5.1. *Let $\delta > 0$ be given, and assume that open loop S-controllability to Σ holds. Then there exists a feedback law $\tilde{k} : \mathbb{R}^n \to U$ and positive reals T_1 and β such that the following hold:*

(a) *For every $b \in (0, \beta)$, there exists $E(b) > 0$ with the property that for any partition π of $[0, T]$ having*

$$\frac{b}{2} \leq t_{i+1} - t_i \leq b \qquad \forall\, i = 0, 1, \ldots, \tag{88}$$

the error bounds

$$\|p_i\| < E(\delta) \qquad \forall\, i = 0, 1, \ldots \tag{89}$$

imply that for any initial state $\alpha \in S$, there exists $t_1(\alpha) \in [0, T_1]$ such that the π-trajectory x_π in the model above, with $x_\pi(0) = \alpha$, satisfies

$$x_\pi(t) \in S \qquad \forall\, t \in [0, t_1(\alpha)] \tag{90}$$

and

$$x(t_1(\alpha)) \in \Sigma^\delta. \tag{91}$$

(b) *If the S-constrained small time controllability hypothesis of Theorem* 4.8 *is posited, then the conclusions of part* (a) *can be strengthened by replacing* (91) *with*

$$x_\pi(t) \in S \cap \{\Sigma^\delta\} \qquad \forall\, t \geq t_1(\alpha). \tag{92}$$

For each $x \in \mathbb{R}^n$, choose $s(x) \in \mathrm{proj}_S(x)$. Then the feedback law featuring in the theorem is simply given by

$$\tilde{k}(x) := k(s(x)) \qquad \forall\, x \in \mathbb{R}^n, \tag{93}$$

where $k(\cdot)$ is the feedback law from Theorem 4.8.

The proof of Theorem 5.1 follows from arguments similar to those employed in section 4.2 of [12]. There a finite time optimal control problem was studied, but the technique needed to extend the proofs of Theorems 4.1 and 4.8 to the above robust versions is provided there. The fact that partitions with sufficiently small diameter are required in Theorems 4.1, 4.8, and 5.1 is to be expected, since this is what is needed in order for the decrease property (as manifested by proximal Hamilton–Jacobi inequalities) to come to bear in a sample-and-hold scheme such as ours. On the other hand, as was pointed out in [12], [7], and Sontag [35] (with the latter two references dealing with robust feedback stabilization via a shell-based approach), the near-uniformity of partitions posited in condition (88) precludes a possible "chattering phenomenon" which could otherwise occur in the presence of state measurement errors.

5.2. *S*-restricted dynamics. Suppose that the function f in the dynamics (1) is defined only for state values $x \in S$, where S is the state constraint set in the problem we have studied. In many problems arising in economics and engineering, for example, such a restricted definition is quite reasonable, since the dynamics might not make sense or break down when $x \notin S$. So suppose that $f(x,u)$ is only defined on $S \times U$, while corresponding versions of (F1)–(F3) hold. In this situation, it is possible to now extend f from $S \times U$ to $\mathbb{R}^n \times U$ in a suitable way.

Let f_i denote the ith component function of f, $i = 1, 2, \ldots, n$. For each fixed $u \in U$, define a function $x \to \hat{f}_i(x,u)$ on $\mathbb{R}^n$ as follows:

$$\hat{f}_i(x,u) = \min_{y \in S} \{ f_i(y,u) + K\|y - x\| \}.$$

It is not difficult to show that $x \to \hat{f}_i(x,u)$ agrees with $f_i(x,u)$ on S and is globally Lipschitz of rank K. We extend f componentwise by setting $f_i(x,u) = \hat{f}_i(x,u)$ for every $(x,u) \in \mathbb{R}^n \times U$. The resulting function $f : \mathbb{R}^n \times U \to \mathbb{R}^n$ satisfies (F1)–(F2), but may not satisfy (F3) since the velocity sets $f(x,U)$ need not be convex for $x \notin S$. This poses no difficulty, however, as the tracking results in Lemma 3.2 still hold, as was pointed out in [12].

5.3. The case of unbounded *S*. The main results in this article (as well as [12]) have been stated for the case of compact S. It is worth noting that if compactness of S is relaxed to mere closedness, corresponding versions can be framed. In particular, in the corresponding versions of Definitions 1.2 and 1.3, the open and closed loop controllability properties to target Σ are provided not for any $\alpha \in S$, but for any α in a specified bounded subset of S. In order to show that this is valid, the essential task (and a somewhat routine one) is to obtain appropriately localized versions of the tracking properties in Lemma 3.2 as well as Lemma 3.3 on modified dynamics; we omit these details. Note, however, that in carrying this out, (F2) is required, unlike the weakened version of this condition mentioned in Remark 1.1.

Acknowledgment. The authors are grateful to the referees for their constructive comments.

REFERENCES

[1] F. Ancona and A. Bressan, *Patchy vector fields and asymptotic stabilization*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 445–471.

[2] A. Bacciotti, *Local Stabilizability of Nonlinear Systems*, World Scientific, River Edge, NJ, 1992.

[3] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Boston, 1997.

[4] R. W. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Birkhäuser Boston, Boston, 1983, pp. 181–191.

[5] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5, SIAM, Philadelphia, 1990.

[6] F. H. Clarke, *Methods of Dynamic and Nonsmooth Optimization*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 57, SIAM, Philadelphia, 1989.

[7] F. H. Clarke, Yu. S. Ledyaev, L. Rifford, and R. J. Stern, *Feedback stabilization and Lyapunov functions*, SIAM J. Control Optim., 39 (2000), pp. 25–48.

[8] F. H. Clarke, Yu. S. Ledyaev, E. D. Sontag, and A. I. Subbotin, *Asymptotic controllability implies control feedback stabilization*, IEEE Trans. Automat. Control, 42 (1997), p. 1394.

[9] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Qualitative properties of trajectories of control systems: A survey*, J. Dynam. Control Systems, 1 (1995), pp. 1–48.

[10] F. H. Clarke, Yu. S. Ledyaev, and R. J. Stern, *Complements, approximations, smoothings and invariance properties*, J. Convex Anal., 4 (1997), pp. 189–219.

[11] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Grad. Texts in Math. 178, Springer–Verlag, New York, 1998.

[12] F. H. Clarke, L. Rifford, and R. J. Stern, *Feedback in state constrained optimal control*, ESAIM Control Optim. Calc. Var., 7 (2002), pp. 97–134.

[13] F. H. Clarke and R. J. Stern, *Inner approximation of state constrained optimal control problems*, in Advances in Convex Analysis and Global Optimization, N. Hadjisavvas and P. M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 1–11.

[14] F. H. Clarke, R. J. Stern, and P. R. Wolenski, *Proximal smoothness and the lower-C^2 property*, J. Convex Anal., 2 (1995), pp. 117–145.

[15] J.-M. Coron, *Global asymptotic stabilization for controllable systems without drift*, Math. Control Signals Systems, 5 (1992), pp. 295–312.

[16] J.-M. Coron and L. Rosier, *A relation between continuous time-varying and discontinuous feedback stabilization*, J. Math. Systems Estim. Control, 4 (1994), pp. 67–84.

[17] F. Forcellini and F. Rampazzo, *On nonconvex differential inclusions whose state is constrained in the closure of an open set. Applications to dynamic programming*, Differential Integral Equations, 12 (1999), pp. 471–497.

[18] H. Frankowska and F. Rampazzo, *Filippov's and Filippov-Wazewski's theorems on closed domains*, J. Differential Equations, 161 (2000), pp. 449–478.

[19] H. Hermes, *Discontinuous vector fields and feedback control*, in Differential Equations and Dynamic Systems, J. K. Hale and J. P. LaSalle, eds., Academic Press, New York, 1967.

[20] H. Hermes, *Resonance, stabilizing feedback controls, and regularity of Hamilton-Jacobi-Bellman equations*, Math. Control Signals Systems, 9 (1996), pp. 59–72.

[21] C. M. Kellett and A. R. Teel, *Uniform asymptotic controllability to a set implies locally Lipschitz control-Lyapunov function*, in Proceedings of the 39th IEEE Conference on Decision and Control, Vol. 4, IEEE Press, Piscataway, NJ, 2000, pp. 3994–3999.

[22] P. V. Kokotovic and H. J. Sussmann, *A positive real condition for stabilization of nonlinear systems*, Systems Control Lett., 13 (1989), pp. 125–134.

[23] N. N. Krasovskiĭ and A. I. Subbotin, *Game-Theoretical Control Problems*, Springer–Verlag, New York, 1988.

[24] P. Loewen, *Optimal Control via Nonsmooth Analysis*, *CRM Proc. Lecture Notes* 2, AMS, Providence, RI, 1993.

[25] C. Prieur, *A robust globally asymptotically stabilizing feedback: The example of the Artstein's circles*, in Nonlinear Control in the Year 2000, A. Isidori, F. Lamnabhi-Lagarrigue, and W. Respondek, eds., Lectures Notes in Control and Inform. Sci. 258, Springer–Verlag, London, 2000, pp. 279–300.

[26] C. Prieur, *Uniting local and global controllers with robustness to vanishing noise*, Math. Control Signals Systems, 14 (2000), pp. 143-172.

[27] L. Rifford, *Problèmes de Stabilisation en Théorie du Contrôle*, Doctoral thesis, Univ. Claude Bernard-Lyon 1, France, 2000.

[28] L. Rifford, *Stabilisation des systèmes globalement asymptotiquement commandables*, C. R. Acad. Sci. Paris Sér. I. Math., 330 (2000), pp. 211–216.

[29] L. Rifford, *Existence of Lipschitz and semiconcave control-Lyapunov functions*, SIAM J. Control Optim., 39 (2000), pp. 1043–1064.

[30] R. T. Rockafellar, *Clarke's tangent cones and boundaries of closed sets in $\mathbb{R}^n$*, Nonlinear Anal., 3 (1979), pp. 145–154.

[31] R. T. Rockafellar, *Favorable classes of Lipschitz continuous functions in subgradient optimization*, in Nondifferentiable Optimization, E. Nurminski, ed., Permagon Press, New York, 1982.

[32] E. P. Ryan, *On Brockett's condition for smooth stabilizability and its necessity in a context of nonsmooth feedback*, SIAM J. Control Optim., 32 (1994), pp. 1597–1604.

[33] H. M. Soner, *Optimal control with state-space constraint* I, SIAM J. Control Optim., 24 (1986), pp. 552–561.

[34] E. D. Sontag, *A Lyapunov-like characterization of asymptotic controllability*, SIAM J. Control Optim., 21 (1983), pp. 462–471.

[35] E. D. Sontag, *Clock and insensitivity to small measurement errors*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 537–557.

[36] E. D. Sontag and H. J. Sussman, *Remarks on continuous feedback*, in Proceedings of the 19th IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 1980, pp. 916–921.

[37] A. R. Teel and L. Praly, *A smooth Lyapunov function from a class-$\mathcal{KL}$ estimate involving two positive semidefinite functions*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 313–367.

[38] R. B. Vinter, *Optimal Control*, Birkhäuser Boston, Boston, 2000.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 442–468

CONSTRAINED AVERAGE COST MARKOV CONTROL PROCESSES IN BOREL SPACES*

ONÉSIMO HERNÁNDEZ-LERMA†, JUAN GONZÁLEZ-HERNÁNDEZ‡, AND RAQUIEL R. LÓPEZ-MARTÍNEZ§

Abstract. This paper considers constrained Markov control processes in Borel spaces, with unbounded costs. The criterion to be minimized is a long-run expected average cost, and the constraints can be imposed on similar average costs, or on average *rewards*, or *discounted* costs or rewards. We give conditions under which the constrained problem (CP) is solvable and equivalent to an equality constrained (EC) linear program. Furthermore, we show that there is no duality gap between EC and the dual program EC* and that in fact the strong duality condition holds. Finally, we introduce an explicit procedure to solve CP in some cases which is illustrated with a detailed example.

Key words. constrained Markov control processes, average cost, discounted cost

AMS subject classifications. 90C40, 93E20

PII. S0363012999361627

1. Introduction.

In this paper, we study constrained Markov control processes (MCPs) in *Borel spaces*, with *unbounded costs*. The criterion to be minimized is a long-run expected *average cost*, and the constraints are imposed on similar average cost functionals. We also consider the cases in which the latter cost functionals are substituted by average *rewards* or by *discounted* costs (or rewards), or even by both average *and* discounted costs. Our results on the *constrained problem* (CP) include the following. First, we give conditions under which CP is *solvable*. Second, by introducing suitable vector spaces of measures and functions, we formulate an "equality constrained" primal linear program EC that is *equivalent* to CP. Third, we consider the dual program EC* and show that there is *no duality gap*, which means EC and EC* both have the same value. Moreover, we show that the *strong duality* condition holds, so both programs are solvable and their optimal values coincide. Finally, we introduce a procedure that can actually solve CP under suitable conditions. This procedure is illustrated with a detailed example.

Constrained MCPs form an important class of stochastic control problems with applications in many areas; see, e.g., [8, 12, 13, 14, 24, 27, 29, 30, 32, 33, 34] as well as the books [1] and [31] and their extensive bibliographies. However, on the other hand, a look at these references shows that virtually all of the related literature is concentrated on constrained MCPs in which the state space is either *countable* or *compact*. This excludes, of course, important control processes for which the state space is as common as $\mathbb{R}^d$, as well as partially observable systems whose solution typically requires transforming the original problem into a control problem on a Borel space of probability measures (p.m.'s) [3, 5, 11, 31, 42]. In fact, as far as we can

*Received by the editors September 14, 1999; accepted for publication (in revised form) September 7, 2002; published electronically May 12, 2003. Research for this work was partially supported by the Consejo Nacional de Ciencia y Tecnología (CONACyT) grant 37355-E.

http://www.siam.org/journals/sicon/42-2/36162.html

†Departamento de Matemáticas, CINVESTAV-IPN, A. Postal 14-740, México D.F. 07000, México (ohernand@math.cinvestav.mx).

‡Departamento de Probabilidad y Estadística, IIMAS-UNAM, A. Postal 20-726, México D.F. 01000, México (juan@sigma.iimas.unam.mx).

§Facultad de Matemáticas, UV, A. Postal, 270 Xalapa, Ver. 91090, México (ralopez@uv.mx). The research of this author was also supported by a CONACyT scholarship.

tell, for *average cost* constrained MCPs, the only case dealt with in the literature of a problem with an *uncountable*, *noncompact* state space X—actually, $X = \mathbb{R}^q$—is for linear-quadratic (LQ) systems in Chapter 4 of [31]. The techniques for the latter case, however, are very specific of LQ systems, and so, in general, they cannot be extended to non-LQ problems or to general Borel spaces.

As can be seen in the above references, there are several techniques to analyze CP. The easiest and most common is the so-called *direct method* (see [1, 8, 12, 13, 31, 37], for instance, or [17, 18, 21, 22] for the unconstrained case). In this method, the idea is to use occupation measures (which in the average cost case are as in (3.8) below) to transform CP into an equivalent optimization problem, say, CP′, in a suitable space of measures, and then one uses the well-known fact that a lower semicontinuous (l.s.c.) function on a compact topological space attains its minimum value. A second approach is to use (either finite- or infinite-dimensional) *linear programming* (LP) techniques (see, e.g., [1, 27, 31, 37]). To do this , one introduces linear spaces of measures and functions on which CP′ can be expressed as a linear program. In contrast, in the *Lagrange* or *convex-analytic* approach, one rewrites CP′ as a convex program, say, CP″ (see, e.g., [1, 8, 12, 13, 29, 31, 33, 36]). The latter method is closely related to the *Pareto* or *multiobjetive-control* approach because solving CP″ using Lagrange multipliers turns out to be the same as finding a special class of Pareto policies for a certain multiobjective-control problem (see, e.g., [1, 12, 13, 31, 36, 38, 39, 40]). Our paper is concerned with the first two of these approaches, namely, the direct method and the LP formulation, which were developed in, for instance, [17, 18, 21, 22] for *unconstrained* MCPs. Thus our results and techniques can be seen as a natural extension of those in the latter references.

We begin in section 2 by introducing some basic terminology and notation and the CP we are concerned with (see (2.8)) as well as some variants of it (Remark 2.2) that can be studied with our present approach. Also, in Remark 2.2(a), we explain the difference between the "minimum pair" problem we study here and the "ergodic" problem analyzed by other authors. In section 3, we state our first main result, Theorem 3.2, which gives (reasonably mild) conditions under which CP is solvable. A key step in proving Theorem 3.2 is that the analysis of CP can be reduced to consider only *stable policies* (Definition 3.4 and Lemma 3.5). The latter fact is also crucial to introducing, in section 4, an "equality constrained" (EC) linear program (see (4.10), (4.11)), which turns out to be equivalent to CP in the sense that both problems have the same optimal value (see (2.9), (3.10), and (4.12)). Also, in section 4, we introduce the dual program EC* (in (4.14), (4.15)) of EC and show that there is no duality gap for these programs (Theorem 4.4). In addition, a Farkas-like result is presented (Theorem 4.5), which, as any other Farkas-like theorem, gives necessary and sufficient conditions for EC to be consistent. In section 5, we consider the issue of strong duality, and a characterization of optimal solutions to EC and EC* (Theorem 5.2). In particular, we show that a "constrained optimality equation" holds *almost everywhere* (see (5.6)), which of course is the "constrained version" of the well-known average cost optimality equation (ACOE). Finally, in section 6, we introduce a procedure to solve CP, using maximizing sequences for EC* and ACOEs. A detailed example illustrates this procedure. In a work in progress, we use the LP formulation of CP to obtain approximations of it by *finite* linear programs, similar to those developed in [20, 21, 22].

Finally, it is worth mentioning that our results include the so-called multichain case, in which an MCP can have several ergodic classes in addition to a possibly

nonempty set of transient states. This is due to the fact that our assumptions on CP ensure that the corresponding optimization problem can be restricted to *stable* policies—see Definition 3.4 and Lemma 3.5. A multichain system is presented in Example 6.9. To the best of our knowledge, multichain CPs have been studied only by Hordijk and Kallenberg [43] for MCPs with *finite* state and action spaces. (For *unconstrained* multichain MCPs in Borel spaces, see [17].)

2. Constrained MCPs. The material in this introductory section is quite standard—see [1, 3, 8, 10, 11, 12, 13, 15, 16, 17, 18, 22, 24, 25, 26, 27, 28, 29], for instance. For the sake of reference, let us first recall that the usual *unconstrained* (discrete-time, time-homogeneous) Markov control model is of the form

$$(\mathrm{X},\ \mathrm{A},\ \{\mathrm{A}(x)\,|\,x\in\ \mathrm{X}\}, Q, c). \tag{2.1}$$

The spaces X and A are the *state space* and the *control* (or *action*) *set*, respectively. We shall assume that X and A are Borel spaces (that is, Borel subsets of complete and separable metric spaces) endowed with the corresponding Borel σ-algebras $\mathcal{B}(\mathrm{X})$, $\mathcal{B}(A)$. For each state $x\in\mathrm{X}$, the nonempty set $A(x)\in\mathcal{B}(A)$ in (2.1) denotes the set of feasible control actions in x. We suppose that the set

$$\mathbb{K} := \{(x,a)\,|\,x\in\mathrm{X},\ a\in\mathrm{A}(x)\} \tag{2.2}$$

of feasible state-action pairs is a *closed* (hence Borel measurable) subset of $\mathrm{X}\times\mathrm{A}$ and that it contains the graph of a measurable function from X to A. (In other words, the set $\mathbb{F}$ in Definition 2.1(b) below is nonempty.) Moreover, Q (or $Q(B|x,a)$ for $B\in\mathcal{B}(\mathrm{X})$ and $(x,a)\in\mathbb{K}$) stands for the *transition law*, and, finally, $c\colon\mathbb{K}\to\mathbb{R}$ is a measurable function that denotes the *cost-per-stage*.

The *constrained* Markov control model, on the other hand, is of the form

$$(\mathrm{X},\ \mathrm{A},\ \{\mathrm{A}(x)\,|\,x\in\mathrm{X}\}, Q, c, \mathbf{d}, \mathbf{k}), \tag{2.3}$$

where the first five components are as in (2.1), and, furthermore, $\mathbf{d}=(d_1,\dots,d_q)\colon\mathbb{K}\to\mathbb{R}^q$ is a given function and $\mathbf{k}=(k_1,\dots,k_q)$ is a given vector in $\mathbb{R}^q$; these are used to define the CP in (2.7) and (2.8) below. To state the latter, we need the following definitions.

DEFINITION 2.1. (a) *A stochastic kernel φ on* A *given* X *is a function $(x,B)\to\varphi(B|x)$ on $\mathrm{X}\times\mathcal{B}(\mathrm{A})$ such that $\varphi(B|\cdot)$ is a measurable function on* X *for each fixed $B\in\mathcal{B}(\mathrm{A})$, and $\varphi(\cdot|x)$ is a probability measure (p.m.) on $\mathcal{B}(\mathrm{A})$ for each fixed $x\in\mathrm{X}$.*

(b) *Φ stands for the family of stochastic kernels φ on* A *given* X *such that $\varphi(\mathrm{A}(x)|x)=1$ for all $x\in\mathrm{X}$, and $\mathbb{F}$ denotes the set of measurable functions f from* X *to* A *such that $f(x)$ is in $\mathrm{A}(x)$ for all $x\in\mathrm{X}$. (A function f in $\mathbb{F}$ may be identified with the stochastic kernel $\varphi\in\Phi$ such that $\varphi(\cdot\,|x)=\delta_{f(x)}(\cdot)$ is the Dirac measure at $f(x)$ for all $x\in\mathrm{X}$. Thus $\mathbb{F}\subset\Phi$.)*

(c) *Let $H_0:=\mathrm{X}$ and $H_n:=\mathbb{K}^n\times\mathrm{X}$ for $n=1,2,\dots$. A control policy is a sequence $\pi=\{\pi_n\}$ of stochastic kernels π_n on* A *given H_n that satisfy the constraint*

$$\pi_n(\mathrm{A}(x_n)\,|\,h_n)=1 \tag{2.4}$$

for every "history" $h_n=(x_0,a_0,\dots,x_{n-1},a_{n-1},x_n)$ in H_n and $n=0,1,\dots$. The set of all control policies is denoted by Π.

(d) *A control policy $\pi=\{\pi_n\}$ is said to be randomized stationary if there exists a stochastic kernel $\varphi\in\Phi$ such that $\pi_n(\cdot\,|h_n)=\varphi(\cdot\,|x_n)$ for each history $h_n\in H_n$ and*

$n = 0, 1, \ldots$. *The family of randomized stationary policies will be identified with the set* Φ. *Moreover,* π *is called deterministic stationary if there exists an* $f \in \mathbb{F}$ *such that* $\pi_n(\cdot|h_n)$ *is the Dirac measure at* $f(x_n)$ *for all* $h_n \in H_n$ *and* $n = 0, 1, \ldots$. *We shall identify* $\mathbb{F}$ *with the set of deterministic stationary policies.*

Let $\mathbb{P}(\mathrm{X})$ be the class of p.m.'s on $\mathcal{B}(\mathrm{X})$. For each policy $\pi \in \Pi$ and each "initial distribution" $\nu \in \mathbb{P}(\mathrm{X})$, there exist a p.m. P_ν^π and a stochastic process $\{(x_n, a_n), n = 0, 1, \ldots\}$ defined on a canonical measurable space $(\Omega, \mathcal{F})$, where x_n and a_n represent the state and the control variables at time n $(n = 0, 1, \ldots)$. The expectation operator with respect to P_ν^π is denoted by E_ν^π. If ν is concentrated at the initial state $x_0 = x$, then we write P_ν^π and E_ν^π as P_x^π and E_x^π, respectively.

Let c and $\mathbf{d} = (d_1, \ldots, d_q)$ be as in (2.1) and (2.3). For each control policy $\pi \in \Pi$ and initial distribution $\nu \in \mathbb{P}(\mathrm{X})$, consider the long-run expected average costs

$$J_0(\pi, \nu) := \limsup_{n\to\infty} \frac{1}{n} E_\nu^\pi \left[\sum_{t=0}^{n-1} c(x_t, a_t)\right] \tag{2.5}$$

and

$$J_i(\pi, \nu) := \limsup_{n\to\infty} \frac{1}{n} E_\nu^\pi \left[\sum_{t=0}^{n-1} d_i(x_t, a_t)\right] \quad \text{for } i = 1, \ldots, q. \tag{2.6}$$

Furthermore, letting $\mathbf{k} = (k_1, \ldots, k_q)$ be the q-vector in (2.3), define

$$\Delta := \{(\pi, \nu) \in \Pi \times \mathbb{P}(\mathrm{X}) | J_0(\pi, \nu) < \infty \text{ and } J_i(\pi, \nu) \le k_i \ (i = 1, \ldots, q)\}. \tag{2.7}$$

With this notation, we may then define the CP we are concerned with as follows:

$$\begin{aligned} \mathrm{CP}: \quad & \text{Minimize } J_0(\pi, \nu) \\ & \text{subject to } (\pi, \nu) \in \Delta. \end{aligned} \tag{2.8}$$

In the following section, we give conditions under which CP is *solvable*; that is, there exists a pair (π^*, ν^*) in Δ such that

$$J_0(\pi^*, \nu^*) = \inf\{J_0(\pi, \nu) \,|\, (\pi, \nu) \in \Delta\} =: \rho^*. \tag{2.9}$$

In addition, in section 4, we rewrite CP as a linear program on suitable vector spaces, which allows us to obtain further results. First, however, we conclude this section with the following comments.

Remark 2.2. (a) As in [17, 18, 22, 28, 29], we may refer to CP as a (constrained) "minimum pair" problem for an obvious reason: the "decision variables" are the *pairs* (π, ν) in Δ. A different problem, which we might call "ergodic," and which usually requires a different approach (see [1, 29, 31, 33]), is obtained if we *fix the initial distribution* $\nu_0 \in \mathbb{P}(\mathrm{X})$. (In particular, we might take $\nu_0 = \delta_x$ for a given initial state x.) In this case, the CP would be

$$\mathrm{CP}(\nu_0): \quad \text{Minimize } J(\pi, \nu_0) \text{ subject to } \pi \in \Pi \text{ and } (\pi, \nu_0) \in \Delta.$$

In the ergodic CP, one imposes ergodicity (or recurrence) hypotheses that typically ensure a "unichain" behavior, and so, at the outset, we expect the optimal value of CP to be a *constant* independent of the initial state, as for *unconstrained* problems

(see [3, 11, 18, 22], for instance). In this paper, however, we make no such hypotheses. On the contrary, our Assumption 3.1 is designed so that we can transform CP into a "classical optimization" problem on a suitable set of measures (see Lemma 3.5), and the constant optimal value that one would expect in the ergodic problem turns out to be precisely the constant ρ^* in (2.9). To see how one can study CP(ν_0) using our present approach, see Remarks 3.8, 4.6, and 5.3.

(b) *Constrained rewards.* Suppose that the functions d_i in (2.3) are "rewards" rather than costs, and define the long-run *expected average rewards*

$$R_i(\pi,\nu) := \liminf_{n\to\infty} \frac{1}{n} E_\nu^\pi \left[\sum_{t=0}^{n-1} d_i(x_t,a_t)\right] \quad \text{for } i = 1,\dots,q.$$

Then, instead of (2.8), we may consider the CP

$$\text{Minimize } J_0(\pi,\nu) \text{ subject to } R_i(\pi,\nu) \geq k_i, \tag{2.10}$$

and our results in the following sections are valid with the obvious changes. See Remark 3.7(a).

(c) *Constrained discounted cost.* Let $\alpha_1,\dots,\alpha_q$ be "discount factors" in (0,1), and consider the α_i-discounted costs

$$D_i(\pi,\nu) := (1-\alpha_i) E_\nu^\pi \left[\sum_{t=0}^{\infty} \alpha_i^t\, d_i(x_t,a_t)\right] \quad \text{for } i = 1,\dots,q.$$

Then our results are again valid if some or all of the constraints in (2.7) and (2.8) are replaced with

$$D_i(\pi,\nu) \leq k_i \qquad (i = 1,\dots,q), \tag{2.11}$$

and the case of discounted "rewards," as in (2.10), is similar. See Remark 3.7(b).

3. Solvability of CP. The following conditions ensure that CP is solvable (see Theorem 3.2).

Assumption 3.1.

(a) CP is *consistent*; that is, the set Δ in (2.7) is nonempty.

(b) $c(x,a)$ is nonnegative and inf-*compact*, which means that, for each $r \in \mathbb{R}$, the set $\{(x,a) \in \mathbb{K} \mid c(x,a) \leq r\}$ is compact.

(c) $d_i(x,a)$ is nonnegative and l.s.c. for $i = 1,\dots,q$.

(d) The transition law Q is *weakly continuous*; that is (denoting by $C_b(\mathrm{S})$ the space of continuous bounded functions on a topological spaces S), Q is such that $\int_{\mathrm{X}} u(y)Q(dy|\cdot\,)$ belongs to $C_b(\mathbb{K})$ for each function u in $C_b(\mathrm{X})$.

Observe that Assumption 3.1(b) yields, in particular, that c is l.s.c.

On the other hand, parts (b) and (c) can be replaced with the following condition: *all of the cost functions* c, $d_1,\dots,d_q$ *are nonnegative and l.s.c., and (at least) one of them is* inf-*compact.* Moreover, the "nonnegativity" condition on c and d_i may be replaced with "boundedness from below." For further comments on Assumption 3.1, see Remark 3.6.

THEOREM 3.2 (solvability of CP). *Under Assumption* 3.1, *CP is solvable. (See* (2.9)*.)*

Our proof of Theorem 3.2 is based on the "direct method," in which, as was already noted in the introduction, one uses the occupation measures in (3.8) to reduce

CP to an equivalent, "static" optimization problem. The key fact about this approach is that *to solve* CP *we may restrict ourselves to considering "stable" policies.* To state this fact precisely (Lemma 3.5), we shall use the following well-known result.

Remark 3.3 (see, for instance, [7], pp. 88–89 in [11], or p. 89 in [23]). If μ is a p.m. on X×A concentrated on $\mathbb{K}$ (i.e., $\mu(\mathbb{K}^c) = 0$, where $\mathbb{K}^c$ denotes the complement of $\mathbb{K}$), then there exists $\varphi \in \Phi$ such that μ can be "disintegrated" as

$$\mu(\mathrm{B} \times \mathrm{C}) = \int_B \varphi(\mathrm{C}|x)\widehat{\mu}(dx) \quad \forall\ \mathrm{B} \in \mathcal{B}(\mathrm{X}), \mathrm{C} \in \mathcal{B}(\mathrm{A}), \tag{3.1}$$

where $\widehat{\mu}(\mathrm{B}) = \mu(\mathrm{B} \times \mathrm{A})$ for all B in $\mathcal{B}(\mathrm{X})$ is the *marginal* (or projection) of μ on X. Conversely, for each $\varphi \in \Phi$ and $\nu \in \mathbb{P}(\mathrm{X})$, the p.m. μ on X × A defined by

$$\mu(\mathrm{B} \times \mathrm{C}) := \int_B \varphi(\mathrm{C}|x)\nu(dx) \qquad \forall\ \mathrm{B} \in \mathcal{B}(\mathrm{X}), \mathrm{C} \in \mathcal{B}(\mathrm{A}) \tag{3.2}$$

is concentrated on $\mathbb{K}$ (by Definition 2.1(b)), and its marginal on X is $\widehat{\mu} = \nu$. The p.m. μ in (3.1) and (3.2) will be written as $\mu = \widehat{\mu} \cdot \varphi$ and $\mu = \nu \cdot \varphi$, respectively.

For each $\varphi \in \Phi$ and $x \in \mathrm{X}$, we write

$$c(x, \varphi) := \int_A c(x, a)\varphi(da|x) \quad \text{and} \quad Q(\cdot|x, \varphi) := \int_A Q(\cdot|x, a)\varphi(da|x), \tag{3.3}$$

and similarly for $d_i(x, \varphi)$. In particular, for $f \in \mathbb{F}$, (3.3) reduces to

$$c(x, f) := c(x, f(x)) \ \text{ and } \ Q(\cdot\,|x, f) := Q(\cdot\,|x, f(x)). \tag{3.4}$$

(Recall the identification $f(x) \mapsto \varphi(\cdot\,|x) := \delta_{f(x)}(\cdot\,)$ in Definition 2.1(b).)

Definition 3.4 (stable policies). *Let $\mu = \widehat{\mu} \cdot \varphi$ be as in* (3.1). *Then the p.m. μ (or the randomized stationary policy $\varphi \in \Phi$) is said to be* stable *if*

(a) $\langle \mu, c \rangle := \int c(x, a)\mu(d(x, a)) = \int c(x, \varphi)\widehat{\mu}(dx) < \infty$, *and*

(b) *the marginal $\widehat{\mu}$ is an* invariant probability measure *(i.p.m.) for the transition kernel $Q(\cdot\,|\cdot\,, \varphi)$; that is,*

$$\widehat{\mu}(\mathrm{B}) = \int_{\mathrm{X}} Q(\mathrm{B}|x, \varphi)\widehat{\mu}(dx) \qquad \forall\ \mathrm{B} \in \mathcal{B}(\mathrm{X}).$$

We shall denote by $\mathbb{P}(\mathbb{K})$ the family of p.m.'s on X×A concentrated on $\mathbb{K}$, and by $\mathbb{P}_{\mathrm{s}}(\mathbb{K}) \subset \mathbb{P}(\mathbb{K})$ the set defined as

$$\mathbb{P}_{\mathrm{s}}(\mathbb{K}) := \{\mu \in \mathbb{P}(\mathbb{K}) |\, \mu \text{ is stable}\}.$$

By the individual ergodic theorem (see, for instance, p. 388 in [35] or Theorem E.11 in [18]), if $\mu = \widehat{\mu} \cdot \varphi$ is stable, then the long-run expected average cost $J_0(\varphi, \widehat{\mu})$ when using the policy $\varphi \in \Phi$ and the initial distribution is $\widehat{\mu}$ is given by

$$J_0(\varphi, \widehat{\mu}) = \lim_{n \to \infty} \frac{1}{n} E_{\mu}^{\varphi} \left[\sum_{t=0}^{n-1} c(x_t, a_t) \right] = \langle \mu, c \rangle.$$

Thus, for $\mu = \widehat{\mu} \cdot \varphi$ in $\mathbb{P}_{\mathrm{s}}(\mathbb{K})$, we have (using the notation (3.3))

$$J_0(\varphi, \widehat{\mu}) = \langle \mu, c \rangle = \int_{\mathrm{X}} c(x, \varphi)\widehat{\mu}(dx) \tag{3.5}$$

and, similarly,

$$J_i(\varphi,\widehat{\mu}) = \langle \mu, d_i\rangle = \int_{\mathrm{X}} d_i(x,\varphi)\widehat{\mu}(dx) \quad \text{for } i = 1,\dots,q. \tag{3.6}$$

To establish the connection between CP and stable policies, in the right-hand side of (2.5), replace $c(x,a)$ by the indicator function I_Γ of a set Γ in $\mathcal{B}(\mathrm{X}\times\mathrm{A})$. Then we can rewrite (2.5) as

$$J_0(\pi,\nu) = \limsup_{n\to\infty}\langle \mu^\pi_{\nu,n}, c\rangle, \tag{3.7}$$

where $\mu^\pi_{\nu,n}$ denotes the n-stage expected *occupation measure* associated to (π,ν), that is,

$$\mu^\pi_{\nu,n}(\Gamma) := \frac{1}{n}\sum_{t=0}^{n-1} P^\pi_\nu\,[(x_t,a_t)\in\Gamma] \quad \forall \quad \Gamma\in\mathcal{B}(\mathrm{X}\times\mathrm{A}). \tag{3.8}$$

Similarly, (2.6) becomes

$$J_i(\pi,\nu) = \limsup_{n\to\infty}\langle \mu^\pi_{\nu,n}, d_i\rangle \quad \text{for} \quad i = 1,\dots,q. \tag{3.9}$$

Having (3.7)–(3.9), the same proof of Theorem 5.7.9(a) in [18] yields the following.

LEMMA 3.5 (reduction of CP to stable policies). *Under Assumption* 3.1, *for each feasible pair* $(\pi,\nu)\in\Delta$ *for CP, there exists a stable p.m.* $\mu = \widehat{\mu}\cdot\varphi$ *such that*

(a) $(\varphi,\widehat{\mu})$ *is in* Δ, *and*

(b) $J_0(\pi,\nu)\ge J_0(\varphi,\widehat{\mu}) = \langle\mu, c\rangle$.

Hence we can write ρ^* *in* (2.9) *as*

$$\rho^* = \inf\{\langle\mu,c\rangle | \mu\in\Delta_{\mathrm{s}}\}, \tag{3.10}$$

where $\Delta_{\mathrm{s}} := \{\mu\in\mathbb{P}_{\mathrm{s}}(\mathbb{K})|$ *if* $\mu = \widehat{\mu}\cdot\varphi$, *then* $(\varphi,\widehat{\mu})\in\Delta\}$, *and, moreover, there is a p.m.* μ^* *in* Δ_{s} *such that* $\langle\mu^*,c\rangle = \rho^*$.

Lemma 3.5 *yields Theorem* 3.2. Indeed, let $\mu^*\in\Delta_{\mathrm{s}}$ be as in Lemma 3.5, and let us disintegrate it as in (3.1), that is, $\mu^* = \widehat{\mu}^*\cdot\varphi^*$. Then the pair $(\varphi^*,\widehat{\mu}^*)$ is optimal for CP because, by (3.5), (3.6), and (3.10), we have

$$J_0(\varphi^*,\widehat{\mu}^*) = \langle\mu^*,c\rangle = \int_{\mathrm{X}} c(x,\varphi^*)\widehat{\mu}^*(dx) = \rho^* \tag{3.11}$$

and

$$J_i(\varphi^*,\widehat{\mu}^*) = \langle\mu^*,d_i\rangle = \int_{\mathrm{X}} d_i(x,\varphi^*)\widehat{\mu}^*(dx) \le k_i \quad \text{for} \quad i = 1,\dots,q. \tag{3.12}$$

Remark 3.6 (examples). Examples that satisfy Assumption 3.1(b), (d) can be seen, for instance, in [1, 10, 11, 13, 15, 16, 18, 22, 24, 25, 26, 31, 33, 41]. In particular, the examples in [16], [18, Chapter 5], [22, Chapter 10], [26], [31], and [41] include queueing systems, linear systems with quadratic costs, inventories, and a cash-balance model, and they all have in common that the state process $\{x_t\}$ follows a discrete-time equation of the form

$$x_{t+1} = G(x_t,a_t,\xi_t), \quad t = 0,1,\dots,$$

where the ξ_t are independently and identically distributed (i.i.d.) disturbances, independent of the initial state x_0, and $G(x, a, s)$ is a given measurable function. Moreover, G is continuous in (x, a) for each s, and so Assumption 3.1(d) holds. On the other hand, in each case, the cost function $c(x, a)$ satisfies Assumption 3.1(b) or an analogous condition when c is a "reward," as in the cash-balance model in [26]. In the latter reference and also in [16, 22], for instance, $c(x, a)$ is required, in addition, to satisfy a "growth" condition

$$0 \leq \sup_{a \in \mathrm{A}(x)} c(x, a) \leq \mathrm{V}(x) \quad \forall\, x \in \mathrm{X}, \tag{3.13}$$

where $\mathrm{V}(\cdot\,) \geq 1$ is a given measurable function which acts as a "bounding" or "weight" function. Furthermore, conditions are given for the state process $\{x_t\}$ to be a V-*geometrically ergodic* Markov chain for each deterministic stationary policy $f \in \mathbb{F}$. This implies, in particular, that the long-run average cost, say, $J_0(f, \nu)$, is finite for any $f \in \mathbb{F}$ and $\nu \in \mathbb{P}(\mathrm{X})$, and similary for $J_i(f, \nu)$ if the costs $d_i(x, a)$ satisfy (3.13) for $i = 1, \ldots, q$. Thus, as Assumption 3.1(c) is bound to hold in most applications, for all of Assumption 3.1 to be true, we would need only to ensure the existence of $f \in \mathbb{F}$ (say) and $\nu \in \mathbb{P}(\mathrm{X})$ such that $J_i(f, \nu) \leq k_i$ for $i = 1, \ldots, q$. For similar examples (using "bounding" functions) when the state space is a *countable* set, see, for instance, [1, 10, 24].

Remark 3.7. (a) Consider the *constrained rewards* problem (2.10). Then it is easily seen that Theorem 3.2 and Lemma 3.5 are valid for (2.10) if Assumption 3.1(c) is replaced with the following: $d_i(x, a)$ *is upper semicontinuous and bounded above for* $i = 1, \ldots, q$.

(b) *A solution to* CP (see (3.10)) *also solves the problem of minimizing the average cost* $J_0(\pi, \nu)$ *subject to the constrained discounted costs* (2.11). To see this, for each $i = 1, \ldots, q$, consider the *discounted* occupation measures (cf. (3.8))

$$\gamma^{\pi}_{\nu,i}(\Gamma) := (1 - \alpha_i) \sum_{t=0}^{\infty} \alpha_i^t P^{\pi}_{\nu}[(x_t, a_t) \in \Gamma] \quad \forall\, \Gamma \in \mathcal{B}(\mathrm{X} \times \mathrm{A}) \tag{3.14}$$

so that the discounted costs D_i can be expressed as

$$D_i(\pi, \nu) = \langle \gamma^{\pi}_{\nu,i}, d_i \rangle. \tag{3.15}$$

Now let $\mu = \widehat{\mu} \cdot \varphi$ be a *stable* p.m. Then, in particular, by the invariance condition in Definition 3.4(b), we have

$$\widehat{\mu}(\cdot) = \int_{\mathrm{X}} Q^t(\cdot|x, \varphi)\widehat{\mu}(dx) \quad \forall\, t = 0, 1, \ldots. \tag{3.16}$$

Therefore, replacing the pair (π, ν) in (3.14) with $(\varphi, \widehat{\mu})$, we get

$$\gamma^{\varphi}_{\widehat{\mu},i}(\cdot) = \mu(\cdot) \quad \forall\, i = 1, \ldots, q \tag{3.17}$$

because, for any measurable rectangle $\Gamma = \mathrm{B} \times \mathrm{C}$ in $\mathcal{B}(\mathrm{X} \times \mathrm{A})$ and $t = 0, 1, \ldots$, it holds that

$$\begin{aligned} P^{\varphi}_{\mu}(x_t \in \mathrm{B}, a_t \in \mathrm{C}) &= \int_{\mathrm{X}} \int_{\mathrm{B}} \varphi(\mathrm{C}|x) Q^t(dx|y, \varphi)\widehat{\mu}(dy) \\ &= \int_{\mathrm{B}} \varphi(\mathrm{C}|x)\widehat{\mu}(dx) \quad \text{(by (3.16))} \\ &= \mu(\mathrm{B} \times \mathrm{C}) \quad \text{(by (3.1))}. \end{aligned}$$

The latter equality and (3.14) give (3.17). In turn, (3.17) yields that *for any stable p.m.* $\mu = \widehat{\mu} \cdot \varphi$ we can write (3.15) as

$$D_i(\varphi, \widehat{\mu}) = \langle \mu, d_i \rangle \quad \forall\ i = 1, \dots q,$$

and so the discounted costs in (2.11) coincide with the *average costs* in (3.6). Hence, by Lemma 3.5(a), (b), we obtain the desired conclusion.

Finally, observe that the above argument also shows that some of the constraints in CP can include both average *and* discounted costs. This can be relevant for applications in which one has long-run (average) constraints as well as short-term (discounted) constraints.

Remark 3.8. Consider $\mathrm{CP}(\nu_0)$ in Remark 2.2(a), where $\nu_0 \in \mathbb{P}(\mathrm{X})$ is a *fixed* initial distribution. Let us suppose that Assumption 3.1 holds except that part (a) is replaced with

(a′) $\mathrm{CP}(\nu_0)$ *is consistent.*

In addition, we suppose that *there exists* $\varphi \in \Phi$ *such that* $\mu := \nu_0 \cdot \varphi$ *is stable.* Then, from the proof of Lemma 3.5 and Theorem 3.2, it can be seen that these results hold when CP is replaced with $\mathrm{CP}(\nu_0)$. In particular, $\mathrm{CP}(\nu_0)$ is solvable.

4. The LP formulation. General references for this section are Chapter 3 in [2], Chapter 6 in [18], or Chapter 12 in [22]; in fact, we shall follow section 12.3 in [22] closely. Additional related material can be found in [1, 17, 21, 27, 29, 31, 36, 37], for instance. Assumption 3.1 is supposed to hold throughout the following.

As in the unconstrained case, the idea in this section is as follows. In Lemma 3.5, we have reduced CP to the minimization of the *linear map* $\mu \mapsto \langle \mu, c \rangle$ over the subset of p.m.'s μ in Δ_{s}, as in (3.10). Thus, as the objective function is already linear, to transform (3.10) into a *linear program*, the idea is simply to imbed Δ_{s} into a suitable vector space of measures. We next show how this is done.

For each (x, a) in $\mathbb{K}$, define

$$w(x,a) := 1 + c(x,a) \ \text{ and } \ \widehat{w}(x) := \inf_{a \in A(x)} w(x,a) = 1 + \widehat{c}(x), \tag{4.1}$$

where $\widehat{c}(x) := \inf_{a \in A(x)} c(x,a)$ is assumed to be measurable. (To develop the LP formulation, in (4.1), we may in fact take *any measurable function* $w(x,a) \geq c(x,a)$ as long as it is bounded away from zero so that (4.3) is well defined.) Let us now consider the dual pair $(\mathrm{M}(\mathbb{K}), \mathrm{F}(\mathbb{K}))$ of vector spaces defined as follows:

- $\mathrm{M}(\mathbb{K})$ is the vector space of finite signed measures μ on $\mathrm{X} \times \mathrm{A}$, concentrated on $\mathbb{K}$, such that

$$\|\mu\|_w := \int w d|\mu| < \infty, \tag{4.2}$$

  where $|\mu| := \mu^+ + \mu^-$ denotes the total variation of μ (see [4] or [6]).
- $\mathrm{F}(\mathbb{K})$ is the vector space of measurable functions $v : \mathbb{K} \to \mathbb{R}$ such that

$$\|v\|_w := \sup_{(x,a) \in \mathbb{K}} \frac{|v(x,a)|}{w(x,a)} < \infty. \tag{4.3}$$

- The bilinear form $\langle \cdot\,, \cdot \rangle$ on $(\mathrm{M}(\mathbb{K}), \mathrm{F}(\mathbb{K}))$ is

$$\langle \mu, v \rangle := \int_{\mathrm{X} \times \mathrm{A}} v(x,a) \mu(d(x,a)). \tag{4.4}$$

  (See Remark 4.1(b).)

In addition, we consider the dual pair $(\mathrm{M}(\mathrm{X}), \mathrm{F}(\mathrm{X}))$ defined exactly as above, but with X and $\widehat{w}$ in lieu of $\mathrm{X} \times \mathrm{A}$ and w, respectively.

Remark 4.1. (a) By (4.1) and (4.3), it is evident that $c(x,a)$ is in $\mathrm{F}(\mathbb{K})$.

(b) We assume that a function v in $\mathrm{F}(\mathbb{K})$ can be measurably extended to all of $\mathrm{X} \times \mathrm{A}$ in an arbitrary way as long as the integral in (4.4) is finite. For instance, if we define $c(x,a) := +\infty$ for (x,a) in the complement $\mathbb{K}^{\mathrm{c}}$ of $\mathbb{K}$ and make the convention that $0 \cdot (+\infty) = 0$, then (4.4) becomes

$$\langle \mu, c \rangle = \int_{\mathbb{K}} c \, d\mu,$$

which is finite because, by (4.1)–(4.3), $|\langle \mu, c \rangle| \le \int c \, d|\mu| \le \int w \, d|\mu| < \infty$. Also observe that $\langle \mu, 1 \rangle = \mu(\mathrm{X} \times \mathrm{A}) = \mu(\mathbb{K})$.

To rewrite (3.10) as a linear program, we need the following additional assumption.

Assumption 4.2.

(a) d_i is in $F(\mathbb{K})$ for $i = 1, \dots, q$.

(b) $\int_{\mathrm{X}} \widehat{w}(y) Q(dy|\cdot)$ is in $F(\mathbb{K})$, i.e., $\sup_{(x,a)} w(x,a)^{-1} \int \widehat{w}(y) Q(dy|x,a) < \infty$.

Now let $\mathrm{L}_0 : \mathrm{M}(\mathbb{K}) \to \mathrm{M}(\mathrm{X})$ be the linear map $\mu \mapsto \mathrm{L}_0\mu$ defined by

$$\mathrm{L}_0\mu(\mathrm{B}) := \widehat{\mu}(\mathrm{B}) - \int_{\mathrm{X} \times \mathrm{A}} Q(\mathrm{B}|x,a) \mu(d(x,a)) \text{ for } \mathrm{B} \in \mathcal{B}(\mathrm{X}). \tag{4.5}$$

The *adjoint* $\mathrm{L}_0^* : \mathrm{F}(\mathrm{X}) \to \mathrm{F}(\mathbb{K})$ of L_0, which is defined by the relation

$$\langle \mathrm{L}_0\mu, u \rangle = \langle \mu, \mathrm{L}_0^* u \rangle \quad \forall \ \mu \in \mathrm{M}(\mathbb{K}), u \in \mathrm{F}(\mathrm{X}),$$

is given by

$$(\mathrm{L}_0^* u)(x,a) = u(x) - \int_{\mathrm{X}} u(y) Q(dy|x,a) \ \ \forall \ (x,a) \in \mathbb{K}, u \in \mathrm{F}(\mathrm{X}). \tag{4.6}$$

Moreover, by (4.3) and Assumption 4.2(b), L_0^* indeed maps $\mathrm{F}(\mathrm{X})$ into $\mathrm{F}(\mathbb{K})$, which is equivalent to saying that

$$\text{the linear map } \mathrm{L}_0 \colon \mathrm{M}(\mathbb{K}) \to \mathrm{M}(\mathrm{X}) \text{ is weakly continuous,} \tag{4.7}$$

that is, continuous in the weak topologies $\sigma(\mathrm{M}(\mathbb{K}), \mathrm{F}(\mathbb{K})), \sigma(\mathrm{M}(\mathrm{X}), \mathrm{F}(\mathrm{X}))$. With this fact we now have all the ingredients to state (3.10) as a linear program.

First note that (by Definition 3.4, (3.5), and (3.6)), the condition "$\mu \in \Delta_{\mathrm{s}}$" in (3.10) can be written as

$$\mathrm{L}_0\mu = 0, \ \langle \mu, 1 \rangle = \mu(\mathbb{K}) = 1, \ \langle \mu, d_i \rangle \le k_i \ \ (i = 1, \dots, q), \ \mu \in \mathrm{M}_+(\mathbb{K}), \tag{4.8}$$

where $\mathrm{M}_+(\mathbb{K})$ denotes the convex cone of nonnegative measures in $\mathrm{M}(\mathbb{K})$. Moreover, the inequalities $\langle \mu, d_i \rangle \le k_i \ \ (i = 1, \dots, q)$ hold if and only if $\langle \mu, d_i \rangle + \alpha_i = k_i$ for some $\alpha_i \ge 0$. Therefore, letting

$$\mathrm{L} \colon \mathrm{M}(\mathbb{K}) \times \mathbb{R}^q \to \mathrm{M}(\mathrm{X}) \times \mathbb{R} \times \mathbb{R}^q$$

be the linear map given by

$$\mathrm{L}(\mu, \boldsymbol{\alpha}) := (\mathrm{L}_0\mu, \langle \mu, 1 \rangle, \langle \mu, d_1 \rangle + \alpha_1, \dots, \langle \mu, d_q \rangle + \alpha_q) \tag{4.9}$$

for each μ in $\mathrm{M}(\mathbb{K})$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)$ in $\mathbb{R}^q$, we can write (4.8) as

$$\mathrm{L}(\mu, \boldsymbol{\alpha}) = (0, 1, \mathbf{k}) \text{ for some } \boldsymbol{\alpha} \in \mathbb{R}^q_+.$$

Thus the nonnegative components α_i of $\boldsymbol{\alpha}$ have an obvious interpretation as "slack variables," and, on the other hand, by (3.10) we see that CP *is equivalent to the* EC *primal linear program*

$$\textbf{EC} \qquad \text{Minimize } \langle \mu, c\rangle = \langle (\mu, \boldsymbol{\alpha}), (c, \mathbf{0})\rangle \tag{4.10}$$

$$\text{subject to } \mathrm{L}(\mu, \boldsymbol{\alpha}) = (0, 1, \mathbf{k}), \quad (\mu, \boldsymbol{\alpha}) \in \mathrm{M}_+(\mathbb{K}) \times \mathbb{R}^q_+. \tag{4.11}$$

Hence, in particular, Theorem 3.2 and Lemma 3.5 yield the following.

COROLLARY 4.3 (solvability of EC). *Under Assumptions* 3.1 *and* 4.2, EC *is solvable.*

In other words, there exists $(\mu^*, \boldsymbol{\alpha}^*)$ that satisfies (4.11), and, in addition, the *value* of EC, namely,

$$\inf(\mathrm{EC}) := \inf\{\langle \mu, c\rangle |\ (4.11) \text{ holds}\},$$

can be written as a "minimum" rather than "infimum," and

$$\langle \mu^*, c\rangle = \langle (\mu^*, \boldsymbol{\alpha}^*), (c, \mathbf{0})\rangle = \min(\mathrm{EC}) = \rho^*. \tag{4.12}$$

To introduce the dual EC^* of EC, let us first note that the adjoint

$$\mathrm{L}^*\colon \mathrm{F}(\mathrm{X}) \times \mathbb{R} \times \mathbb{R}^q \to \mathrm{F}(\mathbb{K}) \times \mathbb{R}^q$$

of L is the linear map

$$\mathrm{L}^*(u, \beta_0, \boldsymbol{\beta}) = \left(\mathrm{L}_0^* u + \beta_0 + \sum_{i=1}^q \beta_i d_i, \boldsymbol{\beta} \right) \tag{4.13}$$

for all $(u, \beta_0, \boldsymbol{\beta})$ in $\mathrm{F}(\mathrm{X}) \times \mathbb{R} \times \mathbb{R}^q$. In particular, by (4.7) and Assumption 4.2(a), L is weakly continuous, and the dual of EC (see (4.10), (4.11)) is

$$\textbf{EC}^* \quad \text{Maximize } \langle (0, 1, \mathbf{k}), (u, \beta_0, \boldsymbol{\beta})\rangle = \beta_0 + \langle \mathbf{k}, \boldsymbol{\beta}\rangle \tag{4.14}$$

$$\text{subject to } \mathrm{L}^*(u, \beta_0, \boldsymbol{\beta}) \le (c, \mathbf{0}),\ (u, \beta_0, \boldsymbol{\beta}) \in \mathrm{F}(\mathrm{X}) \times \mathbb{R} \times \mathbb{R}^q. \tag{4.15}$$

In (4.14), $\langle \mathbf{k}, \boldsymbol{\beta}\rangle$ denotes the usual inner (or scalar) product of q-vectors, i.e.,

$$\langle \mathbf{k}, \boldsymbol{\beta}\rangle = \sum_{i=1}^q k_i \beta_i,$$

whereas the inequality in (4.15) in understood componentwise, i.e., (by (4.13))

$$\mathrm{L}_0^* u + \beta_0 + \sum_{i=1}^q \beta_i d_i \le c \text{ and } \boldsymbol{\beta} \le \mathbf{0}$$

or, more explicitly, by the definition (4.6) of L_0^*,

$$\beta_0 + u(x) \le c(x,a) - \sum_{i=1}^{q} \beta_i d_i(x,a) + \int_{\mathrm{X}} u(y) Q(dy|x,a) \quad \forall\ (x,a) \in \mathbb{K} \tag{4.16}$$

and

$$\beta_i \le 0 \quad \text{for} \quad i = 1, \ldots, q. \tag{4.17}$$

As the cost-per-stage $c(x,a)$ is nonnegative (Assumption 3.1(b)), it is clear that the dual EC* is *consistent.* For instance, the triplet $(u, \beta_0, \boldsymbol{\beta})$=(0,0,$\mathbf{0}$) satisfies (4.15). Therefore, the *weak duality* property yields $\sup(\mathrm{EC}^*) \le \inf(\mathrm{EC})$, where

$$\sup(\mathrm{EC}^*) := \sup\{\langle (0,1,\mathbf{k}), (u, \beta_0,\ \boldsymbol{\beta})\rangle \mid (4.15) \text{ holds}\} \tag{4.18}$$

denotes the value of EC*. Thus, by (4.12), we have

$$\sup(\mathrm{EC}^*) \le \min(\mathrm{EC}),$$

and, in fact, it turns out that the same hypotheses of Corollary 4.3 yield that there is *no duality gap*, that is,

$$\sup(\mathrm{EC}^*) = \min(\mathrm{EC}) \quad (= \rho^*). \tag{4.19}$$

Formally stated, we have the following result, which can be proved as Theorem 12.3.4 in [22] (see also [17] or [37]).

THEOREM 4.4 (absence of duality gap). *If Assumptions* 3.1 *and* 4.2 *are satisfied, then* (4.19) *holds.*

In the next section, we give conditions for the *strong duality* property to be true. This means that the dual EC* is also *solvable*, in which case the value in (4.18) is written as $\max(\mathrm{EC}^*)$, and so (4.19) becomes

$$\max(\mathrm{EC}^*) = \min(\mathrm{EC}) \quad (= \rho^*). \tag{4.20}$$

In the meantime, we conclude this section with a Farkas-like result that gives necessary and sufficient conditions for EC to be consistent—that is, for the existence of a pair $(\mu, \boldsymbol{\alpha})$ that satisfies (4.11).

THEOREM 4.5 (necessary and sufficient conditions for consistency of EC). *Suppose that Assumptions* 3.1(d) *and* 4.2(b) *are satisfied. In addition,*

(i) X *and* $\mathbb{K}$ *are locally compact separable metric spaces, and*

(ii) *for each compact* $\mathrm{K} \subset \mathrm{X}$, *the function* $Q(\mathrm{K}|\cdot\,)$ *vanishes at infinity; that is, for each* $\epsilon > 0$, *there exists a compact set* $\mathrm{K}' = \mathrm{K}'(\epsilon, \mathrm{K}) \in \mathbb{K}$ *such that*

$$Q(\mathrm{K}|x,a) \le \epsilon \quad \forall\ (x,a) \notin \mathrm{K}'.$$

Then the following statements are equivalent:

(a) *There exists a pair* $(\mu, \boldsymbol{\alpha})$ *that satisfies* (4.11).

(b) *If the triplet* $(u, \beta_0, \boldsymbol{\beta})$ *in* $\mathrm{F}(\mathrm{X}) \times \mathbb{R} \times \mathbb{R}^q$ *is such that* $\mathrm{L}^*(u, \beta_0, \boldsymbol{\beta}) \ge (0, \mathbf{0})$, *then*

$$\langle (0,1,\mathbf{k}), (u, \beta_0,\ \boldsymbol{\beta})\rangle = \beta_0 + \langle \mathbf{k}, \boldsymbol{\beta}\rangle \ge 0.$$

The proof of Theorem 4.5 is essentially the same as the proof of Theorem 4.5(a) in [17], which also appears as Theorem 12.3.7 in [22]. In fact, the proof is a straightforward application of a generalized Farkas theorem of Craven and Koliha [9]. (For related Farkas-like theorems, see, for instance, [19].)

Remark 4.6. The results in this section can be expressed for $\mathrm{CP}(\nu_0)$ in Remark 2.2(a) as follows. Let L_0 and L be as in (4.5) and (4.9), and consider the linear maps

$$\mathrm{L}_1 : \mathrm{M}(\mathbb{K}) \to \mathrm{M}(\mathrm{X}) \quad \text{and} \quad \mathrm{T} : \mathrm{M}(\mathbb{K}) \times \mathbb{R}^{\mathrm{q}} \to \mathrm{M}(\mathrm{X})^2 \times \mathbb{R} \times \mathbb{R}^{\mathrm{q}},$$

with $\mathrm{M}(\mathrm{X})^2 := \mathrm{M}(\mathrm{X}) \times \mathrm{M}(\mathrm{X})$, defined by $\mathrm{L}_1\mu := \widehat{\mu}$ and $\mathrm{T}(\mu, \boldsymbol{\alpha}) := (\mathrm{L}_1\mu, \mathrm{L}(\mu, \boldsymbol{\alpha}))$. More explicitly,

$$\mathrm{T}(\mu, \boldsymbol{\alpha}) := (\mathrm{L}_1\mu, \mathrm{L}_0\mu, \langle \mu, 1\rangle, \langle \mu, d_1\rangle + \alpha_1, \ldots, \langle \mu, d_q\rangle + \alpha_q).$$

The adjoint maps $\mathrm{L}_1^* : \mathrm{F}(\mathrm{X}) \to \mathrm{F}(\mathbb{K})$ and $\mathrm{T}^* : \mathrm{F}(\mathrm{X})^2 \times \mathbb{R} \times \mathbb{R}^{\mathrm{q}} \to \mathrm{F}(\mathbb{K}) \times \mathbb{R}^{\mathrm{q}}$ are $(\mathrm{L}_1^* u)(x, a) := u(x)$ and, by (4.13),

$$\mathrm{T}^*(u, v, \beta_0, \boldsymbol{\beta}) := \left(\mathrm{L}_1^* u + \mathrm{L}_0^* v + \beta_0 + \sum_{i=1}^{q} \beta_i d_i, \boldsymbol{\beta} \right).$$

Then the EC linear program corresponding to $\mathrm{CP}(\nu_0)$ is

$$\mathrm{EC}(\nu_0) \quad \text{Minimize} \quad \langle \mu, c\rangle = \langle (\mu, \boldsymbol{\alpha}), (c, \mathbf{0})\rangle$$

$$\text{subject to} \quad \mathrm{T}(\mu, \boldsymbol{\alpha}) = (\nu_0, 0, 1, \mathbf{k}), \quad (\mu, \boldsymbol{\alpha}) \in \mathrm{M}_+(\mathbb{K}) \times \mathbb{R}^{\mathrm{q}}_+. \tag{4.21}$$

(Compare with (4.10), (4.11).) The associated dual program is (cf. (4.14), (4.15))

$$\mathrm{EC}^*(\nu_0) : \quad \text{Maximize } \langle (\nu_0, 0, 1, \mathbf{k}), (u, v, \beta_0, \boldsymbol{\beta})\rangle = \langle \nu_0, u\rangle + \beta_0 + \langle \mathbf{k}, \boldsymbol{\beta}\rangle$$

$$\text{subject to} \quad \mathrm{T}^*(u, v, \beta_0, \boldsymbol{\beta}) \le (c, \mathbf{0}), \quad (u, v, \beta_0, \boldsymbol{\beta}) \in \mathrm{F}(\mathrm{X})^2 \times \mathbb{R} \times \mathbb{R}^{\mathrm{q}}.$$

(a) Suppose that the hypotheses of Remark 3.8 and Assumption 4.2 are satisfied. Then, by Remark 3.8, $\mathrm{EC}(\nu_0)$ is solvable, and, furthermore, there is no duality gap for $\mathrm{EC}(\nu_0)$, i.e., $\sup \mathrm{EC}^*(\nu_0) = \min \mathrm{EC}(\nu_0)$.

(b) Under the hypotheses of Theorem 4.5, the following statements are equivalent:
   (b_1) There exists a pair $(\mu, \boldsymbol{\alpha})$ that satisfies (4.21).
   (b_2) If $(u, v, \beta_0, \boldsymbol{\beta}) \in \mathrm{F}(\mathrm{X})^2 \times \mathbb{R} \times \mathbb{R}^{\mathrm{q}}$ is such that $\mathrm{T}^*(u, v, \beta_0, \boldsymbol{\beta}) \ge (0, \mathbf{0})$, then

$$\langle (\nu_0, 0, 1, \mathbf{k}), (u, v, \beta_0, \boldsymbol{\beta})\rangle = \langle \nu_0, u\rangle + \beta_0 + \langle \mathbf{k}, \boldsymbol{\beta}\rangle \ge 0.$$

5. Strong duality. In view of Theorem 4.4, to prove the strong duality condition (4.20), we need conditions for the dual program EC^* to be *solvable*, that is, for the existence of a triplet $(u^*, \beta_0^*, \boldsymbol{\beta}^*)$ in $\mathrm{F}(\mathrm{X}) \times \mathbb{R} \times \mathbb{R}^q$ that satisfies (4.15) (or (4.16), (4.17)) and attains the supremum in (4.14) (or (4.18)), i.e.,

$$\max(\mathrm{EC}^*) = \beta_o^* + \langle \mathbf{k}, \boldsymbol{\beta}^*\rangle. \tag{5.1}$$

To do this, we first take a *maximizing sequence* for EC*, namely, a sequence of triplets $(u^n, \beta_0^n, \boldsymbol{\beta}^n)$ in $\mathrm{F(X)} \times \mathbb{R} \times \mathbb{R}^q$ that satisfies (4.16) and (4.17), i.e.,

$$\beta_0^n + u^n(x) \le c(x,a) - \sum_{i=1}^{q} \beta_i^n d_i(x,a) + \int_{\mathrm{X}} u^n(y) Q(dy|x,a) \quad \text{and } \boldsymbol{\beta}^n \le \mathbf{0} \tag{5.2}$$

for all $(x,a) \in \mathbb{K}$ and $n = 1, 2, \dots$, and, moreover,

$$\langle (0, 1, \mathbf{k}), (u^n, \beta_0^n, \boldsymbol{\beta}^n) \rangle = \beta_0^n + \langle \mathbf{k}, \boldsymbol{\beta}^n \rangle \uparrow \sup(\mathrm{EC}^*) = \rho^*, \tag{5.3}$$

where the last equality is due to (4.19). The idea now is, of course, to use the maximizing sequence to deduce the existence of an optimal triplet $(u^*, \beta_0^*, \boldsymbol{\beta}^*)$. With this in mind, let us first note that $\beta_0^n + \langle \mathbf{k}, \boldsymbol{\beta}^n \rangle \le \beta_0^n$, which, together with (5.3), yields that

$$\liminf_{n \to \infty} \beta_0^n \ge \rho^* \ge 0.$$

Thus, without loss of generality, we may suppose that the sequence $\{\beta_0^n\}$ is *nonnegative*. We also require the following assumption, which is discussed in the next section together with an example.

Assumption 5.1. There exists a maximizing sequence $\{(u^n, \beta_0^n, \boldsymbol{\beta}^n),\ n = 1, 2, \dots\}$ for EC* such that

(a) the (nonnegative) sequence $\{\beta_0^n\}$ is bounded, and

(b) the sequence $\{u^n\} \subset \mathrm{F(X)}$ is bounded in the $\widehat{w}$-norm (that is, for some constant $\widehat{k}$, $|u^n(x)| \le \widehat{k}\widehat{w}(x)$ for all $x \in \mathrm{X}$ and $n = 1, 2, \dots$).

In Proposition 6.2(a), we show that $\{u^n\}$ can be taken as a *nondecreasing* sequence of *nonnegative* functions. Hence we may replace Assumption 5.1(b) with (6.11).

THEOREM 5.2 (strong duality and the constrained optimality equation). *Suppose that Assumptions* 3.1, 4.2, *and* 5.1 *are satisfied. Then the following hold:*

(a) EC* *is solvable; hence, the strong duality condition* (4.20) *holds.*

(b) *Let* $(\mu^*, \boldsymbol{\alpha}^*)$ and $(u^*, \beta_0^*, \boldsymbol{\beta}^*)$ *be optimal solutions for* EC *and* EC*, *respectively. Disintegrate* μ^* *as in* (3.1), *that is,* $\mu^* = \widehat{\mu}^* \cdot \varphi^*$ *with* φ^* *in* Φ, *and let*

$$c^*(x,a) := c(x,a) - \sum_{i=1}^{q} \beta_i^* d_i(x,a) \quad \textit{for } (x,a) \in \mathbb{K}. \tag{5.4}$$

Then

$$\langle \boldsymbol{\alpha}^*, \boldsymbol{\beta}^* \rangle = 0, \tag{5.5}$$

and, moreover, for $\widehat{\mu}^*$*-almost all (a.a.)* $x \in \mathrm{X}$, *the constrained optimality equation*

$$\beta_0^* + u^*(x) = \inf_{a \in \mathrm{A}(x)} \left[c^*(x,a) + \int_{\mathrm{X}} u^*(y) Q(y|x,a) \right] \tag{5.6}$$

holds, as well as

$$\beta_0^* + u^*(x) = c^*(x, \varphi^*) + \int_{\mathrm{X}} u^*(y) Q(y|x, \varphi^*) \tag{5.7}$$

for $\widehat{\mu}^*$*-a.a.* $x \in \mathrm{X}$. *Hence the following hold:*

(c) *There exists a deterministic stationary policy* $f^* \in \mathbb{F}$ *such that*

$$\beta_0^* + u^*(x) = c^*(x, f^*) + \int_{\mathrm{X}} u^*(y) Q(y|x, f^*) \tag{5.8}$$

for $\widehat{\mu}^*$*-a.a.* $x \in \mathrm{X}$.

(d) *The number* β_0^* *satisfies*

$$\beta_0^* = \langle \mu^*, c^* \rangle = \rho^* - \sum_{i=1}^{q} \beta_i^* \langle \mu^*, d_i \rangle = \rho^* - \langle \mathbf{k}, \boldsymbol{\beta}^* \rangle, \tag{5.9}$$

with $\rho^* = \min(\mathrm{EC}) = \max(\mathrm{EC}^*)$, *as in* (4.20).

We refer to (5.6) as the "constrained optimality equation" because it clearly is the analogue (in the constrained case) of the ACOE for unconstrained MCPs. (See [3, 11, 15, 16, 17, 18, 22, 25, 26].) It should be remarked, though, that (5.6)–(5.8) hold $\widehat{\mu}^*$*-almost everywhere* only. To get the equality for all $x \in \mathrm{X}$ one needs to impose suitable hypotheses, say, as in [29] or [33] (see also the example in the next section). For instance, Example 2.2 in [33] shows a constrained MCP in which every $f \in \mathbb{F}$ induces an irreducible positive recurrent Markov chain, and yet there is no $f \in \mathbb{F}$ that satisfies (5.8) for all $x \in \mathrm{X}$.

Proof of Theorem 5.2. (a) Let $\{(u^n, \beta_0^n, \boldsymbol{\beta}^n), n = 1, 2, \dots\}$ be a maximizing sequence for EC* (that is, as in (5.2) and (5.3)) that satisfies Assumption 5.1. As $\{\beta_0^n\}$ is bounded, it has a convergent subsequence, and, therefore, by (5.3), so does $\{\langle \mathbf{k}, \boldsymbol{\beta}^n \rangle\}$. Hence, as $\langle \mathbf{k}, \boldsymbol{\beta}^n \rangle \le k_i \beta_i^n \le 0$ for all $i = 1, \dots, q$, there exists a subsequence $\{m\}$ of $\{n\}$ such that the limits

$$\beta_i^* := \lim_{m \to \infty} \beta_i^m \text{ for } i = 0, 1, \dots, q \tag{5.10}$$

exist. Now define

$$u^*(x) := \limsup_{m \to \infty} u^m(x) \quad \forall\, x \in \mathrm{X}.$$

By Assumption 5.1(b), $u^*(\cdot)$ is in F(X), and, moreover, by Assumption 4.2(b) and Fatou's lemma,

$$\limsup_{m \to \infty} \int_{\mathrm{X}} u^m(y) Q(dy|x, a) \le \int_{\mathrm{X}} u^*(y) Q(dy|x, a) \quad \forall\, (x, a) \in \mathbb{K}.$$

We can now see that $(u^*, \beta_0^*, \boldsymbol{\beta}^*)$ is an optimal solution for EC*. Indeed, in (5.2), replace n with m, and then take the limit supremum as $m \to \infty$ to get

$$\beta_0^* + u^*(x) \le c(x, a) - \sum_{i=1}^{q} \beta_i^* d_i(x, a) + \int_{\mathrm{X}} u^*(y) Q(dy|x, a) \tag{5.11}$$

for all $(x, a) \in \mathbb{K}$, and

$$\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_q^*) \le \mathbf{0}. \tag{5.12}$$

This means that $(u^*, \beta_0^*, \boldsymbol{\beta}^*)$ is feasible for EC* (see (4.15) or (4.16)–(4.17)), and, furthermore, from (5.10) and (5.3), we see that (5.1) holds.

(b) Let $(\mu^*, \boldsymbol{\alpha}^*)$ and $(u^*, \beta_0^*, \boldsymbol{\beta}^*)$ be optimal solutions for EC and EC*, respectively. By the strong duality condition (4.20) and *complementary slackness* (see [2, 18, 22]),

$$\langle (\mu^*, \boldsymbol{\alpha}^*), (c, \mathbf{0}) - \mathrm{L}^*(u^*, \beta_0^*, \boldsymbol{\beta}^*) \rangle = 0.$$

That is, by (4.13) and (5.4),

$$0 \le \langle \mu^*, c^* - \mathrm{L}_0^* u^* - \beta_0^* \rangle = \langle \boldsymbol{\alpha}^*, \boldsymbol{\beta}^* \rangle \le 0, \tag{5.13}$$

where the first and the second inequalities follow from (5.11) and (5.12), respectively. From (5.13) we obtain (5.5) and also that

$$\int [c^*(x,a) - \mathrm{L}_0^* u^*(x,a) - \beta_0^*] \mu^*(d(x,a)) = 0.$$

The remainder of the proof of part (b) as well as the proof of (c) can be done as the proof of Theorem 12.4.2 in [22, pp. 236, 237]. (Similar arguments are used in [37, Theorem 4.6].)

(d) As μ^* is stable (Definition 3.4), its marginal $\widehat{\mu}^*$ on X is an i.p.m. for the transition kernel $Q(\cdot\,|\,\cdot\,, \varphi^*)$. Thus, integrating both sides of (5.7) with respect to μ^*, we get the first equality in (5.9), i.e.,

$$\beta_0^* = \langle \mu^*, c^* \rangle.$$

In turn, the latter equality combined with (5.4) yields

$$\begin{aligned} \beta_0^* &= \langle \mu^*, c \rangle - \sum_{i=1}^{q} \beta_i^* \langle \mu^*, d_i \rangle \\ &= \rho^* - \sum_{i=1}^{q} \beta_i^* \langle \mu^*, d_i \rangle \text{ (by (4.12))}, \end{aligned}$$

which is the second equality in (5.9). Finally, from (4.20) and (5.1),

$$\rho^* = \beta_0^* + \langle \mathbf{k}, \boldsymbol{\beta}^* \rangle,$$

which gives the third equality in equation (5.9). This completes the proof of Theorem 5.2. □

Remark 5.3. Theorem 5.2 holds for the linear programs EC(ν_0) and EC*(ν_0) in Remark 4.6 with the obvious changes. Indeed, in Assumption 5.1, replace $(u^n, \beta_0^n, \boldsymbol{\beta}^n)$ with a maximizing sequence $(v^n, u^n, \beta_0^n, \boldsymbol{\beta}^n) \in \mathrm{F(X)}^2 \times \mathbb{R} \times \mathbb{R}^{\mathrm{q}}$ for EC*(ν_0); both v^n and u^n are supposed to satisfy the boundedness condition in Assumption 5.1(b). Moreover, suppose that the hypotheses in Remark 4.6(a) hold. Then Theorem 5.2(a) is valid for EC*(ν_0) in lieu of EC*. Similarly, parts (b) and (c) hold when $\widehat{\mu}^*$ and $(u^*, \beta_0^*, \boldsymbol{\beta}^*)$ are replaced with ν_0 and $(v^*, u^*, \beta_0^*, \boldsymbol{\beta}^*)$, respectively. Finally, instead of (5.9), we have the following: β_0^* satisfies $\beta_0^* = \langle \mu^*, c^* \rangle$, with c^* as in (5.8) and $\mu^* = \nu_0 \cdot \varphi^*$ for some φ^* in Φ.

6. Computing optimal policies. By (5.6) and (5.9), it is evident that solving CP via the dual linear program EC* is essentially equivalent to solving an ACOE for a certain cost-per-stage function (for instance, as in (5.4)). In this section, we describe a general procedure for solving CP using ACOEs, assuming of course that they are

well defined, and then we present a detailed example to illustrate this approach. We begin by introducing some notation and useful comments.

For each $i = 0, 1, \ldots, q$, let ρ_i be the *unconstrained minimum* of the cost $J_i(\pi, \nu)$ in (2.5), (2.6). That is, for $i = 0, 1, \ldots, q$,

$$\rho_i := \inf\{J_i(\pi, \nu) \,|\, \pi \in \Pi,\ \nu \in \mathbb{P}(\mathrm{X})\}. \tag{6.1}$$

A stationary policy $f_i \in \mathbb{F}$ is said to be *strictly optimal* for J_i $(i = 0, 1, \ldots, q)$ if

$$\rho_i = J_i(f_i, \nu) =: J_i(f_i) \quad \forall\ \nu \in \mathbb{P}(\mathrm{X}). \tag{6.2}$$

Strict optimality typically holds, for instance, for so-called *canonical policies*, which are obtained as "minimizers" of suitable ACOEs—see [3, 11, 15, 18, 22, 26, 31] and (6.13). On the other hand, from (2.7)–(2.9), we can easily deduce the following fact for a strictly optimal policy for J_0 to be optimal for CP.

PROPOSITION 6.1. *Let $f_0 \in \mathbb{F}$ be a stationary policy such that*
(a) *f_0 is strictly optimal for J_0, i.e., $J_0(f_0) = \rho_0$;*
(b) *$J_i(f_0, \nu) = J_i(f_0)$ for all $\nu \in \mathbb{P}(\mathrm{X})$ and $i = 1, \ldots, q$; and*
(c) *$J_i(f_0) \le k_i$ for all $i = 1, \ldots, q$.*
Then f_0 is an optimal policy for CP *in the sense that $J_0(f_0) = \rho^*$.*

Proposition 6.1 is illustrated (with $q = 1$) in Figure 6.1, in which $\Gamma \subset \mathbb{R}^{q+1}$ is the set of all of the cost vectors $\mathbf{J}(\pi, \nu) := (J_0(\pi, \nu), J_1(\pi, \nu), \ldots, J_q(\pi, \nu))$, i.e.,

$$\Gamma := \{\mathbf{J}(\pi, \nu) \mid \pi \in \Pi,\ \nu \in \mathbb{P}(\mathrm{X})\}. \tag{6.3}$$

This set is called the *performance* or *achievable* set of the *multiobjective control problem* with cost vectors $\mathbf{J}(\pi, \nu)$; see [1, 12, 13, 31, 36, 38, 39, 40, 43]. (Actually, some of the components of $\mathbf{J}(\pi, \nu)$ might be $+\infty$, but this is irrelevant in our present case.)

We next proceed to introduce a sequence of feasible triplets for EC*, that is, as in (5.2).

Choose an *arbitrary* q-vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q) \le 0$, and consider the cost-per-stage

$$c^\beta(x, a) := c(x, a) - \sum_{i=1}^{q} \beta_i d_i(x, a) \quad \forall\ (x, a) \in \mathbb{K} \tag{6.4}$$

and the corresponding long-run expected average cost

$$J^\beta(\pi, \nu) := \limsup_{n \to \infty} \frac{1}{n} V_n^\beta(\pi, \nu),$$

where

$$V_n^\beta(\pi, \nu) := E_\nu^\pi \left[\sum_{t=0}^{n-1} c^\beta(x_t, a_t) \right].$$

By (2.5) and (2.6),

$$J^\beta(\pi, \nu) \le J_0(\pi, \nu) - \sum_{i=1}^{q} \beta_i J_i(\pi, \nu). \tag{6.5}$$

Moreover, by Assumption 3.1(b), (c), the cost function c^β is *nonnegative and* inf-*compact*. Therefore, by standard dynamic programming arguments, the optimal n-stage cost

$$v_n^\beta(x) := \inf_{\pi \in \Pi} V_n^\beta(\pi, x) \text{ for } x \in \mathrm{X} \text{ and } n = 1, 2, \ldots,$$

with $v_0^\beta(\cdot) \equiv 0$, satisfies that

$$v_n^\beta(x) = \inf_{a\in \mathrm{A}(x)} \left[c^\beta(x,a) + \int_{\mathrm{X}} v_{n-1}^\beta(y) Q(dy|x,a) \right] \quad \forall\ x \in \mathrm{X}. \tag{6.6}$$

Now, for $n = 1, 2, \ldots$, let

$$\rho_n^\beta := \inf_{x\in \mathrm{X}} [v_n^\beta(x) - v_{n-1}^\beta(x)], \quad m_n^\beta := \rho_n^\beta + m_{n-1}^\beta, \tag{6.7}$$

with $m_0^\beta := 0$, and

$$u_n^\beta(x) := v_n^\beta(x) - m_n^\beta.$$

Then, for all $x \in \mathrm{X}$ and $n = 1, 2, \ldots$, we may rewrite (6.6) as

$$\rho_n^\beta + u_n^\beta(x) = \inf_{a\in \mathrm{A}(x)} \left[c^\beta(x,a) + \int_{\mathrm{X}} u_{n-1}^\beta(y) Q(dy|x,a) \right], \tag{6.8}$$

and we also have the following proposition.

PROPOSITION 6.2.

(a) $\{u_n^\beta(\cdot)\}$ *is a nondecreasing, sequence of nonnegative functions.*

(b) $\{\rho_n^\beta\}$ *is nondecreasing, and*

$$0 \le \rho_n^\beta \le \rho^* - \langle \mathbf{k}, \boldsymbol{\beta} \rangle \quad \forall\ n. \tag{6.9}$$

Proof. Part (a) follows from (6.7). To prove (b), first note that $\rho_n^\beta \ge 0$ because, as c^β is nonnegative, we have $v_n^\beta \ge v_{n-1}^\beta$. Moreover, by (6.6) and using the fact that $\inf u(\cdot) - \inf v(\cdot) \ge \inf[u(\cdot) - v(\cdot)]$ for any two functions $u(\cdot)$ and $v(\cdot)$, with $v(\cdot)$ bounded below, we get

$$v_n^\beta(x) - v_{n-1}^\beta(x) \ge \inf_{a\in \mathrm{A}(x)} \int_{\mathrm{X}} [v_{n-1}^\beta(y) - v_{n-2}^\beta(y)] Q(dy|x,a) \ge \rho_{n-1}^\beta,$$

and so $\{\rho_n^\beta\}$ is nondecreasing. On the other hand, by (6.8) and part (a),

$$\rho_n^\beta + u_n^\beta(x) \le c^\beta(x,a) + \int_{\mathrm{X}} u_n^\beta(y) Q(dy|x,a) \quad \forall\ (x,a) \in \mathbb{K}. \tag{6.10}$$

By Assumption 4.2, together with (6.6) and a straightforward induction argument, the function u_n^β is in $\mathbb{F}(\mathrm{X})$ for each n. Now let $\mu^* \in \Delta_{\mathrm{s}}$ be as in (3.11) and (3.12), and integrate both sides of (6.10) with respect to μ^*. Then (6.4) and the invariance property in Definition 3.4(b) give that

$$\rho_n^\beta \le \rho^* - \sum_{i=1}^q \beta_i \langle \mu^*, d_i \rangle \le \rho^* - \sum_{i=1}^q \beta_i k_i,$$

and (6.9) follows. □

By (6.10), the triplets $(u_n^\beta, \rho_n^\beta, \boldsymbol{\beta})$ satisfy (5.2) for each $n = 1, 2, \ldots$ and $\boldsymbol{\beta} \le \mathbf{0}$. Furthermore, by Proposition 6.2(b), there exists a nonnegative number $\rho^\beta \le \rho^* - \langle \mathbf{k}, \boldsymbol{\beta} \rangle$ such that $\rho_n^\beta \uparrow \rho^\beta$. Now let us suppose that there is a number $\widehat{k} \ge 0$ such that

$$u_n^\beta(\cdot) \le \widehat{k}\widehat{w}(\cdot) \quad \forall\ n. \tag{6.11}$$

Then, by Proposition 6.2(a), there is a nonnegative function u^β in $\mathbb{F}(\mathrm{X})$ such that $u_n^\beta(x) \uparrow u^\beta(x)$ for all $x \in \mathrm{X}$. Therefore, letting $n \to \infty$ in (6.10), monotone convergence yields

$$\rho^\beta + u^\beta(x) \le c^\beta(x,a) + \int_{\mathrm{X}} u^\beta(y) Q(dy|x,a) \quad \forall\ (x,a) \in \mathbb{K},$$

which in turn gives

$$\rho^\beta + u^\beta(x) \le \inf_{a \in \mathrm{A}(x)} \left[c^\beta(x,a) + \int_{\mathrm{X}} u^\beta(y) Q(dy|x,a) \right] \quad \forall\ (x,a) \in \mathrm{X}. \tag{6.12}$$

Summarizing, we have described a general procedure to obtain feasible triplets $(u^\beta, \rho^\beta, \boldsymbol{\beta})$ for EC*. However, to actually solve CP, we need stronger hypotheses, as in the following result. (There are several ways in which one can get the *equality* in (6.12) and the policy f^β in (6.13): see [1, 3, 10, 11, 15, 18, 22, 26, 28, 31, 33, 41].)

THEOREM 6.3. *Suppose Assumptions* 3.1 *and* 4.2 *are satisfied. In addition, there exist a q-vector* $\boldsymbol{\beta} \le \mathbf{0}$ *and a stationary policy* $f^\beta \in \mathbb{F}$ *for which the equality holds in* (6.12), *and* $f^\beta(x) \in \mathrm{A}(x)$ *attains the minimum in* (6.12) *for all* $x \in \mathrm{X}$, *i.e.,*

$$\rho^\beta + u^\beta(x) = c^\beta(x, f^\beta) + \int_{\mathrm{X}} u^\beta(y) Q(dy|x, f^\beta) \quad \forall\ x \in \mathrm{X}. \tag{6.13}$$

Moreover, u^β *is such that*

$$\lim_{n \to \infty} \frac{1}{n} E_x^{f^\beta}[u^\beta(x_n)] = 0, \tag{6.14}$$

and, furthermore,

$$J_i(f^\beta, x) = k_i \quad \forall\ x \in \mathrm{X} \text{ and } i = 1, \ldots, q. \tag{6.15}$$

Then $J_0(f^\beta, \cdot)$ *is a constant, say,* $J_0(f^\beta, x) = J_0(f^\beta)$ *for all* $x \in \mathrm{X}$, *and* f^β *is optimal for* CP *in the sense that* $J_0(f^\beta) = \rho^*$, *and so*

$$\rho^\beta = \rho^* - \langle \mathbf{k}, \boldsymbol{\beta} \rangle.$$

Proof. As in the proof of (6.9),

$$\rho^\beta \le \rho^* - \langle \mathbf{k}, \boldsymbol{\beta} \rangle.$$

On the other hand, iteration of (6.13) and using (6.14) and (6.15) give

$$\rho^\beta = J_0(f^\beta) - \langle \mathbf{k}, \boldsymbol{\beta} \rangle.$$

Hence, as $J_0(f^\beta) \ge \rho^*$, we get

$$\rho^\beta \ge \rho^* - \langle \mathbf{k}, \boldsymbol{\beta} \rangle,$$

and the theorem follows. □

The following LQ example, which is similar to the "stochastic stabilization problem" in [11, 31] and the "mold level control problem" in [41], illustrates the ACOE approach in the previous paragraphs.

Example 6.4. Consider the scalar linear system

$$x_{t+1} = \theta_1 x_t + \theta_2 a_t + \xi_t, \quad t = 0, 1, \ldots, \tag{6.16}$$

with state and control spaces $\mathrm{X} = \mathbb{R}$ and $\mathrm{A}(x) \equiv \mathrm{A} = \mathbb{R}$ for all $x \in \mathrm{X}$, and nonzero coefficients θ_1, θ_2. The random disturbances ξ_t are i.i.d. with

$$E(\xi_0) = 0 \quad \text{and} \quad \sigma^2 := E(\xi_0^2) < \infty$$

and independent of the initial state x_0. The CP we are concerned with is given by (2.5)–(2.8) with $q = 1$, constraint constant $k_1 \geq 0$, and costs

$$c(x, a) := c_1 x^2 + c_2 a^2 \quad \text{and} \quad d_1(x, a) := (x - a)^2, \tag{6.17}$$

where c_1 and c_2 are positive constants. By the "quadratic" nature of the costs, we will consider only initial distributions $\nu \in \mathbb{P}(\mathrm{X})$ with a finite second moment, i.e., $\int x^2 \nu(dx) < \infty$.

Verification of Assumption 3.1. Parts (b), (c), and (d) in Assumption 3.1 trivially hold (concerning (d), see Remark 3.6). To verify (a), observe that the *unconstrained minimum* ρ_1 in (6.1)–(6.2) is given by

$$\rho_1 = J_1(f_1) = 0, \tag{6.18}$$

which is attained by the *strictly optimal* policy $f_1(x) := x$ for all $x \in \mathrm{X}$. Thus $J_1(f_1) \leq k_1$. In fact, for any given constraint constant $k_1 \geq 0$, the stationary policy, say, $f(x) := x + k_1^{\frac{1}{2}}$, satisfies that

$$J_1(f) := J_1(f, x) = k_1 \quad \forall\, x \in \mathrm{X} \text{ (cf. (6.15))}. \tag{6.19}$$

Now let C_0 be the unique positive solution of the quadratic equation

$$(c_2 + \theta_2^2 C)C = c_1 c_2 + (c_2 \theta_1^2 + c_1 \theta_2^2)C, \tag{6.20}$$

and let

$$\widehat{f}_0 := (c_2 + C_0 \theta_2^2)^{-1} C_0 \theta_1 \theta_2.$$

If the constants θ_1, θ_2, and $\widehat{f}_0$ are such that

$$|\theta_1 - \theta_2 \widehat{f}_0| < 1, \tag{6.21}$$

then it is well known that the stationary policy

$$f_0(x) := -\widehat{f}_0 x \quad \forall\, x \in \mathrm{X} \tag{6.22}$$

is *strictly optimal* for J_0 and that the *unconstrained minimum* for J_0 (as in (6.1), (6.2)) is

$$\rho_0 = J_0(f_0) = C_0 \sigma^2; \tag{6.23}$$

see, for instance, [11, 18, 31]. Hence Assumption 3.1(a) holds *at least* when (6.21) is satisfied.

Verification of Assumption 4.2. Let $w(x,a)$ and $\widehat{w}(x)$ be as in (4.1); in our present case, they are given by

$$w(x,a) := 1 + c_1x^2 + c_2a^2 \quad \text{and} \quad \widehat{w}(x) := 1 + c_1x^2. \tag{6.24}$$

As $d_1(x,a) \le 2(x^2+a^2)$, it is clear that part (a) in Assumption 4.2 holds. To verify part (b), note that, by (6.16), for any nonnegative measurable function u on X, we have

$$\int_{\mathrm{X}} u(y)Q(dy|x,a) = E[u(x_{t+1}) \mid x_t = x, a_t = a] = E[u(\theta_1 x + \theta_2 a + \xi_0)].$$

Therefore, *if u is of the form $u(x) = u_1x^2 + u_2$, then* (as $E(\xi_0) = 0$ and $E(\xi_0^2) = \sigma^2$)

$$\int_{\mathrm{X}} u(y)Q(dy|x,a) = u_1(\theta_1 x + \theta_2 a)^2 + u_1\sigma^2 + u_2. \tag{6.25}$$

It follows that Assumption 4.2(b) holds for $\widehat{w}$ in (6.24); that is, for some constant m sufficiently large,

$$\int_{\mathrm{X}} \widehat{w}(y)Q(dy|x,a) \le m \cdot w(x,a) \quad \forall\ (x,a) \in \mathbb{K}.$$

Illustration of Proposition 6.1. First, observe the following lemma.

LEMMA 6.5. *Let $f \in \mathbb{F}$ be a stationary policy given by $f(x) := -\widehat{f}x$ for all $x \in \mathrm{X}$, and let $\widehat{\theta} := \theta_1 - \theta_2\widehat{f}$, where θ_1, θ_2 are the coefficients in* (6.16). *If $|\widehat{\theta}| < 1$, then for all $x \in \mathrm{X}$*

$$J_0(f,x) \equiv J_0(f) = (c_1 + c_2\widehat{f}^2)\sigma^2/(1-\widehat{\theta}^2), \tag{6.26}$$

$$J_1(f,x) \equiv J_1(f) = (1+\widehat{f})^2\sigma^2/(1-\widehat{\theta}^2). \tag{6.27}$$

In particular, for $f_0(x) := -\widehat{f}_0x$ and $f_1(x) := x$ in (6.22) *and* (6.18), *respectively, we have*

$$J_1(f_0) = (1+\widehat{f}_0)^2\sigma^2/(1-\widehat{\theta}_0^{\,2}), \quad \textit{with} \quad \widehat{\theta}_0 := \theta_1 - \theta_2\widehat{f}_0, \tag{6.28}$$

$$J_0(f_1) = (c_1+c_2)\sigma^2/(1-\widehat{\theta}_1^{\,2}), \quad \textit{with} \quad \widehat{\theta}_1 := \theta_1 + \theta_2. \tag{6.29}$$

Proof. Replacing a_t in (6.16) with $a_t := f(x_t) = -\widehat{f}x_t$, we obtain

$$x_t = (\theta_1 - \theta_2\widehat{f})x_{t-1} + \xi_{t-1} = \widehat{\theta}x_{t-1} + \xi_{t-1} \quad \forall\ t = 1,2,\ldots.$$

Hence, for all $t = 1,2,\ldots,$

$$x_t = \widehat{\theta}^t x_0 + \sum_{j=0}^{t-1} \widehat{\theta}^j \xi_{t-1-j},$$

and so

$$E_x^f(x_t^2) = \widehat{\theta}^{2t}x^2 + (\sigma^2(1-\widehat{\theta}^{2t}))/(1-\widehat{\theta}^2).$$

This yields that

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} E_x^f(x_t^2) = \sigma^2/(1-\widehat{\theta}^2) \quad \forall\, x \in \mathrm{X}. \tag{6.30}$$

Now note that, using $a = f(x) = -\widehat{f}x$ in (6.17), we get

$$c(x,a) = (c_1 + c_2\widehat{f}^2)x^2 \quad \text{and} \quad d_1(x,a) = (1+\widehat{f})^2 x^2 \tag{6.31}$$

for all $x \in \mathrm{X}$. Finally, inserting (6.31) and (6.30) in (2.5), (2.6), we obtain (6.26) and (6.27). □

From (6.28) and Proposition 6.1, we conclude the following (see (6.50) and Figure 6.1).

COROLLARY 6.6. *If the constraint constant k_1 is such that $k_1 \geq J_1(f_0)$, then the policy $f_0 \in \mathbb{F}$ in* (6.22) *is optimal for* CP*, and the optimal value ρ^* of* CP *(see* (2.9)*) is the unconstrained minimum $\rho_0 = C_0\sigma^2$ in* (6.23).

Illustration of (6.4)–(6.8) *and Theorem* 6.3. Consider the cost function $c^\beta(x,a)$ in (6.4) with $q = 1$, $\beta \leq 0$, and $c(x,a)$ and $d_1(x,a)$ as in (6.17), i.e.,

$$c^\beta(x,a) := c_1x^2 + c_2a^2 - \beta(x-a)^2. \tag{6.32}$$

From (6.32), (6.6), and a straightforward induction argument (using (6.25)), we obtain the following.

LEMMA 6.7. *Let $v_0(\beta) = m_0^\beta = 0$. For each $x \in X$ and $n = 1, 2, \ldots$, the stationary policy*

$$f_n^\beta(x) := -f_n(\beta) \cdot x, \tag{6.33}$$

with coefficient

$$f_n(\beta) := (\beta + v_{n-1}(\beta)\theta_1\theta_2)/(c_2 - \beta + v_{n-1}(\beta)\theta_2^2), \tag{6.34}$$

realizes the minimum in (6.6)*, and the "value iteration" function $v_n^\beta(\cdot)$ is given by*

$$v_n^\beta(x) = v_n(\beta)x^2 + m_n^\beta \tag{6.35}$$

with coefficients

$$v_n(\beta) = c_1 + c_2 f_n(\beta)^2 - \beta[1+f_n(\beta)]^2 + v_{n-1}(\beta)[\theta_1 - f_n(\beta)\theta_2]^2, \tag{6.36}$$

$$m_n^\beta := v_{n-1}(\beta)\sigma^2 + m_{n-1}^\beta. \tag{6.37}$$

From (6.34) and (6.36), it follows, in particular, that (6.6) can be expressed as in (6.8) with

$$\rho_n^\beta := v_{n-1}(\beta)\sigma^2 \quad \text{and} \quad u_n^\beta(x) := v_n(\beta)x^2 \quad \forall\, x \in \mathrm{X},\ n = 1, 2, \ldots. \tag{6.38}$$

By Proposition 6.2, we know that

$$\rho_n^\beta \uparrow \rho^\beta \quad \text{and} \quad u_n^\beta(x) \uparrow u^\beta(x) \quad \forall\, x \in \mathrm{X}, \quad \text{as} \quad n \to \infty. \tag{6.39}$$

We now wish to show that the limiting values satisfy (6.13) with *equality*, that is, the ACOE

$$\rho^\beta + u^\beta(x) = \min_{a\in \mathrm{A}(x)} \left[c^\beta(x,a) + \int_X u^\beta(y) Q(dy|x,a) \right] \tag{6.40}$$

holds for all $x \in \mathrm{X}$ (note that we have written "min" rather than "inf"), and, in addition, that u^β is in $\mathbb{F}(\mathrm{X})$. To do this, we will use (6.38) to compute the limits in (6.39) as follows.

Inserting (6.34) in (6.36), a direct calculation shows that we can express $v_n(\beta)$ as a linear-fractional transformation of $v_{n-1}(\beta)$, namely,

$$v_n(\beta) = \frac{P + Q v_{n-1}(\beta)}{R + S v_{n-1}(\beta)}$$

with coefficients

$$P := c_1 c_2 - (c_1 + c_2)\beta, \quad Q := c_1\theta_2^2 + c_2\theta_1^2 - \beta(\theta_1 - \theta_2)^2,$$

$$R := c_2 - \beta, \quad \text{and} \quad S := \theta_2^2.$$

Therefore, standard arguments (see [11] or [18], for instance) show that

$$v_n(\beta) \to v(\beta) \quad \text{as} \quad n \to \infty, \tag{6.41}$$

where $v(\beta)$ is the unique positive solution of the quadratic equation

$$v(\beta) = \frac{P + Q v(\beta)}{R + S v(\beta)}.$$

Therefore, by (6.38) and (6.39), the ACOE (6.40) holds with

$$\rho^\beta = v(\beta)\sigma^2 \quad \text{and} \quad u^\beta(x) = v(\beta)x^2, \tag{6.42}$$

and, moreover, the right-hand side of (6.40) is minimized by the stationary policy

$$f^\beta(x) = -f(\beta)x, \quad \text{with} \quad f(\beta) := [\beta + v(\beta)\theta_1\theta_2]/[c_2 - \beta + v(\beta)\theta_2^2]. \tag{6.43}$$

Observe that this policy can also be obtained from (6.41), letting $n \to \infty$ in (6.34). On the other hand, from the calculations leading to (6.30), we see that (6.14) holds provided that

$$|\theta(\beta)| < 1 \quad \text{with} \quad \theta(\beta) := \theta_1 - \theta_2 f(\beta). \tag{6.44}$$

Summarizing, (6.13) and (6.14) hold for the CP related to (6.16)–(6.17) provided that $\beta \le 0$ satisfies (6.44). Assumming that the latter is true, we can now use Theorem 6.3 to find an optimal policy for CP when the nonnegative constraint constant k_1 is such that $k_1 < J_1(f_0)$; otherwise, we can use Corollary 6.6. With this in mind, observe that (6.26) and (6.27) yield

$$J_0(f^\beta) = [c_1 + c_2 f(\beta)^2]\sigma^2/[1 - \theta(\beta)^2] \tag{6.45}$$

and

$$J_1(f^\beta) = [1 + f(\beta)]^2\sigma^2/[1 - \theta(\beta)^2], \tag{6.46}$$

respectively. Hence Theorem 6.3 yields the following corollary.

COROLLARY 6.8. *Suppose that* $0 \le k_1 < J_1(f_0)$. *If* $\beta \le 0$ *satisfies* (6.44) *and that*

$$J_1(f^\beta) = k_1, \tag{6.47}$$

then the policy $f^\beta \in \mathbb{F}$ *in* (6.43) *is optimal for the* CP *associated to* (6.16)–(6.17), *and the optimal value* ρ^* *of* CP *is given by*

$$\rho^* = J_0(f^\beta) = \rho^\beta + k_1\beta, \tag{6.48}$$

with $J_0(f^\beta)$ *and* ρ^β *as in* (6.45) *and* (6.42), *respectively.*

Observe that, for given values of the coefficients θ_i and c_i in (6.16) and (6.17), one can actually find $\beta \le 0$ that satisfies (6.44) and (6.47).

Graphical interpretation of Corollaries 6.6 *and* 6.8. Let $\Gamma \subset \mathbb{R}^{q+1}$ be as in (6.3), that is, the set of cost vectors

$$\mathbf{J}(\pi,\nu) := (J_0(\pi,\nu), J_1(\pi,\nu), \dots, J_q(\pi,\nu)).$$

It can be shown that Γ is a convex set—see [36, 38], for instance. Given a q-vector $\boldsymbol{\beta} \le \mathbf{0}$, let $\widehat{\boldsymbol{\beta}} := (1, -\boldsymbol{\beta}) \in \mathbb{R}^{q+1}_+$. Then we may write (6.5) as

$$J^\beta(\pi,\nu) \le \langle \widehat{\boldsymbol{\beta}}, \mathbf{J}(\pi,\nu)\rangle. \tag{6.49}$$

If a pair (π^*,ν^*) is a minimum pair for J^β, then the equality holds in (6.49), and (π^*,ν^*) is said to be a $\boldsymbol{\beta}$-*Pareto pair* for the multiobjetive control problem associated to $\mathbf{J}(\pi,\nu)$. The set $\Gamma^* \subset \Gamma$ of all of the cost vectors $\mathbf{J}(\pi^*,\nu^*)$ corresponding to $\boldsymbol{\beta}$-Pareto pairs for $\boldsymbol{\beta} \le \mathbf{0}$ is called the *Pareto set* of the multiobjective problem. In particular, for the problem (6.16), (6.17), the Pareto set Γ^* consists of the vectors

$$\mathbf{j}(\beta) := (J_0(f^\beta), J_1(f^\beta)) \in \mathbb{R}^2 \tag{6.50}$$

with $\beta \le 0$ and $J_i(f^\beta)$ as in (6.45), (6.46); see Figure 6.1. The"extreme points" of Γ^* are the vectors

$$\mathbf{j}_0 := (J_0(f_0), J_1(f_0)) = (\rho_0, J_1(f_0)), \quad \mathbf{j}_1 := (J_0(f_1), J_1(f_1)) = (J_0(f_1), 0); \tag{6.51}$$

see (6.18), (6.23), (6.28), and (6.29). The graphical interpretation of Corollaries 6.6 and 6.8 is now obvious. In particular, in the context of Corollary 6.8, finding an optimal policy f^β for CP amounts to determining $\beta \le 0$ for which the second component of $\mathbf{j}(\beta)$ coincides with the constraint constant k_1, that is, $J_1(f^\beta) = k_1$, as in (6.47).

We conclude the paper with an example of a control system which is "multichain" in the sense that, for each policy π (even if $\pi = f$ is stationary), the average costs in (2.5), (2.6) may vary with the initial distribution ν.

Example 6.9. Consider the following two-dimensional variant of Example 6.4:

$$(x_{t+1}, y_{t+1}) = (\theta_1 x_t + \theta_2 a_t + \theta_3(y_t)\xi_t, y_t) \quad \forall\ t = 0, 1, \dots$$

with state space $\mathrm{X} = \mathbb{R} \times [\lambda, \infty)$, where $\lambda > 0$ is a constant and $\theta_3(y)$ is a nondecreasing, nonnegative, and continuous function on $[\lambda, \infty)$. The constants θ_1, θ_2 and the disturbances ξ_t are as in Example 6.4, except that now the common distribution of the ξ_t is also assumed to be symmetric with respect to the origin. The cost-per-stage functions are similar to those in (6.17):

$$c((x,y),a) := c_1 x^2 + c_2 a^2 + c_3 y^3, \quad d_1((x,y),a) := (x-a)^2, \tag{6.52}$$

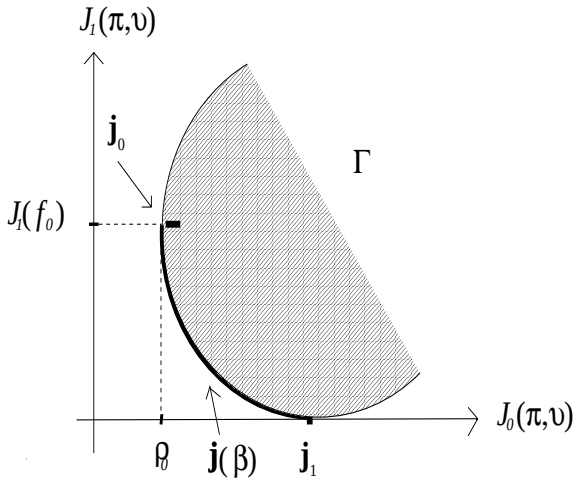


FIG. 6.1. *See* (6.3), (6.50), *and* (6.51).

where c_1, c_2, c_3 are positive constants. Finally, we suppose that (6.21) is satisfied, and so Assumption 3.1 holds.

For each initial state (x_0, y_0) with $y_0 = y \geq \lambda$, let ν_y be an initial distribution concentrated on the set $\{(x, y)|x \in \mathbb{R}\}$. Moreover, let $k_1 \geq 0$ be the constraint constant, and define $f_0 \in \mathbb{F}$ as

$$f_0(x, y) := f_0(x) = -\widehat{f}_0 x \quad \forall\ (x, y) \in \mathrm{X} \quad \text{(see (6.22))}.$$

Then, as in Corollary 6.6, one can see that, if $J_1(f_0, \nu_\lambda) \leq k_1$, then f_0 is optimal for the CP with cost functions in (6.52), and the CP's optimal value is

$$\rho^* = \rho_0 = C_0(\theta_3(\lambda)\sigma)^2 + c_3\lambda^3$$

with C_0 as in (6.23).

REFERENCES

[1] E. ALTMAN, *Constrained Markov Decision Processes*, Chapman and Hall/CRC, Boca Raton, FL, 1999.
[2] E. J. ANDERSON AND P. NASH, *Linear Programming in Infinite-Dimensional Spaces*, Wiley, Chichester, U.K., 1987.
[3] A. ARAPOSTATHIS, V. S. BORKAR, E. FERNÁNDEZ-GAUCHERAND, M. K. GHOSH, AND S. I. MARCUS, *Discrete-time controlled Markov processes with average cost criterion: A survey*, SIAM J. Control Optim., 31 (1993), pp. 282–344.
[4] R. B. ASH, *Real Analysis and Probability*, Academic Press, New York, 1972.
[5] D. P. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.
[6] P. BILLINGSLEY, *Convergence of Probability Measures*, Wiley, New York, 1968.
[7] D. BLACKWELL, *Memoryless strategies in finite-stage dynamic programming*, Ann. Math. Statist., 33 (1964), pp. 863–865.
[8] V. S. BORKAR, *Ergodic control of Markov chains with constraints—the general case*, SIAM J. Control Optim., 32 (1994), pp. 176–186.
[9] B. D. CRAVEN AND J. J. KOLIHA, *Generalizations of Farkas' theorem*, SIAM J. Math. Anal. Appl., 8 (1977), pp. 983–997.
[10] R. DEKKER AND A. HORDIJK, *Average, sensitive and Blackwell optimal policies in denumerable Markov decision chains with unbounded rewards*, Math. Oper. Res., 13 (1988), pp. 395–420.

[11] E. B. Dynkin and A. A. Yushkevich, *Controlled Markov Processes*, Springer-Verlag, Berlin, 1979.
[12] E. Feinberg and A. Shwartz, *Constrained discounted dynamic programming*, Math. Oper. Res., 21 (1996), pp. 922–945.
[13] E. Feinberg and A. Shwartz, *Constrained dynamic programming with two discount factors: Applications and an algorithm*, IEEE Trans. Automat. Control, 44 (1999), pp. 628–631.
[14] K. Golabi, R. B. Kulkarni, and G. B. Way, *A statewide pavement management system*, Interfaces, 12 (1982), pp. 5–21.
[15] E. Gordienko and O. Hernández-Lerma, *Average cost Markov control processes with weighted norms: Existence of canonical policies*, Appl. Math. (Warsaw), 23 (1995), pp. 199–218.
[16] E. Gordienko and O. Hernández-Lerma, *Average cost Markov control processes with weighted norms: Value iteration*, Appl. Math. (Warsaw), 23 (1995), pp. 219–237.
[17] O. Hernández-Lerma and J. González-Hernández, *Infinite linear programming and multichain Markov control processes in uncountable spaces*, SIAM J. Control Optim., 36 (1998), pp. 313–335.
[18] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, New York, 1996.
[19] O. Hernández-Lerma and J. B. Lasserre, *Cone-constrained linear equations in Banach spaces*, J. Convex Anal., 4 (1997), pp. 149–164.
[20] O. Hernández-Lerma and J. B. Lasserre, *Approximation schemes for infinite linear programs*, SIAM J. Optim., 8 (1998), pp. 973–988.
[21] O. Hernández-Lerma and J. B. Lasserre, *Linear programming approximations for Markov control processes in metric spaces*, Acta Appl. Math., 51 (1998), pp. 123–139.
[22] O. Hernández-Lerma and J. B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.
[23] K. Hinderer, *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter*, Lecture Notes in Operations Research and Mathematical Systems 33, Springer-Verlag, Berlin, 1970.
[24] A. Hordijk and F. Spieksma, *Constrained admission control to a queueing system*, Adv. in Appl. Probab., 21 (1989), pp. 409–431.
[25] A. Hordijk and A. A. Yushkevich, *Blackwell optimality in the class of stationary policies in Markov decision chains with a Borel state space and unbounded rewards*, Math. Methods Oper. Res., 49 (1999), pp. 1–39.
[26] A. Hordijk and A. A. Yushkevich, *Blackwell optimality in the class of all policies in Markov decision chains with a Borel state space and unbounded rewards*, Math. Methods Oper. Res., 50 (1999), pp. 421–448.
[27] Y. Huang and M. Kurano, *The LP approach in average rewards MDPs with multiple cost constraints: The countable state case*, J. Inform. Optim. Sci., 18 (1997), pp. 33–47.
[28] M. Kurano, *The existence of a minimum pair of state and policy for Markov decision processes under the hypothesis of Doeblin*, SIAM J. Control Optim., 27 (1989), pp. 296–307.
[29] M. Kurano, J.-I. Nakagami, and Y. Huang, *Constrained Markov decision processes with compact state and action spaces: The average case*, Optimization, 48 (2000), pp. 255–269.
[30] A. Lazar, *Optimal flow control of a class of queueing networks in equilibrium*, IEEE Trans. Automat. Control, 28 (1983), pp. 1001–1007.
[31] A. B. Piunovskiy, *Optimal Control of Random Sequences in Problems with Constraints*, Kluwer, Boston, 1997.
[32] K. Ross and R. Varadarajan, *Multichain Markov decision processes with a sample path constraint: A decomposition approach*, Math. Oper. Res., 16 (1991), pp. 195–207.
[33] L. I. Sennott, *Constrained average cost Markov decision chains*, Probab. Engrg. Inform. Sci., 7 (1993), pp. 69–83.
[34] F. Vakil and A. A. Lazar, *Flow control protocols for integrated networks with partially observed voice traffic*, IEEE Trans. Automat. Control, 32 (1987), pp. 2–14.
[35] K. Yosida, *Functional Analysis*, 5th ed., Springer-Verlag, Berlin, 1978.
[36] R. R. López-Martínez and O. Hernández-Lerma, *The Lagrange and Pareto Approaches to Constrained Markov Control Processes*, Internal Report 283, Mathematics Department, CINVESTAV-IPN.
[37] O. Hernández-Lerma and J. González-Hernández, *Constrained Markov control processes in Borel spaces: The discounted case*, Math. Methods. Oper. Res., 52 (2000), pp. 271–285.
[38] O. Hernández-Lerma and R. Romera, *Pareto Optimality in Multiobjective Markov Control Processes*, Internal Report 278, Mathematics Department, CINVESTAV-IPN.

[39] O. Hernández-Lerma and L. F. Hoyos-Reyes, *A multiobjective control approach to priority queues*, Math. Methods Oper. Res., 53 (2001), pp. 265–277.

[40] O. Hernández-Lerma and L. F. Hoyos-Reyes, *A Multiobjective Formulation of Optimal Control Problems with Additive Costs*, Internal Report 286, Mathematics Department, CINVESTAV-IPN.

[41] J. I. González-Trejo, O. Hernández-Lerma, and L. F. Hoyos-Reyes, *Minimax control of discrete-time stochastic systems*, SIAM J. Control Optim., 41 (2003), pp. 1626–1659.

[42] O. Hernández-Lerma and R. Romera, *Limiting discounted-cost control of partially observable stochastic systems*, SIAM J. Control Optim., 40 (2001), pp. 348–369.

[43] A. Hordijk and L. C. M. Kallenberg, *Constrained undiscounted stochastic dynamic programming*, Math. Oper. Res., 9 (1984), pp. 276–289.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 469–494

A NONLINEAR FILTERING APPROACH TO CHANGEPOINT DETECTION PROBLEMS: DIRECT AND DIFFERENTIAL-GEOMETRIC METHODS*

M. H. VELLEKOOP† AND J. M. C. CLARK‡

Abstract. A benchmark change detection problem is considered which involves the detection of a change of unknown size at an unknown time. Both unknown quantities are modelled by stochastic variables, which allows the problem to be formulated within a Bayesian framework. It turns out that the resulting nonlinear filtering problem is much harder than the well-known detection problem for *known* sizes of the change, and in particular that it can no longer be solved in a recursive manner. An approximating recursive filter is therefore proposed, which is designed using differential-geometric methods in a suitably chosen space of unnormalized probability densities. The new nonlinear filter can be interpreted as an adaptive version of the celebrated Shiryayev–Wonham equation for the detection of a priori known changes, combined with a modified Kalman filter structure to generate estimates of the unknown size of the change. This intuitively appealing interpretation of the nonlinear filter and its excellent performance in simulation studies indicate that it may be of practical use in realistic change detection problems.

Key words. change detection, nonlinear filtering, differential geometry

AMS subject classification. 60G35

PII. S0363012900375950

1. Introduction.

The problem of detecting parameter changes in dynamical systems on the basis of noisy observations has been researched extensively over the last twenty years. Successful applications in many fields have guaranteed a wide interest in the subject and the literature dealing with it is extensive. For good recent surveys of the field and further references, the reader is referred to the papers by Basseville [2], Iserman [13], Lai [21], and Willsky [26], and the book by Basseville and Nikiforov [3]. As is pointed out in [2], the basic method proposed in most of the literature on change detection consists of two steps. First, the problem is transformed into a standard problem by generating certain *residuals*: change indicating signals which are ideally close to zero when no change occurs. Then, in a separate second step, sophisticated statistical methods are developed to solve the resulting detection problem in terms of these residuals. In this paper we will provide a contribution to the second step; the first step will very much depend on the particular application that one wishes to consider, and it is therefore not treated here.

The statistical tools used in the second step usually originate in the field of sequential hypothesis testing, and a wide variety of results concerning their use in change detection problems is now available [21]. Typically these tests compare a certain functional of the observations with a threshold, and an alarm is raised as soon as this threshold is reached. Important examples of such schemes include the celebrated

*Received by the editors July 31, 2000; accepted for publication (in revised form) October 26, 2002; published electronically May 12, 2003. A preliminary short version of this paper appeared as *Changepoint detection using nonlinear filters*, in Proceedings of the 4th European Control Conference, Brussels, 1997.

http://www.siam.org/journals/sicon/42-2/37595.html

†Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE, Enschede, The Netherlands (m.h.vellekoop@math.utwente.nl).

‡Centre for Process Systems Engineering, Imperial College, Exhibition Road, London SW7 2BT, United Kingdom (j.m.c.clark@ic.ac.uk).

CUSUM and generalized likelihood ratio (GLR) schemes.

In this paper we want to propose a different approach, in which change detection is considered to be an on-line estimation problem in which a dynamical system possesses certain parameters which may exhibit sudden changes that need to be detected [8]. In our Bayesian formulation of the problem we assume that *both* the time *and* the size of the change are unknown a priori, thus acknowledging the fact that in many practical situations the behavior of the residual after the change is not completely known and detection is thus necessarily linked to estimation. In practical detection problems, one does not only want to know *that* a change has occurred; one also wants to obtain on-line estimates of relevant statistics *after* the change.

We do not consider the problem in which one tries to detect changes off-line, or where one tries to estimate the *time* of the change. GLR methods and maximum likelihood estimators have been defined for such problems; see, for example, the analysis in [20]. In the recent paper [6] results concerning such off-line methods are derived under the asumption that one does not exactly know the correct model after the change (although the assumed model should be "close" to the correct model in a predescribed sense). Those results on the off-line detection problem are in that sense complementary to the methods we will propose here, but since their goal and assumptions differ from ours, we redirect the reader to the reference given above for further information.

In a continuous-time framework, we can define a basic change detection problem concerning a simple jump process, which is equal to zero up to a certain random time τ, then jumps to a random value X, after which it stays constant again. We assume that such a signal can be observed in white noise, and the purpose is to study the conditional distribution of the signal given the σ-algebra generated by all the observations up to the current time t and relevant statistics generated by this conditional distribution. If the value after the change X is known a priori, one can find an explicit stochastic differential equation for the Bayesian a posteriori probability that a change has occurred—the celebrated Shiryayev–Wonham equation [22, 28]. In fact, this statistic can still be calculated recursively if X is known to belong to a finite set of possible values. The problem can then be solved using the theory of hidden Markov models [10].

However, if there exists an infinity of possible jump sizes X, then the problem becomes much harder, since the detection and estimation problem now become closely interrelated. The problem can be addressed using the nonlinear filtering theory for discontinuous stochastic processes, and the optimal nonlinear filter for this case has been derived in [11]. As is often the case in nonlinear filtering problems [7], this filter does not admit a finite-dimensional recursive implementation. However, since the conditional probability distribution of the process based on the noisy observations can be derived explicitly, this may be used as a starting point for approximations which are suboptimal yet can be implemented recursively.

In this paper we will formulate and study such an approximation to the optimal filter for processes containing a jump of unknown size. Our approach extends a powerful statistical projection technique, which was recently introduced by Brigo, Hanzon, and LeGland in order to filter nonlinear diffusions [4, 5, 12], and which is based on differential-geometric methods in statistical information theory [1, 17, 18]. We will show that the resulting filter can be parametrized as a modified Kalman filter which feeds an adaptive version of the Shiryayev–Wonham filter for *known* changes that we mentioned earlier. This interpretation may help to explain its excellent performance when compared to other detection schemes, as will be illustrated in a number of

simulation studies.

The structure of this paper is as follows. In the next section we introduce the stochastic change detection model, and derive the nonlinear filter equations for such models. In sections 3 and 4 we formulate two recursive filtering algorithms, which are based on information-theoretic approximations and approximation of conditional moments, respectively. In section 5 we discuss the relationship between these two filters. In sections 6 and 7 we introduce and analyze a three-dimensional nonlinear filter based on the results derived in earlier sections, and we illustrate the performance of this filter in some simulation studies in section 8. We finish with conclusions and suggestions for further research in the last section.

2. The change detection model and optimal filter equations. Let $(\Omega, \mathcal{F}, P)$ be the complete canonical probability space for Brownian motion, that is, $\Omega = C([0,\infty[)$, the set of all scalar continuous functions on $\mathbb{R}^+$, $\mathcal{F}$ the usual σ-algebra generated by the topology of uniform convergence on compacta, and P the Wiener measure on $\mathcal{F}$. Let $\{\mathcal{F}_t, t \geq 0\}$ be a filtration satisfying the usual conditions, i.e., an increasing family of σ-algebras which is right-continuous and such that $\mathcal{F}_0$ contains all P-null sets. We will use $\mathbb{P}(A)$ as a shorter notation for $P(\{\omega \in \Omega : A(\omega)\})$ in this paper, where $A(\omega)$ is a condition on ω, and we will denote the expectation operator by $\mathbb{E}$, so for a stochastic variable Z we use $\mathbb{E}Z$ to denote $\int_\Omega Z(\omega) dP(\omega)$.

Consider the signal

$$S_t = \begin{cases} 0, & 0 \leq t < \tau, \\ X, & t \geq \tau, \end{cases} \tag{2.1}$$

where $X \in \mathbb{R}$ and $\tau \in \mathbb{R}^+$ are two independent finite random variables on Ω with distribution functions F, G, respectively. We will assume that X and τ have probability densities f and g, so $\mathbb{P}(X \leq x) = F(x) = \int_{-\infty}^{x} f(u)du$ for all $x \in \mathbb{R}$ and $\mathbb{P}(\tau \leq r) = G(r) = \int_0^r g(u)du$ for all $r \in \mathbb{R}^+$. We assume that f and g are both strictly positive on their domains $\mathbb{R}$ and $\mathbb{R}^+$ in this paper, unless we explicitly state otherwise. We will use $E_t = \mathbf{1}_{\{t \geq \tau\}}$ to denote a unit jump process, so $S_t = XE_t$ for all $t \geq 0$.

We will suppose that the signal S_t can be observed in additive white noise. We therefore define a scalar observation process $\{Y_t^\epsilon, t \geq 0\}$ by

$$dY_t^\epsilon \;=\; S_t\, dt \;+\; \epsilon\, dW_t, \qquad\qquad Y_0^\epsilon = 0, \tag{2.2}$$

where $\{(W_t, \mathcal{F}_t), t \geq 0\}$ is a standard Brownian motion process on $(\Omega, \mathcal{F}, P)$, which is independent of both X and τ, and where ϵ is a real positive parameter representing the noise intensity.

Let $\mathcal{S}_t$ be a second filtration which is contained in $\mathcal{F}_t$ and satisfies the usual conditions as well, such that both X and $\mathbf{1}_{\{t \geq \tau\}}$ are $\mathcal{S}_t$-measurable for all $t \geq 0$, i.e., X is $\mathcal{S}_0$-measurable and τ is a stopping time with respect to $\mathcal{S}_t$. The σ-algebra $\mathcal{S}_t$ then represents the *state information* up to time t. Likewise, we define $\mathcal{Y}_t^\epsilon$ as the σ-algebra generated by the observation process up to time t:

$$\mathcal{Y}_t^\epsilon \;\stackrel{\text{def}}{=}\; \sigma(\{Y_s^\epsilon, 0 \leq s \leq t\}) \;\subset\; \mathcal{F}_t.$$

We are interested in the analysis of the conditional laws of the signal S_t, given the observations record up to time t. In particular, we would like to estimate the magnitude of the jump X at time t and the probability that the jump has already occurred before time t:

$$\mathbb{E}[X \mid \mathcal{Y}_t^\epsilon], \qquad \mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon).$$

Since such statistics can be calculated from the conditional distribution of S_t given the observation record $\mathcal{Y}_t^\epsilon$, we will study this conditional law of the signal on a fixed finite time interval $[0, T]$. Our results concerning the postjump time period are therefore conditioned on the set $\{\omega \in \Omega : \tau(\omega) \leq T\}$.

We derive an expression for the conditional distribution of the signal through the Kallianpur–Striebel formula and Girsanov's theorem. One may show that the necessary conditions for these methods to be applicable are indeed satisfied [14] if there exists a $\delta > 0$ such that

$$\mathbb{E}\ \exp[\,\delta X^2\,]\ <\ \infty, \tag{2.3}$$

and we will assume this condition to be satisfied in the rest of the paper. We then find for the conditional distribution of S_t given the observations [14, 27]:

$$\mathbb{P}(S_t \in B \mid \mathcal{Y}_t^\epsilon) = (\rho_t^\epsilon)^{-1} \int_{\mathbb{R}} \int_0^\infty \mathbf{1}_B(x\mathbf{1}_{\{t \geq r\}})\ e^{\frac{Z(x,r,t)}{\epsilon^2}}\, dG(r)dF(x), \tag{2.4}$$

where B is a Borel-measurable set, $\mathbf{1}_B$ is the indicator function for the set B, ρ_t^ϵ is a normalization factor which is equal to the double integral of the right-hand side of this expression for $B = \mathbb{R}$, and

$$\begin{aligned} Z(x,r,t) &= \int_0^t x\,\mathbf{1}_{\{s \geq r\}}\, dY_s^\epsilon\ -\ \frac{1}{2}\int_0^t (x\,\mathbf{1}_{\{s \geq r\}})^2\, ds \\ &= \left[\, x\,(Y_t^\epsilon - Y_r^\epsilon)\ -\ \frac{x^2}{2}\,(t-r)\,\right] \mathbf{1}_{\{r \leq t\}}. \end{aligned}$$

After decomposing the inner integral in (2.4) into the intervals $[0, t[$ and $[t, \infty[$ we find

$$\begin{aligned} &\mathbb{P}(S_t \in B \mid \mathcal{Y}_t^\epsilon) \\ &=\ (\rho_t^\epsilon)^{-1} \int_B \int_0^t\ e^{\frac{Z(x,r,t)}{\epsilon^2}}\, dG(r)dF(x)\ +\ (\rho_t^\epsilon)^{-1}\ (1 - G(t)) \int_B \delta_0(x)\, dx. \end{aligned}$$

Here and in what follows we will allow the slight abuse of notation which represents the Dirac measure with its unit mass in the origin as an integral over a Dirac density $\delta_0(x)$, i.e., $\int_B \delta_0(x)dx = \mathbf{1}_{\{0 \in B\}}$. For $B = \mathbb{R}$ we obtain an expression for the normalization factor:

$$\rho_t^\epsilon = \int_{\mathbb{R}} \int_0^t\ e^{\frac{Z(x,r,t)}{\epsilon^2}}\, dG(r)dF(x)\ +\ (1 - G(t)).$$

We summarize the derived results in the following theorem [11].

THEOREM 2.1. *Under the assumptions mentioned above, the conditional probability density of the signal S_t, given the observations $\{Y_s^\epsilon, 0 \leq s \leq t\}$, is given by*

$$(\rho_t^\epsilon)^{-1}\,[\ (1 - G(t))\,\delta_0(x)\ +\ q_t^\epsilon(x)\ ], \tag{2.5}$$

where

$$q_t^\epsilon(x) = f(x) \int_0^t g(r)\ \exp\left[\frac{x}{\epsilon^2}\,(Y_t^\epsilon - Y_r^\epsilon) - \frac{x^2}{2\epsilon^2}\,(t-r)\right] dr,$$

$$\rho_t^\epsilon = 1 - G(t) + \int_{\mathbb{R}} q_t^\epsilon(x)\,dx.$$

Using Îto's differentiation rule, one may easily check that the density $q_t^\epsilon(x)$ satisfies the following Îto stochastic differential equation,

$$dq_t^\epsilon(x) \;=\; f(x)g(t)\,dt \;+\; \frac{x}{\epsilon^2} q_t^\epsilon(x)\,dY_t^\epsilon\,, \tag{2.6}$$

with initial value $q_0^\epsilon(x) = 0$ for all $x \in \mathbb{R}$. This is the Duncan–Mortensen–Zakai equation of nonlinear filtering for the conditional distribution outside the origin, and it may be derived directly using the infinitesimal generator of the Markov process $\{S_t,\, t \geq 0\}$. The Duncan–Mortensen–Zakai equation suggests that to calculate the optimal filter estimates we have to solve a stochastic partial differential equation online. It can indeed be shown that no finite-dimensional sufficient statistic exists for this problem. Since we need such a finite-dimensional statistic, which can be updated online in a recursive manner for practical implementation, we will propose and analyze finite-dimensional approximations to the infinite-dimensional optimal filter objects in the following sections.

3. Differential-geometric approximations. The first finite-dimensional approximation that we wish to consider uses projection operators in a space of unnormalized probability densities to map the infinite-dimensional optimal filtering objects onto fixed finite-dimensional structures. The appropriate framework for this approximation method is given by the differential-geometrical theory of statistical information and in particular the theory of statistical manifolds. For an excellent introduction to these relatively new fields, the reader is referred to the book by Amari [1] for the general theory and to the papers by Kulhavý [17, 18, 19] for its application to parameter estimation problems. Most important for the approach we wish to take here is the recent application of differential-geometric methods to the filtering problem for nonlinear diffusions [4, 5, 12]. Our analysis forms an extension of the work reported there, and we have therefore tried to keep our notation consistent with these papers whenever possible.

The main idea of our approach will be that we define a finite-dimensional statistical manifold $\mathcal{H}^{1/2}$ in the infinite-dimensional space of unnormalized probability densities. A basis will be derived for the tangent space in every point of this manifold, and we can use these to define a local projection operator which maps the infinitesimal increments generated by the nonlinear filtering equations onto such tangent spaces. The resulting stochastic vector field on $\mathcal{H}^{1/2}$ then defines our nonlinear filter.

In order to use a Hilbert space structure, we will work in the space of square roots of unnormalized probability densities. Let $\mathcal{M}$ be the set of all (not necessarily normalized) finite nonnegative measures κ on $\mathbb{R}$ which are absolutely continuous with respect to Lebesgue measure and have Radon–Nikodym derivatives p which are strictly positive Lebesgue-almost everywhere. Then we have that the function $\sqrt{p}:\ x \mapsto \sqrt{p(x)}$ is an element of $\mathcal{L}^2$, the Hilbert space of Lebesgue-square integrable functions from $\mathbb{R}$ to $\mathbb{R}^+ \setminus \{0\}$. Denote the subspace of $\mathcal{L}^2$ consisting of such square roots of strictly positive densities by $\mathcal{R}$. We define on it a metric $d_{\mathcal{R}}$ induced by the norm

$\|\cdot\|_{\mathcal{L}^2}$, which in turn defines the *Hellinger metric* $d_{\mathcal{M}}$ on the set of measures κ we started with:

$$d_{\mathcal{M}}(\kappa_1,\kappa_2) \;=\; d_{\mathcal{R}}(\sqrt{p_1},\sqrt{p_2}) = \|\sqrt{p_1}-\sqrt{p_2}\|_{\mathcal{L}^2} = \sqrt{\int_{\mathbb{R}} (\,\sqrt{p_1(x)}-\sqrt{p_2(x)}\,)^2\,dx}.$$

To find a recursive approximation for the infinite-dimensional optimal filter, we have to define finite-dimensional structures in the infinite-dimensional space $\mathcal{R}$. We will therefore consider an $m+1$-dimensional manifold N (with $m\in\mathbb{N}$), which as a subset of $\mathcal{R}$ is imbedded in the larger Hilbert space $\mathcal{L}^2$. This means that N is locally homeomorphic to $\mathbb{R}^{m+1}$ and is thus described locally by a *chart*: if $\sqrt{p}\in N$, then there exists an open neighborhood $\mathcal{H}^{1/2}$ of $\sqrt{p}$ in N and a homeomorphism $\varphi:\mathcal{H}^{1/2}\to\Theta$ onto an open and convex subset Θ of $\mathbb{R}^{m+1}$. We will assume that there exists in fact one global and smooth coordinate chart for the entire manifold, so we will consider manifolds $\mathcal{H}^{1/2}$ defined by

$$\mathcal{H}^{1/2} \;=\; \{\sqrt{p(\cdot,\theta)}:\ \theta=(\theta_0,\theta_1,\dots,\theta_m)\in\Theta\} \;=\; \varphi^{-1}(\Theta),$$

where

$$\left\{\frac{\partial\varphi^{-1}(\theta)}{\partial\theta_0},\frac{\partial\varphi^{-1}(\theta)}{\partial\theta_1},\dots,\frac{\partial\varphi^{-1}(\theta)}{\partial\theta_m}\right\}$$

is assumed to be a set of linearly independent vectors in $\mathcal{L}^2$ for all $\theta\in\Theta$. To find the differential-geometric structure of such manifolds $\mathcal{H}^{1/2}$ around a point $\sqrt{p}\in\mathcal{H}^{1/2}$, we consider smooth maps $\alpha:\,]-\nu,\nu[\,\to\mathcal{H}^{1/2}$ ($\nu>0$) such that $\alpha(0)=\sqrt{p}$. The Fréchet derivative of α in zero $D\alpha(0)$, defined by

$$\lim_{t\to0}\frac{\|\alpha(t)-\alpha(0)-D\alpha(0)\cdot t\|_{\mathcal{L}^2}}{t} \;=\; 0,$$

can be interpreted as a tangent vector to the curve α on the manifold $\mathcal{H}^{1/2}$. We therefore define the tangent vector space $\mathcal{T}_{\sqrt{p}}\mathcal{H}^{1/2}$ in $\sqrt{p}$ to $\mathcal{H}^{1/2}$ as the set of *all* possible Fréchet derivatives $D\alpha(0)$ for all such maps α:

$$\mathcal{T}_{\sqrt{p}}\mathcal{H}^{1/2} \;=\; \{\,D\alpha(0):\ \alpha \text{ smooth map }]-\nu,\nu[\,\to\mathcal{H}^{1/2} \text{ with } \alpha(0)=\sqrt{p}\,\}.$$

This is a linear subspace of $\mathcal{L}^2$, which we may calculate more explicitly. Let $\alpha=\varphi^{-1}\circ\bar\alpha$, where $t\to\bar\alpha(t)$ is a smooth map from $]-\nu,\nu[$ to Θ with $\bar\alpha(0)=\theta$ for a fixed $\theta\in\Theta$. Then we may apply the chain rule to $\alpha:t\to\sqrt{p(\cdot,\bar\alpha(t))}$ to find

$$\begin{aligned} D\alpha(0) = D\sqrt{p(\cdot,\bar\alpha(t))}\,\Big|_{t=0} \;&=\; \sum_{k=0}^{m}\frac{\partial\sqrt{p(\cdot,\theta)}}{\partial\theta_k}\,\bar\alpha_k'(0) \\ &= \sum_{k=0}^{m}\frac{1}{2\sqrt{p(\cdot,\theta)}}\,\frac{\partial p(\cdot,\theta)}{\partial\theta_k}\,\bar\alpha_k'(0), \end{aligned}$$

which shows that

$$\mathcal{T}_{\sqrt{p(\cdot,\theta)}}\mathcal{H}^{1/2} \;=\; \operatorname{span}\bigcup_{k=0}^{m}\{B_k(\cdot,\theta)\},\qquad B_k(\cdot,\theta) \;=\; \frac{1}{2\sqrt{p(\cdot,\theta)}}\frac{\partial p(\cdot,\theta)}{\partial\theta_k}. \tag{3.1}$$

The functions $B_k(\cdot,\theta)$ are linearly independent since φ was assumed to be a chart, so they form a basis for the $m+1$-dimensional tangent space in the point $\sqrt{p(\cdot,\theta)}$ on the manifold. The inner products of the basis elements in $\mathcal{L}^2$ generate a matrix function $H(\theta)$:

$$\langle B_i(\cdot,\theta)\,,\, B_j(\cdot,\theta)\rangle_{\mathcal{L}^2} \;=\; \int_{\mathbb{R}} \frac{1}{4p(x,\theta)} \frac{\partial p(x,\theta)}{\partial \theta_i} \frac{\partial p(x,\theta)}{\partial \theta_j}\, dx \;\stackrel{\text{def}}{=}\; \tfrac{1}{4}\, H_{ij}(\theta).$$

In all points of the manifold $\sqrt{p(\cdot,\theta)} \in \mathcal{H}^{1/2}$ where this matrix is invertible, we can define an orthogonal projection operator Π_θ which maps linear subspaces of $\mathcal{L}^2$ containing the finite-dimensional tangent vector space (3.1) onto this tangent vector space, using the formula

$$v \;\stackrel{\Pi_\theta}{\longmapsto}\; \sum_{i=0}^{m} \left[\sum_{j=0}^{m} 4\,[H(\theta)]^{-1}_{ij}\, \langle v,\, B_j(\cdot,\theta)\rangle_{\mathcal{L}^2} \right] B_i(\cdot,\theta). \tag{3.2}$$

In this paper we will use a special class of parametrized families of densities, the finite-dimensional *unnormalized exponential families*. An unnormalized exponential family is given by

$$\mathcal{H}^{1/2} = \{\sqrt{p(\cdot,\theta)},\ \theta \in \Theta\}, \qquad\qquad p(x,\theta) \;=\; f(x)\, \exp\left[\sum_{k=0}^{m} \theta_k c_k(x) \right], \tag{3.3}$$

where m is a strictly positive integer, $\{c_0,\ldots,c_m\}$ is a set of linearly independent scalar functions on $\mathbb{R}$, and f is the probability density of the jump size X, as introduced in the previous section. The parameter vector $\theta = (\theta_0,\theta_1,\ldots,\theta_m)$ is restricted to lie in the parameter set Θ, which is an open nonempty convex subset of $\mathbb{R}^{m+1}$ satisfying

$$\Theta \subseteq \Theta_0 \;\stackrel{\text{def}}{=}\; \left\{ \theta \in \mathbb{R}^{m+1} \,:\, \int_{\mathbb{R}} f(x)\, \exp\left[\sum_{k=0}^{m} \theta_k c_k(x) \right] dx \,<\, \infty \right\}.$$

Throughout this paper we will use the manifold generated by $c_k(x) = x^k$ for $k = 0,1,\ldots,m$, with m an even strictly positive integer, and $\Theta = \{\theta \in \mathbb{R}^{m+1},\, \theta_m < 0\}$. On such manifolds, the differential-geometric structure turns out to be a particularly transparent one. The basis vectors of the tangent space in $\sqrt{p(\cdot,\theta)}$ are given by

$$B_k(x,\theta) \;=\; \frac{1}{2\sqrt{p(x,\theta)}} \frac{\partial p(x,\theta)}{\partial \theta_k} \;=\; \tfrac{1}{2} x^k \sqrt{p(x,\theta)} \tag{3.4}$$

for $k = 0,1,\ldots,m$, and if we define

$$\eta^k(\theta) \;\stackrel{\text{def}}{=}\; \int_{\mathbb{R}} c_k(x) p(x,\theta)\, dx \;=\; \int_{\mathbb{R}} x^k p(x,\theta)\, dx,$$

we find that the earlier defined inner product matrix $H(\theta)$ for the basis elements of the tangent space in a point $\sqrt{p(\cdot,\theta)}$ on the manifold is equal to

$$H_{ij}(\theta) = 4\, \langle B_i(\cdot,\theta), B_j(\cdot,\theta)\rangle_{\mathcal{L}^2} \;=\; \eta^{i+j}(\theta). \tag{3.5}$$

The matrix $H(\theta)$ will be differentiable with respect to θ for all $\theta \in \Theta$ if all finite order moments of the jump size X exist, since

$$\frac{\partial \eta^i}{\partial \theta_j} \;=\; \int_{\mathbb{R}} x^i \,\frac{\partial p(x,\theta)}{\partial \theta_j}\, dx \;=\; \int_{\mathbb{R}} x^{i+j}\, p(x,\theta)\, dx \;=\; \eta^{i+j}(\theta) \;=\; H_{ij}(\theta). \tag{3.6}$$

For $\theta \in \Theta$ the matrix $H(\theta)$ will also be invertible, because if $H(\theta)y = 0$ for some vector $y \in \mathbb{R}^{m+1}$, then

$$\begin{aligned} 0 = \sum_{i=0}^{m}\sum_{j=0}^{m} y_i H_{ij}(\theta) y_j \;&=\; \sum_{i=0}^{m}\sum_{j=0}^{m} \int_{\mathbb{R}} y_i x^{i+j} y_j p(x,\theta)\, dx \\ &= \int_{\mathbb{R}} \left(\sum_{i=0}^{m} y_i\, x^i\right)^2 p(x,\theta)\, dx, \end{aligned}$$

which implies that y is the zero vector in $\mathbb{R}^{m+1}$ since p is strictly positive Lebesgue-almost everywhere. We remark that the matrix $H(\theta)$ coincides with the Fisher information matrix for our class of problems, since we can write it as

$$H_{ij}(\theta) \;=\; \int_{\mathbb{R}} \frac{\partial \ln p(x,\theta)}{\partial \theta_i}\, \frac{\partial \ln p(x,\theta)}{\partial \theta_j}\, p(x,\theta)\, dx.$$

The most important structural property is (3.6). It is exploited repeatedly in [4, 5], and it will play a central role in our analysis as well. A density from the exponential family may be characterized in terms of the θ-coordinate system, or the η-coordinate system, and on Θ the two are related by a diffeomorphism $\eta = \eta(\theta)$, which has the Fisher information matrix H as its Jacobian. In terms of Amari [1], the pair (θ, η) forms a *dual coordinate system*. However, our particular choice for this exponential family is not just motivated by this important property but also by other information theoretic considerations, since it may be shown that it is in fact the class of densities which maximize the *entropy* of a density with respect to Lebesgue measure once its $m+1$ moments $\{\eta_0, \ldots \eta_m\}$ have been specified.

The difference between our problem and the nonlinear filtering problem for diffusions treated in [5] lies mainly in the fact that our state equation does not evolve smoothly (in fact, not even continuously) and that its evolution depends on two stochastic variables (the jump size X and the jump time τ). We have seen in the previous section that the conditional distribution of the signal $\{S_t,\, t \geq 0\}$ consists of a Dirac measure in the origin and a smooth density outside the origin, and for reasons which will become clear later on we do not want to project that part of the conditional distribution which is represented by the Dirac measure. It will therefore be more convenient to apply the projection method to the Duncan–Mortensen–Zakai equation (2.6) for the absolutely continuous part of the density $q_t^\epsilon(x)$ which we defined in Theorem 2.1:

$$dq_t(x) \;=\; f(x)g(t)\, dt \;+\; \frac{x}{\epsilon^2}\, q_t(x)\, dY_t^\epsilon, \tag{3.7}$$

with initial condition $q_0(x) = 0$ for all $x \in \mathbb{R}$. Note that we will suppress the ϵ-dependency of this conditional density in our notation from now on.

Our definition of the exponential family also differs from the manifolds used for diffusion processes in the sense that the densities in our manifold are *not normalized*. In fact the differential-geometric structure takes the form of a cone: all scalar multiples

of a certain density on the manifold also lie on the manifold because of the introduction of the extra parameter θ_0. This is important, since the Duncan–Mortensen–Zakai equation (3.7) provides an unnormalized version of the conditional density outside the origin, and the normalization constant turns out to have a particular significance in our case. Indeed,

$$\mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon) = \mathbb{P}(S_t \neq 0 \mid \mathcal{Y}_t^\epsilon) \;=\; \frac{\int_{\mathbb{R}} q_t(x)\,dx}{\int_{\mathbb{R}} q_t(x)\,dx + 1 - G(t)}, \tag{3.8}$$

so the normalization constant is linked to the probability that a jump has occurred, and estimation of its value using the parameter θ_0 will thus be essential.

Note that alternatively we could have directly defined a projection filter without these modifications, when using projections on measures consisting of convex combinations of the Dirac delta measure and members of the family of exponential distributions

$$\tilde{p}(dx,\theta) = \gamma\,\delta_0(dx) \;+\; (1-\gamma)\frac{f(x)e^{\theta_1 c_1(x)+\cdots+\theta_k c_k(x)}}{\int_{\mathbb{R}} f(u)e^{\theta_1 c_1(u)+\cdots+\theta_k c_k(u)}du}\,dx,$$

where the parameter γ replaces the old parameter θ_0:

$$\gamma = \frac{1-G(t)}{1-G(t)+e^{\theta_0}\int_{\mathbb{R}} f(u)e^{\theta_1 c_1(u)+\cdots+\theta_k c_k(u)}du} \;\in\; [0,1].$$

However, our present formulation involves only measures which are absolutely continuous with respect to Lebesgue measures, and since this allows us to work directly with density functions, it is slightly more convenient.

To simplify the calculations on our statistical manifold we will work with the Stratonovich form of the Duncan–Mortensen–Zakai equation:

$$dq_t(x) \;=\; \left[f(x)g(t) - \frac{x^2}{2\epsilon^2}q_t(x)\right]\,dt \;+\; \frac{x}{\epsilon^2}q_t(x) \circ dY_t^\epsilon,$$

and the differential equation for $\sqrt{q_t} \in \mathcal{L}^2$ thus becomes

$$d\sqrt{q_t(x)} \;=\; \left[\frac{f(x)g(t)}{2\sqrt{q_t(x)}} - \frac{x^2}{4\epsilon^2}\sqrt{q_t(x)}\right]\,dt \;+\; \frac{x}{2\epsilon^2}\sqrt{q_t(x)} \circ dY_t^\epsilon.$$

To simplify notation we rewrite this as

$$d\sqrt{q_t} \;=\; \mathcal{P}_1(\sqrt{q_t})\,dt \;+\; \mathcal{P}_2(\sqrt{q_t}) \circ dY_t^\epsilon,$$

with the nonlinear operators $\mathcal{P}_i$ $(i = 1, 2)$ on $\mathcal{L}^2$ defined in an obvious way. To make sure that these operators do indeed map back into $\mathcal{L}^2$ when we apply them to our approximate densities $p(\cdot,\theta)$, we need the following condition:

For all $\theta \in \Theta$ we have that

$$\text{(A)} \qquad \int_{\mathbb{R}} x^4\,p(x,\theta)\,dx \;<\; \infty \qquad \text{and} \qquad \int_{\mathbb{R}} \frac{f(x)^2}{p(x,\theta)}\,dx \;<\; \infty.$$

The first part of this condition is rather mild, and the second part will be satisfied if the tails of the density f vanish rapidly enough. We will see that both parts of

condition (A) are *not necessary* to formulate our approximate filter, but they are needed if one wants to interpret the filter as the result of a projection in $\mathcal{L}^2$.

The operators $\Pi_\theta \circ \mathcal{P}_i$, with Π_θ as defined in (3.2), now generate a stochastic vector field on the manifold $\mathcal{H}^{1/2}$:

$$d\sqrt{p(\cdot,\theta_t)} = \left[\Pi_{\theta_t} \circ \mathcal{P}_1(\sqrt{p(\cdot,\theta_t)})\right] dt \; + \; \left[\Pi_{\theta_t} \circ \mathcal{P}_2(\sqrt{p(\cdot,\theta_t)})\right] \circ dY_t^\epsilon. \tag{3.9}$$

Note that we will always use the notation q for the real unnormalized conditional density outside the origin, and p for its projection.

Our aim is now to describe the evolution of the density in terms of our parameter vector θ_t, i.e., we want to find a stochastic differential equation for the result of the inverse mapping from the trajectory of projected densities on the manifold $\mathcal{H}^{1/2}$ into our parameter set $\Theta \subseteq \mathbb{R}^{m+1}$. It turns out that we can easily extend the analysis that was carried out in [4] for diffusion processes.

THEOREM 3.1. *Let the conditions of the previous section and condition* (A) *be satisfied. Then the parameter vector θ_t describing the filter* (3.9) *on the manifold generated by the exponential family* (3.3) *with $c_k(x) = x^k$ satisfies the Stratonovich stochastic differential equation*

$$d\theta_t = g(t)\; [H(\theta_t)]^{-1} \begin{pmatrix} 1 \\ \mathbb{E}X \\ \mathbb{E}X^2 \\ \vdots \\ \mathbb{E}X^m \end{pmatrix} dt \; - \; \frac{1}{2\epsilon^2} \begin{pmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} dt \; + \; \frac{1}{\epsilon^2} \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \circ dY_t^\epsilon, \tag{3.10}$$

with the matrix function H defined as before by

$$H_{ij}(\theta_t) \;=\; \int_{\mathbb{R}} x^{i+j} p(x,\theta_t)\, dx \;=\; \theta_{i+j}.$$

This stochastic differential equation has a unique solution up to the (possibly infinite) almost surely strictly positive exit time $\inf\{t \geq 0 : \theta_t \notin \Theta\}$.

Proof of Theorem 3.1. We deal with the two terms in (3.9) separately. For the first one we find, using (3.2),

$$\begin{aligned}
&\Pi_{\theta_t} \circ \mathcal{P}_1(\sqrt{p(\cdot,\theta_t)}) \\
&= \sum_{i=0}^m \sum_{j=0}^m 4\,[\,H(\theta_t)\,]^{-1}_{ij} \left[\int_{\mathbb{R}} \mathcal{P}_1(\sqrt{p})(x)\, B_j(x,\theta_t)\, dx\right] B_i(\cdot,\theta_t) \\
&= \sum_{i=0}^m \sum_{j=0}^m 4\,[H(\theta_t)]^{-1}_{ij} \left[\int_{\mathbb{R}} \left(\frac{f(x)g(t)}{2\sqrt{p(x,\theta_t)}} - \frac{x^2}{4\epsilon^2}\sqrt{p(x,\theta_t)}\right) \tfrac{1}{2} x^j \sqrt{p(x,\theta_t)}\, dx\right] B_i(\cdot,\theta_t) \\
&= \sum_{i=0}^m \sum_{j=0}^m [H(\theta_t)]^{-1}_{ij} \left[g(t)\mathbb{E}X^j - \frac{\eta_t^{j+2}}{2\epsilon^2}\right] B_i(\cdot,\theta_t).
\end{aligned}$$

Analogously, the second vector field can be shown to satisfy

$$\Pi_{\theta_t} \circ \mathcal{P}_2(\sqrt{p(\cdot,\theta_t)}) \;=\; \sum_{i=0}^m \sum_{j=0}^m [H(\theta_t)]^{-1}_{ij} \left[\frac{\eta_t^{j+1}}{\epsilon^2}\right] B_i(\cdot,\theta_t).$$

But since

$$d\sqrt{p(\cdot,\theta_t)} \;=\; \sum_{i=0}^{m}\left[\frac{1}{2\sqrt{p(\cdot,\theta)}}\,\frac{\partial p(\cdot,\theta)}{\partial\theta_i}\bigg|_{\theta=\theta_t}\circ\, d(\theta_t)_i\right] \;=\; \sum_{i=0}^{m} B_i(\cdot,\theta_t)\circ d(\theta_t)_i,$$

equating the coefficients in front of the basis vectors $B_i(\cdot,\theta_t)$ of the tangent space in $\sqrt{p(\cdot,\theta_t)}$ then gives that

$$\begin{aligned}(d\theta_t)_i &= g(t)\sum_{j=0}^{m}[H(\theta_t)]_{ij}^{-1}\,\mathbb{E}X^j\,dt \;-\; \frac{1}{2\epsilon^2}\sum_{j=0}^{m}[H(\theta_t)]_{ij}^{-1}\,\eta_t^{j+2}\,dt\\ &\quad+\frac{1}{\epsilon^2}\sum_{j=0}^{m}[H(\theta_t)]_{ij}^{-1}\,\eta_t^{j+1}\circ dY_t^{\epsilon}\\ &= g(t)\sum_{j=0}^{m}[H(\theta_t)]_{ij}^{-1}\,\mathbb{E}X^j\,dt \;-\; \frac{1}{2\epsilon^2}\mathbf{1}_{\{i=2\}}\,dt \;+\; \frac{1}{\epsilon^2}\,\mathbf{1}_{\{i=1\}}\circ dY_t^{\epsilon},\end{aligned}$$

because of (3.5). Existence and uniqueness of a solution of this equation up to the almost surely positive exit time $\inf\{t>0:\theta_t\notin\Theta\}$ is guaranteed since we showed before that $[H(\theta)]^{-1}$ exists for all $\theta\in\Theta$ and since $H(\theta)$ is infinitely many times differentiable with respect to θ on this set, its inverse certainly satisfies a local Lipschitz condition. □

Some care must be taken when defining the initial conditions for the stochastic differential equation for θ_t. At time $t=0$ the density outside the origin is equal to $q_0(x)=0$ for all x, which would mean that $\theta_0=-\infty$ and that the other values in the θ-vector can be chosen arbitrarily. We can overcome this problem by looking at the moments vector η instead of θ. We have remarked before that on the domain Θ the θ-vectors and η-vectors are related by a diffeomorphism. If we look at a small time $\delta>0$, we see from the Duncan–Mortensen–Zakai equation (3.7) that $q_\delta(x)$ approximately equals $f(x)g(0)\delta$. By (3.6), we have $H(\theta_t)\circ d\theta_t = d\eta_t$, so rewriting (3.10) in terms of moments gives

$$d\eta_t \;=\; g(t)\begin{pmatrix}1\\ \mathbb{E}X\\ \mathbb{E}X^2\\ \vdots\\ \mathbb{E}X^m\end{pmatrix}dt \;-\; \frac{1}{2\epsilon^2}\,[H(\theta_t)]\begin{pmatrix}0\\0\\1\\\vdots\\0\end{pmatrix}dt \;+\; \frac{1}{\epsilon^2}\,[H(\theta_t)]\begin{pmatrix}0\\1\\0\\\vdots\\0\end{pmatrix}\circ dY_t^{\epsilon},$$

so the moments η_δ at time δ are approximately equal to $g(0)\delta$ times the moments of X, as the expression for q_δ confirms. These moments will then uniquely determine the value of the parameter vector θ_δ, which may then be used as the initial condition for the stochastic differential equation for θ_t. Note that this is the only place where our assumption that $g(0)$ be strictly positive is explicitly needed, and if one is prepared to formulate alternative initial conditions for the approximate filter this assumption can be weakened.

Equation (3.10) for the evolution of θ_t has a remarkably simple structure. In particular, since it has a constant diffusion coefficient, the Stratonovich and Îto forms of the stochastic differential equation coincide, and every Euler scheme to find numerical approximations to its solution will coincide with a Milstein scheme, guaranteeing strong convergence of order one [16]. Moreover, it is quite easy to

give a clear interpretation of the stochastic differential equation. Since q_t approximates the conditional density of S_t outside the origin, i.e., the conditional density of X, we can interpret the stochastic differential equation for θ_t as the sum of two separate vector fields. The first one keeps the conditional density of X close to the prior density of X: since $d\eta_t = H(\theta_t) \circ d\theta_t$, the solution of $d\theta_t = g(t)\ [H(\theta_t)]^{-1}\,[1 \quad \mathbb{E}X \quad \dots \quad \mathbb{E}X^m\,]^T\,dt$ would simply be $G(t) = \mathbb{P}(t < \tau)$ times that density on the manifold which has the same first m moments as X. The second vector field $d\theta_t = \ -\frac{1}{2\epsilon^2}\,[0\ 0\ 1\ \dots\ 0]^T\,dt\ +\ \frac{1}{\epsilon^2}\,[0\ 1\ 0\ \dots\ 0]^T \circ dY_t^\epsilon$ describes the evolution of the Kalman filter for a Gaussian distributed random variable X observed in white noise of intensity ϵ^2. Before the jump, $Y_t^\epsilon = \epsilon W_t$ and the influence of the stochastic increment dY_t^ϵ will be small, while after the jump it will become significant due to the nonzero drift in Y_t^ϵ.

The fact that the diffusion coefficient vector in the stochastic differential equation for θ_t is a constant vector is a consequence of our choice of the basis functions $\{c_0(x), \dots, c_m(x)\}$ which generate the exponential family. The diffusion coefficient vector will always be constant if the function j in the observation equation $dY_t^\epsilon = j(S_t)\,dt + \epsilon\,dW_t$ (in our case simply $j(x) = x$) and its square (in our case $j(x)^2 = x^2$) are both in the linear space spanned by the functions $\{c_0(x), \dots, c_m(x)\}$. A proof is given in [4] for nonlinear filtering problems where the signal S_t is a diffusion instead of a jump process, and this result carries over directly to our case.

4. Statistical approximations. In the previous section, the conditional probability distribution of our original signal process $\{S_t,\, t \geq 0\}$ was approximated by a member of a finite-dimensional family of distributions. Another possible approximation to the optimal filter can be found by applying the Kushner–Stratonovich equation of nonlinear filtering. This equation describes the evolution of conditional statistics in time by means of a stochastic differential equation driven by the observation process $\{Y_t^\epsilon, t \geq 0\}$. We will use it to find such stochastic differential equations for the evolution of the moments of our conditional density and then use these equations to define another approximate filter. To do so, we first state the Kushner–Stratonovich equation (for the special case where the state noise and observation noise are independent) and then derive a stochastic differential equation for the process $\{S_t,\, t \geq 0\}$ which makes it possible to apply it to our particular filtering problem.

Let $\{V_t,\, t \in [0,T]\}$ be a scalar stochastic process such that V_0 is $\mathcal{S}_0$-measurable, with $\mathbb{E}|V_0| < \infty$ and

$$dV_t = D_t\,dt\ +\ dM_t, \tag{4.1}$$

$$dY_t^\epsilon = S_t\,dt\ +\ \epsilon\,dW_t. \tag{4.2}$$

We assume that (see section 2 for the definition of the state filtration $\mathcal{S}_t$)

- $\{M_t,\, t \geq 0\}$ is a right-continuous square integrable $\mathcal{S}_t$-martingale with left-hand limits, which is independent of the Wiener process $\{W_t,\, t \geq 0\}$,
- $\{D_t,\, t \geq 0\}$ is an $\mathcal{S}_t$-adapted process with $\mathbb{E}\int_0^T D_u^2\,du < \infty$, and
- $\{V_t,\, t \geq 0\}$ is such that $\mathbb{E}\int_0^T (S_u V_u)^2\,du\ <\ \infty$.

We will use the notation $\widehat{\alpha}_t = \mathbb{E}\,[\alpha_t \mid \mathcal{Y}_t^\epsilon]$ for the conditional expectation of stochastic processes $\{\alpha_t,\, t \geq 0\}$ with respect to the observations σ-algebra $\mathcal{Y}_t^\epsilon$. The Kushner–Stratonovich equation then states that for $t \in [0,T]$ we have [14, 27]

$$d\,\widehat{V}_t = \widehat{D}_t\,dt\ +\ \frac{1}{\epsilon^2}\left(\widehat{S_t V_t} - \widehat{S}_t\,\widehat{V}_t\right)\,d\nu_t^\epsilon, \tag{4.3}$$

with initial condition $\widehat{V}_0 = \mathbb{E}V_0$. The process

$$\nu_t^\epsilon \;=\; Y_t^\epsilon - \int_0^t \widehat{S}_u \, du \tag{4.4}$$

is called the *innovation process*, and under the conditions stated it is a Brownian motion with respect to the observations filtration $\{\mathcal{Y}_t^\epsilon,\, t \geq 0\}$.

In order to be able to apply the Kushner–Stratonovich equation to our problem, we will now derive a description for the signal $\{S_t,\, t \geq 0\}$ of the form (4.1). Let $E_t = \mathbf{1}_{\{t \geq \tau\}}$ denote, as before, the right-continuous $\mathcal{S}_t$-measurable process which jumps from zero to one at time τ. The probability that the jump occurs in the time interval $[t, t+dt]$ given that it has not occurred before time t equals $\lambda(t)dt + o(dt)$, where $\lambda(t)$ is the *hazard rate* at time t, defined by

$$\lambda(t) \;=\; \frac{g(t)}{1 - G(t)}.$$

Define the process M_t as E_t minus the integral of this hazard rate up to time $t \wedge \tau$ (where we introduce the usual notation $a \wedge b$ for the minimum of a and b):

$$M_t \;=\; E_t - K_t, \qquad\qquad K_t \;=\; \int_0^{t\wedge\tau} \lambda(s)\, ds \;=\; -\ln(1 - G(t \wedge \tau)).$$

Tedious but straightforward calculations show that M_t is an $\mathcal{S}_t$-martingale (for details, see, for example, [9]). But since

$$t \wedge \tau \;=\; \int_0^t (1 - E_u)\, du,$$

we have that

$$M_t \;=\; E_t \;+\; \ln[\, 1 - G(\, \textstyle\int_0^t (1 - E_u) du \,)\,],$$

and we thus find the following representation for E_t:

$$\begin{aligned} dE_t &= \lambda(t \wedge \tau)\,(1 - E_t)\, dt \;+\; dM_t \\ &= \lambda(t)\,(1 - E_t)\, dt \;+\; dM_t, \end{aligned} \tag{4.5}$$

where we have used the fact that $\lambda(t \wedge \tau)(1 - E_t) = \lambda(t)(1 - E_t)$, since if $t \wedge \tau = \tau$, then $1 - E_t = 0$. Our original process may now be represented as $S_t = XE_t$ so it satisfies

$$dS_t \;=\; \lambda(t)\,(X - S_t)\, dt \;+\; X\, dM_t, \tag{4.6}$$

and in fact for arbitrary $k \in \mathbb{N} \setminus \{0\}$

$$d(S_t)^k \;=\; \lambda(t)\,(X^k - (S_t)^k)\, dt \;+\; X^k\, dM_t. \tag{4.7}$$

We can now apply the Kushner–Stratonovich equation to this representation of our signal process, but we first prove a lemma that will be used to simplify the equations which it generates.

LEMMA 4.1. *For all $t \in [0, T]$ and $k \in \mathbb{N} \setminus \{0\}$ we have that, almost surely,*

$$\mathbb{E}\,[X^k - (S_t)^k \mid \mathcal{Y}_t^\epsilon] \;=\; (1 - \widehat{E}_t)\, \mathbb{E}X^k. \tag{4.8}$$

Proof of Lemma 4.1. Let B be any set in $\mathcal{Y}_t^\epsilon$. Then by definition

$$\begin{aligned}\int_B \mathbb{E}\,[X^k - (S_t)^k \mid \mathcal{Y}_t^\epsilon](\omega)\,dP(\omega) &= \int_B (\,X^k(\omega) - (S_t)^k(\omega)\,)\,dP(\omega)\\ &= \int_B X^k(\omega)\,\mathbf{1}_{\{t<\tau(\omega)\}}\,dP(\omega)\\ &= \int_{B\cap\{\omega:t<\tau(\omega)\}} X^k(\omega)\,dP(\omega).\end{aligned}$$

But we have that the σ-algebra generated by sets of the form $B\cap\{\omega : t < \tau(\omega)\}$ (with $B \in \mathcal{Y}_t^\epsilon$) is independent of sets in the σ-algebra generated by X^k, since $Y_t^\epsilon = \epsilon W_t$ on $\{\omega : t < \tau(\omega)\}$ and the process $\{W_t,\, t \geq 0\}$ is independent of X, so

$$\begin{aligned}\int_{B\cap\{\omega:t<\tau(\omega)\}} X^k(\omega)\,dP(\omega) &= \int_{B\cap\{\omega:t<\tau(\omega)\}} \mathbb{E}X^k\,dP(\omega)\\ &= \big(\mathbb{E}X^k\big)\int_B \mathbf{1}_{\{t<\tau(\omega)\}}\,dP(\omega)\\ &= \big(\mathbb{E}X^k\big)\int_B \mathbb{E}\,[1-\mathbf{1}_{\{t\geq\tau\}} \mid \mathcal{Y}_t^\epsilon](\omega)\,dP(\omega)\\ &= \big(\mathbb{E}X^k\big)\int_B (1-\widehat{E}_t(\omega))\,dP(\omega),\end{aligned}$$

and we may now conclude that (4.8) holds by the almost sure uniqueness property of conditional expectations. □

THEOREM 4.2. *Let the random variables X and τ and the stochastic processes $\{S_t,\, t \geq 0\}$ and $\{Y_t^\epsilon,\, t \geq 0\}$ be defined as in section* 2, *and let X and τ satisfy all conditions mentioned in that section. Then the optimal filter estimate $\widehat{S}_t = \mathbb{E}\,[S_t \mid \mathcal{Y}_t^\epsilon]$ and higher order moments for $t \in [0,T]$ are generated by the following Îto stochastic differential equations ($k \in \mathbb{N}\setminus\{0\}$):*

$$d\widehat{E}_t = \lambda(t)(1-\widehat{E}_t)\,dt \;+\; \frac{1}{\epsilon^2}\widehat{S}_t(1-\widehat{E}_t)\,d\nu_t^\epsilon, \tag{4.9}$$

$$\begin{aligned}d\,\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon] &= \lambda(t)\,\mathbb{E}X^k(1-\widehat{E}_t)\,dt\\ &\quad+\frac{1}{\epsilon^2}\Big(\,\mathbb{E}\,[(S_t)^{k+1} \mid \mathcal{Y}_t^\epsilon] - \widehat{S}_t\,\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon]\Big)\,d\nu_t^\epsilon,\end{aligned} \tag{4.10}$$

with initial conditions $\widehat{E}_0 = \mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon]_{t=0} = 0$ for all $k \in \mathbb{N}\setminus\{0\}$, and where the innovation process $\{\nu_t^\epsilon,\, t \in [0,T]\}$ is defined by (4.4).

Proof of Theorem 4.2. The conditions for application of the Kushner–Stratonovich equation are obviously satisfied for the process E_t since

$$\mathbb{E}\int_0^T [\lambda(u)(1-E_u)]^2 du \;\leq\; \int_0^T \lambda(u)^2 du < \infty$$

(note that $\lambda(t)$ is finite for all $t \geq 0$ and continuous since we assumed that $g(t)$ is continuous and $G(t) < 1$ for all $t \geq 0$) and $\mathbb{E}\int_0^T |E_u S_u| du < T\cdot\mathbb{E}|X| < \infty$. Here and in the rest of the proof we use the fact that all finite order moments of X exist because of condition (2.3), which implies that $\mathbb{E}|X|^k < \infty$ for all $k \geq 0$.

Since $\{W_t,\, t \geq 0\}$ was assumed to be independent of X and τ, it is independent of $\{M_t,\, t \geq 0\}$. The Kushner–Stratonovich equation applied to (4.5) thus results in

$$\begin{aligned} d\widehat{E}_t &= \lambda(t)(1-\widehat{E}_t)\,dt \;+\; \frac{1}{\epsilon^2}(\,\widehat{S_tE_t} - \widehat{S}_t\widehat{E}_t\,)\,d\nu_t^\epsilon \\ &= \lambda(t)(1-\widehat{E}_t)\,dt \;+\; \frac{1}{\epsilon^2}\widehat{S}_t(1-\widehat{E}_t)\,d\nu_t^\epsilon, \end{aligned} \tag{4.11}$$

where we have used the fact that $S_tE_t = S_t$. The initial condition is $\widehat{E}_0 = \mathbb{E}(E_0) = 0$. To find the conditional moments $\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon]$ for $k \in \mathbb{N} \setminus \{0\}$, we use (4.7). Since X is $\mathcal{S}_t$-measurable and independent of τ, and $\mathbb{E}|X|^{2k} < \infty$, the process $\{X^kM_t,\, t \geq 0\}$ is again a square integrable $\mathcal{S}_t$-martingale which is independent of $\{W_t,\, t \geq 0\}$. The two other conditions for the Kushner–Stratonovich formula are satisfied as well, since

$$\mathbb{E}\int_0^T (X^k - (S_u)^k)^2\lambda(u)^2 du \leq \mathbb{E}|X|^{2k}\int_0^T \lambda(u)^2 du \;<\; \infty,$$

$$\mathbb{E}\int_0^T |(S_u)^k S_u| du \leq T\cdot\mathbb{E}|X|^{k+1} \;<\; \infty.$$

We therefore have that for $k \in \mathbb{N} \setminus \{0\}$

$$\begin{aligned} d\,\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon] &= \lambda(t)\,\mathbb{E}\,[X^k - (S_t)^k \mid \mathcal{Y}_t^\epsilon]\,dt \\ &\quad + \frac{1}{\epsilon^2}\Big(\,\mathbb{E}\,[(S_t)^{k+1} \mid \mathcal{Y}_t^\epsilon] - \widehat{S}_t\,\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon]\Big)\,d\nu_t^\epsilon, \end{aligned} \tag{4.12}$$

with initial condition $\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon]_{t=0} = \mathbb{E}[(S_t)^k]_{t=0} = 0$.

Using the result of Lemma 4.1, we see that (4.12) can be simplified to

$$\begin{aligned} d\,\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon] &= \lambda(t)\,\mathbb{E}X^k(1-\widehat{E}_t)\,dt \\ &\quad + \frac{1}{\epsilon^2}\Big(\,\mathbb{E}\,[(S_t)^{k+1} \mid \mathcal{Y}_t^\epsilon] - \widehat{S}_t\,\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon]\Big)\,d\nu_t^\epsilon. \end{aligned} \tag{4.13}$$

This proves Theorem 4.2. □

Note that $\mathbb{E}\,[(S_t)^k \mid \mathcal{Y}_t^\epsilon]$, the conditional moment of order k, depends on the conditional moment of order $k+1$, $\mathbb{E}\,[(S_t)^{k+1} \mid \mathcal{Y}_t^\epsilon]$, so (4.9)–(4.10) do not form a closed set of equations. If we want to use these equations to define a finite-dimensional approximation to the optimal filter, we need to use an appropriate closure formula to approximate higher order moments in terms of lower order moments. One possible closure formula, which was proposed in [11], assumes the third order central moment to be zero at all times, i.e., $\mathbb{E}\,[(S_t - \widehat{S}_t)^3 \mid \mathcal{Y}_t^\epsilon] = 0$ for all $t \in [0,T]$.

This closing of the infinite set of moment equations that has now been generated, by expressing higher order moments in terms of lower order moments, means that we restrict our densities to belong to a specific family of distributions. As was pointed out earlier in [4], the *a priori assumption* that the conditional density will belong to this family at every time instant is often incorrect. But it was shown in the same paper that a sound mathematical basis can be given for this so-called *assumed density principle* in some cases which involve the filtering of nonlinear diffusions, by showing that the resulting filter is equivalent to a *projection* in probability density space, like the one we described in the preceding section. In the next section we will show that this idea can be applied to our change detection problem as well, and that we can gain considerable insight into the nature of such problems in doing so.

5. The assumed density principle. To formulate our differential-geometric approximate filter of section 3 in terms of the conditional moments it generates, we define, bearing in mind the interpretation of the normalization constant given in (3.8), the following statistics (where $\approx$ means approximates):

$$
\begin{aligned}
\check{E}_t &= \frac{\int_{\mathbb{R}} p(x,\theta_t)dx}{\int_{\mathbb{R}} p(x,\theta_t)dx + 1 - G(t)} &&\approx \mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon),\\
\check{X}_t &= \frac{\int_{\mathbb{R}} xp(x,\theta_t)dx}{\int_{\mathbb{R}} p(x,\theta_t)dx} &&\approx \mathbb{E}[X \mid \mathcal{Y}_t^\epsilon], \qquad (5.1)\\
\check{S}_t^n &= \frac{\int_{\mathbb{R}} x^n\, p(x,\theta_t)dx}{\int_{\mathbb{R}} p(x,\theta_t)dx + 1 - G(t)} &&\approx \mathbb{E}[(S_t)^n \mid \mathcal{Y}_t^\epsilon],
\end{aligned}
$$

with $n \in \mathbb{N}$, so $\check{S}_t^0 = E_t$ and $\check{S}_t^1 = \check{S}_t$.

We remark that this implies that $\check{S}_t = \check{X}_t\check{E}_t$, i.e., the conditional estimate of the signal $\check{S}_t$ naturally splits into two statistics $\check{E}_t$ and $\check{X}_t$, which approximate the conditional probability of a jump having occurred and the best estimate of the jump size, respectively. We have shown in (4.8) that the optimal filter estimates satisfy, for $\widehat{X}_t = \mathbb{E}\,[X \mid \mathcal{Y}_t^\epsilon] \neq 0$,

$$
\widehat{S}_t = \widehat{X}_t \left(1 - (1 - \widehat{E}_t)\frac{\mathbb{E}X}{\widehat{X}_t}\right),
$$

so the *optimal* filter estimates will in general *not* satisfy the equation $\widehat{S}_t = \widehat{X}_t\widehat{E}_t$.

However, this does not imply that the estimate $\check{S}_t$ which is generated by the differential-geometric approximation is different from the filter estimate $\widehat{S}_t$ generated by closing the Kushner–Stratonovich equations, as we did in the previous section. We now show that they are in fact the same if we use (4.9) and (4.10) to calculate the first $m+1$ moments and then close the equations *in an appropriate way*, by choosing $\check{S}_t^{m+1}$ appropriately.

THEOREM 5.1. *Let the conditions of Theorem* 3.1 *be satisfied, and let the process* $\{\theta_t,\, t \geq 0\}$ *be defined as in* (3.10). *Define* $\check{E}_t$, $\check{S}_t$, *and* $\check{S}_t^n$ *as in* (5.1) *for* $n = 0 \ldots m{+}1$. *Then* $\check{E}_t = \check{S}_t^0$, $\check{S}_t = \check{S}_t^1$, *and* $\check{S}_t^n$ $(n = 2 \ldots m)$ *satisfy*

$$
d\check{S}_t^n = \lambda(t)\,(1 - \check{E}_t)\,\mathbb{E}X^n\,dt \;+\; \frac{1}{\epsilon^2}\,(\,\check{S}_t^{n+1} - \check{S}_t\,\check{S}_t^n\,)\,(\,dY_t^\epsilon - \check{S}_t\,dt\,),
$$

with initial conditions $\check{S}_0^n = 0$ *for all* $n = 0 \ldots m$.

Proof of Theorem 5.1. To find the stochastic differential equations for the $\check{S}_t^n$ $(n = 0 \ldots m)$ we must first find the equations for the approximated conditional moments η_t^k $(k = 0 \ldots m)$, but this is relatively simple since (3.6) implies that

$$
d\eta_t^k \;=\; [\,\eta_t^k \;\; \eta_t^{k+1} \;\; \cdots \;\; \eta_t^{k+m}\,] \circ d\theta_t,
$$

and the result of Theorem 3.1 then gives

$$
d\eta_t^k \;=\; g(t)\,\mathbb{E}X^k\,dt \;-\; \frac{\eta_t^{k+2}}{2\epsilon^2}\,dt \;+\; \frac{\eta_t^{k+1}}{\epsilon^2} \circ dY_t^\epsilon. \qquad (5.2)
$$

Using the Îto form of (5.2),

$$
d\eta_t^k \;=\; g(t)\,\mathbb{E}X^k\,dt \;+\; \frac{\eta_t^{k+1}}{\epsilon^2}\,dY_t^\epsilon,
$$

we find by Îto's differentiation rule that for all $k = 0 \dots m$,

$$\begin{aligned}
d\check{S}_t^k &= \frac{d\eta_t^k}{\eta_t^0 + 1 - G(t)} - \frac{\eta_t^k\, d(\eta_t^0 + 1 - G(t))}{(\eta_t^0 + 1 - G(t))^2} - \frac{d\eta_t^k\, d(\eta_t^0 + 1 - G(t))}{(\eta_t^0 + 1 - G(t))^2} \\
&\quad + \frac{\eta_t^k\, d(\eta_t^0 + 1 - G(t)) d(\eta_t^0 + 1 - G(t))}{(\eta_t^0 + 1 - G(t))^3} \\
&= \frac{\mathbb{E}X^k g(t)dt + \eta_t^{k+1} dY_t^\epsilon/\epsilon^2}{\eta_t^0 + 1 - G(t)} - \frac{\eta_t^k\, \eta_t^1 dY_t^\epsilon/\epsilon^2}{(\eta_t^0 + 1 - G(t))^2} - \frac{\eta_t^{k+1}\eta_t^1 dt/\epsilon^2}{(\eta_t^0 + 1 - G(t))^2} \\
&\quad + \frac{\eta_t^k\, (\eta_t^1)^2 dt/\epsilon^2}{(\eta_t^0 + 1 - G(t))^3} \\
&= \frac{g(t)}{1 - G(t)} \left(1 - \frac{\eta_t^0}{\eta_t^0 + 1 - G(t)}\right) \mathbb{E}X^k\, dt \\
&\quad + \frac{1}{\epsilon^2} \left(\frac{\eta_t^{k+1}}{\eta_t^0 + 1 - G(t)} - \frac{\eta_t^k \eta_t^1}{(\eta_t^0 + 1 - G(t))^2} \right) \left(dY_t^\epsilon - \frac{\eta_t^1}{\eta_t^0 + 1 - G(t)}\, dt \right) \\
&= \lambda(t)(1 - \check{E}_t)\, \mathbb{E}X^k\, dt \;+\; \frac{1}{\epsilon^2} \left(\check{S}_t^{k+1} - \check{S}_t\, \check{S}_t^k \right) \left(dY_t^\epsilon - \check{S}_t\, dt \right),
\end{aligned}$$

which proves the theorem. □

These equations are precisely the same as the ones we derived for the filter of the previous section, (4.9) and (4.10), if we replace $\mathbb{E}\,[S_t^k \mid \mathcal{Y}_t^\epsilon]$ by $\check{S}_t^k$. It thus follows that if we close these equations by choosing $\check{S}_t^{m+1}$ appropriately, then the two filters generate the same estimates almost surely. However, some care must be taken in finding the appropriate closure formula. For example, in the Gaussian case ($m = 2$), we must *not* choose the third central moment to be equal to zero, as we proposed in the end of section 4. Since $p(x, \theta_t) / \int_{\mathbb{R}} p(x, \theta_t) dx$ is assumed to be Gaussian, and since a Gaussian variable A satisfies $\mathbb{E}A^3 = [\mathbb{E}A] \cdot [3\mathbb{E}A^2 - 2(\mathbb{E}A)^2]$, we have

(5.3)
$$\frac{\eta_t^3}{\eta_t^0} = \frac{\eta_t^1}{\eta_t^0} \left(3\frac{\eta_t^2}{\eta_t^0} - 2\left(\frac{\eta_t^1}{\eta_t^0}\right)^2 \right) \qquad \Rightarrow \qquad \frac{\check{S}_t^3}{\check{E}_t} = \frac{\check{S}_t}{\check{E}_t} \left(3\frac{\check{S}_t^2}{\check{E}_t} - 2\left(\frac{\check{S}_t}{\check{E}_t}\right)^2 \right).$$

Only when this more complicated closure formula for $\mathbb{E}\,[(S_t)^3 \mid \mathcal{Y}_t^\epsilon]$ in terms of the lower order moments $\mathbb{E}\,[(S_t)^2 \mid \mathcal{Y}_t^\epsilon]$, $\mathbb{E}\,[S_t \mid \mathcal{Y}_t^\epsilon]$, and $\mathbb{E}\,[E_t \mid \mathcal{Y}_t^\epsilon]$ is used will the estimates generated by the Kushner–Stratonovich equation be the same, almost surely, as those generated by our differential-geometric approximation.

6. A three-dimensional filter. Although the filter derived in section 3 using differential-geometric methods is thus equivalent to the filter derived in section 4 *when the correct closure formula is used*, there are certain advantages of the first parametrization. We already mentioned the fact that better schemes can be used to calculate numerical approximations of (3.10). Another advantage is the much more intuitive structure of the filter. If we define the a priori moments of the jump size X as $P^n = \mathbb{E}(X - \mathbb{E}X)^n$, and the approximate filter estimates

$$\check{P}_t^n = \frac{\int_{\mathbb{R}} (x - \check{X}_t)^n p(x, \theta_t) dx}{\int_{\mathbb{R}} p(x, \theta_t) dx} \qquad \approx \qquad \mathbb{E}[(X - \mathbb{E}\,[X \mid \mathcal{Y}_t^\epsilon])^n \mid \mathcal{Y}_t^\epsilon],$$

then one may show by a tedious but straightforward exercise in Stratonovich calculus [23] that for $n = 2 \dots m$,

$$d\check{E}_t = \lambda(t)(1-\check{E}_t)\,dt \;+\; \check{E}_t(1-\check{E}_t)\frac{\check{X}_t}{\epsilon^2}\,(\,dY_t^\epsilon - \check{S}_t\,dt\,), \tag{6.1}$$

$$d\check{X}_t = \lambda(t)\frac{1-\check{E}_t}{\check{E}_t}\,(\mathbb{E}X - \check{X}_t)\,dt \;+\; \frac{\check{P}_t^2}{\epsilon^2}\,(\,dY_t^\epsilon - \check{X}_t\,dt\,), \tag{6.2}$$

$$\begin{aligned} d\check{P}_t^n = \lambda(t)\frac{1-\check{E}_t}{\check{E}_t}\Bigg[& P^n - \check{P}_t^n \;+\; n(P^{n-1} - \check{P}_t^{n-1})\,(\mathbb{E}X - \check{X}_t) \\ & + \sum_{k=0}^{n-2}\binom{n}{k} P^k(\mathbb{E}X - \check{X}_t)^{n-k}\Bigg]\,dt \\ - \frac{n\,\check{P}_t^2}{\epsilon^2}\,\big[\,\check{P}_t^n - \tfrac{1}{2}(n-1)\,\check{P}_t^2\check{P}_t^{n-2}\,\big]\,dt & \\ + \frac{1}{\epsilon^2}\,[\,\check{P}_t^{n+1} - n\check{P}_t^2\check{P}_t^{n-1}\,]\,(dY_t^\epsilon - \check{X}_t\,dt). & \end{aligned}$$

These stochastic differential equations can be interpreted as the sum of vector fields which drive the conditional density to the a priori density of X (and these dominate before the jump when $\check{E}_t$ will be close to zero), vector fields which resemble those of the Kalman filter for a constant signal (which dominate after the jump when $\check{E}_t$ will be close to one), and some extra terms which make sure we do not leave the manifold that we project upon.

Note that the terms involving the innovation process in the equations for $\widehat{P}_t^n$ ($n = 2 \dots m$) will all be zero if and only if the central moments satisfy the equation $\widehat{P}_t^{n+1} = n\widehat{P}_t^2\widehat{P}_t^{n-1}$ for $n = 0 \dots m$. Since we have that $\widehat{P}_t^0 = 1$ and $\widehat{P}_t^1 = 0$ for all $t \geq 0$, this will be the case if for all $n \in \mathbb{N}$,

$$\widehat{P}_t^{2n} \;=\; (\widehat{P}_t^2)^n \cdot 2^{-n}\,\frac{(2n)!}{n!}, \qquad \widehat{P}_t^{2n+1} \;=\; 0,$$

i.e., the first $m+1$ central moments should be the same as those of a Gaussian distribution. This suggests that the equations will become even simpler if both X and our manifold are Gaussian, which is exactly the exponential family we get if we take $m = 2$ and the parameter set $\Theta = \{(\theta_0, \theta_1, \theta_2) : \theta_2 < 0\}$. In the rest of this section we will analyze the detection and estimation scheme when such a manifold of unnormalized Gaussian densities is used.

If we substitute the relation $\check{S}_t = \check{X}_t\check{E}_t$ into the stochastic differential equation for $\check{E}_t$ given by (6.1), we see the close connection with the celebrated Shiryayev–Wonham detector for *known* jump sizes. As we remarked in section 2, if we assume that the jump size X is known a priori, say $X = a$, then the conditional probability that the jump has occurred $\pi_t = \mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon)$ is finite-dimensionally computable. In fact, it follows the Shiryayev–Wonham equation [22, 28] that

$$d\pi_t \;=\; \lambda(t)(1-\pi_t)\,dt \;+\; \pi_t(1-\pi_t)\frac{a}{\epsilon^2}(dY_t^\epsilon - a\pi_t\,dt), \qquad \pi_0 \;=\; 0. \tag{6.3}$$

Our estimate of the probability that a jump has occurred satisfies a modified version of this Shiryayev–Wonham equation, where a *known* jump size $X = a$ in the equation is replaced by a time-varying *estimated* jump size $\check{X}_t$. For $m = 2$ our differential-geometric approximation thus becomes a mixture of modified Kalman filter equations and this *adaptive* Shiryayev–Wonham equation:

$$
\begin{aligned}
\check{S}_t &= \check{E}_t\,\check{X}_t, \\
d\check{E}_t &= \lambda(t)(1-\check{E}_t)\,dt \;+\; \check{E}_t(1-\check{E}_t)\frac{\check{X}_t}{\epsilon^2}\,(\,dY_t^{\epsilon} - \check{X}_t\check{E}_t\,dt\,), \\
d\check{X}_t &= \lambda(t)\frac{1-\check{E}_t}{\check{E}_t}\,(\mathbb{E}X - \check{X}_t)\,dt \;+\; \frac{\check{P}_t^2}{\epsilon^2}\,(\,dY_t^{\epsilon} - \check{X}_t\,dt\,), \\
d\check{P}_t^2 &= \lambda(t)\frac{1-\check{E}_t}{\check{E}_t}\,[\,(\mathbb{E}X - \check{X}_t)^2 \;+\; \mathrm{Var}\;X \;-\; \check{P}_t^2\,]\,dt \;-\; \frac{(\check{P}_t^2)^2}{\epsilon^2}\,dt,
\end{aligned}
$$

with $\check{E}_0 = 0$, $\check{P}_0^2 = \mathrm{Var}\;X$, and $\check{X}_0 = \mathbb{E}X$.

In Figure 1, a block diagram of the filter is given, which highlights the decomposition of the problem in a detection and an estimation part, which communicate through the jump size estimate $\check{X}_t$ and the conditional probability ratio $(1-\check{E}_t)/\check{E}_t$. We remark that the original optimal detection and estimation problem as we formulated it cannot be solved recursively because we want to perform detection and estimation *simultaneously*. If the estimation problem would be trivial (i.e., if we would know the jump size X immediately after the jump), the detection problem could be solved recursively, since we can then use a Shiryayev–Wonham filter tuned at $a = X$. If the detection problem would be trivial (i.e., we would know immediately after time τ that a jump has occurred), then the estimation problem could be solved recursively, since we could simply start a Kalman filter at time τ.

In our combined problem, however, we must make sure that the Kalman filter does not start filtering too early, since it would then filter the zero signal for some time while its conditional variance $\check{P}_t^2$ would decrease, making its reaction too slow when the jump does indeed occur. The likelihood ratio term in the equation for $\check{P}_t^2$, which

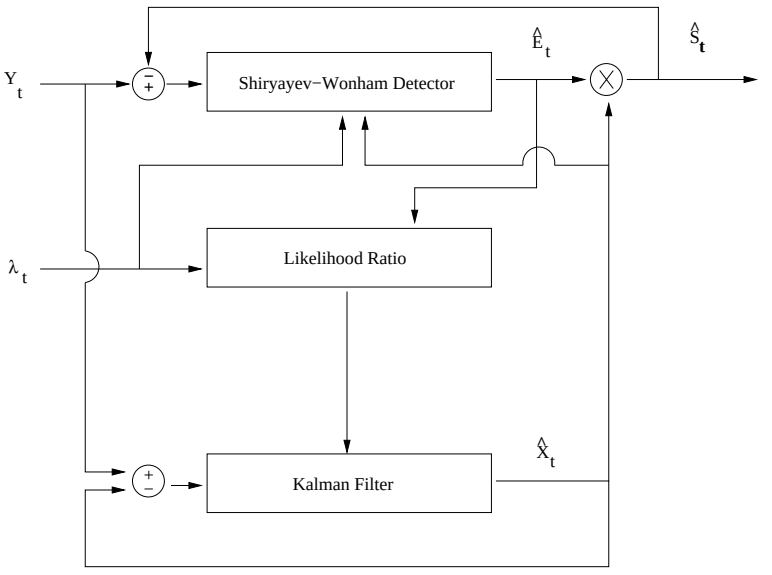


FIG. 1. *Structure of the three-dimensional approximating filter.*

pulls the conditional variance $\check{P}_t^2$ back to the variance of X as long as $\check{E}_t$ is small, prevents this from happening. After the jump has occurred, a good estimate of X should quickly become available, and the stochastic differential equation for $\check{E}_t$ will then resemble the Shiryayev–Wonham equation. We therefore expect the estimate for the conditional probability that a change has occurred to converge to one quite quickly after that. We will see in the simulation studies of section 8 that this will indeed be the case. But first we must make sure that the three-dimensional filter that we have just defined has a finite, nonexploding global solution for all $t \geq 0$. This will be the subject of the next section.

7. Existence and uniqueness of filter estimates. When we derived the equations for our new nonlinear filter in section 3, we defined it only up to the exit time $\inf\{t \geq 0 : \theta_t \notin \Theta\}$. In this section we will show that for the three-dimensional filter that we defined in the preceding section, this exit time will be equal to infinity almost surely under rather mild conditions. This means that a unique, well-defined, finite, nonexploding solution for the stochastic differential equations exists for *all* $t \geq 0$, a fact which has not yet been proven for the diffusion case [4].

The filter equations suggest that it will be convenient to work with the scaled likelihood ratio process defined by

$$\check{Z}_t^\epsilon \;=\; \epsilon^2 \cdot \check{Z}_t \;=\; \epsilon^2 \cdot \lambda(t) \frac{1 - \check{E}_t}{\check{E}_t}.$$

We remark that, strictly speaking, $\check{Z}_t$ is not defined at $t = 0$ since $\check{E}_0 = 0$ but as we remarked before, it can be defined after an arbitrarily small time step, in the same way as we suggested in the discussion after the proof of Theorem 3.1. A simple application of Îto's differentiation rule shows that we may rewrite the filter equations in the following form:

$$d\check{Z}_t^\epsilon = \check{Z}_t^\epsilon \left[-\check{Z}_t^\epsilon \;+\; \epsilon^2 \frac{g'(t)}{g(t)} \;+\; \check{X}_t(\check{X}_t - S_t) \right] \frac{dt}{\epsilon^2} \;-\; \check{Z}_t^\epsilon \check{X}_t \frac{dW_t}{\epsilon}, \tag{7.1}$$

$$d\check{X}_t = [\, \check{Z}_t^\epsilon (\mathbb{E}X - \check{X}_t) \;+\; \check{P}_t^2 (S_t - \check{X}_t) \,] \frac{dt}{\epsilon^2} \;+\; \check{P}_t^2 \frac{dW_t}{\epsilon}, \tag{7.2}$$

$$d\check{P}_t^2 = [\, \check{Z}_t^\epsilon \,(\, (\mathbb{E}X - \check{X}_t)^2 \;+\; \mathrm{Var}\; X \;-\; \check{P}_t^2 \,) \;-\; (\check{P}_t^2)^2 \,] \frac{dt}{\epsilon^2}. \tag{7.3}$$

This shows that the natural time scale is of order ϵ^2.

Using this representation of the filter, we can now prove that the filter equations will indeed have a solution. To do so we shall use the following result.

THEOREM 7.1. *Consider the* $m + 1$*-dimensional stochastic differential system defined by*

$$D_t = D_{t_0} \;+\; \int_{t_0}^t b(s, D_s)\, ds \;+\; \int_{t_0}^t \sigma(s, D_s)\, dW_s, \tag{7.4}$$

with $\{W_t,\, t \geq t_0\}$ *a standard Wiener process. Denote the generator of this process by*

$$L \;=\; \tfrac{\partial}{\partial t} + \sum_i b_i \tfrac{\partial}{\partial x_i} + \tfrac{1}{2} \sum_{i,j} \sigma_i \sigma_j \tfrac{\partial^2}{\partial x_i \partial x_j},$$

and suppose the following:

1. *D_{t_0} is independent of $\{W_t, t \geq t_0\}$.*
2. *Both b and σ are functions $\mathbb{R} \times \mathbb{R}^{m+1} \to \mathbb{R}^{m+1}$ such that for all $R > t_0$ and every x, y in the compact set $\mathcal{U}_R \stackrel{\text{def}}{=} \{d : \|d\| \leq R\}$ and all $t_0 \leq s \leq R$, there exists a constant M_R such that*

$$\begin{aligned} \|b(s,x) - b(s,y)\| \ + \ \|\sigma(s,x) - \sigma(s,y)\| &\leq M_R \|x - y\|, \\ \|b(s,x)\| \ + \ \|\sigma(s,x)\| &\leq M_R (1 + \|x\|). \end{aligned}$$

3. *There exists a nonnegative function $V(t,d) : \mathbb{R} \times \mathbb{R}^{m+1} \to \mathbb{R}^+$, twice differentiable with respect to d, such that*

$$\liminf_{R \to \infty} \inf_{\|d\| > R} V(t,d) \ = \ \infty,$$

which satisfies the linear inequality $LV(t,d) \leq c\,V(t,d)$ for some $c > 0$, for all d outside a compact set $\mathcal{G} \subseteq \mathbb{R}^{m+1}$ and all $t \geq t_0$.

Then there exists a solution to (7.4) for every $t \geq t_0$ (i.e., the solution does not explode in finite time). This solution is unique up to equivalence, and an almost surely continuous stochastic process.

A proof of this theorem can be found in [15]; we just remark that the first two conditions guarantee that the solution is well defined up to the stopping times

$$t_R \ = \ \inf\{t \geq t_0 : \ \|D_t\| \geq R\},$$

while the last condition guarantees that

$$\mathbb{P}\Big(\lim_{R \to \infty} t_R = \infty \Big) \ = \ 1,$$

which means we can extend the solution to all times $t \geq t_0$. We now use this theorem to conclude that (7.1)–(7.3) have a unique nonexploding solution.

THEOREM 7.2. *If the function g is differentiable on its entire domain $\mathbb{R}^+$ and satisfies*

$$\sup_{t \geq 0} \ \frac{g'(t)}{g(t)} \ \stackrel{\text{def}}{=} \ \breve{g} \ < \ \infty,$$

and the stochastic variable X is bounded, i.e., $|X| \leq \bar{X}$ almost surely for some constant $\bar{X} > 0$, then (7.1)–(7.3) have an almost surely continuous solution for all $t \geq 0$, which is unique up to equivalence.

We remark that the condition that X is bounded means that one of the conditions for the derivation of these equations as a *projection* is violated, since the density f is no longer strictly positive Lebesgue-almost everywhere in this case. Although the equations will have a well-defined solution, they can strictly speaking no longer be interpreted as a projection in $\mathcal{L}^2$.

Proof of Theorem 7.2. The first part of the conditions in Theorem 7.1 is trivially satisfied, and so is the second part, since all terms may be linearly bounded in the state variables on a compact domain $\mathcal{U}_R$. We have therefore proven the result if we show that the function

$$V(\check{X}_t, \check{P}_t^2, \check{Z}_t^\epsilon) \ = \ (\check{X}_t)^2 \ + \ \check{P}_t^2 \ + \ \tfrac{1}{2}\check{Z}_t^\epsilon \ + \ (\check{Z}_t^\epsilon)^{-\frac{1}{2}} \tag{7.5}$$

satisfies the third condition. We remark that we have included the term $(\check{Z}_t^\epsilon)^{-\frac{1}{2}}$ to show that $\check{Z}_t^\epsilon$ cannot become zero in finite time and that $\check{Z}_t^\epsilon$ will thus stay strictly positive for all times $t \geq 0$ if its initial value is chosen to be strictly positive, as we suggested in the discussion after the proof of Theorem 3.1. Since the differential equation for $\check{P}_t^2$ shows that $\check{P}_t^2$ will stay strictly positive as well, this will then imply that $V(\check{X}_t, \check{P}_t^2, \check{Z}_t^\epsilon)$ is strictly positive for all $t \geq 0$. Using (7.1)–(7.3), we find that

$$\begin{aligned}
LV &= \tfrac{2\check{X}_t}{\epsilon^2} \cdot [\, \check{Z}_t(\mathbb{E}X - \check{X}_t) \;+\; \check{P}_t^2(S_t - \check{X}_t)\,] \;+\; \tfrac{(\check{P}_t^2)^2}{\epsilon^2} \;+\; \tfrac{\check{Z}_t^\epsilon}{\epsilon^2}[\,(\mathbb{E}X - \check{X}_t)^2 + \text{Var}\; X - \check{P}_t^2\,] \\
&\quad - \tfrac{(\check{P}_t^2)^2}{\epsilon^2} \;+\; \tfrac{\check{Z}_t^\epsilon}{2\epsilon^2}\,[\,-\check{Z}_t^\epsilon + \epsilon^2\tfrac{g'(t)}{g(t)} + \check{X}_t(\check{X}_t - S_t)\,] \\
&\quad - \tfrac{1}{2}(\check{Z}_t^\epsilon)^{-\frac{3}{2}} \cdot \tfrac{\check{Z}_t^\epsilon}{\epsilon^2}\,[\,-\check{Z}_t^\epsilon + \epsilon^2\tfrac{g'(t)}{g(t)} + \check{X}_t(\check{X}_t - S_t)\,] \;+\; \tfrac{1}{2}\cdot\tfrac{3}{4}(\check{Z}_t^\epsilon)^{-\frac{5}{2}} \cdot \tfrac{(\check{Z}_t^\epsilon \check{X}_t)^2}{\epsilon^2} \\
&= \tfrac{\check{Z}_t^\epsilon}{\epsilon^2}\,[\,(\mathbb{E}X)^2 - \tfrac{1}{4}(\check{X}_t)^2 + \text{Var}\; X - \check{P}_t^2 - \tfrac{1}{2}\check{Z}_t^\epsilon + \tfrac{1}{2}\epsilon^2\tfrac{g'(t)}{g(t)} - \tfrac{1}{2}\check{X}_t S_t - \tfrac{1}{4}(\check{X}_t)^2\,] \\
&\quad - 2\tfrac{\check{P}_t^2}{\epsilon^2}\check{X}_t(\check{X}_t - S_t) \;-\; \tfrac{1}{2}\tfrac{(\check{Z}_t^\epsilon)^{-\frac{1}{2}}}{\epsilon^2}\,[\,-\check{Z}_t^\epsilon + \epsilon^2\tfrac{g'(t)}{g(t)} + \tfrac{1}{4}(\check{X}_t)^2 - \check{X}_t S_t\,] \\
&\leq \tfrac{\check{Z}_t^\epsilon}{\epsilon^2}\,[\,(\mathbb{E}X)^2 + \text{Var}\; X + \tfrac{1}{2}\epsilon^2\check{g} - \tfrac{1}{4}\,\big[\,(\check{X}_t)^2 + \check{P}_t^2 + \tfrac{1}{2}\check{Z}_t^\epsilon\big] - \tfrac{1}{2}\check{X}_t S_t - \tfrac{1}{4}(\check{X}_t)^2\,] \\
&\quad - 2\tfrac{\check{P}_t^2}{\epsilon^2}\check{X}_t(\check{X}_t - S_t) \;+\; \tfrac{(\check{Z}_t^\epsilon)^{\frac{1}{2}}}{2\epsilon^2} \;+\; \tfrac{(\check{Z}_t^\epsilon)^{-\frac{1}{2}}}{\epsilon^2}\,[\,\tfrac{1}{2}\epsilon^2\,|\check{g}| - \tfrac{1}{8}(\check{X}_t)^2 + \tfrac{1}{2}\check{X}_t S_t\,].
\end{aligned}$$

We have that $(\check{Z}_t^\epsilon)^{\frac{1}{2}} \leq \check{Z}_t^\epsilon + (\check{Z}_t^\epsilon)^{-\frac{1}{2}}$ and we use the inequality $by - ay^2 \leq \frac{b^2}{4a}$ (all $y \in \mathbb{R}$) to find that

$$\begin{aligned}
-\tfrac{1}{2}\check{X}_t S_t - \tfrac{1}{4}(\check{X}_t)^2 \leq \tfrac{1}{4}(S_t)^2 \;&<\; \tfrac{1}{4}\bar{X}^2, \\
-2\check{X}_t(\check{X}_t - S_t) \leq \tfrac{1}{2}(S_t)^2 \;&<\; \tfrac{1}{2}\bar{X}^2, \\
-\tfrac{1}{8}(\check{X}_t)^2 + \tfrac{1}{2}\check{X}_t S_t \leq \tfrac{1}{2}(S_t)^2 \;&<\; \tfrac{1}{2}\bar{X}^2,
\end{aligned}$$

and therefore

$$\begin{aligned}
LV \leq \frac{\check{Z}_t^\epsilon}{\epsilon^2}\,&\left[\,(\mathbb{E}X)^2 + \text{Var}\; X + \frac{1}{2}\epsilon^2\check{g} - \frac{1}{4}\,\left[\,(\check{X}_t)^2 + \check{P}_t^2 + \frac{1}{2}\check{Z}_t^\epsilon\right] + \frac{1}{4}\bar{X}^2\,\right] \\
&+ \frac{\check{P}_t^2\,\bar{X}^2}{2\epsilon^2} \;+\; \frac{1}{2\epsilon^2}(\,\check{Z}_t^\epsilon + (\check{Z}_t^\epsilon)^{-\frac{1}{2}}\,) \;+\; \frac{(\check{Z}_t^\epsilon)^{-\frac{1}{2}}}{2\epsilon^2}\,[\,\epsilon^2\,|\check{g}| + \bar{X}^2\,],
\end{aligned}$$

so we can take, for example, $c = \frac{1}{2\epsilon^2}(2 + \bar{X}^2 + \epsilon^2|\check{g}|)$ in the last condition of Theorem 7.1, and

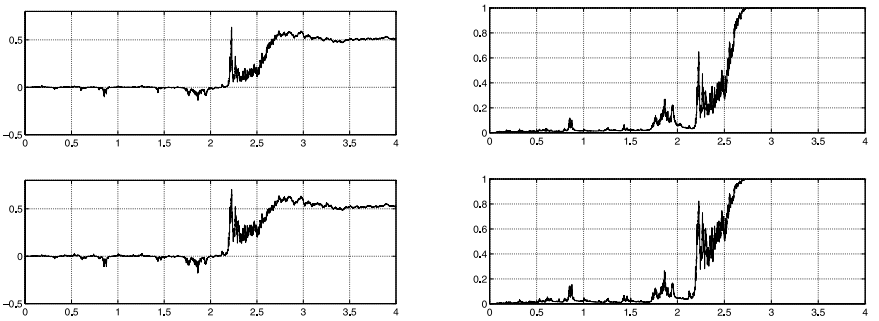
$$\mathcal{G} \;=\; \{\,(x,p,z) :\; x^2 + p + \tfrac{1}{2}z \;\leq\; 4(\mathbb{E}X)^2 + 4\,\text{Var}\; X + 2\epsilon^2\check{g} + \bar{X}^2\,\}.$$

All conditions of Theorem 7.1 are satisfied, and the filter equations have a unique, nonexploding solution. □

We have thus proven that the algorithm that we derived will be well behaved for all times $t \geq 0$ almost surely, and we can now analyze it in simulation studies in the next section.

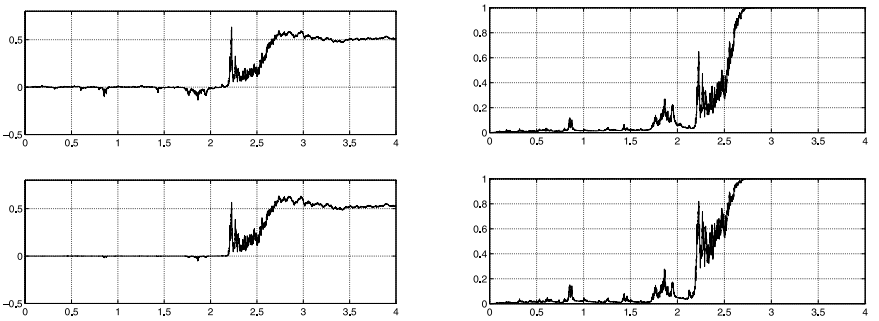
8. Simulation results. In this section we will investigate the performance of the approximating filters that we defined in previous sections, by means of simulation studies.

Experiment 1. For the first simulation study we took τ to be an exponentially distributed stochastic variable with mean 15.0, and the jump size X normally distributed with zero mean and unit variance. We let the actual jump take place at



(a) Estimate of $\mathbb{E}[S_t \mid \mathcal{Y}_t^\epsilon]$ (b) Estimate of $\mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon)$

FIG. 2. *Comparison between optimal and approximate filter, using* (3.10).

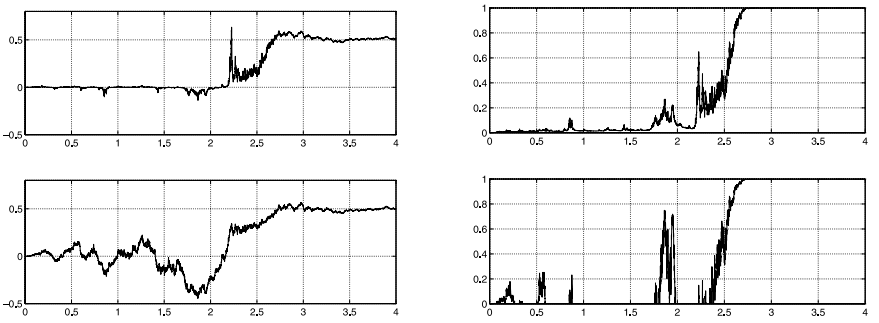


(a) Estimate of $\mathbb{E}[S_t \mid \mathcal{Y}_t^\epsilon]$ (b) Estimate of $\mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon)$

FIG. 3. *Comparison between optimal and approximate filter, using moments.*

$\tau = 2.0$, and the jump size was taken to be $X = 0.5$ exactly. The noise parameter ϵ was taken 0.10. All filters estimates were calculated on a time interval $t \in [0.0, 4.0]$, using an Euler scheme with step size $4.0 \cdot 10^{-5}$.

We first simulated the differential-geometric approximation as formulated in the previous section, i.e., we took the dimension of the filter $m + 1 = 3$, which means we project upon a manifold of Gaussian densities. In Figures 2 and 3 the results are shown for two different implementations of our filter. The top graphs show the real conditional estimate of the signal $\mathbb{E}[S_t \mid \mathcal{Y}_t^\epsilon]$ and the conditional probability that a jump has occurred $\mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon)$. These were obtained by solving the Duncan–Mortensen–Zakai equation on a grid which divided the interval $[-3.0, 3.0]$ for possible values of X in 1500 equidistant points. In Figure 2 the filter was implemented by (3.10), the stochastic differential equation for the parameter vector θ_t, while in Figure 3 the direct equations for the moments which we derived in the previous section were used. There are some small differences between the two, which should be attributed to inaccuracies in the calculation of $[H(\theta_t)]^{-1}$ in (3.10) and in the numerical method we use.



(a) Estimate of $\mathbb{E}[S_t \mid \mathcal{Y}_t^\epsilon]$.

(b) Estimate of $\mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon)$.

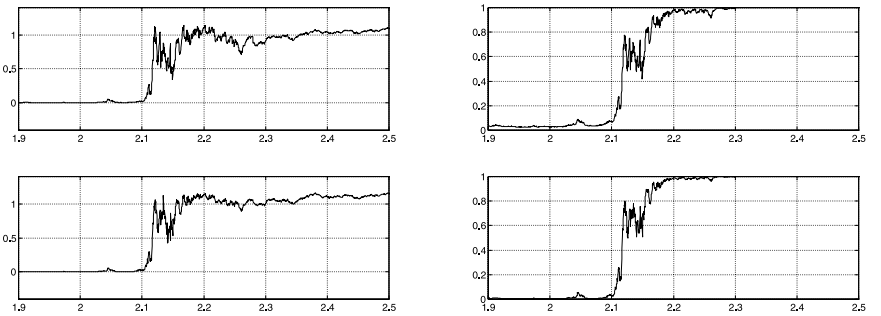
FIG. 4. *Comparison between optimal filter and filter of section* 4.

However, in both cases the filter estimates show an excellent behavior both before and after the change point. Both implementations slightly overestimate the conditional probability of a jump having occurred, but only after the jump. Around $t = 2.6$ the approximate and the exact conditional signal estimate are already indistinguishable. More importantly, the small delay in detection of the optimal filter (seen to be approximately 0.10 here) is the same for the approximate filter. For an extensive analysis of such detection delays in the optimal filter and its suboptimal approximations, the reader is referred to [11] and [23].

For comparison, Figure 4 shows a simulation of the same model setup for the approximating filter which we derived in section 4, where conditional moments are generated by using the Kushner–Stratonovich equation and the assumption that the third order central conditional moment is equal to zero. We showed in (5.3) that this filter, which was proposed in [11], is not equivalent to our filter and its behavior is seen to be a lot worse. Although it will estimate both the signal and the conditional probability correctly in the long run, its behavior before the change is totally unacceptable. Indeed, the conditional probability is negative most of the time, and the estimates of the signal before the change are not close to the true value zero at all.

Experiment 2. To show that the excellent results for our differential-geometric approximation are not just a consequence of X being Gaussian, we performed a second set of simulations in which X was taken to be uniformly distributed on $[0, 2]$. The jump time was given the same distribution as in the first set of experiments, and the actual jump time was again taken to be $\tau = 2.0$. The jump size was taken equal to $X = 1.0$, and $\epsilon = 0.10$.

Figure 5 shows the estimates generated by our approximate filter, implemented by (3.10). The detection delay of 0.10 is almost exactly the same as for the optimal filter, and good filter estimates are produced almost directly after that. Apparently the algorithm works quite well for a jump size X with a uniform distribution, even though this distribution cannot be approximated very well on the exponential manifold that we project upon. In practice this is not important, since after the jump the behavior in the center of the state space can be shown to be asymptotically Gaussian in a large deviations sense. We again refer to [11] and [23] where an exact statement of this result is given, which helps to explain the good performance of our filter.



(a) Estimate of $\mathbb{E}\left[S_t \mid \mathcal{Y}_t^\epsilon\right]$ (b) Estimate of $\mathbb{P}(t \geq \tau \mid \mathcal{Y}_t^\epsilon)$

FIG. 5. *Comparison between optimal and approximate filter.*

9. Conclusions. In this paper, we have argued that nonlinear filtering theory can be used to characterize and approximate relevant conditional statistics in those change detection problems where the size of the change is not known a priori. We have shown that a simple three-dimensional nonlinear filter can be defined which has a global and unique solution under mild conditions, and which performs well in simulation studies. Apart from an interpretation in terms of information geometry and in terms of an assumed density principle, we may view the equations for this filter as an adaptive version of the Shiryayev–Wonham equation, fed by estimates from a modified Kalman filter.

Some interesting problems are still open at the moment. These include, for example, the design of adaptive change detectors for discrete time problems, the design of detectors for more complicated signal changes such as the changing slope process [25]

$$R_t = \begin{cases} 0, & 0 \leq t < \tau, \\ X(t-\tau), & t \geq \tau, \end{cases}$$

and the derivation of further theoretical properties of the filters that we defined in this paper. We hope to address these problems in future research.

REFERENCES

[1] S. AMARI, *Differential-Geometric Methods in Statistics*, Lecture Notes in Statist. 28, Springer–Verlag, Berlin, 1985.

[2] M. BASSEVILLE, *Detecting changes in signals and systems, a survey*, Automatica, 24 (1988), pp. 309–326.

[3] M. BASSEVILLE AND I. V. NIKIFOROV, *Detection of Abrupt Changes: Theory and Application*, Prentice–Hall, Englewood Cliffs, NJ, 1993.

[4] D. BRIGO, B. HANZON, AND F. LEGLAND, *A differential geometric approach to nonlinear filtering: The projection filter*, IEEE Trans. Automat. Control, 43 (1998), pp. 247–252.

[5] D. BRIGO, B. HANZON, AND F. LEGLAND, *Approximate filtering by projection on the manifold of exponential densities*, Bernoulli, 5 (1999), pp. 495–534.

[6] F. CAMPILLO, Y. KUTOYANTS, AND F. LEGLAND, *Small noise asymptotics of the GLR test for off-line change detection in misspecified diffusion processes*, Stochastics Stochastics Rep., 70 (2000), pp. 109–129.

[7] M. Chaleyat-Maurel and D. Michel, *Des résultats de non existence de filtre de dimension finie*, Stochastics, 13 (1984), pp. 83–102.
[8] M. H. A. Davis, *The application of nonlinear filtering to fault detection in linear systems*, IEEE Trans. Automat. Control, 20 (1975), pp. 257–259.
[9] M. H. A. Davis, *Markov Models and Optimization*, Chapman and Hall, London, 1993.
[10] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*, Springer–Verlag, New York, 1995.
[11] P. G. Fotopoulos, *Estimation and Detection of Jump Processes with Small Observation Noise*, Ph.D. Thesis, Imperial College, London, 1994.
[12] B. Hanzon, *A differential-geometric approach to nonlinear filtering*, in Geometrization of Statistical Theory, C. T. J. Dodson, ed., ULDM Publications, University of Lancaster, Lancaster, UK, 1987.
[13] R. Iserman, *Process fault detection based on modeling and estimation methods—A survey*, Automatica, 20 (1984), pp. 387–404.
[14] G. Kallianpur, *Stochastic Filtering Theory*, Springer–Verlag, New York, 1980.
[15] R. Z. Khas'minskii, *Stochastic Stability of Differential Equations*, Sijthoff & Noordhoff, Alphen aan den Rijn, The Netherlands, 1980.
[16] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Appl. Math. 23, Springer–Verlag, Berlin, 1992.
[17] R. Kulhavý, *Recursive nonlinear estimation: A geometric approach*, Automatica, 26 (1990), pp. 545–555.
[18] R. Kulhavý, *Recursive nonlinear estimation: Geometry of a space of posterior densities*, Automatica, 28 (1992), pp. 313–323.
[19] R. Kulhavý, *System identification: From matching data to matching probabilities*, in Plenary Lectures and Minicourses, G. Bastin and M. Gevers, eds., European Control Conference, Brussels, 1997, pp. 131–160.
[20] Y. A. Kutoyants, *Parameter Estimation for Stochastic Processes*, Heldermann Verlag, Berlin, 1984.
[21] T. L. Lai, *Sequential changepoint detection in quality control and dynamical systems (with discussion)*, J. Roy. Statist. Soc. Ser. B, 57 (1995), pp. 613–658.
[22] A. N. Shiryayev, *On optimum methods in quickest detection problems*, Theory Probab. Appl., 8 (1963), pp. 22–46.
[23] M. H. Vellekoop, *Rapid Detection and Estimation of Abrupt Changes by Nonlinear Filtering*, Ph.D. Thesis, Imperial College, London, 1997.
[24] M. H. Vellekoop and J. M. C. Clark, *Asymptotic behaviour of the optimal filter of jump and slope jump processes*, in Proceedings of the 35th IEEE Conference on Decision and Control, Vol. 2, IEEE Press, Piscataway, NJ, 1996, pp. 1163–1168.
[25] M. H. Vellekoop and J. M. C. Clark, *Changepoint detection using nonlinear filters*, in Proceedings of the 4th European Control Conference, Brussels, 1997.
[26] A. S. Willsky, *A survey of design methods for failure detection in dynamic systems*, Automatica, 12 (1976), pp. 601–611.
[27] E. Wong and B. Hajek, *Stochastic Processes in Engineering Systems*, Springer–Verlag, New York, 1984.
[28] W. M. Wonham, *Some applications of stochastic differential equations to optimal nonlinear filtering*, SIAM J. Control, 2 (1965), pp. 347–369.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 495–512

A FORMULA FOR THE DERIVATIVE WITH RESPECT TO DOMAIN VARIATIONS IN NAVIER–STOKES FLOW BASED ON AN EMBEDDING DOMAIN METHOD*

THOMAS SLAWIG†

Abstract. Fréchet differentiability and a formula for the derivative with respect to domain variation of a general class of cost functionals under the constraint of the two-dimensional stationary incompressible Navier–Stokes equations are shown. An embedding domain technique provides an equivalent formulation of the problem on a fixed domain and leads to a simple and computationally cheap line integral formula for the derivative of the cost functional with respect to domain variation. Existence of a solution to the corresponding domain optimization problems is proved. A numerical example shows the effectivity of the derivative formula.

Key words. domain optimization, Navier–Stokes equations, embedding domain technique

AMS subject classifications. 49Q10, 49J50, 35Q30, 76D05

PII. S0363012901385708

1. Introduction.

In this paper we present an explicit formula for the Fréchet derivative of a certain class of cost functionals with respect to variations in the shape of the domain under the constraint of the stationary, two-dimensional, incompressible Navier–Stokes equations. The derivative takes the form of a one-dimensional integral and thus is computationally cheap. It is based on the adjoint equation technique and on an embedding domain method, which allows it to formulate the state equations given on a complicated shaped domain in an equivalent way on a simple-shaped domain, e.g., a square. Moreover this method significantly reduces the rediscretization and re-assembling effort of the finite-dimensional system of the state equations on complicated-shaped and varying domains that occur during an iterative domain optimization process.

Embedding (or "fictitious") domain techniques have been widely applied in the treatment of PDEs. For Navier–Stokes equations on complicated-shaped domains, they were studied by, e.g., Glowinski, Pan, and Periaux [1]. Our Lagrange multiplier approach is similar to Glowinski's. Haslinger et al. used a slightly different one by introducing a distributed Lagrange multiplier and applied it on domain optimization problems; see, e.g., [2]. Domain optimization for the Navier–Stokes equations were studied, for example, by Pironneau [3], who computed the shape of body with minimum drag. Gunzburger and Kim [4] showed existence of an optimal shape for a minimum drag problem in a channel flow. Bello et al. [5] proved differentiability of the drag with respect to domain variations in Navier–Stokes flow.

The emphasis of this work is not to prove differentiability but to obtain a fast and effective numerical algorithm to solve domain optimization problems by iterative, gradient-based methods.

The same technique was used by Kunisch and Peichl [6] to obtain a derivative formula for the scalar Poisson problem. This paper extends a former work by Slawig [7] for the Stokes equations to the full nonlinear Navier–Stokes equations.

*Received by the editors February 28, 2001; accepted for publication (in revised form) October 4, 2002; published electronically May 12, 2003.
http://www.siam.org/journals/sicon/42-2/38570.html
†Technical University Berlin, MA 4-5, 10623 Berlin, Germany (slawig@math.tu-berlin.de).

The outline of the paper is the following: in the next two sections we define the geometric model configuration and summarize the needed results for the stationary Navier–Stokes equations. Then the considered class of domain optimization problems and the embedding domain technique are presented. Later on we show the continuous dependence of the solution of the Navier–Stokes equations with respect to the variation of the domain. The presentation of the explicit formula for the Fréchet derivative is followed by a brief presentation of the numerical solution techniques for the state and adjoint equations and optimization problems. At the end we show an inverse problem as a numerical example.

2. The geometric model configuration. Our model geometry is determined by two requirements for the derivative formula and the embedding domain method:

(a) We need sufficient regularity of the solutions to the state and adjoint equations, i.e., H^2 and H^1 for velocity and pressure, respectively. Classical results (see the next section) require either a smooth (C^2) or a polygonal boundary with convex corners. As can be deduced from [9], a combination of both is sufficient also.

(b) In the embedding domain method we treat the problem on a fixed domain $\hat{\Omega}$ which satisfies $\Omega_\gamma \subset \hat{\Omega}$ for all admissible domains $\Omega_\gamma := \Omega(\gamma) \subset \mathbb{R}^2$, where γ is a parameter describing the shape of Ω_γ. To ensure existence of a solution to the state and adjoint equations on the "fictitious" part $\Omega_\gamma^c := \hat{\Omega} \setminus \bar{\Omega}_\gamma$ (see Figure 1 and section 5) we have to guarantee that Ω_γ^c is Lipschitz.

To ensure that a variable polygonal boundary retains its convex corners is rather technical. Thus we choose the following model configuration, which of course may be generalized according to the two points above. As in [7] the boundary $\partial\Omega_\gamma$ shall consist of the following:

- A fixed part Γ, which is the union of the two lateral sides and the top side of the unit square, i.e., the three segments $[(0,0),(0,1)], [(0,1),(1,1)], [(1,1),(1,0)]$. Thus Γ is a polygon with convex angles.
- A variable part Γ_γ, which is the graph of a function $\gamma : [0,1] \to [0,1)$ with $\gamma(0) = \gamma(1) = 0$; compare Figure 1, left. To guarantee that Ω_γ^c is Lipschitz we assume that γ is linear in neighborhoods of the two end points $(0,0)$ and $(1,0)$. Working in Sobolev spaces we choose $\gamma \in H^3(I)$ with $I := (0,1)$, which by classical embedding theorems ensures $\gamma \in C^2(I)$. To show existence

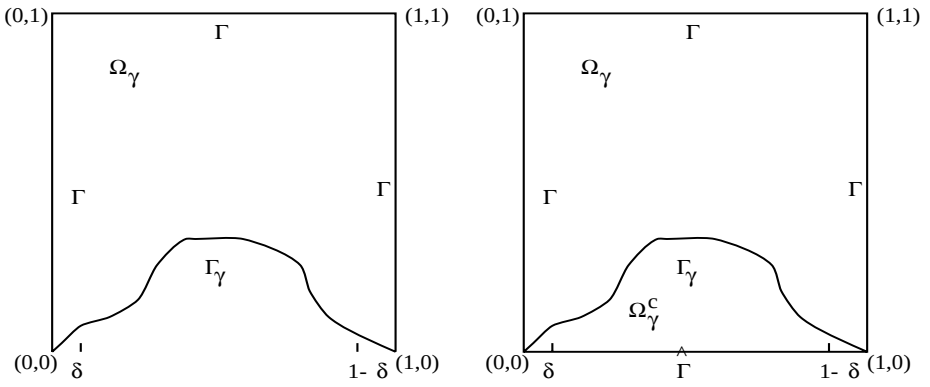


FIG. 1. *The domain Ω_γ in original version (left) and embedded into $\hat{\Omega}$ (right).*

of a solution of the considered domain optimization problems we assume boundedness in $H^3(I)$. The set of admissible functions γ is chosen as

$$\begin{aligned}(2.1)\quad \mathcal{S} := \{\gamma \in H^3(I) : \|\gamma\|_{H^3(I)} \le c_0, \gamma(0) = \gamma(1) = 0,\\ c_1 \le \gamma|_{(\delta,1-\delta)} \le c_2, \gamma'|_{(0,\delta)} = c_3, \gamma'|_{(1-\delta,1)} = c_4\},\end{aligned}$$

where $c_1, c_2 \in (0,1), \delta \in (0, \frac{1}{2}), c_0, c_3 \in \mathbb{R}^+, c_4 \in \mathbb{R}^-$ are fixed. The fact that we choose constant slopes c_3, c_4 here is just for simplicity.

Throughout this paper all considered functions γ shall be in $\mathcal{S}$ except where noted.

3. The Navier–Stokes equations. The stationary incompressible Navier–Stokes equations on a domain $\Omega_\gamma \subset \mathbb{R}^2$ in variational formulation read as follows: find the pair of velocity vector and pressure $(\mathbf{u}_\gamma, p_\gamma) \in H^1(\Omega_\gamma)^2 \times L^2(\Omega_\gamma)$ such that

$$\begin{aligned}\nu(\nabla \mathbf{u}_\gamma, \nabla \mathbf{v})_{\Omega_\gamma} + (\mathbf{u}_\gamma \cdot \nabla \mathbf{u}_\gamma, \mathbf{v})_{\Omega_\gamma} - (p_\gamma, \operatorname{div} \mathbf{v})_{\Omega_\gamma} &= (\mathbf{f}_\gamma, \mathbf{v})_{\Omega_\gamma} && \text{for all } \mathbf{v} \in H_0^1(\Omega_\gamma)^2,\\ (\operatorname{div} \mathbf{u}_\gamma, q)_{\Omega_\gamma} &= 0 && \text{for all } q \in L_0^2(\Omega_\gamma),\\ \mathbf{u}_\gamma &= \Phi && \text{on } \Gamma,\\ \mathbf{u}_\gamma &= \mathbf{0} && \text{on } \Gamma_\gamma,\end{aligned} \tag{3.1}$$

with $L_0^2(\Omega_\gamma) := \{q \in L^2(\Omega_\gamma) : \int_{\Omega_\gamma} q\,dx = 0\}$. The parameter $\nu > 0$ represents the inverse of the Reynolds number. For scalar-valued functions $(\cdot,\cdot)_{\Omega_\gamma}$ denotes the $L^2(\Omega_\gamma)$ inner product; for vector-valued functions we define $(\mathbf{u}, \mathbf{v})_{\Omega_\gamma} := \sum_{i=1}^2 (u_i, v_i)_{\Omega_\gamma}$ and $(\nabla \mathbf{u}, \nabla \mathbf{v})_{\Omega_\gamma} := \sum_{i,j=1}^2 (\frac{\partial u_i}{\partial x_j}, \frac{\partial v_i}{\partial x_j})_{\Omega_\gamma}$. By $|\mathbf{v}|_{H^1(\Omega_\gamma)^2} := (\nabla \mathbf{v}, \nabla \mathbf{v})_{\Omega_\gamma}^{1/2}$ we denote the H^1 seminorm. The nonlinearity is defined by the operator $(\mathbf{u} \cdot \nabla) := \sum_{j=1}^2 u_j \frac{\partial}{\partial x_j}$.

To obtain the needed regularity of the solution we assume $\mathbf{f}_\gamma \in L^2(\Omega_\gamma)^2$; to prove the formula for the derivative in section 7 we will need $\mathbf{f}_\gamma \in L^p(\Omega_\gamma)^2$ with $p > 2$. The Dirichlet boundary conditions indicate that the variable boundary part Γ_γ is a wall with no-slip conditions, whereas the fixed part Γ may be either a wall (if $\Phi = \mathbf{0}$) or a region with prescribed in or outflow velocity. The function Φ shall be in the space

$$H(\Gamma) := \left\{\Phi \in L^2(\Gamma)^2 : \text{there is } \bar{\mathbf{u}}_\gamma \in H^2(\Omega_\gamma)^2 : \operatorname{div} \bar{\mathbf{u}}_\gamma = 0 \text{ in } \Omega_\gamma, \bar{\mathbf{u}}_\gamma|_{\Gamma_\gamma} = \mathbf{0}, \bar{\mathbf{u}}_\gamma|_\Gamma = \Phi\right\}.$$

By our geometric definitions this space is independent of γ. As a consequence Φ satisfies $\int_\Gamma \Phi \cdot \mathbf{n}\,ds = 0$, where $\mathbf{n}$ is the normal vector on Γ.

Existence and uniqueness of the pressure (viewed as Lagrange multiplier corresponding to the constraint of zero divergence) rely on the surjectivity of the weak divergence operator as mapping from $H_0^1(\Omega_\gamma)^2$ onto $L_0^2(\Omega_\gamma)$; see [8, Lemma I.4.1].

LEMMA 3.1. *For every velocity vector* $\mathbf{u}_\gamma \in H^1(\Omega_\gamma)^2$ *solving* (3.1) *the corresponding pressure* p_γ *is unique in* $L^2(\Omega_\gamma)/\mathbb{R}$.

Proof. See [8, Theorem IV.1.4, Corollary I.2.4]. □

The additive constant in the pressure regarded as a function in $L^2(\Omega_\gamma)$ now can be chosen such that $p_\gamma \in L_0^2(\Omega_\gamma)$. This space is often used since—endowed with the L^2 norm—it can be identified isomorphically with $L^2(\Omega_\gamma)/\mathbb{R}$.

Uniqueness of the velocity component of a solution to (3.1) depends on a property of the nonlinear term.

LEMMA 3.2. *Let* $\Omega \subset \mathbb{R}^n, n \le 4$, *be bounded. Then*

$$(\mathbf{u} \cdot \nabla \mathbf{v}, \mathbf{w})_\Omega \le k|\mathbf{u}|_{H^1(\Omega)^2}|\mathbf{v}|_{H^1(\Omega)^2}|\mathbf{w}|_{H^1(\Omega)^2} \quad \textit{for all } \mathbf{u}, \mathbf{w} \in H_0^1(\Omega)^2, \mathbf{v} \in H^1(\Omega)^2,$$

where the constant $k = \frac{1}{2}|\Omega|^{1/2}$ *depends only on* Ω *(*$|\Omega|$ *denotes the measure of* Ω*).*

Proof. See [13, Lemma VIII.1.1]. □

This leads to the following result.

THEOREM 3.3. *Problem* (3.1) *has a unique solution* $(\mathbf{u}_\gamma, p_\gamma) \in H^1(\Omega_\gamma)^2 \times L^2(\Omega_\gamma)/\mathbb{R}$ *if* $\nu > \nu_1 := |\mathbf{u}_\gamma|_{H^1(\Omega_\gamma)^2}/k_\gamma$ *for* $k_\gamma = \frac{1}{2}|\Omega_\gamma|^{1/2}$. *In the case* $\mathbf{f}_\gamma = \mathbf{0}$ *and* $\Phi = \mathbf{0}$, *uniqueness is given without any restriction on* ν.

Proof. See [13, Theorem VIII.2.1]; the result is also valid with weaker assumptions on $\mathbf{f}_\gamma$ and Φ. □

Regularity results for the Navier–Stokes equations are based on those for the Stokes equations by treating the nonlinearity as an additional inhomogeneity and using embedding and function space interpolation theorems; see, e.g., [9, Theorem pp. 403–404] and [10, Prop. II.1.1, Remarks II.1.4, II.1.6]. This implies that the geometric requirements for the regularity and uniform boundedness of the solution (with respect to γ) are the same as for the Stokes equations; see [7, Theorem 3.1].

THEOREM 3.4. *Every solution of* (3.1) *satisfies* $(\mathbf{u}_\gamma, p_\gamma) \in H^2(\Omega_\gamma)^2 \times H^1(\Omega_\gamma)$. *Moreover there exists* $C > 0$ *independent of* $\gamma, \mathbf{f}_\gamma$, *and* Φ *such that*

$$\|\mathbf{u}_\gamma\|_{H^2(\Omega_\gamma)^2} + \|p_\gamma\|_{H^1(\Omega_\gamma)} \le C \left(\|\mathbf{f}_\gamma\|_{L^2(\Omega_\gamma)^2} + \|\Phi\|_{L^\infty(\Gamma)^2} \right).$$

Using this result a lower bound $\nu_0(\gamma, \Phi, \mathbf{f}_\gamma)$ for the constant $\nu_1 = \nu_1(\mathbf{u}_\gamma)$ defined in Theorem 3.3 can be given. Thus uniqueness of the velocity is given if $\nu > \nu_0$, where ν_0 depends on γ, Φ, and $\mathbf{f}_\gamma$. See also [8, Theorem IV.2.4].

4. A class of domain optimization problems. We consider domain optimization problems of the form

$$\min_{\gamma \in \mathcal{S}} \mathcal{J}(\gamma) \;:=\; \min_{\gamma \in \mathcal{S}} \frac{1}{2} \|\mathcal{A}(\mathbf{u}_\gamma) - \mathbf{u}_d\|^2_{L^2(\Omega_C)^k} \quad \text{subject to} \quad \mathbf{u}_\gamma \text{ solves (3.1)} \tag{4.1}$$

with $\mathcal{A} \in \mathcal{L}(H^1(\Omega_\gamma)^2, L^2(\Omega_C)^2), \mathbf{u}_d \in L^2(\Omega_C)^2$ or $\mathcal{A} \in \mathcal{L}(H^1(\Omega_\gamma)^2, L^2(\Omega_C)^{2\times 2}), \mathbf{u}_d \in L^2(\Omega_C)^{2\times 2}$, where $L^2(\Omega_C)^{2\times 2}$ denotes the space of (2×2) matrix-valued functions. The set $\Omega_C \subset \Omega_\gamma$ is assumed to satisfy $\operatorname{dist}(\Gamma_\gamma, \Omega_C) > 0$ for all $\gamma \in \mathcal{S}$. The dependence of $\mathcal{J}$ on γ is implicit due to the fact that $\mathcal{J}$ depends on $\mathbf{u}_\gamma$, which itself depends on γ. The above definition of the cost functional includes typical choices such as the *tracking type* functional

$$\mathcal{J}(\gamma) := \frac{1}{2} \|\mathbf{u}_\gamma - \mathbf{u}_d\|^2_{L^2(\Omega_C)^2}$$

or the *minimum drag problem*; compare Pironneau [3], Gunzburger and Kim [4]:

$$\mathcal{J}(\gamma) := \frac{\nu}{2} \|\nabla \mathbf{u}_\gamma + (\nabla \mathbf{u}_\gamma)^T\|^2_{L^2(\Omega_C)^{2\times 2}} \;:=\; \frac{\nu}{2} (\nabla \mathbf{u}_\gamma + (\nabla \mathbf{u}_\gamma)^T, \nabla \mathbf{u}_\gamma + (\nabla \mathbf{u}_\gamma)^T)_{\Omega_C}.$$

A regularization term penalizing the L^2- or H^1-norm of γ may be added to the cost functional. We skip it in our theoretical investigations since its differentiability is obtained in a straightforward way.

5. The embedding domain method. To solve the domain optimization problem (4.1) by a gradient-based iterative scheme it is necessary to discretize the domain, assemble the system matrices and nonlinear operators, and solve the outgoing nonlinear system in each iteration step. To reduce this effort an equivalent formulation of the Navier–Stokes equations (3.1) on a fixed domain is used: a so-called *fictitious domain* $\hat{\Omega}$ is introduced. It is chosen in such a way that all admissible domains can be embedded in it, i.e., $\Omega_\gamma \subset \hat{\Omega}$ for all $\gamma \in \mathcal{S}$. Furthermore, the fixed boundary part

Γ shall remain a part of $\partial\hat{\Omega}$ whereas Γ_γ is replaced by a partition called $\hat{\Gamma}$, which now is also fixed. Thus $\partial\hat{\Omega} = \bar{\Gamma} \cup \bar{\hat{\Gamma}}$; compare Figure 1 (right). The "fictitious" part of $\hat{\Omega}$ is denoted by $\Omega_\gamma^c := \hat{\Omega} \setminus \bar{\Omega}_\gamma$.

Now we introduce an equivalent formulation of (3.1) on $\hat{\Omega}$. The boundary condition on Γ_γ in the Navier–Stokes equations now becomes a constraint on an inner line of the fictitious domain $\hat{\Omega}$ and is treated similarly to the constraint of zero divergence in (3.1). Analogously to the pressure, which can be regarded as a Lagrange multiplier corresponding to this constraint, an additional multiplier g_γ corresponding to the former boundary condition $\mathbf{u}_\gamma = \mathbf{0}$ on Γ_γ is introduced.

The resulting *fictitious domain formulation of the Navier–Stokes equations* reads as follows: find $(\hat{\mathbf{u}}_\gamma, \hat{p}_\gamma, g_\gamma) \in H^1(\hat{\Omega})^2 \times L_0^2(\hat{\Omega}) \times H_\gamma^*$ such that

$$
\begin{aligned}
\left.\begin{aligned} \nu(\nabla\hat{\mathbf{u}}_\gamma, \nabla\hat{\mathbf{v}})_{\hat{\Omega}} + (\hat{\mathbf{u}}_\gamma \cdot \nabla\hat{\mathbf{u}}_\gamma, \hat{\mathbf{v}})_{\hat{\Omega}} \\ -(\hat{p}_\gamma, \operatorname{div}\hat{\mathbf{v}})_{\hat{\Omega}} - \langle g_\gamma, \tau_\gamma\hat{\mathbf{v}}\rangle_{H_\gamma^*, H_\gamma} \end{aligned}\right\} &= (\tilde{\mathbf{f}}_\gamma, \hat{\mathbf{v}})_{\hat{\Omega}} \qquad \text{for all } \hat{\mathbf{v}} \in H_0^1(\hat{\Omega})^2, \\
(\operatorname{div}\hat{\mathbf{u}}_\gamma, \hat{q})_{\hat{\Omega}} &= 0 \qquad \text{for all } \hat{q} \in L_0^2(\hat{\Omega}), \\
\hat{\mathbf{u}}_\gamma &= \mathbf{0} \qquad \text{on } \Gamma_\gamma, \\
\hat{\mathbf{u}}_\gamma &= \tilde{\Phi} \qquad \text{on } \partial\hat{\Omega}.
\end{aligned} \tag{5.1}
$$

Here τ_γ is the inner trace operator $\hat{\mathbf{v}} \mapsto \hat{\mathbf{v}}|_{\Gamma_\gamma}$, which is linear and continuous from $H^1(\hat{\Omega})^2$ into $H^{1/2}(\Gamma_\gamma)^2$ and from $H_0^1(\hat{\Omega})^2$ onto

$$
H_\gamma := H_{00}^{1/2}(\Gamma_\gamma)^2 = \big\{\mathbf{h} \in H^{1/2}(\Gamma_\gamma)^2 : \text{there is } \tilde{\mathbf{h}} \in H^{1/2}(\partial\Omega_\gamma)^2 : \tilde{\mathbf{h}}|_{\Gamma_\gamma} = \mathbf{h}, \tilde{\mathbf{h}}|_\Gamma = \mathbf{0}\big\};
$$

compare with [11, VII, section 2.1, Remark 1]. To guarantee equivalence between (3.1) and (5.1)—in a sense that is necessary for the derivative formula—inhomogeneity and boundary values have to be extended *by zero* onto the fictitious part of the domain and its boundary, respectively. Moreover the regularity of Theorem 3.4 is needed. In domain optimization problems such as (3.1) it makes sense to assume that the inhomogeneity does not explicitly depend on the shape of the domain but is given by a function $\mathbf{f}$ defined on a domain containing all admissible Ω_γ, let us say $\hat{\Omega}$. Summarizing, we thus assume that

$$
\tilde{\mathbf{f}}_\gamma := \left\{ \begin{array}{lll} \mathbf{f}_\gamma := \mathbf{f}|_{\Omega_\gamma} & \text{in } \Omega_\gamma \\ \mathbf{0} & \text{in } \Omega_\gamma^c \end{array} \right\} \quad \text{and} \quad \tilde{\Phi} := \left\{ \begin{array}{ll} \Phi & \text{on } \Gamma \\ \mathbf{0} & \text{on } \Gamma_\gamma \end{array} \right\}, \tag{5.2}
$$

where $\mathbf{f} \in L^2(\hat{\Omega})^2$. As a consequence $\tilde{\mathbf{f}}_\gamma \in L^2(\hat{\Omega})^2$ and $\tilde{\Phi} \in H^{3/2}(\partial\hat{\Omega})^2$, $\int_{\partial\hat{\Omega}} \tilde{\Phi}\cdot\mathbf{n}\, ds = 0$.

To obtain uniqueness of the Lagrange multipliers $\hat{p}_\gamma, g_\gamma$ we introduce the space

$$
L_*^2(\hat{\Omega}) := \big\{\hat{q} \in L_0^2(\hat{\Omega}) : \hat{q}|_\Sigma = 0\big\} \quad \text{for } \Sigma := \{(x,y) \in \hat{\Omega} : 0 < x < \delta, 0 < y < c_3 x\}
$$

with δ, c_3 defined in (2.1). We have $\Sigma \subset \Omega_\gamma^c$ and $|\Sigma| > 0$ for all $\gamma \in \mathcal{S}$. Thus if $\hat{p} \in L_*^2(\hat{\Omega})$ satisfies $\hat{p}|_{\Omega_\gamma^c} = 0$ in $L^2(\Omega_\gamma^c)/\mathbb{R}$, then $\hat{p}|_{\Omega_\gamma^c} = 0$ even in $L^2(\Omega_\gamma^c)$. We use this fact to show the equivalence of the fictitious domain formulation and (3.1).

THEOREM 5.1. *Let $\mathbf{n}_\gamma$ denote the outer (with respect to Ω_γ) normal vector on Γ_γ. Then $(\hat{\mathbf{u}}_\gamma, \hat{p}_\gamma, g_\gamma) \in H^1(\hat{\Omega})^2 \times L_*^2(\hat{\Omega}) \times H_\gamma^*$ is a solution to* (5.1) *if and only if*

(a) $(\mathbf{u}_\gamma, p_\gamma) := (\hat{\mathbf{u}}_\gamma, \hat{p}_\gamma)|_{\Omega_\gamma} \in H^1(\Omega_\gamma)^2 \times L_0^2(\Omega_\gamma)$ *is a solution to* (3.1),

(b) $(\hat{\mathbf{u}}_\gamma, \hat{p}_\gamma)|_{\Omega_\gamma^c} = (\mathbf{0}, 0)$ *in* $H^1(\Omega_\gamma^c)^2 \times L^2(\Omega_\gamma^c)$, *and*

$$
\langle g_\gamma, \mathbf{h}\rangle_{H_\gamma^*, H_\gamma} = \left(\nu\frac{\partial\mathbf{u}_\gamma}{\partial\mathbf{n}_\gamma} - p_\gamma\mathbf{n}_\gamma, \mathbf{h}\right)_{\Gamma_\gamma} \qquad \textit{for all } \mathbf{h} \in H_\gamma. \tag{5.3}
$$

Proof. (a) Clearly $\mathbf{u}_\gamma := \hat{\mathbf{u}}_\gamma|_{\Omega_\gamma}$ satisfies the boundary conditions in (3.1). We take any $(\mathbf{v}, q) \in H_0^1(\Omega_\gamma)^2 \times L_0^2(\Omega_\gamma)$ and denote by $(\tilde{\mathbf{v}}, \tilde{q})$ its extension by zero onto $\hat{\Omega}$, which clearly is in $H_0^1(\hat{\Omega})^2 \times L_0^2(\hat{\Omega})$. Testing (5.1) with this pair we obtain that $(\mathbf{u}_\gamma, p_\gamma) \in H^1(\Omega_\gamma)^2 \times L^2(\Omega_\gamma)$ with $p_\gamma := \hat{p}_\gamma|_{\Omega_\gamma}$ is a solution to (3.1).

Taking the extensions by zero of any $(\mathbf{v}, q) \in H_0^1(\Omega_\gamma^c)^2 \times L_0^2(\Omega_\gamma^c)$ and proceeding in the same way (5.1) implies that $(\hat{\mathbf{u}}_\gamma, \hat{p}_\gamma)|_{\Omega_\gamma^c} \in H^1(\Omega_\gamma^c)^2 \times L^2(\Omega_\gamma^c)$ is a solution to the homogeneous Navier–Stokes equations on Ω_γ^c with homogeneous boundary conditions. By Theorem 3.3 we have uniqueness and $(\hat{\mathbf{u}}_\gamma, \hat{p}_\gamma)|_{\Omega_\gamma^c} = (\mathbf{0}, 0)$ in $H^1(\Omega_\gamma^c)^2 \times L^2(\Omega_\gamma^c)/\mathbb{R}$. Here it is required that also Ω_γ^c is Lipschitz which is guaranteed by (2.1). Now $\hat{p}_\gamma \in L_*^2(\hat{\Omega})$ implies $\hat{p}_\gamma|_{\Omega_\gamma^c} = 0$ and $p_\gamma \in L_0^2(\Omega_\gamma)$. Thus the first equation of (5.1) reads as

$$\left.\begin{array}{l} \nu(\nabla \mathbf{u}_\gamma, \nabla \hat{\mathbf{v}})_{\Omega_\gamma} + (\mathbf{u}_\gamma \cdot \nabla \mathbf{u}_\gamma, \hat{\mathbf{v}})_{\Omega_\gamma} \\ -(p_\gamma, \operatorname{div} \hat{\mathbf{v}})_{\Omega_\gamma} - \langle g_\gamma, \tau_\gamma \hat{\mathbf{v}} \rangle_{H_\gamma^*, H_\gamma} \end{array}\right\} = (\mathbf{f}_\gamma, \hat{\mathbf{v}})_{\Omega_\gamma} \quad \text{for all } \hat{\mathbf{v}} \in H_0^1(\hat{\Omega})^2. \tag{5.4}$$

Applying Green's formula (see [8, Lemma I.1.4 and eq. I.(2.17)]) on Ω_γ we get

$$\nu(\nabla \mathbf{u}_\gamma, \nabla \hat{\mathbf{v}})_{\Omega_\gamma} - (p_\gamma, \operatorname{div} \hat{\mathbf{v}})_{\Omega_\gamma} = (-\nu \triangle \mathbf{u}_\gamma + \nabla p_\gamma, \hat{\mathbf{v}})_{\Omega_\gamma} + \left(\nu \frac{\partial \mathbf{u}_\gamma}{\partial \mathbf{n}_\gamma} - p_\gamma \mathbf{n}_\gamma, \hat{\mathbf{v}} \right)_{\Gamma_\gamma}$$

for all $\hat{\mathbf{v}} \in H_0^1(\hat{\Omega})^2$ and thus

$$\left.\begin{array}{l} (-\nu \triangle \mathbf{u}_\gamma + \mathbf{u}_\gamma \cdot \nabla \mathbf{u}_\gamma + \nabla p_\gamma - \mathbf{f}_\gamma, \hat{\mathbf{v}})_{\Omega_\gamma} \\ + \left(\nu \dfrac{\partial \mathbf{u}_\gamma}{\partial \mathbf{n}_\gamma} - p_\gamma \mathbf{n}_\gamma, \hat{\mathbf{v}} \right)_{\Gamma_\gamma} - \langle g_\gamma, \tau_\gamma \hat{\mathbf{v}} \rangle_{H_\gamma^*, H_\gamma} \end{array}\right\} = 0 \quad \text{for all } \hat{\mathbf{v}} \in H_0^1(\hat{\Omega})^2. \tag{5.5}$$

Testing this equation with the extension of $\mathbf{v} \in H_0^1(\Omega_\gamma)^2$ by zero onto $\hat{\Omega}$ we get

$$(-\nu \triangle \mathbf{u}_\gamma + \mathbf{u}_\gamma \cdot \nabla \mathbf{u}_\gamma + \nabla p_\gamma - \mathbf{f}_\gamma, \mathbf{v})_{\Omega_\gamma} = 0 \quad \text{for all } \mathbf{v} \in H_0^1(\Omega_\gamma)^2,$$

which means $-\nu \triangle \mathbf{u}_\gamma + \mathbf{u}_\gamma \cdot \nabla \mathbf{u}_\gamma + \nabla p_\gamma = \mathbf{f}_\gamma$ in $H^{-1}(\Omega_\gamma)^2$ and by regularity even in $L^2(\Omega_\gamma)^2$. Now we test (5.5) with any $\hat{\mathbf{v}} \in H_0^1(\hat{\Omega})^2$. Thus $\tau_\gamma \hat{\mathbf{v}} \in H_\gamma$ and by the surjectivity of τ_γ we obtain (5.3).

(b) Obviously $\hat{\mathbf{u}}_\gamma$ satisfies the boundary conditions and the second equation in (3.1). Since $(\hat{\mathbf{u}}_\gamma, \hat{p}_\gamma)|_{\Omega_\gamma^c} := (\mathbf{0}, 0)$, the first equation in (5.1) is equivalent to (5.4) and, again with Green's formula on Ω_γ, to (5.5). Now (3.1) implies

$$-\nu \triangle \mathbf{u}_\gamma + \mathbf{u}_\gamma \cdot \nabla \mathbf{u}_\gamma + \nabla p_\gamma - \mathbf{f}_\gamma = \mathbf{0} \quad \text{in } L^2(\Omega_\gamma)^2,$$

which together with (5.3) gives (5.5). Obviously $p_\gamma \in L_0^2(\Omega_\gamma)$ implies $\hat{p}_\gamma \in L_*^2(\hat{\Omega})$. Equation (5.3) defines $g_\gamma \in H_\gamma^*$ since

$$\begin{aligned} \left(\nu \frac{\partial \mathbf{u}_\gamma}{\partial \mathbf{n}_\gamma} - p_\gamma \mathbf{n}_\gamma, \mathbf{h} \right)_{\Gamma_\gamma} &\le \left(\nu \left\| \frac{\partial \mathbf{u}_\gamma}{\partial \mathbf{n}_\gamma} \right\|_{L^2(\Gamma_\gamma)^2} + \|p_\gamma \mathbf{n}_\gamma\|_{L^2(\Gamma_\gamma)} \right) \|\mathbf{h}\|_{L^2(\Gamma_\gamma)^2} \\ &\le C \left(\nu |\mathbf{u}_\gamma|_{H^1(\Omega_\gamma)^2} + \|p_\gamma\|_{L^2(\Gamma_\gamma)} \right) \|\mathbf{h}\|_{H^{1/2}(\Gamma_\gamma)^2}. \end{aligned}$$

□

Remark 5.2. Formula (5.3) is due to the assumptions (5.2) and $\hat{p}_\gamma \in L_*^2(\hat{\Omega})$ (as a correction to [12] where $\hat{p}_\gamma \in L_0^2(\hat{\Omega})$ was used). Without (5.2) the Lagrange multiplier g_γ equals the jump of the right-hand side of $\nu \frac{\partial \mathbf{u}_\gamma}{\partial \mathbf{n}_\gamma} - p_\gamma \mathbf{n}_\gamma$ along Γ_γ; see [1].

We state two direct consequences of the result above and Lemma 3.1 and Theorem 3.3, respectively.

COROLLARY 5.3. *For every solution component* $\hat{\mathbf{u}}_\gamma$ *in* (5.1) *the corresponding pair* $(\hat{p}_\gamma, g_\gamma)$ *is unique in* $L^2_*(\hat{\Omega}) \times H^*_\gamma$.

COROLLARY 5.4. *The families* $\{(\hat{\mathbf{u}}_\gamma, \hat{p}_\gamma)\}_{\gamma\in\mathcal{S}}$ *and* $\{\|g_\gamma\|_{H^*_\gamma}\}_{\gamma\in\mathcal{S}}$ *are uniformly bounded in* $H^1(\hat{\Omega})^2 \times L^2(\hat{\Omega})$ *and* $\mathbb{R}$, *respectively. By* (5.3) *the functional* g_γ *can be extended onto* $L^2(\Gamma_\gamma)^2$. *The family* $\{\|g_\gamma\|_{L^2(\Gamma_\gamma)^2}\}_{\gamma\in\mathcal{S}}$ *is bounded.*

6. Continuous dependence of the solution on the shape of the domain. As a direct consequence of (5.2) the following equation holds for $\gamma, \bar{\gamma} \in \mathcal{S}$ and $I_+ := \{x \in I : \bar{\gamma}(x) \geq \gamma(x)\}, I_- := \{x \in I : \bar{\gamma}(x) < \gamma(x)\}$:

$$\tilde{\mathbf{f}}_{\bar{\gamma}}(x,y) - \tilde{\mathbf{f}}_\gamma(x,y) = \begin{cases} \mathbf{f}(x,y), & x \in I_+, \gamma(x) \leq y \leq \bar{\gamma}(x), \\ -\mathbf{f}(x,y), & x \in I_-, \bar{\gamma}(x) \leq y \leq \gamma(x), \\ \mathbf{0} & \text{otherwise.} \end{cases} \tag{6.1}$$

To show continuity of solutions to (5.1) with respect to γ we need the following results.

LEMMA 6.1. *Let* $\hat{\mathbf{u}}_\gamma, \hat{\mathbf{u}}_{\bar{\gamma}}$ *denote components of solutions to* (5.1). *Then*

$$(\tilde{\mathbf{f}}_{\bar{\gamma}} - \tilde{\mathbf{f}}_\gamma, \hat{\mathbf{u}}_{\bar{\gamma}} - \hat{\mathbf{u}}_\gamma)_{\hat{\Omega}} \leq \sqrt{2}\,\|\mathbf{f}\|_{L^2(\hat{\Omega})^2} |\hat{\mathbf{u}}_{\bar{\gamma}} - \hat{\mathbf{u}}_\gamma|_{H^1(\hat{\Omega})^2} \|\bar{\gamma} - \gamma\|_{L^\infty(I)}.$$

Proof. Using (6.1) we obtain for $\hat{\mathbf{u}} := \hat{\mathbf{u}}_{\bar{\gamma}} - \hat{\mathbf{u}}_\gamma$

$$(\tilde{\mathbf{f}}_{\bar{\gamma}} - \tilde{\mathbf{f}}_\gamma, \hat{\mathbf{u}})_{\hat{\Omega}} = \int_{I_+} \int_\gamma^{\bar{\gamma}} \mathbf{f}(x,y) \cdot \hat{\mathbf{u}}(x,y)\, dydx - \int_{I_-} \int_{\bar{\gamma}}^{\gamma} \mathbf{f}(x,y) \cdot \hat{\mathbf{u}}(x,y)\, dydx =: A - B.$$

We get with $\hat{\mathbf{u}}(x,\gamma(x)) = \mathbf{0}$ a.e. in I_+

$$\begin{aligned} |A| &\leq \int_{I_+} \int_\gamma^{\bar{\gamma}} |\mathbf{f}(x,y) \cdot [\hat{\mathbf{u}}(x,y) - \hat{\mathbf{u}}(x,\gamma)]|\, dydx \\ &\leq \|\mathbf{f}\|_{L^2(\hat{\Omega})^2} \left(\int_{I_+} \int_\gamma^{\bar{\gamma}} \left\| \int_\gamma^y \frac{\partial \hat{\mathbf{u}}}{\partial y}(x,\xi)\, d\xi \right\|_2^2 dydx \right)^{1/2} \\ &\leq \|\mathbf{f}\|_{L^2(\hat{\Omega})^2} \left(\int_{I_+} \int_\gamma^{\bar{\gamma}} |y-\gamma| \int_\gamma^y \left\| \frac{\partial \hat{\mathbf{u}}}{\partial y}(x,\xi) \right\|_2^2 d\xi dydx \right)^{1/2} \\ &\leq \|\mathbf{f}\|_{L^2(\hat{\Omega})^2} \left(\int_{I_+} \int_\gamma^{\bar{\gamma}} |y-\gamma| \int_\gamma^{\bar{\gamma}} \left\| \frac{\partial \hat{\mathbf{u}}}{\partial y}(x,\xi) \right\|_2^2 d\xi dydx \right)^{1/2} \\ &\leq \frac{1}{\sqrt{2}} \|\mathbf{f}\|_{L^2(\hat{\Omega})^2} |\hat{\mathbf{u}}|_{H^1(\hat{\Omega})^2} \|\bar{\gamma} - \gamma\|_{L^\infty(I)}. \end{aligned}$$

Note that $\frac{\partial \hat{\mathbf{u}}}{\partial y} = -\frac{\partial \hat{\mathbf{u}}_\gamma}{\partial y}$ on $I_+ \times (\gamma, \bar{\gamma})$, which by Theorem 3.4 is an H^1 function. Thus the innermost integral exists. The integral B is estimated in a similar way using $\hat{\mathbf{u}}(x, \bar{\gamma}(x)) = \mathbf{0}$ a.e. in I_-. □

LEMMA 6.2. *Let* $\hat{\mathbf{u}}_\gamma, g_\gamma, \hat{\mathbf{u}}_{\bar{\gamma}}$ *be components of solutions to* (5.1) *for* $\gamma, \bar{\gamma}$, *respectively. Then there exists* L *independent of* $\gamma, \bar{\gamma}$ *with*

$$|\langle g_\gamma, \tau_\gamma(\hat{\mathbf{u}}_{\bar{\gamma}} - \hat{\mathbf{u}}_\gamma)\rangle_{H^*_\gamma, H_\gamma}| \leq L |\hat{\mathbf{u}}_{\bar{\gamma}} - \hat{\mathbf{u}}_\gamma|_{H^1(\hat{\Omega})^2} \|\bar{\gamma} - \gamma\|_{L^\infty(I)}.$$

Proof. Since $\tau_\gamma \hat{\mathbf{u}}_\gamma, \tau_\gamma \hat{\mathbf{u}}_{\bar{\gamma}} \in H_\gamma$, we have for $\hat{\mathbf{u}} := \hat{\mathbf{u}}_{\bar{\gamma}} - \hat{\mathbf{u}}_\gamma$ that

$$|\langle g_\gamma, \tau_\gamma \hat{\mathbf{u}}\rangle_{H^*_\gamma, H_\gamma}| = |(g_\gamma, \tau_\gamma \hat{\mathbf{u}})_{\Gamma_\gamma}| \;\leq\; \|g_\gamma\|_{L^2(\Gamma_\gamma)^2} \|\tau_\gamma \hat{\mathbf{u}}\|_{L^2(\Gamma_\gamma)^2}.$$

The first term on the right-hand side is bounded independently of γ by Corollary 5.4. For the last term we use $\hat{\mathbf{u}}(x,\bar\gamma(x)) = \mathbf{0}$ a.e. in I_- and $\hat{\mathbf{u}}(x,\gamma(x)) = \mathbf{0}$ a.e. in I_+:

$$\begin{aligned}\|\tau_\gamma\hat{\mathbf{u}}\|^2_{L^2(\Gamma_\gamma)^2} &= \int_{I_-} \|\hat{\mathbf{u}}(x,\bar\gamma(x)) - \hat{\mathbf{u}}(x,\gamma(x))\|_2^2 \sqrt{1+\gamma'(x)^2}\,dx \\ &= \int_{I_-} \left\| \int_\gamma^{\bar\gamma} \frac{\partial\hat{\mathbf{u}}}{\partial y}(x,\xi)\,d\xi \right\|_2^2 \sqrt{1+\gamma'(x)^2}\,dx \\ &\le \|\bar\gamma-\gamma\|^2_{L^\infty(I)} |\hat{\mathbf{u}}|^2_{H^1(\hat\Omega)^2} (1+\|\gamma\|^2_{W^{1,\infty}(I)}).\end{aligned}$$

Because $\mathcal{S}$ is bounded in $H^3(I) \hookrightarrow W^{1,\infty}(I)$ the lemma is proved. ∎

The following property of the nonlinear term is a generalization of [8, Lem. IV.2.2].

LEMMA 6.3. *Let $\Omega \subset \mathbb{R}^n$, $\mathbf{u},\mathbf{v},\mathbf{w} \in H^1(\Omega)^n$ with $\operatorname{div}\mathbf{w} = 0$ in Ω. Then*

$$(\mathbf{w}\cdot\nabla\mathbf{u},\mathbf{v})_\Omega = -(\mathbf{w}\cdot\nabla\mathbf{v},\mathbf{u})_\Omega + (\mathbf{w}\cdot\mathbf{n},\mathbf{u}\cdot\mathbf{v})_{\partial\Omega},$$

where $\mathbf{n}$ denotes the outer normal vector on $\partial\Omega$.

Proof. Since $(\mathbf{w}\cdot\nabla\mathbf{u})\cdot\mathbf{u} = \frac12\mathbf{w}\cdot\nabla(\mathbf{u}\cdot\mathbf{u})$ Green's formula and $\operatorname{div}\mathbf{w} = 0$ give

$$(\mathbf{w}\cdot\nabla\mathbf{u},\mathbf{u})_\Omega = \tfrac12\left[-(\operatorname{div}\mathbf{w},\mathbf{u}\cdot\mathbf{u})_\Omega + (\mathbf{w}\cdot\mathbf{n},\mathbf{u}\cdot\mathbf{u})_{\partial\Omega}\right] = \tfrac12(\mathbf{w}\cdot\mathbf{n},\mathbf{u}\cdot\mathbf{u})_{\partial\Omega}$$

and thus $(\mathbf{w}\cdot\nabla(\mathbf{u}-\mathbf{v}),\mathbf{u}-\mathbf{v})_\Omega = \frac12(\mathbf{w}\cdot\mathbf{n},\mathbf{u}\cdot\mathbf{u}-2\mathbf{u}\cdot\mathbf{v}+\mathbf{v}\cdot\mathbf{v})_{\partial\Omega}$. On the other hand

$$\begin{aligned}(\mathbf{w}\cdot\nabla(\mathbf{u}-\mathbf{v}),\mathbf{u}-\mathbf{v})_\Omega &= (\mathbf{w}\cdot\nabla\mathbf{u},\mathbf{u})_\Omega + (\mathbf{w}\cdot\nabla\mathbf{v},\mathbf{v})_\Omega - (\mathbf{w}\cdot\nabla\mathbf{u},\mathbf{v})_\Omega - (\mathbf{w}\cdot\nabla\mathbf{v},\mathbf{u})_\Omega \\ &= \tfrac12(\mathbf{w}\cdot\mathbf{n},\mathbf{u}\cdot\mathbf{u}+\mathbf{v}\cdot\mathbf{v})_{\partial\Omega} - (\mathbf{w}\cdot\nabla\mathbf{u},\mathbf{v})_\Omega - (\mathbf{w}\cdot\nabla\mathbf{v},\mathbf{u})_\Omega.\end{aligned}$$

Taking the difference of both equations the claim follows. ∎

Assuming the sufficient condition on ν for uniqueness of the velocity in (3.1) we now can show Lipschitz continuity.

THEOREM 6.4. *Let $\nu > \nu_1$ be as in Theorem 3.3. Then the velocity part of the solution to (5.1) is Lipschitz continuous with respect to γ, i.e., there exists L independent of $\gamma,\bar\gamma$ with*

$$|\hat{\mathbf{u}}_{\bar\gamma} - \hat{\mathbf{u}}_\gamma|_{H^1(\hat\Omega)^2} \le L\|\bar\gamma-\gamma\|_{L^\infty(I)} \qquad \text{for all } \bar\gamma,\gamma \in \mathcal{S}.$$

Proof. For $\hat{\mathbf{u}} := \hat{\mathbf{u}}_{\bar\gamma} - \hat{\mathbf{u}}_\gamma \in H^1_0(\hat\Omega)^2$ we have using Lemma 6.3

$$(\hat{\mathbf{u}}_{\bar\gamma}\cdot\nabla\hat{\mathbf{u}}_{\bar\gamma} - \hat{\mathbf{u}}_\gamma\cdot\nabla\hat{\mathbf{u}}_\gamma,\hat{\mathbf{u}})_{\hat\Omega} = (\hat{\mathbf{u}}_{\bar\gamma}\cdot\nabla\hat{\mathbf{u}},\hat{\mathbf{u}})_{\hat\Omega} + (\hat{\mathbf{u}}\cdot\nabla\hat{\mathbf{u}}_\gamma,\hat{\mathbf{u}})_{\hat\Omega} = (\hat{\mathbf{u}}\cdot\nabla\hat{\mathbf{u}}_\gamma,\hat{\mathbf{u}})_{\hat\Omega}.$$

Testing the first equation of (5.1), once for γ and another time for $\bar\gamma$, with $\hat{\mathbf{u}}$ and subtracting both equations thus leads to

$$\nu|\hat{\mathbf{u}}|^2_{H^1(\hat\Omega)^2} = (\tilde{\mathbf{f}}_{\bar\gamma} - \tilde{\mathbf{f}}_\gamma,\hat{\mathbf{u}})_{\hat\Omega} + \langle g_{\bar\gamma},\tau_{\bar\gamma}\hat{\mathbf{u}}\rangle_{H^*_{\bar\gamma},H_{\bar\gamma}} - \langle g_\gamma,\tau_\gamma\hat{\mathbf{u}}\rangle_{H^*_\gamma,H_\gamma} - (\hat{\mathbf{u}}\cdot\nabla\hat{\mathbf{u}}_\gamma,\hat{\mathbf{u}})_{\hat\Omega}.$$

Lemmas 3.2, 6.1, and 6.2 give the estimate

$$\nu|\hat{\mathbf{u}}|^2_{H^1(\hat\Omega)^2} \le L\|\bar\gamma-\gamma\|_{L^\infty(I)}|\hat{\mathbf{u}}|_{H^1(\hat\Omega)^2} + k_\gamma|\hat{\mathbf{u}}_\gamma|_{H^1(\hat\Omega)^2}|\hat{\mathbf{u}}|^2_{H^1(\hat\Omega)^2}$$

with L independent of $\gamma,\bar\gamma$ and k_γ as in Theorem 3.3. This implies

$$\left(\nu - k_\gamma|\hat{\mathbf{u}}_\gamma|_{H^1(\hat\Omega)^2}\right)|\hat{\mathbf{u}}|_{H^1(\hat\Omega)^2} \le L\|\bar\gamma-\gamma\|_{L^\infty(I)}$$

and thus Lipschitz continuity if $(\nu - k_\gamma |\hat{\mathbf{u}}_\gamma|_{H^1(\hat\Omega)^2}) > 0$, which together with $|\hat{\mathbf{u}}_\gamma|_{H^1(\hat\Omega)^2} = |\mathbf{u}_\gamma|_{H^1(\Omega_\gamma)^2}$ exactly is $\nu > \nu_1$ as in Theorem 3.3. ∎

To show continuity of the pressure and the Lagrange multipliers g_γ with respect to γ we introduce the isomorphism

$$\begin{aligned} \mathcal{I}_\gamma &: L^2(\Gamma_\gamma)^2 \to L^2(I)^2, \\ \mathcal{I}_\gamma \mathbf{h}(x) &:= \mathbf{h}(x, \gamma(x)), \qquad \mathbf{h} \in L^2(\Gamma_\gamma)^2, x \in I, \end{aligned}$$

which is also an isomorphism between $H^{1/2}(\Gamma_\gamma)^2$ and $H^{1/2}(I)^2$ and between H_γ and

$$H_I := H_{00}^{1/2}(I)^2 := \left\{ g \in H^{1/2}(I)^2 : \int_I \frac{\|g(x)\|_2^2}{x(1-x)}\, dx < \infty \right\};$$

see [12, Theorem 2.4]. For the latter case the adjoint of $\mathcal{I}_\gamma^{-1}$ is defined as

$$\begin{aligned} \left(\mathcal{I}_\gamma^{-1}\right)^* &: H_\gamma^* \to H_I^*, \\ \langle \left(\mathcal{I}_\gamma^{-1}\right)^* g, \mathbf{h} \rangle_{H_I{}^*, H_I} &:= \langle g, \mathcal{I}_\gamma^{-1} \mathbf{h} \rangle_{H_\gamma^*, H_\gamma}, \qquad g \in H_\gamma^*, \mathbf{h} \in H_I. \end{aligned}$$

We need a result concerning convergence of the transformed trace operators.

LEMMA 6.5. *If $\gamma_n \to \gamma$ in $W^{1,\infty}(I)$, then $\mathcal{I}_{\gamma_n} \tau_{\gamma_n} \to \mathcal{I}_\gamma \tau_\gamma$ as linear operators from $\{\hat{\mathbf{v}} \in H^1(\hat\Omega)^2 : \hat{\mathbf{v}}|_{\hat\Gamma} = \mathbf{0}\}$ into H_I.*

Proof. See [12, Lemma 2.11]. ∎

Now we can show continuity of the pair $(\hat p_\gamma, g_\gamma)$ in (5.1).

THEOREM 6.6. *Let $\gamma_n \to \gamma$ in $W^{1,\infty}(I)$ and $\nu > \nu_1$ as in Theorem* 3.3. *Then the corresponding components of solutions to* (5.1) *satisfy*

$$\begin{aligned} \hat p_{\gamma_n} &\to \hat p_\gamma && \text{in } L^2_*(\hat\Omega), \\ \left(\mathcal{I}_{\gamma_n}^{-1}\right)^* g_{\gamma_n} &\overset{*}{\rightharpoonup} \left(\mathcal{I}_\gamma^{-1}\right)^* g_\gamma && \text{in } \left(L^2(I)^2\right)^*. \end{aligned}$$

Proof. By Corollary 5.4, $\hat p_{\gamma_n} \rightharpoonup \hat p \in L^2(\hat\Omega)$ for a subsequence. With the strong convergence of the velocities (Theorem 6.4) the first equation of (5.1) gives

$$\begin{aligned} \langle g_{\gamma_n}, \tau_{\gamma_n} \hat{\mathbf{v}} \rangle_{H_{\gamma_n}^*, H_{\gamma_n}} &= \nu(\nabla \hat{\mathbf{u}}_{\gamma_n}, \nabla \hat{\mathbf{v}})_{\hat\Omega} + (\hat{\mathbf{u}}_{\gamma_n} \cdot \nabla \hat{\mathbf{u}}_{\gamma_n}, \hat{\mathbf{v}})_{\hat\Omega} - (\hat p_{\gamma_n}, \operatorname{div} \hat{\mathbf{v}})_{\hat\Omega} - (\tilde{\mathbf{f}}_{\gamma_n}, \hat{\mathbf{v}})_{\hat\Omega} \\ &\to \nu(\nabla \hat{\mathbf{u}}, \nabla \hat{\mathbf{v}})_{\hat\Omega} + (\hat{\mathbf{u}} \cdot \nabla \hat{\mathbf{u}}, \hat{\mathbf{v}})_{\hat\Omega} - (\hat p, \operatorname{div} \hat{\mathbf{v}})_{\hat\Omega} - (\tilde{\mathbf{f}}_\gamma, \hat{\mathbf{v}})_{\hat\Omega} \end{aligned}$$

for all $\hat{\mathbf{v}} \in H_0^1(\hat\Omega)^2$. Taking any $\mathbf{h} \in H_I$ there exists $\hat{\mathbf{v}} \in H_0^1(\hat\Omega)^2$ such that $\tau_\gamma \hat{\mathbf{v}} = \mathcal{I}_\gamma^{-1} \mathbf{h}$. We define $G \in H_I^*$ as

$$G : \mathbf{h} \mapsto \nu(\nabla \hat{\mathbf{u}}, \nabla \hat{\mathbf{v}})_{\hat\Omega} + (\hat{\mathbf{u}} \cdot \nabla \hat{\mathbf{u}}, \hat{\mathbf{v}})_{\hat\Omega} - (\hat p, \operatorname{div} \hat{\mathbf{v}})_{\hat\Omega} - (\tilde{\mathbf{f}}_\gamma, \hat{\mathbf{v}})_{\hat\Omega},$$

and $g := \mathcal{I}_\gamma^* G \in H_\gamma^*$. Thus $(\hat{\mathbf{u}}_\gamma, \hat p, g)$ solves (5.1) and due to the uniqueness this implies weak convergence for the whole sequence and $(\hat p, g) = (\hat p_\gamma, g_\gamma)$. Since H_I is dense in $L^2(I)^2$ the second claim follows; see, e.g., [8, Corollary I.2.5, eq. I.(2.14)].

Since the divergence operator is an isomorphism (see, e.g., [8, Corollary I.2.4]), for every n there exists a *unique* function $\hat{\mathbf{v}}_n \in V^\perp \subset H_0^1(\hat\Omega)^2$ such that

$$\operatorname{div} \hat{\mathbf{v}}_n = \hat p_{\gamma_n} - \hat p_\gamma \text{ in } \hat\Omega, \quad \|\mathbf{v}_n\|_{H^1(\hat\Omega)^2} \le c(\hat\Omega) \|\hat p_{\gamma_n} - \hat p_\gamma\|_{L^2(\hat\Omega)} \le C(\hat\Omega),$$

where $V := \{\mathbf{v} \in H_0^1(\hat\Omega)^2 : \operatorname{div} \mathbf{v} = 0 \text{ in } \hat\Omega\}$ and the last inequality holds due to the boundedness of $\{\hat p_\gamma\}_{\gamma \in \mathcal{S}}$ in $L^2(\hat\Omega)$. Testing (5.1) with this function gives

$$\begin{aligned} \|\hat p_{\gamma_n} - \hat p_\gamma\|_{L^2(\hat\Omega)}^2 &= \nu\big(\nabla(\hat{\mathbf{u}}_{\gamma_n} - \hat{\mathbf{u}}_\gamma), \nabla \hat{\mathbf{v}}_n\big)_{\hat\Omega} + (\hat{\mathbf{u}}_{\gamma_n} \cdot \nabla \hat{\mathbf{u}}_{\gamma_n} - \hat{\mathbf{u}}_\gamma \cdot \nabla \hat{\mathbf{u}}_\gamma, \hat{\mathbf{v}}_n)_{\hat\Omega} \\ &\quad - (\tilde{\mathbf{f}}_{\gamma_n} - \tilde{\mathbf{f}}_\gamma, \hat{\mathbf{v}}_n)_{\hat\Omega} - \langle g_{\gamma_n}, \tau_{\gamma_n} \hat{\mathbf{v}}_n \rangle_{H_{\gamma_n}^*, H_{\gamma_n}} + \langle g_\gamma, \tau_\gamma \hat{\mathbf{v}}_n \rangle_{H_\gamma^*, H_\gamma}. \end{aligned}$$

The first term on the right tends to zero because of the Lipschitz continuity of $\{\hat{\mathbf{u}}_\gamma\}_{\gamma\in\mathcal{S}}$ and the boundedness of $\{\hat{\mathbf{v}}_n\}$; the third one tends to zero because of $\|\tilde{\mathbf{f}}_{\gamma_n}-\tilde{\mathbf{f}}_\gamma\|_{L^2(\hat\Omega)^2}\to 0$ due to (5.2). The second term can be estimated using $\hat{\mathbf{u}}_n := \hat{\mathbf{u}}_{\gamma_n}-\hat{\mathbf{u}}_\gamma$ and Lemma 3.2:

$$\begin{aligned}(\hat{\mathbf{u}}_{\gamma_n}\cdot\nabla\hat{\mathbf{u}}_{\gamma_n}-\hat{\mathbf{u}}_\gamma\cdot\nabla\hat{\mathbf{u}}_\gamma,\hat{\mathbf{v}}_n)_{\hat\Omega} &= (\hat{\mathbf{u}}_{\gamma_n}\cdot\nabla\hat{\mathbf{u}}_n,\hat{\mathbf{v}}_n)_{\hat\Omega}+(\hat{\mathbf{u}}_n\cdot\nabla\hat{\mathbf{u}}_\gamma,\hat{\mathbf{v}}_n)_{\hat\Omega}\\ &\le k\big(|\hat{\mathbf{u}}_{\gamma_n}|_{H^1(\hat\Omega)^2}+|\hat{\mathbf{u}}_\gamma|_{H^1(\hat\Omega)^2}\big)|\hat{\mathbf{u}}_n|_{H^1(\hat\Omega)^2}|\hat{\mathbf{v}}_n|_{H^1(\hat\Omega)^2}\end{aligned}$$

and thus tends to zero for the same reasons. The remaining terms we write as

$$\begin{aligned}\langle g_\gamma,\tau_\gamma\hat{\mathbf{v}}_n\rangle_{H^*_\gamma,H_\gamma}-\langle g_{\gamma_n},\tau_{\gamma_n}\hat{\mathbf{v}}_n\rangle_{H^*_{\gamma_n},H_{\gamma_n}} &= \langle\big(\mathcal{I}_\gamma^{-1}\big)^*g_\gamma-\big(\mathcal{I}_{\gamma_n}^{-1}\big)^*g_{\gamma_n},\mathcal{I}_\gamma\tau_\gamma\hat{\mathbf{v}}_n\rangle_{H_I^*,H_I}\\ &\quad+\langle\big(\mathcal{I}_{\gamma_n}^{-1}\big)^*g_{\gamma_n},(\mathcal{I}_\gamma\tau_\gamma-\mathcal{I}_{\gamma_n}\tau_{\gamma_n})\hat{\mathbf{v}}_n\rangle_{H_I^*,H_I}.\end{aligned}$$

Boundedness of $\{\hat{\mathbf{v}}_n\}$ together with its uniqueness now implies $\hat{\mathbf{v}}_n\rightharpoonup\mathbf{0}$ weakly in $H^1_0(\hat\Omega)^2$, and thus $\tau_\gamma\hat{\mathbf{v}}_n\to\mathbf{0}$ in $L^2(\Gamma_\gamma)^2$ and $\mathcal{I}_\gamma\tau_{\gamma_n}\hat{\mathbf{v}}_n\to\mathbf{0}$ in $L^2(I)^2$. Since $\big(\mathcal{I}_{\gamma_n}^{-1}\big)^*g_{\gamma_n}\overset{*}{\rightharpoonup}\big(\mathcal{I}_\gamma^{-1}\big)^*g_\gamma$ in $\big(L^2(I)^2\big)^*$, the first term on the right tends to zero. The second one does so because $\{\big(\mathcal{I}_\gamma^{-1}\big)^*g_\gamma\}_{\gamma\in\mathcal{S}}$ and $\{\hat{\mathbf{v}}_n\}$ are bounded and $\mathcal{I}_{\gamma_n}\tau_{\gamma_n}\to\mathcal{I}_\gamma\tau_\gamma$ strongly due to Lemma 6.5. Thus $\hat p_{\gamma_n}\to\hat p_\gamma$ in $L^2(\hat\Omega)$ and since $L^2_*(\hat\Omega)$ is closed also in this space. ∎

As a consequence the boundedness of $\mathcal{S}$ in $H^3(I)$ now guarantees existence of a solution to the domain optimization problem.

COROLLARY 6.7. *Problem* (4.1) *has at least one solution* $\gamma\in\mathcal{S}$.

Proof. Choose a minimizing sequence and use the compactness of $H^3(I)\hookrightarrow C^2(\bar I)$. ∎

7. Fréchet differentiability and derivative formula. To show differentiability we make use of the adjoint equation of the Navier–Stokes system (3.1). Its derivation is standard and we thus refer to [7, section 7] or [14, section 3.2]: Find $(\lambda_\gamma,\mu_\gamma)\in H^1_0(\Omega_\gamma)^2\times L^2(\Omega_\gamma)$ such that

$$\begin{aligned}\left.\begin{aligned}\nu(\nabla\lambda_\gamma,\nabla\mathbf{v})_{\Omega_\gamma}-(\mu_\gamma,\operatorname{div}\mathbf{v})_{\Omega_\gamma}\\ +(\lambda_\gamma,\mathbf{u}_\gamma\cdot\nabla\mathbf{v}+\mathbf{v}\cdot\nabla\mathbf{u}_\gamma)_{\Omega_\gamma}\end{aligned}\right\} &= -D_u\mathcal{J}(\gamma)\mathbf{v} && \text{for all } \mathbf{v}\in H^1_0(\Omega_\gamma)^2,\\ (\operatorname{div}\lambda_\gamma,q)_{\Omega_\gamma} &= 0 && \text{for all } q\in L^2_0(\Omega_\gamma).\end{aligned}\tag{7.1}$$

Here $\mathbf{u}_\gamma$ is the velocity component of a solution to (3.1), and $D_u\mathcal{J}(\gamma)\mathbf{v}$ denotes the derivative of $\mathcal{J}$ with respect to $\mathbf{u}$ in direction $\mathbf{v}$. Note that the linearized nonlinear term can be rewritten using Green's formula as, e.g., in [14]. As in section 5 we derive a fictitious domain formulation: Find $(\hat\lambda_\gamma,\hat\mu_\gamma,\chi_\gamma)\in H^1_0(\hat\Omega)^2\times L^2(\hat\Omega)\times H^*_\gamma$ such that

$$\begin{aligned}\left.\begin{aligned}\nu(\nabla\hat\lambda_\gamma,\nabla\hat{\mathbf{v}})_{\hat\Omega}+(\hat\lambda_\gamma,\hat{\mathbf{u}}_\gamma\cdot\nabla\hat{\mathbf{v}}+\hat{\mathbf{v}}\cdot\nabla\hat{\mathbf{u}}_\gamma)_{\hat\Omega}\\ -(\hat\mu_\gamma,\operatorname{div}\hat{\mathbf{v}})_{\hat\Omega}-\langle\chi_\gamma,\tau_\gamma\hat{\mathbf{v}}_\gamma\rangle_{H^*_\gamma,H_\gamma}\end{aligned}\right\} &= -D_u\mathcal{J}(\gamma)\hat{\mathbf{v}} && \text{for all } \hat{\mathbf{v}}\in H^1_0(\hat\Omega)^2,\\ (\operatorname{div}\hat\lambda,\hat q)_{\hat\Omega} &= 0 && \text{for all } \hat q\in L^2_0(\hat\Omega),\\ \hat\lambda_\gamma &= \mathbf{0} && \text{on } \Gamma_\gamma.\end{aligned}\tag{7.2}$$

Some main results for the above two systems are summarized in the following.

THEOREM 7.1. *Problem* (7.1) *has a solution* $(\lambda_\gamma,\mu_\gamma)\in[H^2(\Omega_\gamma)^2\cap H^1_0(\Omega_\gamma)^2]\times[H^1(\Omega_\gamma)\cap L^2_0(\Omega_\gamma)]$ *satisfying*

$$\|\lambda_\gamma\|_{H^2(\Omega_\gamma)^2}+\|\mu_\gamma\|_{H^1(\Omega_\gamma)}\le C\|\mathbf{u}_\gamma\|_{L^2(\Omega_\gamma)^2}$$

for some $C>0$ *independent of* γ. *For every* λ_γ *the corresponding* μ_γ *is unique. If* $\nu>\nu_1$ *as in Theorem* 3.3, *also* λ_γ *is unique. Moreover* $(\hat\lambda_\gamma,\hat\mu_\gamma,\chi_\gamma)\in H^1_0(\hat\Omega)^2\times L^2_*(\hat\Omega)\times H^*_\gamma$ *is a solution to* (7.2) *if and only if*

- $(\lambda_\gamma, \mu_\gamma) := (\hat\lambda_\gamma, \hat\mu_\gamma)|_{\Omega_\gamma}$ *is a solution to* (7.1),
- $(\hat\lambda_\gamma, \hat\mu_\gamma)|_{\Omega_\gamma^c} = (\mathbf{0}, 0)$,
- $\langle \chi_\gamma, \mathbf{h}\rangle_{H_\gamma^*, H_\gamma} = (\nu \frac{\partial \lambda_\gamma}{\partial \mathbf{n}_\gamma} - \mu_\gamma \mathbf{n}_\gamma, \mathbf{h})_{\Gamma_\gamma}$ *for all* $\mathbf{h} \in H_\gamma$.

The functional χ_γ *can be extended onto* $L^2(\Gamma_\gamma)^2$. *The families* $\{(\hat\lambda_\gamma, \hat\mu_\gamma)\}_{\gamma\in\mathcal{S}}$ *and* $\{\|\chi_\gamma\|_{H_\gamma^*}\}_{\gamma\in\mathcal{S}}, \{\|\chi_\gamma\|_{L^2(\Gamma_\gamma)^2}\}_{\gamma\in\mathcal{S}}$ *are bounded in* $H_0^1(\hat\Omega)^2 \times L_0^2(\hat\Omega)$ *and* $\mathbb{R}$, *respectively.*

Proof. Existence is shown in [14, Theorem 3.1]; uniqueness follows from continuity and ellipticity of the bilinear form

$$a(\lambda, \mathbf{v}) := \nu(\nabla\lambda, \nabla\mathbf{v})_{\Omega_\gamma} + (\lambda, \mathbf{u}_\gamma \cdot \nabla\mathbf{v} + \mathbf{v}\cdot\nabla\mathbf{u}_\gamma)_{\Omega_\gamma}.$$

Continuity is obvious; for the ellipticity we use Lemmas 3.2 and 6.3 to estimate

$$a(\lambda,\lambda) = \nu|\lambda|^2_{H^1(\Omega_\gamma)^2} - (\lambda\cdot\nabla\lambda, \mathbf{u}_\gamma)_{\Omega_\gamma} \;\geq\; \left(\nu - k_\gamma|\mathbf{u}_\gamma|_{H^1(\Omega_\gamma)^2}\right)|\lambda|^2_{H^1(\Omega_\gamma)^2}$$

and obtain ellipticity under the same condition as for the uniqueness of $\mathbf{u}_\gamma$. For the regularity of the solution of (7.1) see [14, Thm. 3.2]. The generalization for the combination of smooth and convex boundary and the uniform regularity with respect to γ can be obtained as for the (Navier–)Stokes equations; see Theorem 3.4. Equivalence of (7.1) and (7.2) is proved analogously to Theorem 5.1, and the uniform boundedness of the solution is a direct consequence. ∎

For the differentiability of $\mathcal{J}$ with respect to variations in γ we consider $\gamma \in \operatorname{int}\mathcal{S}$, the interior of $\mathcal{S}$, and define the set of admissible directions as

$$\mathcal{S}' := \left\{\bar\gamma \in H^3(I) : \bar\gamma|_{[0,\delta]\cup[1-\delta,1]} = 0\right\}.$$

For every $\gamma \in \operatorname{int}\mathcal{S}$, $\bar\gamma \in \mathcal{S}'$ there exists $t_0 > 0$ such that $\gamma + t\bar\gamma \in \operatorname{int}\mathcal{S}$ for all $t \in [0, t_0)$. Thus we can properly define a directional derivative. We now define $I_+ := \{x \in I : \bar\gamma(x) \geq 0\}, I_- := \{x \in I : \bar\gamma(x) < 0\}$ and state three results needed to prove the derivative formula. The first one is a relaxation of [7, Lemma 7.2].

LEMMA 7.2. *Let* $\hat\lambda_\gamma$ *denote a solution to* (7.2) *and* $\mathbf{f} \in L^p(\hat\Omega)^2$, $p \in (2, \infty]$. *Then there exists* $c = c(\gamma, p)$ *such that for* $\alpha = 2(p-1)/p > 1$ *we have*

$$(\tilde{\mathbf{f}}_{\gamma+t\bar\gamma} - \tilde{\mathbf{f}}_\gamma, \hat\lambda_\gamma)_{\hat\Omega} \leq c\,\|\mathbf{f}\|_{L^p(\hat\Omega)^2}|\lambda_\gamma|_{H^2(\Omega_\gamma)^2}\left(t\|\bar\gamma\|_{L^\infty(I)}\right)^\alpha.$$

Proof. Using (6.1) we obtain, similarly to the proof of Lemma 6.1 for I_+ defined there and $1/p + 1/q = 1$,

$$\begin{aligned}
(\tilde{\mathbf{f}}_{\gamma+t\bar\gamma} - \tilde{\mathbf{f}}_\gamma, \hat\lambda_\gamma)_{\hat\Omega} &\leq \int_{I_+}\int_\gamma^{\gamma+t\bar\gamma} |\mathbf{f}(x,y)\cdot[\lambda_\gamma(x,y) - \lambda_\gamma(x,\gamma)]|\,dydx \\
&\leq \|\mathbf{f}\|_{L^p(\hat\Omega)^2}\left(\int_{I_+}\int_\gamma^{\gamma+t\bar\gamma}\left\|\int_\gamma^y \frac{\partial\lambda_\gamma}{\partial y}(x,\xi)\,d\xi\right\|_2^q dydx\right)^{1/q} \\
&\leq \|\mathbf{f}\|_{L^p(\hat\Omega)^2}\left(\int_{I_+}\int_\gamma^{\gamma+t\bar\gamma}|y-\gamma|\int_\gamma^{\gamma+t\bar\gamma}\left\|\frac{\partial\lambda_\gamma}{\partial y}(x,\xi)\right\|_2^q d\xi dydx\right)^{1/q} \\
&\leq 2^{-1/q}\|\mathbf{f}\|_{L^p(\hat\Omega)^2}|\lambda_\gamma|_{W^{1,q}(\hat\Omega)^2}\left(t\|\bar\gamma\|_{L^\infty(I)}\right)^{2/q}.
\end{aligned}$$

Because of the continuous embedding $H^2(\Omega_\gamma) \hookrightarrow W^{1,q}(\Omega_\gamma)$ the claim follows. ∎

The following two results were already proved in [7, Lemmas 7.3 and 7.4].

LEMMA 7.3. *Let $\hat{\mathbf{u}}_{\gamma+t\bar\gamma}$ and χ_γ be components of solutions to* (5.1) *and* (7.2)*, respectively. Then*

$$\lim_{t\to 0}\frac{1}{t}\langle \chi_\gamma, \tau_\gamma\hat{\mathbf{u}}_{\gamma+t\bar\gamma}\rangle_{H^*_\gamma,H_\gamma} = -\int_{I_+}\chi_\gamma(x,\gamma)\cdot\mathbf{u}_{\gamma,y}(x,\gamma)\bar\gamma\sqrt{1+\gamma'^2}\,dx.$$

The integral exists because $\mathcal{S}'\subset L^\infty(I)$, $\chi_\gamma\in L^2(\Gamma_\gamma)^2$, and $\tau_\gamma\mathbf{u}_{\gamma,y}=\mathbf{u}_{\gamma,y}(x,\gamma)$ and also its restriction on the set $\Gamma_\gamma^+:=\{(x,\gamma(x)):x\in I_+\}\subset\Gamma_\gamma$ are L^2 functions since $\mathbf{u}_\gamma\in H^2(\Omega_\gamma)^2$. The same arguments hold for the integral in the next lemma.

LEMMA 7.4. *Let $g_{\gamma+t\bar\gamma}$ and $\hat\lambda_\gamma$ be components of solutions to* (5.1) *and* (7.2)*, respectively. Then*

$$\lim_{t\to 0}\frac{1}{t}\langle g_{\gamma+t\bar\gamma}, \tau_{\gamma+t\bar\gamma}\hat\lambda_\gamma\rangle_{H^*_{\gamma+t\bar\gamma},H_{\gamma+t\bar\gamma}} = \int_{I_-} g_\gamma(x,\gamma)\cdot\lambda_{\gamma,y}(x,\gamma)\bar\gamma\sqrt{1+\gamma'^2}\,dx.$$

Now differentiability of the cost functional with respect to γ and a formula for the derivative are shown under the restriction on ν given in Theorem 3.3 and a slightly stronger assumption on the inhomogeneity (due to Lemma 7.2).

THEOREM 7.5. *Let $\gamma\in\operatorname{int}\mathcal{S}$,* (5.2) *hold for $\mathbf{f}\in L^p(\hat\Omega)^2$, $p>2$, and $\nu>\nu_1$ as in Theorem* 3.3*. Then $\mathcal{J}$ is Fréchet differentiable and the derivative in γ in direction $\bar\gamma\in\mathcal{S}'$ satisfies*

$$D_\gamma\mathcal{J}(\gamma)\bar\gamma = \frac{1}{\nu}\int_I\big[g_\gamma\big(x,\gamma(x)\big)\cdot\chi_\gamma\big(x,\gamma(x)\big) - p_\gamma\big(x,\gamma(x)\big)\mu_\gamma\big(x,\gamma(x)\big)\big]\,\bar\gamma(x)\,dx. \tag{7.3}$$

Proof. Denoting $\hat{\mathbf{u}}:=\hat{\mathbf{u}}_{\gamma+t\bar\gamma}-\hat{\mathbf{u}}_\gamma\in H_0^1(\hat\Omega)^2$ we have

$$\mathcal{J}(\gamma+t\bar\gamma)-\mathcal{J}(\gamma)=\frac{1}{2}\|\mathcal{A}\hat{\mathbf{u}}\|^2_{L^2(\Omega_C)^k}+D_u\mathcal{J}(\gamma)\hat{\mathbf{u}}$$

and by the Lipschitz continuity of $\hat{\mathbf{u}}_\gamma$ with respect to γ (Theorem 6.4)

$$\lim_{t\to 0}\frac{1}{t}\left(\mathcal{J}(\gamma+t\bar\gamma)-\mathcal{J}(\gamma)\right)=\lim_{t\to 0}\frac{1}{t}D_u\mathcal{J}(\gamma)\hat{\mathbf{u}}.$$

Testing the first equation in (7.2) with $\hat{\mathbf{u}}$ gives, using $\tau_\gamma\hat{\mathbf{u}}_\gamma=\mathbf{0}$ and the fact that $\hat{\mathbf{u}}_{\gamma+t\bar\gamma}$ and $\hat{\mathbf{u}}_\gamma$ are weakly divergence free,

$$D_u\mathcal{J}(\gamma)\hat{\mathbf{u}}=-\nu(\nabla\hat\lambda_\gamma,\nabla\hat{\mathbf{u}})_{\hat\Omega}-(\hat\lambda_\gamma,\hat{\mathbf{u}}_\gamma\cdot\nabla\hat{\mathbf{u}}+\hat{\mathbf{u}}\cdot\nabla\hat{\mathbf{u}}_\gamma)_{\hat\Omega}+\langle\chi_\gamma,\tau_\gamma\hat{\mathbf{u}}_{\gamma+t\bar\gamma}\rangle_{H^*_\gamma,H_\gamma}.$$

Because of

$$\begin{aligned}-\big[\hat{\mathbf{u}}_\gamma\cdot\nabla\hat{\mathbf{u}}+\hat{\mathbf{u}}\cdot\nabla\hat{\mathbf{u}}_\gamma\big] &= -\hat{\mathbf{u}}_\gamma\cdot\nabla\hat{\mathbf{u}}_{\gamma+t\bar\gamma}+\hat{\mathbf{u}}_\gamma\cdot\nabla\hat{\mathbf{u}}_\gamma-\hat{\mathbf{u}}\cdot\nabla\hat{\mathbf{u}}_\gamma\\ &= \hat{\mathbf{u}}\cdot\nabla\hat{\mathbf{u}}_{\gamma+t\bar\gamma}-\hat{\mathbf{u}}_{\gamma+t\bar\gamma}\cdot\nabla\hat{\mathbf{u}}_{\gamma+t\bar\gamma}+\hat{\mathbf{u}}_\gamma\cdot\nabla\hat{\mathbf{u}}_\gamma-\hat{\mathbf{u}}\cdot\nabla\hat{\mathbf{u}}_\gamma\\ &= \hat{\mathbf{u}}\cdot\nabla\hat{\mathbf{u}}-\hat{\mathbf{u}}_{\gamma+t\bar\gamma}\cdot\nabla\hat{\mathbf{u}}_{\gamma+t\bar\gamma}+\hat{\mathbf{u}}_\gamma\cdot\nabla\hat{\mathbf{u}}_\gamma,\end{aligned}$$

we obtain

$$\begin{aligned}D_u\mathcal{J}(\gamma)\hat{\mathbf{u}} = &-\big[\nu(\nabla\hat{\mathbf{u}}_{\gamma+t\bar\gamma},\nabla\hat\lambda_\gamma)+(\hat{\mathbf{u}}_{\gamma+t\bar\gamma}\cdot\nabla\hat{\mathbf{u}}_{\gamma+t\bar\gamma},\hat\lambda_\gamma)_{\hat\Omega}\big]+(\hat{\mathbf{u}}\cdot\nabla\hat{\mathbf{u}},\hat\lambda_\gamma)_{\hat\Omega}\\ &+\big[\nu(\nabla\hat{\mathbf{u}}_\gamma,\nabla\hat\lambda_\gamma)_{\hat\Omega}+(\hat{\mathbf{u}}_\gamma\cdot\nabla\hat{\mathbf{u}}_\gamma,\hat\lambda_\gamma)_{\hat\Omega}\big]+\langle\chi_\gamma,\tau_\gamma\hat{\mathbf{u}}_{\gamma+t\bar\gamma}\rangle_{H^*_\gamma,H_\gamma}.\end{aligned}$$

The two terms in parentheses can be expressed using the first equation in (5.1) for γ and $\gamma + t\bar{\gamma}$, respectively, with the adjoint variable $\hat{\lambda}_\gamma \in H_0^1(\hat{\Omega})^2$ as a test function. Thus

$$D_u\mathcal{J}(\gamma)\hat{\mathbf{u}} = (\tilde{\mathbf{f}}_\gamma - \tilde{\mathbf{f}}_{\gamma+t\bar{\gamma}}, \hat{\lambda}_\gamma)_{\hat{\Omega}} - \langle g_{\gamma+t\bar{\gamma}}, \tau_{\gamma+t\bar{\gamma}}\hat{\lambda}_\gamma\rangle_{H^*_{\gamma+t\bar{\gamma}},H_{\gamma+t\bar{\gamma}}} + \langle \chi_\gamma, \tau_\gamma \hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}\rangle_{H^*_\gamma,H_\gamma} + (\hat{\mathbf{u}}\cdot\nabla\hat{\mathbf{u}}, \hat{\lambda}_\gamma)_{\hat{\Omega}}.$$

Divided by t the first term on the right tends to zero because of Lemma 7.2, the last one because of Lemma 3.2 and the Lipschitz continuity of the velocities:

$$(\hat{\mathbf{u}}\cdot\nabla\hat{\mathbf{u}}, \hat{\lambda}_\gamma)_{\hat{\Omega}} \le k|\hat{\mathbf{u}}|^2_{H^1(\hat{\Omega})^2}|\hat{\lambda}_\gamma|_{H^1(\hat{\Omega})^2} \;\le\; kt^2\|\bar{\gamma}\|^2_{L^\infty(I)}|\hat{\lambda}_\gamma|_{H^1(\hat{\Omega})^2}$$

with k independent of $\gamma, \bar{\gamma}$. Using Lemmas 7.3 and 7.4 the remaining terms give (7.3) as directional derivative. To show that it is actually a Fréchet derivative we proceed exactly as in the proof of the same result for the Stokes case in [7, Theorem 7.5]. □

Note that the formula for the derivative does not include any normal derivatives of state or adjoint variables along the boundary.

8. Numerical methods. To validate the derivative formula (7.3) a numerical example is presented in the next section. To ensure the needed regularity of the boundary partition Γ_γ, i.e., to guarantee that $\gamma \in H^3(I)$, we considered boundaries Γ_γ which were generated by a cubic spline function γ. This function was generated as a linear combination of B-splines on an equidistant grid in the interval $(0, 1)$. The end points $(0, 0)$ and $(1, 0)$ and the slopes at these points were fixed. This conforms with our definition (2.1) for the admissible functions γ; specifically, the regularity and the convexity of the boundary part Γ_γ at the left and right end of the interval $(0, 1)$ is guaranteed. By these conditions the first and last two coefficients of the spline were determined. The values of the remaining ones were used as control parameters. One coefficient c_i in the B-spline representation contributes to the y-value of the generated spline at the point x_i with the weight $\frac{2c_i}{3}$ and to the y-value at the two neighbors x_{i-1} and x_{i+1} with the weight $\frac{c_i}{6}$.

The B-spline representation has the advantages that for nonnegative control parameters the y-values of the points of the generated curve are always nonnegative, too. This is crucial to obtain a feasible boundary $\Gamma_\gamma \subset \hat{\Omega} = (0, 1)^2$. Additional upper bounds on γ were realized as linear inequality constraints of the control parameters, i.e., the B-spline coefficients.

We used the SQP implementation in MATLAB's optimization toolbox version 2, namely the routine `fmincon` (see [15]). It allows the user to provide gradient information which we obtained by evaluating (7.3) for the splines corresponding to independent variations in each control parameter.

We also want to emphasize that the restriction on the parameter ν that we had made to obtain formula (7.3), see Theorems 7.5 and 3.3, namely

$$\nu > \frac{|\mathbf{u}_\gamma|_{H^1(\Omega_\gamma)^2}}{k_\gamma},$$

is restrictive. In our geometric configuration we have $k_\gamma = \frac{1}{2}|\Omega_\gamma|^{1/2} < \frac{1}{2}$ and thus $\nu > 2|\mathbf{u}_\gamma|_{H^1(\Omega_\gamma)^2}$ is required. The standard estimates for dependence of the solution on the inhomogeneity and boundary values then give

$$\nu > 2C\left(\|\mathbf{f}\|_{H^{-1}(\hat{\Omega})^2} + \|\Phi\|_{H^{1/2}(\Gamma)^2}\right),$$

where the constant C depends on the domain only. This makes clear that formula (7.3) is only appropriate for high values of ν, i.e., low Reynolds numbers.

In each step of a gradient-based iterative optimization process (as, e.g., SQP) usually one gradient and several function evaluations for different control variables γ are necessary. Each gradient evaluation requires one solution of both systems (5.1) and (7.2); a function evaluation implies the solution of (5.1).

The velocity and pressure variables and their adjoint counterparts in (5.1) and (7.2) are discretized by stabilized linear finite elements (see [16]), whereas the Lagrange multipliers g_γ and χ_γ are by piecewise constant elements. To satisfy the inf-sup condition for the latter we use a coarser discretization for the Lagrange multipliers as suggested in [17]. This means that the support of a constant basis function for g_γ and χ_γ is at least twice as long as the minimal length of the triangle edges in the discretization of $\hat{\Omega}$. It turned out that this was sufficient to avoid oscillations of the Lagrange multipliers.

The nonlinear system (5.1) was solved by the semi-implicit algorithm given in [8, eq. IV(2.25)] (here in continuous form):

(i) Choose $\hat{\mathbf{w}}$ with $\operatorname{div} \hat{\mathbf{w}} = 0$.

(ii) Compute $(\hat{\mathbf{u}}, \hat{p}, g)$ from

$$\left\{ \begin{array}{rcl} -\nu \triangle \hat{\mathbf{u}} + \hat{\mathbf{w}} \cdot \nabla \hat{\mathbf{u}} + \nabla \hat{p} - \tau_\gamma^* g & = & \tilde{\mathbf{f}}, \\ \operatorname{div} \hat{\mathbf{u}} & = & 0. \end{array} \right.$$

(iii) If "convergence," stop, else set $\hat{\mathbf{w}} := \hat{\mathbf{u}}$ and go back to (ii).

The discrete system to be solved in step (ii) then reads as

$$\begin{pmatrix} \nu A + N(W_\gamma) & B^T & D_\gamma^T \\ B - Q(W_\gamma) & C & 0 \\ D_\gamma & 0 & 0 \end{pmatrix} \begin{pmatrix} U_\gamma \\ P_\gamma \\ G_\gamma \end{pmatrix} = \begin{pmatrix} F_\gamma \\ H \\ 0 \end{pmatrix}, \tag{8.1}$$

where the matrices C, Q and the vector H appear due to the stabilization. The matrix N represents the discretized and linearized convective term and W_γ the discretized last iterate $\hat{\mathbf{w}}$ in (ii). Solving the Navier–Stokes system requires several solutions of (8.1), depending on the necessary number of iterations of the semi-implicit algorithm, in our case at most 8. For (7.2) one linear system of the same structure as (8.1) with different matrices N, Q and a changed right-hand side has to be solved.

Since (8.1) has to be solved quite often for different control parameters γ it is important to note that only the entities with subscript γ change when the control parameter is modified. All other matrices can be assembled in advance and kept fixed. Here D_γ represents a one-dimensional trace operator and thus is very sparse. The inhomogeneity F_γ has to be modified since for the validity of the derivative formula it has to be set to zero in the fictitious part Ω_γ^c. This adjustment requires the integration of the basis functions on triangle partitions whenever Γ_γ intersects the interior of a triangle. To simplify this quadrature the use of linear basis functions, which can be integrated exactly by low-order Gaussian rules, is preferable.

A technical part is to find the intersection points between the two-dimensional grid for U_γ, P_γ and the one-dimensional grid along Γ_γ for G_γ (and for their adjoint counterparts). Here the decisive idea is to exploit the neighborhood information of the triangulation already used to assemble the finite element matrices. This avoids a time-consuming search over all triangles that would be even worse in three space dimensions. After the intersection points are found, the mass matrix D_γ and the

inhomogeneity F_γ can be easily computed by interpolation of the values of the piecewise linear or constant finite element basis functions. Over all, the numerical effort of this assembling and modification is negligible compared to the one needed for the solution of the state and adjoint equations.

In contrast to Stokes problems (see [7]) the discrete systems now change in every iteration of the optimization since the system matrix depends on the last iterate W_γ. Thus one advantage, namely the possibility of one factorization of the system matrix in the beginning of the optimization loop that was possible for the Stokes case (see [7, section 8]), is lost. It can be retained, e.g., by using a projected *cg* algorithm that solves (3.1) by a sequence of Stokes problems; see [18].

Once the discrete counterparts of $p_\gamma, g_\gamma, \mu_\gamma, \chi_\gamma$ are computed, the evaluation of the derivative via (7.3) requires only the evaluation of a one-dimensional integral. This can be done exactly by appropriate numerical quadrature formulas and thus no additional discretization error is introduced by the derivative evaluation. Normal derivatives of state and adjoint velocity along the boundary usually occurring here are implicitly included in g_γ, χ_γ. This fact is due to the embedding domain technique by which these two Lagrange multipliers were introduced.

TABLE 1
Convergence behavior for 5 *control parameters.*

It.	$\mathcal{J}$	Control parameters (B-spline coefficients)				
1	1.3825e-03	0.0010	0.0010	0.0010	0.0010	0.0010
2	4.5050e-04	0.0685	0.2901	0.2307	0.2900	0.0687
5	1.5073e-05	0.0612	0.4560	0.3044	0.4894	0.1332
25	1.4691e-06	0.0252	0.4121	0.3837	0.4013	0.0421
45	3.8337e-07	0.0010	0.3494	0.4144	0.3654	0.0010

TABLE 2
Convergence behavior for 9 *control parameters.*

It.	$\mathcal{J}$	Control parameters (B-spline coefficients)								
1	1.4e-03	.0010	.0010	.0010	.0010	.0010	.0010	.0010	.0010	.0010
4	1.0e-04	.0426	.0901	.2946	.4674	.4262	.4679	.2882	.0364	.0010
5	1.2e-05	.0596	.0926	.2746	.4271	.4006	.4291	.2650	.0127	.0024
7	2.1e-06	.0732	.0944	.2585	.3948	.3800	.3960	.2451	.0010	.0032
15	2.7e-07	.0760	.0812	.2470	.3860	.4017	.3944	.2230	.0010	.0010
27	5.3e-08	.0769	.0784	.2463	.3863	.3995	.3913	.2204	.0010	.0015

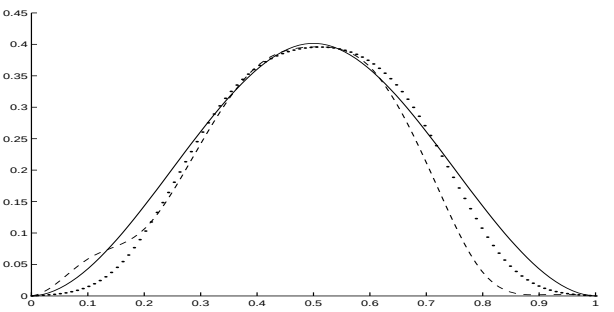


FIG. 2. *Inverse problem. Comparison between desired (solid line) and optimized curves for* 5 *(dotted) and* 9 *(dashed) control parameters.*

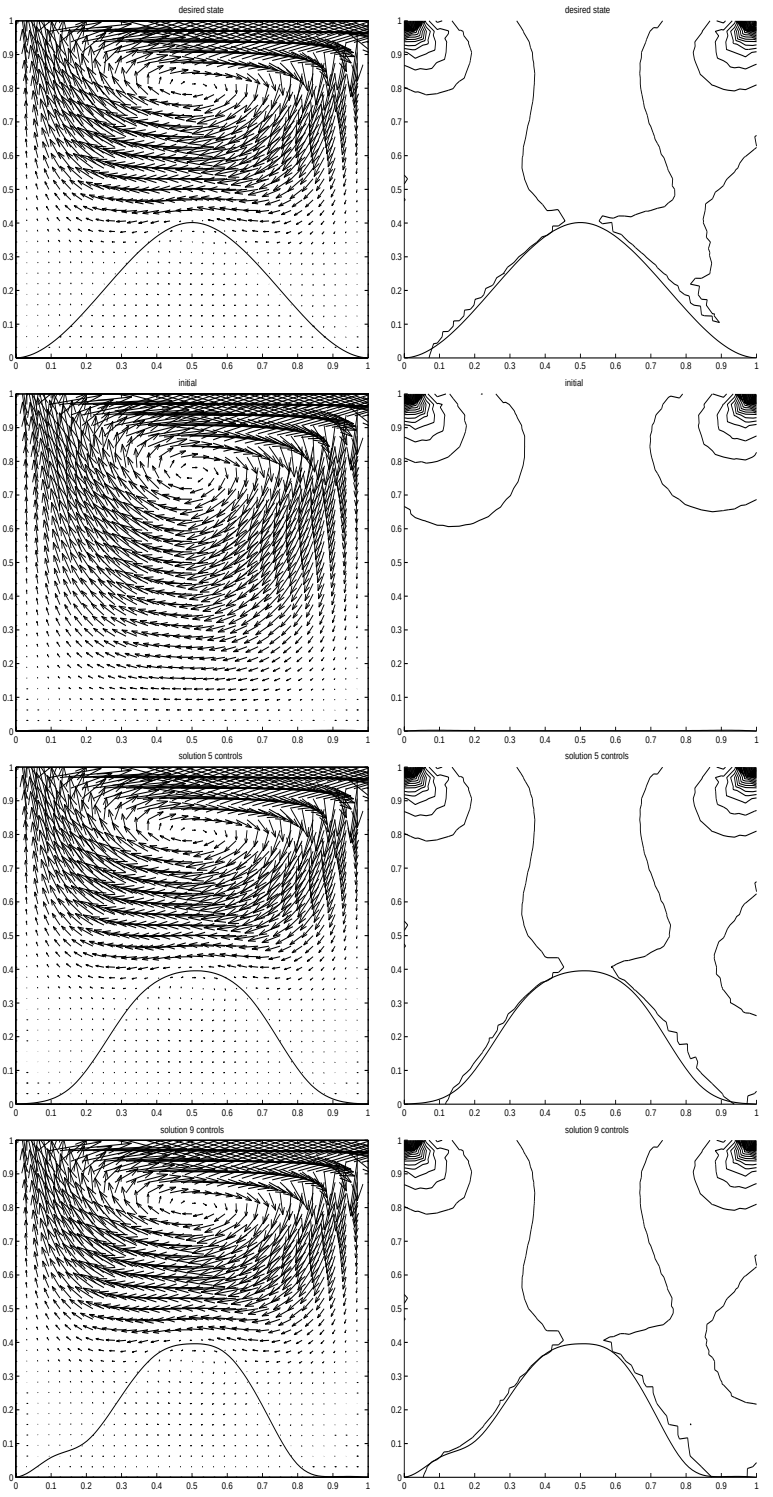


FIG. 3. *Velocity vectors (left) and pressure distribution (right) for (from top to bottom) desired state, initial, and optimized curve with* 5 *and* 9 *control parameters.*

9. Numerical example. We present an inverse problem for a driven cavity flow. The computational domain is the unit square. On one edge a constant tangential velocity is prescribed, the other edges are regarded as walls with homogeneous Dirichlet boundary conditions for the velocity. The function Φ representing the boundary values was smoothed at the corners. The inhomogeneity $\mathbf{f}$ was set to zero and the parameter $\nu = 1$. In both examples the bottom edge of the cavity was variable between the fix points $(0,0)$ and $(1,0)$. As initial curves for the optimization, we used a straight line as bottom edge. Lower box constraints for the control parameters to ensure $\gamma \geq 0$ and a linear inequality constraint to guarantee $\Gamma_\gamma \cap \Omega_C = \emptyset$ were used. No regularizations were necessary. State and adjoint equations were solved on a triangular grid with 1089 velocity nodes.

9.1. Inverse problem for the driven cavity. Here the function $\Phi = (\Phi_1, \Phi_2) = (1,0)$ was used at the top edge of the cavity. We considered a tracking type cost functional with $\Omega_C := (0,1) \times (0.5,1)$ and the desired state $\mathbf{u}_d := \mathbf{u}_{\gamma_d}$, where γ_d was a cubic spline interpolating the points $(0,0), (0.5,0.4), (1,0)$.

We performed two optimization runs with 5 and 9 control parameters. In Tables 1 and 2 the convergence behavior is documented. It can be seen that the optimization reduces the cost functional rather fast up to one percent. Further reduction affords more iterations, specifically a very low stopping criterion for the accuracy in the control parameters and function value. The convergence for the higher number of controls was faster (concerning the number of iterations). Even if the changes in the controls are in the range of 10^{-4} and are hardly visible, further optimization steps lead to a better value of the cost and to a curve close to the one used to compute the desired state.

Figure 2 shows the boundary curves of the desired state and the two optimized ones. In Figure 3 we depicted the velocity vectors and pressure distribution for the desired state, the initial, and the two obtained optimized curves.

Acknowledgment. The author would like to thank Prof. Karl Kunisch from Karl-Franzens University Graz, Austria, for his support.

REFERENCES

[1] R. Glowinski, T.-W. Pan, and J. Periaux, *A fictitious domain method for external incompressible viscous flow modeled by Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 112 (1994), pp. 133–148.

[2] J. Dankova and J. Haslinger, *Numerical realization of a fictitious domain approach used in shape optimization.* I: *Distributed controls*, Appl. Math., 41 (1996), pp. 123–147.

[3] O. Pironneau, *On optimum design in fluid mechanics*, J. Fluid Mech., 64 (1974), pp. 97–110.

[4] M. D. Gunzburger and H. Kim, *Existence of an optimal solution of a shape control problem for the stationary Navier–Stokes equations*, SIAM J. Control. Optim., 36 (1998), pp. 895–909.

[5] J. A. Bello, E. Fernandez-Cara, J. Lemoine, and J. Simon, *The differentiability of the drag with respect to the variations of a Lipschitz domain in a Navier–Stokes flow*, SIAM J. Control. Optim., 35 (1997), pp. 626–640.

[6] K. Kunisch and G. Peichl, *Shape optimization for mixed boundary value problems based on an embedding domain method*, Dynam. Contin. Discrete Impuls. Systems, 4 (1998), pp. 439–478.

[7] T. Slawig, *An explicit formula for the derivative of a class of cost functionals with respect to domain variations in Stokes flow*, SIAM J. Control. Optim., 39 (2000), pp. 141–158.

[8] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer Ser. Comput. Math., Springer–Verlag, Berlin, New York, 1986.

[9] R. B. Kellogg and J. E. Osborn, *A regularity result for the Stokes problem in a convex polygon*, J. Funct. Anal., 21 (1976), pp. 397–431.

[10] R. Temam, *Navier-Stokes Equations*, North–Holland, Amsterdam, 1979.
[11] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 2, Springer–Verlag, Berlin, 1988.
[12] T. Slawig, *Domain Optimization for the Stationary Stokes and Navier-Stokes Equations by an Embedding Domain Technique*, Ph.D. thesis, TU Berlin, Berlin, 1998.
[13] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations, Vol.* 2, Springer–Verlag, New York, 1994.
[14] M. D. Gunzburger, L. Hou, and T. P. Svobodny, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with distributed and Neumann control*, Math. Comput., 57 (1991), pp. 123–151.
[15] The Mathworks Inc., *Optimization Toolbox—For Use with* MATLAB*, User's Guide Version* 2, The Mathworks Inc., Natick, MA, 1998.
[16] T. E. Tezduyar, *Stabilized finite element formulations for incompressible flow computations*, Adv. Appl. Math. 28, Academic Press, Boston, 1992, pp. 1–44.
[17] V. Girault and R. Glowinski, *Error analysis of a fictitious domain method applied to a Dirichlet problem*, Japan J. Indust. Appl. Math., 12 (1995), pp. 487–514.
[18] M. C. Desai and K. Ito, *Optimal controls of Navier–Stokes equations*, SIAM J. Control Optim., 32 (1994), pp. 1428–1446.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 513–531

ON THE LOCAL STRUCTURE OF OPTIMAL TRAJECTORIES IN R^3*

ANDREI A. AGRACHEV† AND MARIO SIGALOTTI‡

Abstract. We analyze the structure of a control function $u(t)$ corresponding to an optimal trajectory for the system $\dot{q} = f(q) + u\,g(q)$ in a three-dimensional manifold, near a point where some nondegeneracy conditions are satisfied. The kind of optimality which is studied includes time-optimality. The control turns out to be the concatenation of some bang and some singular arcs. Studying the index of the second variation of the switching times, the number of such arcs is bounded by four.

Key words. optimal control, Lie brackets

AMS subject classifications. 49K15, 49K30

PII. S0363012902409246

1. Introduction.

Consider the time-optimal control problem

$$\dot{q} = f(q) + u\,g(q), \qquad u \in [-1, 1], \tag{1}$$

on a three-dimensional manifold M^3, where f and g are two smooth vector fields on M^3 and the admissible controls are all measurable functions $u : t \mapsto u(t) \in [-1, 1]$. The kind of results we are interested in are local regularity properties for the control function corresponding to optimal trajectories. Namely, we fix $q_0 \in M^3$ and we study whether *locally at* q_0 the control function corresponding to a time-optimal trajectory is piecewise smooth, meaning by this that there exist a neighborhood U of q_0 and a time $T > 0$ such that any control function corresponding to a time-optimal trajectory of the system (1) contained in U and defined on a time-interval of length less than or equal to T is piecewise smooth. We are interested in giving an upper bound to the number of smooth pieces, called *arcs*, and in describing the possible concatenations. (For instance, we want to know how many arcs are bang, i.e., such that the control restricted to them is the constant function $+1$ or -1.)

As is well known (see [19]), any irregularity of the optimal control is possible; that is, for any measurable control function $u(\cdot)$; there exists a control system of type (1) such that the trajectory corresponding to $u(\cdot)$ is time-optimal. The correct question is, What kind of behavior can we expect for time-optimal trajectories of a generic system?

A major motivation for the study of this topic is the following: to give a priori restrictions on the local structure of optimal trajectories is a crucial step in the direction of the description of the local optimal synthesis (see, for instance, [22]). The problem is known to be deep: Fuller [7] first proposed a polynomial system of the kind studied here such that the switching time moments of the optimal control form a convergent sequence. (We call *chattering* a control of this kind.) Since then it has been an important issue to understand whether this phenomenon is structurally stable (i.e., cannot

*Received by the editors June 5, 2002; accepted for publication (in revised form) December 2, 2002; published electronically May 12, 2003.
http://www.siam.org/journals/sicon/42-2/40924.html
†SISSA-ISAS, via Beirut 2-4, 34014 Trieste, Italy and Steklov Mathematical Institute, ul. Gubkina 8, Moscow, Russia (agrachev@sissa.it).
‡SISSA-ISAS, via Beirut 2-4, 34014 Trieste, Italy (sigalott@sissa.it).

be eliminated by a small perturbation of the system) or not. In big enough dimensions it most likely is stable: stable chattering extremals were constructed in [10, 23], though the optimality of these extremals is not proved.

The language in which the genericness of the system will be expressed, in order to concretely tackle the problem, is the one of "nonresonance conditions" on the configuration of iterated Lie brackets between f and g evaluated at the fixed point q_0. This is natural since the family of Lie bracket relations form a set of differential invariants of the pair of vector fields (f, g), which is complete for analytic systems (see [16, section 4]).

The first step in this direction is to study the structure of the trajectories near the point where some prescribed triples of iterated Lie brackets are linearly independent. If a local regularity property is proved under such conditions, by standard transversality considerations it follows that for a generic pair of vector fields (f, g), for a generic point $q_0 \in M^3$, the regularity property holds locally at q_0. The first result of this kind for a three-dimensional manifold appeared in [18] and asserts that if each of the tensor fields $f \wedge g \wedge [f, g]$, $g \wedge [f, g] \wedge [f + g, [f, g]]$, and $g \wedge [f, g] \wedge [f - g, [f, g]]$ does not vanish at q_0, then locally at q_0 any time-optimal bang-bang trajectory has at most two switchings.

A satisfactory understanding of the three-dimensional problem would be for us to individuate a regularity property, which holds locally at every point of M^3 for a generic pair of vector fields (f, g), in analogy with what has already been done for two-dimensional manifolds by Sussmann (see [17, 20] and also [11]). By Thom's transversality theorem we know that this reduces to the study of all Lie bracket configurations with up to three nontrivial independent relations of linear dependency between its elements. Up to now, in addition to the cited result by Sussmann, there have been similar upper bounds proved in other situations: by Bressan [6] in the case of trajectories steering to an equilibrium point for f under some extra nondependency conditions; by Schättler [12] for bang-bang trajectories for a generic pair (f, g), locally at all points in which $f \wedge g \wedge [f, g] \neq 0$. (For trajectories which are not a priori bang-bang Schättler [13] requires that some extra nondependency conditions hold at the point.) In [3] the following complementary result is stated, which generalizes the one in [6]: if $g \wedge [f, g] \wedge [f \pm g, [f, g]](q_0) \neq 0$, then locally at q_0 a bang-bang time-optimal trajectory has no more than two switches. Part of the present paper is devoted to giving a detailed proof of that statement. The same result had also been proved by Krener and Schättler [9] and Sussmann [21] by different methods.

Unfortunately, the cases studied in the literature do not cover all Lie bracket configurations with up to one dependency relation. In the present paper we give the first result concerning all such configurations: we furnish a common finite local bound on the number of arcs, which is valid not only for bang-bang trajectories, but also for all trajectories with no a priori restriction on the control. We also discuss in which sense the bound is sharp. The main results are contained in Theorem 3 in section 3 and in Lemma 2 in subsection 4.3 and can be summarized as follows, in terms of generic properties for (f, g) and of properties holding at points of M^3, in the local sense introduced above.

THEOREM 1. *Let $V_2^4(M^3)$ be the space of all pairs of $\mathcal{C}^4$ vector fields on M^3, endowed with the Whitney topology. For any element (f, g) from an open everywhere dense subset of $V_2^4(M^3)$ there exist $W_1 \subset W_2 \subset M^3$, where W_1 and W_2 are, respectively, a one-dimensional and a two-dimensional stratified set in M^3, such that*

- *for any point q_0 in $M^3 \setminus W_2$, locally at q_0, any time-optimal trajectory of the system (1) is a finite concatenation of at most three between bang and singular arcs;*

- *for any point q_0, in $W_2 \setminus W_1$, locally at q_0, any time-optimal trajectory of the system* (1) *is a finite concatenation of at most four between bang and singular arcs.*

PROPOSITION 1. *For any (f, g) from an open subset of $V_2^4(M^3)$ there exists a two-dimensional submanifold $U_2 \subset M^3$ such that for any $q_0 \in U_2$ and any $T > 0$ there exists a trajectory passing through q_0 of time-length less than T which is locally (in the $\mathcal{C}^0$ topology) time-optimal and is the concatenation of four bang arcs (of positive length).*

As we mentioned, it is the possible occurrence of the chattering phenomenon that makes the problem deep and hard to study. Our paper is one more step in the demarcation of the nonchattering territory, at least in dimension three. It is not excluded that structurally stable chattering phenomena occur, but it is given a stronger a priori bound to their "dimension."

Another problem is whether the Fuller phenomenon is the worst possible stable behavior. Proposition 2 in section 6 gives a partial answer to this question, stating that under some very weak conditions (which hold everywhere for a generic pair (f, g) in any dimension) an optimal control with values in $\{-1, 1\}$ either has a finite number of switches or is such that its restriction to a subinterval is chattering. This means that any possible bad behavior is built up, in some sense, by chattering modes. If we prove, in particular, that in a certain region chattering does not occur, then a time-optimal trajectory passing through the region (and whose control function takes value in $\{-1, 1\}$) must be bang-bang.

2. A second order optimality condition. Let M be a smooth manifold and f, g two smooth vector fields on M. It is natural and costless to assume throughout this section that the dimension of M is equal to $n \in \mathbf{N}$, with no further restriction on n, since the second order condition we are going to state is independent on the dimension.

A *trajectory* of the system

$$\dot{q} = f(q) + u\, g(q), \qquad u \in [-1, 1], \tag{2}$$

is an absolutely continuous curve $t \mapsto q(t) \in M$ for which there exists a measurable *control function* $t \mapsto u(t) \in [-1, 1]$ such that (2) is verified for almost every t in the domain of $q(\cdot)$. For all $T > 0$ and $q_0 \in M$ we define the *attainable set* from q_0 at time T:

$$A(T, q_0) = \{q(T) \mid q : [0, T] \to M \text{ is a trajectory of (2) such that } q(0) = q_0\}.$$

Let $q : [0, T] \to M$ be a trajectory of (2) and u the corresponding control function. By the Pontryagin maximum principle we know that if $q(T)$ belongs to $\partial A(T, q(0))$, then $q(\cdot)$ is *extremal*; that is, there exist $c \in \mathbf{R}$ and an absolutely continuous covector trajectory $p : [0, T] \to T^*M$ such that $p(t) \in T^*_{q(t)}M \setminus \{0\}$ for every $t \in [0, T]$, which verifies for almost every t the equation

$$\dot{p}(t) = -p(t)(Df(x(t)) + u(t)Dg(x(t))) \tag{3}$$

and the relation

$$\langle p(t), (f + u(t)g)(q(t))\rangle = \min_{v \in [-1,1]} \langle p(t), (f + vg)(q(t))\rangle \equiv c. \tag{4}$$

We say that $q : [0, T] \to M$ is *bang-bang* if the control function u takes values in $\{-1, 1\}$ and there exists a finite number of *switching times* $0 < t_1 < t_2 < \cdots < t_{k-1} < t_k < T$, splitting $[0, T]$ in intervals on which u is alternately the constant $+1$ and -1. In general we say that a piece of trajectory defined on a time subinterval is a bang arc if the corresponding control is constantly equal to $+1$ or -1 (we speak, respectively, of a $+$ arc or a $-$ arc), whereas it is a singular arc if it is not a bang arc and the corresponding control is smooth. We will describe the structure of a trajectory by the standard agreement that, for instance, a $+-$S trajectory is a concatenation of a $+$, a $-$, and a singular arc.

Since we are interested in local results it is justified and convenient for us to assume that all the vector fields involved are complete. Given a complete smooth vector field h on M and a time $t \in \mathbf{R}$, we can associate the flow of h at a time t, which we will denote by $e^{th} : q \mapsto e^{th}(q)$. Both the vector field h and the diffeomorphism e^{th} have a natural interpretation as operators on $\mathcal{C}^\infty(M)$: given a smooth function a on M and a point $q \in M$, $ha(q)$ is defined as the derivative of a in the direction $h(q)$ at the point q, whereas

$$(e^{th}a)(q) = a(e^{th}(q)).$$

Given two smooth vector fields h_1 and h_2, it is always possible to define their commutator (Lie bracket) according to the formula

$$((\mathrm{ad}h_1)h_2)a = [h_1, h_2]a = h_1(h_2a) - h_2(h_1a).$$

In operator terms the action $e_*^{-th_1}$ of the diffeomorphism e^{-th_1} on the vector fields has the form

$$e_*^{-th_1}h_2 = e^{th_1} \circ h_2 \circ e^{-th_1}.$$

The formula

$$\frac{\mathrm{d}}{\mathrm{d}t} e^{th_1} \circ h_2 \circ e^{-th_1}(q) = [h_1, e^{th_1} \circ h_2 \circ e^{-th_1}](q)$$

justifies the notation

$$e^{t\mathrm{ad}h_1}h_2 = e^{th_1} \circ h_2 \circ e^{-th_1}.$$

Notice that, for every t and h, the following relation holds:

$$e^{t\mathrm{ad}h}h = h. \tag{5}$$

Remark. In the case of vector fields which are not smooth but just $\mathcal{C}^k$, the above definitions extend to the case where they are still licit; in particular, if h is $\mathcal{C}^k$, then $e^{t\mathrm{ad}h}$ is a well-defined transformation of $\mathcal{C}^k$ vector fields while $k + 1$ iterated Lie brackets between $\mathcal{C}^k$ vector fields are well-defined $\mathcal{C}^0$ vector fields.

In order to formulate the second order optimality condition it is useful to remark the following fact: Let $q(\cdot)$ be an extremal trajectory of (2) and $p(\cdot)$ an associated covector trajectory. Fix a vector field h and two time instants t and τ in the same bang interval of $q(\cdot)$ on which the control is equal to ν. By (3) we have

$$\langle p(t), h(q(t))\rangle = \langle p(\tau), e^{(t-\tau)\mathrm{ad}(f+\nu g)}h(q(t))\rangle. \tag{6}$$

THEOREM 2. *Let $q : [0,T] \to M$, $T > 0$, be a bang-bang trajectory of (2) and let $u(t)$ be the corresponding control function with k switching times $0 < \tau_1 < \tau_2 < \cdots < \tau_k < T$. Denote by ν the value of u in $(0,\tau_1)$. Let $q(\cdot)$ be extremal and $p(\cdot)$ a corresponding covector trajectory. Assume that $p(\cdot)$ is uniquely defined (up to multiplication by a positive scalar) by (3) and (4). Let $p_0 = p(0)$ and*

$$\begin{aligned} h_0 &= f + \nu g, \\ h_i &= e^{\tau_1 \mathrm{ad}(f+\nu g)} \circ e^{(\tau_2-\tau_1)\mathrm{ad}(f-\nu g)} \circ e^{(\tau_3-\tau_2)\mathrm{ad}(f+\nu g)} \circ \cdots \\ &\quad \circ e^{(\tau_i-\tau_{i-1})\mathrm{ad}(f-(-1)^i \nu g)}(f + (-1)^i \nu g), \qquad i = 1, \ldots, k. \end{aligned}$$

Take the quadratic form

$$Q(\alpha) = \sum_{0 \le i < j \le k} \alpha_i \alpha_j \langle p_0, [h_i, h_j](q(0)) \rangle \tag{7}$$

defined on the space

$$\left\{ \alpha = (\alpha_0, \ldots, \alpha_k) \in \mathbf{R}^{k+1} \;\middle|\; \sum_{i=0}^{k} \alpha_i = 0,\ \sum_{i=0}^{k} \alpha_i h_i(q(0)) = 0 \right\}. \tag{8}$$

If Q is not nonnegative definite, then $q(T) \in \mathrm{int} A(T, q(0))$. Moreover, for every system (f', g') which is close enough to (f, g) in the $\mathcal{C}^1$ topology, the point $q(T)$ is in the interior of the corresponding attainable set $A_{(f',g')}(T, q(0))$.

Remark. The sign condition on Q is usually rephrased in terms of the *index* of Q; that is, the dimension of the maximal subspace on which Q is negative definite. The quadratic form Q is not nonnegative definite if and only if its index is strictly positive.

Remark. Stated as above, the theorem seems to acknowledge a special role for the tangent space $T_{q(0)}M$. If we fix a time t in $[0,T]$, however, an equivalent version of the theorem can be easily set in $T_{q(t)}M$. Let m be such that $\tau_{m-1} \le t < \tau_m$ (with the agreement that $\tau_0 = 0$ and $\tau_{k+1} = T$). Define

$$h_i^t = e^{-(t-\tau_{m-1})\mathrm{ad}(f-(-1)^m \nu g)} \circ \cdots \circ e^{-(\tau_2-\tau_1)\mathrm{ad}(f-\nu g)} \circ e^{-\tau_1 \mathrm{ad}(f+\nu g)} h_i$$

for every $i = 0, \ldots, k$. The equivalent formulation of the theorem follows: the space defined by (8) is unaltered if we replace $h_i(q(0))$ by $h_i^t(q(t))$, and the quadratic form

$$Q^t(\alpha) = \sum_{0 \le i < j \le k} \alpha_i \alpha_j \langle p(t), [h_i^t, h_j^t](q(t)) \rangle$$

is actually independent of t, as follows from an iterated use of (6). We will find useful, in order to simplify the computations, to apply the theorem choosing t between the switching times of the trajectory.

This theorem, in a much more general setting, has been proved in [3]. We will give here just a brief sketch of the proof. Let $q_0 = q(0)$ and let $F : w(\cdot) \mapsto F(w)$ be the endpoint mapping at time T for the system

$$\begin{cases} \dot{x} = f(x) + w g(x), \\ x(0) = q_0, \end{cases}$$

which is defined in an L^1 neighborhood of the reference control $u(\cdot)$. Take

$$G(v) = e^{-\tau_1(f+\nu g)} \circ e^{(\tau_1-\tau_2)(f-\nu g)} \circ \cdots \circ e^{(\tau_k - T)(f+(-1)^k \nu g)} F(u + v),$$

which is just F, simply pulled back by the flow generated by the control u, in order to have $G(0) = q_0$. Our aim is to prove that G is locally open at 0 and, moreover, that this property is stable with respect to $\mathcal{C}^1$ perturbations of our system. Let

$$h(t) = e^{\tau_1 \mathrm{ad}(f+\nu g)} \circ e^{(\tau_2-\tau_1)\mathrm{ad}(f-\nu g)} \circ e^{(\tau_3-\tau_2)\mathrm{ad}(f+\nu g)} \circ \dots \\ \circ e^{(t-\tau_{i-1})\mathrm{ad}(f-(-1)^i \nu g)}(f + (-1)^i \nu g)$$

for all $t \in [\tau_{i-1}, \tau_i]$ and all $i = 1, \dots, k+1$. The extremality condition (4) implies that $\langle p_0, h(t)(q_0)\rangle = c$ for all $t \in [0, T]$, as we immediately deduce from (6). The uniqueness of $p(\cdot)$ means, moreover, that the closed convex cone generated by $\{h(t)(q_0) \mid t \in [0, T]\}$ is an half-space of $T_{q_0}M$, which we denote by H.

For any $\alpha = (\alpha_0, \dots, \alpha_k) \in \mathbf{R}^{k+1}$ with $\sum_{i=0}^k \alpha_i = 0$ and for $s > 0$ small enough, we can define $w_s^\alpha(\cdot)$ as the bang-bang control with switching times

$$0 < \tau_1 + s\alpha_0 < \tau_2 + s(\alpha_0 + \alpha_1) < \dots < \tau_k + s\sum_{i=0}^{k-1} \alpha_i < T.$$

Let $v_s^\alpha = u - w_s^\alpha$. We have that

$$\frac{\mathrm{d}}{\mathrm{d}s} G(v_s^\alpha)\Big|_{s=0+} = \sum_{i=0}^k \alpha_i h_i(q_0)$$

and

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2} G(v_s^\alpha)\Big|_{s=0+} = \sum_{i=0}^k \sum_{j=i+1}^k \alpha_i \alpha_j [h_i, h_j](q_0).$$

The positiveness of Q can thus be read as follows: the elements of the space defined by (8) correspond to the variations of the switching times which preserve the total time T and which produce a zero first order variation in the endpoint. Assume that a variation α exists such that $Q(\alpha)$ is negative. Then the convex cone generated by

$$v_0 = \frac{\mathrm{d}^2}{\mathrm{d}s^2} G(v_s^\alpha)\Big|_{s=0+}$$

and H is the whole $T_{q_0}M$. Fix $t_1, \dots, t_n \in]0, T[\setminus\{\tau_1, \dots, \tau_k\}$ such that the convex cone generated by v_0 and

$$v_i = h(t_i)(q_0), \qquad i = 1, \dots, n,$$

is equal to $T_{q_0}M$. Define the admissible control

$$w_{(s_0,s_1,\dots,s_n)}(t) = \begin{cases} -u(t) & \text{if } t \in \cup_{i=1}^n [t_i, t_i + s_i], \\ w_{s_0}^\alpha(t) & \text{otherwise} \end{cases}$$

for $s_0, s_1, \dots, s_n > 0$ small enough and let $v_{(s_0,s_1,\dots,s_n)} = u - w_{(s_0,s_1,\dots,s_n)}$. Fix $a_0, a_1, \dots, a_n > 0$ and remark that

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} G(v_{(a_0\varepsilon, a_1\varepsilon^2/2, \dots, a_n\varepsilon^2/2)})\Big|_{\varepsilon=0+} = 0$$

and

$$\frac{\mathrm{d}^2}{\mathrm{d}\varepsilon^2} G(v_{(a_0\varepsilon, a_1\varepsilon^2/2, \ldots, a_n\varepsilon^2/2)})\Big|_{\varepsilon=0+} = \sum_{i=0}^{n} a_i v_i.$$

Now a standard application of the Brower fixed point theorem implies that for any continuous mapping $\mathcal{C}^0$ close to $(s_0, s_1, \ldots, s_n) \mapsto G(v_{(\sqrt{s_0}, s_1, \ldots, s_n)})$, the image of any neighborhood of 0 in $[0, +\infty)^{n+1}$ contains a neighborhood of q_0 in M. This concludes the sketch of the proof of Theorem 2.

The theorem above naturally suggests the following definition, which makes sense for any smooth control system.

DEFINITION. *We say that a trajectory $q : [0, T] \to M$ of $\dot{q} = f(q, u)$, where f is $\mathcal{C}^1$ with respect to both the state q and the control u, is essential if $q(T)$ belongs to the interior of the attainable set from $q(0)$ at time T for any $\mathcal{C}^1$-close system. Vice versa, we call quasi-optimal a trajectory which is not essential.*

Remark. Given an essential trajectory $q : [0, T] \to M$, if we consider time-rescaled systems (which, up to some technical use of cut-off functions, can be seen as $\mathcal{C}^1$-close to the original one, if the scaling factor is close to 1), we get that $q(T) \in A(t, q(0))$ for t close to T. In particular, $q(\cdot)$ is neither the fastest trajectory connecting $q(0)$ to $q(T)$ (i.e., it is not time-optimal) nor the slowest. Theorem 2, giving necessary conditions for quasi-optimality, furnishes a unified approach for a wide range of phenomena, including time-optimality.

Remark. A property that quasi-optimality shares with time-optimality (and, in general, with any optimality defined by an integral cost) is the fact that the time-reversed of a quasi-optimal trajectory is quasi-optimal for the time-reversed system. Indeed, if $q : [0, T] \to M$ is essential for the control system $\dot{q} = f(q, u)$, then there exist $\delta > 0$ and a neighborhood U of $q(0)$ such that for any f' δ-close to f and for any $q \in U$, $q(T)$ belongs to the attainable set from q at time T of the system $\dot{q} = f'(q, u)$, as we can derive by a reparameterization argument similar to the one above. Thus $q(0)$ belongs to the interior of the attainable set from $q(T)$ at time T for any system δ-close to $\dot{q} = -f(q, u)$.

3. Statement of the results and some general considerations. In this section we explicitly furnish a partition of all Lie bracket configurations with zero or one relation of linear dependency between its elements and we state the corresponding bound on the number of switches that we will prove in the next sections. For the sake of conciseness we set

$$X_{\pm} = [f \pm g, [f, g]],$$
$$X_{\pm\mp} = [f \pm g, [f \mp g, [f, g]]],$$

and, in general, if w is a word with letters in $\{+, -\}$, we set $X_{\pm w} = [f \pm g, X_w]$. The two classes of Lie bracket configurations we will consider are characterized as follows:

Case 0.
$g \wedge [f, g] \wedge X_+ \neq 0,$
$g \wedge [f, g] \wedge X_- \neq 0.$

Case 1.
$g \wedge [f, g] \wedge X_+ = 0,$
$g \wedge [f, g] \wedge X_{++} \neq 0,$
$g \wedge [f, g] \wedge X_- \neq 0.$

Every set of equalities and inequalities has to be interpreted as evaluated at q_0. We say that a Lie bracket configuration is of type 0 or 1 to mean that it satisfies the

conditions of, respectively, Case 0 or 1. This partition is complete since, if in (1) we substitute g by $-g$, we obtain exactly the same system (due to the symmetry of the control set), but the roles of $+$ and $-$ are transposed.

DEFINITION. *Given a point q_0 of type 0 (respectively, 1), we say that a neighborhood U of q_0 is adapted if it is precompact and the relations in inequality form characterizing Case 0 (respectively, Case 1) hold throughout $\overline{U}$.*

Our main result is the following.

THEOREM 3. *Let q_0 be a point of M^3 whose Lie bracket configuration is of type 0 or 1 and fix an adapted neighborhood U of q_0. Then there exists a time $T > 0$ for which a quasi-optimal trajectory of the system (1) contained in U and of time-length less than or equal to T is a finite concatenation of bang and singular arcs. In particular, for T small enough, if we are in the conditions of Case 0, a quasi-optimal trajectory is a concatenation of at most three bang arcs or of a bang, a singular, and a bang arc; if we are in the conditions of Case 1, a quasi-optimal trajectory is a concatenation of at most four bang arcs or of two bang, a singular, and a bang arc and the only possible maximal concatenations including singular arcs are of the type $-+S\pm$ or $\pm S+-$.*

Given an extremal trajectory $q(\cdot)$ and an associated covector trajectory $p(\cdot)$ we can define the so-called *switching function*

$$\varphi(t) = \langle p(t), g(q(t))\rangle.$$

The minimality condition (4) in the Pontryagin maximum principle implies that φ assumes the value zero at the switching times of $q(\cdot)$. From Lemma 1 in [1] we know that, given a smooth vector field X, for almost every t in the domain of $q(\cdot)$, we have

$$\frac{d}{dt}\langle p(t), X(q(t))\rangle = \langle p(t), [f + u(t)g, X](q(t))\rangle. \tag{9}$$

In particular, for almost every t we have

$$\dot\varphi(t) = \langle p(t), [f, g](q(t))\rangle.$$

This equality holds, moreover, for every t since φ is absolutely continuous and $t \mapsto \langle p(t), [f, g](q(t))\rangle$ is (absolutely) continuous. Therefore φ is a $\mathcal{C}^1$ function, its derivative is absolutely continuous, and

$$\ddot\varphi(t) = \langle p(t), [f + u(t)g, [f, g]](q(t))\rangle$$

almost everywhere. We stress that these considerations on the switching function do not depend on the dimension of the manifold. A technical consequence, which we will state in dimension three, is the following.

LEMMA 1. *Let X_i, $i = 1, 2, 3$, be three smooth vector fields on M^3. Let U be a precompact subset of M^3 such that $X_1 \wedge X_2 \wedge X_3 \neq 0$ in $\overline{U}$. Then there exists $T > 0$ such that for each interval I of time-length less than or equal to T, for each extremal trajectory $q : I \to U$ and for each corresponding covector trajectory $p(\cdot)$, if both $x_1(\cdot)$ and $x_2(\cdot)$ have at least one zero in I, then the sign of $x_3(\cdot)$ is constant on I, where $x_i(t) := \langle p(t), X_i(q(t))\rangle$, $i = 1, 2, 3$.*

Proof. Define an Euclidean structure on the cotangent bundle T^*U as follows: for every $q \in U$ and every $p \in T^*_q(M^3)$, let

$$\|p\|^2 = \langle p, X_1(q)\rangle^2 + \langle p, X_2(q)\rangle^2 + \langle p, X_3(q)\rangle^2.$$

From the Gronwall inequality applied to (3), it follows that, fixed $0 < c < 1 < C$, there exists $T > 0$ such that, for every $q(\cdot)$ and $p(\cdot)$ as in the hypothesis of the lemma, with $p(\cdot)$ normalized in such a way that $\|p(t_0)\| = 1$ at the middle point of I, we have $c \leq \|p(t)\| \leq C$ for every $t \in I$. From the expressions of $\dot{x}_i$ given in (9) we easily deduce that there exists a constant K independent of T such that

$$|x_i(t)| \leq TCK\,, \qquad i = 1, 2\,,$$

for every $t \in I$. Thus

$$|x_3(t)|^2 = \|p(t)\|^2 - |x_1(t)|^2 - |x_2(t)|^2 \geq c^2 - 2T^2C^2K^2\,.$$

Taking a small enough T the lemma is proved. □

4. The bang-bang case. In this section we aim to prove Theorem 3 for trajectories which are already known to be bang-bang. Fix a point $q_0 \in M^3$ of type 0 or 1 and an adapted neighborhood U of q_0. Let $q(\cdot)$ be an extremal $-+-+$ trajectory in U with consecutive switching times $0, t_1, t_1+t_2$. This means that $q : [-\eta, t_1+t_2+\eta] \to U$ for some $\eta > 0$ and that u takes value $+1$ on $(0, t_1) \cup (t_1 + t_2, t_1 + t_2 + \eta)$ and -1 on $(-\eta, 0) \cup (t_1, t_1 + t_2)$. We will always assume that $t_1 + t_2 < T$, where T is independent of the trajectory.

Since a switching time is a zero of φ, we have

$$0 = \langle p_0, g(\bar{q})\rangle, \tag{10}$$

$$0 = \langle p_0, e^{t_1 \mathrm{ad}(f+g)} g(\bar{q})\rangle, \tag{11}$$

$$0 = \langle p_0, e^{t_1 \mathrm{ad}(f+g)} e^{t_2 \mathrm{ad}(f-g)} g(\bar{q})\rangle, \tag{12}$$

where $p_0 = p(0)$ and $\bar{q} = q(0)$. We show in the following the uniqueness of the covector p_0, at least if T is small enough. This will allow us to apply Theorem 2 and to formulate necessary conditions for the quasi-optimality of the studied trajectory in terms of the index of

$$Q(\alpha) = \sum_{0 \leq i < j \leq 3} \alpha_i \alpha_j \langle p_0, [h_i, h_j](\bar{q})\rangle,$$

defined for all $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3) \in \mathbf{R}^4$ such that

$$\sum_{i=0}^{3} \alpha_i h_i(\bar{q}) = 0, \tag{13}$$

$$\sum_{i=0}^{3} \alpha_i = 0, \tag{14}$$

where

$$\begin{aligned} h_0 &= f - g, \\ h_1 &= f + g, \\ h_2 &= e^{t_1 \mathrm{ad}(f+g)}(f - g), \\ h_3 &= e^{t_1 \mathrm{ad}(f+g)} e^{t_2 \mathrm{ad}(f-g)}(f + g). \end{aligned}$$

Remark that the extremality condition (4) implies that $\langle p_0, h_i(\bar{q})\rangle = c$ for $i = 0, \ldots, 3$. For the sake of conciseness we write

$$\sigma_{ij} = \langle p_0, [h_i, h_j](\bar{q})\rangle. \tag{15}$$

Since we want to apply the second order condition to all trajectories of the bang-bang type described above which are contained in U, we will consider Q as a family of quadratic forms parameterized also by the point $\bar{q}$.

Choose a triple $\xi_1,\ \xi_2,\ \xi_3$ of 1-forms on M^3 such that their restriction to U is a dual (moving) basis to $g,\ [f,g],\ -X_-$ (that is, $\langle \xi_i(q), g(q)\rangle = \delta_{1i},\ q \in U,\ i = 1,2,3$, and so on). Equality (10) implies that p_0 is a linear combination of $\xi_2(\bar{q})$ and $\xi_3(\bar{q})$. Moreover, the case in which p_0 is proportional to $\xi_2(\bar{q})$ can be excluded choosing T small enough, as we can easily deduce from (11). Since $p(\cdot)$ is defined up to multiplication by a positive scalar we may normalize:

$$p_0 = \pm(\varepsilon\xi_2(\bar{q}) + \xi_3(\bar{q})). \tag{16}$$

For every word w with letters in $\{+,-\}$, we set

$$\lambda_w = \langle \xi_3, X_w\rangle.$$

Let us stress that these are functions defined on U and that, by definition, $\lambda_- \equiv -1$. Remark, moreover, that, due to our choice of U, λ_+ is separated from zero in Case 0, while the same is true for λ_{++} in Case 1.

In the following we will evaluate the asymptotics of various quantities as the time-length of the trajectory goes to 0. To do it we will consider functions of t and q defined on $(0,T)\times U$ and we will say that $\chi(t,q)$ is of order r with respect to t (we will write $\chi(t,q) = O(t^r)$) if it is actually of order r uniformly in U as t goes to 0.

From (11) we have

$$0 = t_1\langle \varepsilon\xi_2(\bar{q}) + \xi_3(\bar{q}), [f,g](\bar{q})\rangle + \langle \varepsilon\xi_2(\bar{q}) + \xi_3(\bar{q}), O(t_1^2)\rangle = \varepsilon(t_1 + O(t_1^2)) + O(t_1^2),$$

and thus for T small enough we can think at ε as a function of t_1 and $\bar{q}$. Remark, moreover, that, as a function on $(0,T)\times U$, ε is of order 1 with respect to t_1.

Take T, the bound on the length of the trajectory, small enough to apply Lemma 1 to the triple of vector fields g, $[f,g]$, and X_-. Let us check that $q(\cdot)$ actually fits the hypothesis of the lemma: the roles of x_1 and x_2 are played here by φ and $\dot{\varphi}$ and φ takes the value zero at all switching times, while between two switching times $\dot{\varphi}$ must have at least one zero, corresponding to a maximum or a minimum of the switching function. The conclusion that we derive from the lemma is that $\langle p(t), X_-(q(t))\rangle$ has constant sign. This sign has to be equal to -1, since the function under consideration is the second derivative of φ in the interval (t_1, t_1+t_2). Since $\langle \varepsilon\xi_2(\bar{q}) + \xi_3(\bar{q}), X_-(\bar{q})\rangle$ depends continuously on ε, it is clear that, possibly for a further smaller T, we can solve the sign uncertainty in (16) and write

$$p_0 = \varepsilon(t_1, \bar{q})\xi_2(\bar{q}) + \xi_3(\bar{q}).$$

This proves the stated uniqueness of p_0. Remark that

$$\varepsilon = \langle p_0, [f,g](\bar{q})\rangle = \dot{\varphi}(0) \le 0, \tag{17}$$

since φ is nonpositive in a right neighborhood of 0.

Remark. It turns out that the subspace of $\mathbf{R}^4$ defined by (13) and (14) is one-dimensional. Equivalently, let us show that

$$V = \left\{ \sum_{i=0}^{3} \alpha_i h_i(\bar{q}) \;\middle|\; \sum_{i=0}^{3} \alpha_i = 0,\ (\alpha_0,\alpha_1,\alpha_2,\alpha_3) \in \mathbf{R}^4 \right\}$$

is two-dimensional. Let $\lambda \in T^*_{\bar{q}} M^3$ such that $\langle \lambda, V \rangle = 0$. In particular,

$$\langle \lambda, h_i(\bar{q}) - h_{i-1}(\bar{q}) \rangle = 0$$

for $i = 1, 2, 3$, and then λ verifies the relations (10), (11), and (12), as can be derived from (5). Thus λ is proportional to p_0, which means that the orthogonal space to V is one-dimensional and the remark is proved. Moreover, the space of admissible α is given by the solutions of the system

$$(18) \quad \begin{cases} \alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 = 0, \\ -\alpha_0 + \alpha_1 + \langle \xi_1(\bar{q}), (h_2 - f)(\bar{q}) \rangle \alpha_2 + \langle \xi_1(\bar{q}), (h_3 - f)(\bar{q}) \rangle \alpha_3 = 0, \\ \langle \xi_2(\bar{q}), (h_2 - f)(\bar{q}) \rangle \alpha_2 + \langle \xi_2(\bar{q}), (h_3 - f)(\bar{q}) \rangle \alpha_3 = 0. \end{cases}$$

4.1. Case 0. Let q_0 be of type 0. From (11) and recalling that $\varepsilon = O(t_1)$, we have

$$0 = t_1 \langle p_0, [f, g](\bar{q}) \rangle + \frac{t_1^2}{2} \langle p_0, X_+(\bar{q}) \rangle + O(t_1^3) = \varepsilon t_1 + \frac{t_1^2}{2} \lambda_+ + O(t_1^3).$$

Thus

$$(19) \qquad \varepsilon = -t_1 \frac{\lambda_+}{2} + O(t_1^2).$$

Similarly, from (12) we get

$$t_2 = 2(\varepsilon + t_1 \lambda_+) + O(t_2^2) = \lambda_+ t_1 + O(t_1^2).$$

In particular, t_2 is of the same order as t_1 and we have

$$\begin{aligned} h_2 &= f - g - 2t_1 [f, g] + O(t_1^2), \\ h_3 &= f + g + 2t_2 [f, g] + O(t_1^2), \end{aligned}$$

and so

$$\begin{aligned} \sigma_{01} &= 2\varepsilon, \\ \sigma_{02} &= 2t_1 + O(t_1^2), \\ \sigma_{12} &= -2\varepsilon - 2t_1 \lambda_+ + O(t_1^2), \\ \sigma_{03} &= 2\varepsilon - 2t_2 + O(t_1^2), \\ \sigma_{13} &= 2t_2 \lambda_+ + O(t_1^2), \\ \sigma_{23} &= 2\varepsilon - 2t_2 + 2t_1 \lambda_+ + O(t_1^2). \end{aligned}$$

The system (18) has the following form:

$$\begin{cases} \alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 = 0, \\ -\alpha_0 + \alpha_1 - (1 + O(t_1^2))\alpha_2 + (1 + O(t_1^2))\alpha_3 = 0, \\ - (2t_1 + O(t_1^2))\alpha_2 + (2t_2 + O(t_1^2))\alpha_3 = 0, \end{cases}$$

from which we obtain

$$\begin{cases} \alpha_0 = -(\frac{t_2}{t_1} + O(t_1))\alpha_3, \\ \alpha_1 = -(1 + O(t_1^2))\alpha_3, \\ \alpha_2 = (\frac{t_2}{t_1} + O(t_1))\alpha_3. \end{cases}$$

Finally, after some calculations,

$$Q = Q(\alpha_3) = \left(-2t_1\lambda_+^2 + O(t_1^2)\right)\alpha_3^2.$$

Recalling that λ_+ is separated from zero we have that Q is negative definite (for small T), and so $q(\cdot)$ is essential.

This completes the proof that in the hypothesis of Case 0 locally at q_0 a bang-bang quasi-optimal trajectory has no more than two switches. We remark that the choice we made between the order of $+$ and $-$ arcs is irrelevant since the hypotheses we were starting from are symmetric in the two signs.

4.2. Case 1. Let now q_0 be of type 1. We will find it useful to introduce the following notation: given a function χ of t and q defined on $(0,T)\times U$ we write $\chi(t,q) = \Omega(t)$ if $\chi(t,q) = \lambda_+(q)O(1) + O(t)$. We also set $\pi_\star = \langle p_0, X_\star\rangle$, $\star = +, -, \dots$.

From (11) we have

$$\begin{aligned} 0 &= t_1\varepsilon + \frac{t_1^2}{2}\pi_+ + \frac{t_1^3}{6}\pi_{++} + O(t_1^4) \\ &= t_1\left(\varepsilon(1+O(t_1)) + t_1\frac{\lambda_+}{2} + t_1^2\frac{\lambda_{++}}{6} + O(t_1^3)\right). \end{aligned}$$

Thus

$$\varepsilon = -t_1\frac{\lambda_+}{2} - t_1^2\frac{\lambda_{++}}{6} + \lambda_+O(t_1^2) + O(t_1^3) = -t_1\frac{\lambda_+}{2} - t_1^2\frac{\lambda_{++}}{6} + t_1^2\Omega(t_1).$$

From (12) we get similarly

$$t_2 = t_1\lambda_+ + \frac{2}{3}\lambda_{++}t_1^2 + \lambda_+O(t_1^2) + O(t_1^3) = t_1\lambda_+ + \frac{2}{3}\lambda_{++}t_1^2 + t_1^2\Omega(t_1).$$

Remark that, in our notation, $t_2, \varepsilon = t_1\Omega(t_1)$. We have

$$\begin{aligned} h_2 &= f - g - 2t_1[f,g] - t_1^2X_+ + O(t_1^3), \\ h_3 &= f + g + 2t_2[f,g] + 2t_1t_2X_+ + t_1t_2\Omega(t_1), \end{aligned}$$

and so

$$\begin{aligned} \sigma_{01} &= 2\varepsilon, \\ \sigma_{02} &= -2t_1\pi_- + O(t_1^2), \\ \sigma_{12} &= -2\varepsilon - 2t_1\pi_+ - t_1^2\pi_{++} + O(t_1^3), \\ \sigma_{03} &= 2\varepsilon + 2t_2\pi_- + t_2O(t_1), \\ \sigma_{13} &= 2t_2\pi_+ + 2t_1t_2\pi_{++} + t_1t_2\Omega(t_1), \\ \sigma_{23} &= 2\varepsilon + 2t_2\pi_- + 2t_1\pi_+ + t_1^2\pi_{++} + t_1^2\Omega(t_1). \end{aligned}$$

The space of admissible α is determined by the system (18), which has the form

$$\begin{cases} \alpha_0 & +\alpha_1 & +\alpha_2 & & +\alpha_3 & = & 0, \\ -\alpha_0 & +\alpha_1 & -(1+O(t_1^2))\alpha_2 & & +(1+t_2O(t_1))\alpha_3 & = & 0, \\ & & -(2t_1+O(t_1^2))\alpha_2 & & +(2t_2+t_2O(t_1))\alpha_3 & = & 0, \end{cases}$$

from which we obtain

$$\begin{cases} \alpha_0 &= -(\frac{t_2}{t_1} + t_2O(1))\alpha_3, \\ \alpha_1 &= -(1+t_2O(t_1))\alpha_3, \\ \alpha_2 &= (\frac{t_2}{t_1} + t_2O(1))\alpha_3. \end{cases} \tag{20}$$

Thus,

$$\begin{aligned} Q(\alpha_3) = \Bigg[& 2\varepsilon\frac{t_2}{t_1} + 2t_1\pi_-\frac{t_2^2}{t_1^2} + (2\varepsilon + 2t_1\pi_+ + t_1^2\pi_{++})\frac{t_2}{t_1} - (2\varepsilon + 2t_2\pi_-)\frac{t_2}{t_1} \\ & - (2t_2\pi_+ + 2t_1t_2\pi_{++}) + (2\varepsilon + 2t_2\pi_- + 2t_1\pi_+ + t_1^2\pi_{++})\frac{t_2}{t_1} \\ & + t_1t_2\Omega(t_1)\Bigg]\alpha_3^2 \\ = & \left[2\frac{t_2}{t_1}(2\varepsilon + t_1\lambda_+ + t_2\lambda_-) + t_1t_2\Omega(t_1)\right]\alpha_3^2 \\ = & -2t_2(\lambda_+ + t_1\lambda_{++} + t_1\Omega(t_1))\alpha_3^2. \end{aligned}$$

If Q is nonnegative definite, then, since $\varepsilon \le 0$, the following system of inequalities is satisfied:

$$\begin{cases} \lambda_+ + t_1\lambda_{++} & + & t_1\Omega(t_1) & \le & 0, \\ -t_1(\lambda_+ + t_1\frac{\lambda_{++}}{3}) & + & t_1^2\Omega(t_1) & \le & 0, \end{cases}$$

and so

$$\begin{cases} \lambda_+(1 + O(t_1)) & + & t_1\lambda_{++}(1 + O(t_1)) & \le & 0, \\ -\lambda_+(1 + O(t_1)) & - & t_1\lambda_{++}\left(\frac{1}{3} + O(t_1)\right) & \le & 0, \end{cases}$$

from which we deduce

$$\frac{2}{3}t_1\lambda_{++} + O(t_1^2) \le 0.$$

Since λ_{++} is uniformly bounded away from 0 on U, a necessary condition for Q to be nonnegative definite (and even more so a necessary condition for $q(\cdot)$ to be quasi-optimal) is that $\lambda_{++} < 0$. In particular, (recall that $\lambda_- = -1$) if

$$(\operatorname{sign} g \wedge [f,g] \wedge X_-(q_0))(\operatorname{sign} g \wedge [f,g] \wedge X_{++}(q_0)) < 0$$

(that is, if the orientation of the two triples of vectors does not coincide), then a $-+-+$ trajectory of small enough length contained in U is not quasi-optimal. If the above signs coincide, then they are different for the same system with reversed time, that is, where f and g are substituted, respectively, by $-f$ and $-g$. Since the time reversed of a quasi-optimal trajectory of the old system is quasi-optimal for the new one, we have that a $+-+-$ trajectory of the initial system of small enough length contained in U is not quasi-optimal.

We conclude that a short enough quasi-optimal bang-bang trajectory lying in U has no more than three switches, in both Cases 0 and 1. In particular, we have proved that a quasi-optimal trajectory in U has no chattering control.

4.3. Sharpness of the result. The bound given in Case 0 is clearly sharp, as we can realize by a purely dimensional reasoning. To investigate whether the extra switch that we add in Case 1 is needed or not, we refer to the sufficiency condition for optimality proved in [4, Theorem 2.6]: the same reasoning applied to the present case implies that if the quadratic form Q is positive definite, then the corresponding

trajectory $q : [0,T] \to M^3$ is locally time-optimal, in the $\mathcal{C}^0$ sense; that is, there exists a neighborhood U of the graph $\{(t,q(t)) \mid t \in [0,T]\}$ in $[0,T] \times M^3$ such that $q(\cdot)$ is time-optimal between all the admissible trajectories whose graph is contained in U. Our aim is to use this result to show that the bound we gave is sharp in the following sense.

LEMMA 2. *Let $\mathbf{J}^4_{q_0}$ be the space of 4-jets at q_0 of pairs of smooth vector fields on M^3 and let $C_1 \subset \mathbf{J}^4_{q_0}$ be the set of all $J^4_{q_0}(f,g)$ such that (f,g) is of type 1 at q_0. Then there exists an open nonempty subset A_1 of C_1 such that if the 4-jet of (f,g) belongs to A_1, then for any $T > 0$ there exists a trajectory with three switches, passing through q_0 and of time length smaller than T, which is locally time-optimal in the $\mathcal{C}^0$ sense.*

Assume that q_0 is a point of type 1. Let $q(\cdot)$ be a $+-+-$ extremal trajectory with switching times $0, t_1, t_1+t_2$ and such that $q(0) = q_0$. In this subsection we will write $O(t^\gamma)$, $\gamma > 0$, simply to denote functions of $t \in (0,T)$ which are of order t^γ as t goes to 0 (the first switching point is now fixed). The equations for the switching times are

$$0 = \langle p_0, g(q_0)\rangle, \tag{21}$$

$$0 = \langle p_0, e^{t_1 \mathrm{ad}(f-g)} g(q_0)\rangle, \tag{22}$$

$$0 = \langle p_0, e^{t_1 \mathrm{ad}(f-g)} e^{t_2 \mathrm{ad}(f+g)} g(q_0)\rangle. \tag{23}$$

In terms of $\xi_1,\ \xi_2,\ \xi_3$, defined as before, we have

$$p_0 = \varepsilon \xi_2(q_0) + \xi_3(q_0),$$

with $\varepsilon \geq 0$. Remark that now

$$\begin{aligned} h_0 &= f+g, \\ h_1 &= f-g, \\ h_2 &= e^{t_1 \mathrm{ad}(f-g)}(f+g), \\ h_3 &= e^{t_1 \mathrm{ad}(f-g)} e^{t_2 \mathrm{ad}(f+g)}(f-g). \end{aligned}$$

We follow the same procedure as in the previous subsections: we use (22) and (23) to establish the asymptotics of ε and t_2; from (13) and (14) we find the relations among the α_i and we compute Q. The calculations are here more lengthy since the first two computable orders of Q—order $t_1^{3/2}$ and order t_1^2—both annihilate.

In detail, from (22) and (23) we have

$$\varepsilon = \frac{t_1}{2} - \frac{t_1^2}{6}\lambda_{--} + O(t_1^3) \tag{24}$$

and

$$t_2^2 = \frac{3}{\lambda_{++}} t_1 + O(t_1^{3/2}).$$

We can now compute

$$\begin{aligned} h_2 &= f + g + 2t_1[f,g] + t_1^2 X_- + O(t_1^3), \\ h_3 &= f - g - 2t_2[f,g] - t_2^2 X_+ - 2t_1 t_2 X_- - \frac{t_2^3}{3} X_{++} + O(t_1^2), \end{aligned}$$

and

$$\begin{aligned}
\sigma_{01} &= -2\varepsilon,\\
\sigma_{02} &= -2t_1 + t_1^2\pi_{+-} + O(t_1^3),\\
\sigma_{12} &= 2\varepsilon - 2t_1 + t_1^2\pi_{--} + O(t_1^3),\\
\sigma_{03} &= -2\varepsilon - 2t_2\pi_{+} - t_2^2\pi_{++} - 2t_1t_2\pi_{+-} - \frac{t_2^3}{3}\pi_{+++} + O(t_1^2),\\
\sigma_{13} &= 2t_2 - t_2^2\pi_{-+} - 2t_1t_2\pi_{--} - \frac{t_2^3}{3}\pi_{-++} + O(t_1^2),\\
\sigma_{23} &= \sigma_{03} - \sigma_{12} + 2\varepsilon - 2t_1t_2^2\pi_{\times} + O(t_1^{5/2}),
\end{aligned}$$

where

$$\pi_{\times} = \langle p_0, [[f,g], X_+](q_0)\rangle.$$

The space on which Q is defined turns out to be described by

$$\begin{aligned}
\alpha_0 &= -(1 + O(t_1^{3/2}))\alpha_2,\\
\alpha_1 &= \left(-\frac{t_1}{t_2} + t_1\frac{\delta_+}{2} + \frac{t_1t_2}{2}\left(\frac{\delta_{++}}{3} - \frac{\delta_+^2}{2} - \gamma_+\right) + O(t_1^2)\right)\alpha_2,\\
\alpha_3 &= \left(\frac{t_1}{t_2} - t_1\frac{\delta_+}{2} - \frac{t_1t_2}{2}\left(\frac{\delta_{++}}{3} - \frac{\delta_+^2}{2}\right) + O(t_1^2)\right)\alpha_2,
\end{aligned}$$

where

$$\begin{aligned}
\gamma_+ &= \langle \xi_1(q_0), X_+(q_0)\rangle\,,\\
\delta_\star &= \langle \xi_2(q_0), X_\star(q_0)\rangle\,, \qquad \star = +, ++\,.
\end{aligned}$$

Finally, we get

$$\begin{aligned}
Q = Q(\alpha_2) = \Bigg[t_1^2t_2\Bigg(&\frac{\lambda_{-++} + \delta_{++}}{3} - \delta_+^2 - \lambda_{-+}\delta_+ - \gamma_+\\
&- 2\lambda_{\times} + \frac{2}{9}\lambda_{++}\lambda_{--}\Bigg) + O(t_1^3)\Bigg]\alpha_2^2.
\end{aligned}$$

To conclude the proof of Lemma 2 it suffices to exhibit a system on $\mathbf{R}^3$ for which at the point $q_0 = 0$ the Lie bracket configuration is of type 1 and the following sign conditions hold: $\lambda_{++} > 0$ and

$$\frac{\lambda_{-++} + \delta_{++}}{3} - \delta_+^2 - \lambda_{-+}\delta_+ - \gamma_+ - 2\lambda_{\times} + \frac{2}{9}\lambda_{++}\lambda_{--} > 0.$$

This is the case, for instance, of the control system:

$$f(x,y,z) = \begin{bmatrix} 0\\ -x\\ -\frac{x^2}{4} - \frac{x^3}{4} + \frac{y}{2} + y^2 - z\end{bmatrix}, \qquad g(x,y,z) = \begin{bmatrix}1\\0\\0\end{bmatrix},$$

for which we get $\lambda_{++} = 2$ and

$$Q = Q(\alpha_2) = \left(\frac{17}{6}t_1^2t_2 + O(t_1^3)\right)\alpha_2^2.$$

It happens that not only the bang-bang part of Theorem 3 is sharp but also the one involving singular arcs. The results in [15] imply, indeed, that the presence of short time-optimal concatenations of two bang, one singular, and one bang arc is structurally stable in Case 1. We stress [5] that in this case the optimality is proved to be not only local, but also global.

5. Allowing singular arcs. Let $q_0 \in M^3$ be a point of type 0 or 1 and U an adapted neighborhood of q_0. Consider a quasi-optimal trajectory $q : (T_1, T_2) \to U$ with no preliminary restriction on the structure of its corresponding control function u. Let $p : (T_1, T_2) \to T^*M^3$ be an associated covector trajectory.

It is known from Proposition 1 in [1] that, after possibly a modification on a set of measure zero, u is $\mathcal{C}^\infty$ on a open dense subset O of (T_1, T_2). Assume that O is maximal (also with respect to further modifications of u on sets of measure zero). An *arc* is a piece of trajectory corresponding to a connected component of O; it is *bang* if u takes value in $\{-1, 1\}$ on it, otherwise it is *singular*. We will use the word arc also to refer to the connected component of O itself. Remark that on an arc also φ is $\mathcal{C}^\infty$ and all its derivatives can be computed iterating (9). We say that two distinct arcs (τ_1, τ_2) and (t_1, t_2) are *concatenated* if $\tau_2 = t_1$ or $\tau_1 = t_2$. Let $\Sigma = \partial O$. We say that two distinct points of Σ are *subsequent* if the open interval which they identify does not intersect Σ (that is, by density of O, if it is an arc).

In what follows we always assume that $T_2 - T_1 < T$, for $T > 0$, which will be considered as small as needed. In particular, we will assume T to be small enough to deduce from Lemma 1 the following property: if both φ and $\dot\varphi$ have at least one zero in (T_1, T_2), then $\psi_+(\cdot)$ and $\psi_-(\cdot)$ in Case 0 ($\psi_-(\cdot)$ and $\psi_{++}(\cdot)$ in Case 1) do not change sign on (T_1, T_2), where

$$\psi_\star(t) = \langle p(t), X_\star(q(t))\rangle.$$

Remark that if, for instance, $q(\cdot)$ has a bang arc compactly contained in (T_1, T_2), then both φ and $\dot\varphi$ have, indeed, at least one zero in (T_1, T_2).

We start from the situation in which there is no bang arc. If this is the case, then the switching function is identically equal to zero: to prove it, by density of O, we just need to show that if I is a singular arc, then $\varphi|_I \equiv 0$. This is indeed the case: assume that I is singular and that $\varphi|_I$ is not identically equal to zero. Then

$$J = \mathrm{int}\{\tau \in I | \varphi(\tau) = 0\}$$

is a proper nonempty subset of I. Let t be in the boundary of J and in the interior of I. By continuity we obtain that both $|u(t)| = 1$ and $\varphi^{(n)}(t) = 0$ for every $n \geq 0$. As remarked, however, $\varphi^{(n)}(t)$ can be computed iterating (9). It follows, from the nondegeneracy conditions of both Cases 0 and 1, that $p(t)$ is equal to 0, which is impossible.

Thus, if there is no bang arc, φ is identically equal to zero on (T_1, T_2). We want to deduce that the trajectory is made of a single singular arc, by proving that if $\varphi \equiv 0$ on an open interval I, then u is smooth on I. Indeed, both $\dot\varphi$ and its further derivatives are also identically equal to zero on I and so $p(t)$ is orthogonal to both $g(q(t))$ and $[f, g](q(t))$ and, for almost all $t \in I$,

$$\langle p(t), [f, [f, g]](q(t))\rangle + u(t)\langle p(t), [g, [f, g]](q(t))\rangle = 0\,. \tag{25}$$

We remark that in both Cases 0 and 1

$$\mathrm{span}\{g(q), [f, g](q), [f, [f, g]](q), [g, [f, g]](q)\} = T_qM^3$$

for every $q \in U$. Thus $\langle p(t), [g,[f,g]](q(t))\rangle \neq 0$ for every t for which (25) holds, otherwise $p(t) \neq 0$ would annihilate $T_{q(t)}M^3$. If $\langle p(\bar t), [g,[f,g]](q(\bar t))\rangle = 0$ for some $\bar t \in I$, however, we would have that near $\bar t$ the function $t \mapsto \langle p(t), [f,[f,g]](q(t))\rangle$ is bounded away from zero and consequently $|u(t)| > 1$ for some t at which (25) holds. Thus, for every $t \in I$, $\langle p(t), [g,[f,g]](q(t))\rangle \neq 0$ and

$$u(t) = -\frac{\langle p(t), [f,[f,g]](q(t))\rangle}{\langle p(t), [g,[f,g]](q(t))\rangle}. \tag{26}$$

Substituting the last expression in the Hamiltonian, we find that $p|_I$ is a solution of the smooth (autonomous) Hamiltonian system generated by the Hamiltonian

$$h(p) = \langle p, f\rangle - \frac{\langle p, [f,[f,g]]\rangle}{\langle p, [g,[f,g]]\rangle}\langle p, g\rangle,$$

and, in particular, it is smooth. According to (26), the same is true for $u|_I$.

If no bang arc $(\tau_1, \tau_2) \subset\subset (T_1, T_2)$ exists, then, by the same reasoning, the trajectory is a concatenation of at most a bang, a singular, and a bang arc. Let now $(\tau_1, \tau_2) \subset\subset (T_1, T_2)$ be a bang arc. We can associate with it the smaller point t_2 of Σ which satisfies the following conditions: $t_2 \geq \tau_2$, t_2 is the upper bound of a bang arc, and t_2 is not the lower bound of any bang arc. (This is possible since the results of the previous section exclude the existence of an infinite sequence of subsequent bang arcs.) Analogously we can define $t_1 \leq \tau_1$. If the trajectory is not bang-bang, then either $t_1 \neq T_1$ or $t_2 \neq T_2$. Assume that $t_2 \neq T_2$. Then $p(t_2)$ is orthogonal to both $g(q(t_2))$ and $[f,g](q(t_2))$. Indeed, by definition of t_2 in each of its (right) neighborhoods lies the interior point of a singular arc or an entire bang arc, on which $\varphi(t)$ necessarily assumes a zero and a maximum or minimum. By continuity we deduce the stated orthogonality. Denote by I the arc having t_2 as upper bound and by ν be the control corresponding to such arc. Since $\varphi(t_2) = \dot\varphi(t_2) = 0$ and I is compactly contained in (T_1, T_2), it is clear that $\psi_\nu(\cdot)$, the second derivative of $\varphi(\cdot)$ along I, must have a zero in I. Due to our assumptions on T we can exclude that $\nu = -1$, since $\psi_-(\cdot)$ has constant sign along the trajectory in both Cases 0 and 1. For the same reason, the Lie bracket configuration at q_0 cannot be of type 0, and we can restrict our attention to Case 1: therefore $\psi_{++}(\cdot)$ has constant sign along (T_1, T_2) and, since $\psi_{++}(\cdot) = \varphi^{(3)}(\cdot)$ on I, such sign must be -1 (think of the restrictions on the changes of concavity of φ on I). Analogous considerations could be carried out in t_1 in the case $t_1 \neq T_1$, leading up to the opposite sign of ψ_{++}. Thus $t_1 = T_1$. For the same reason there cannot be any bang arc compactly contained in the interval (t_2, T_2). Thus (t_2, T_2) is a singular arc itself or the union of a singular and a bang arc. Up to now we proved that a short enough quasi-optimal trajectory lying in U is the concatenation of at most four bang arcs, a singular arc, and a bang arc. Actually, if the considered trajectory contains a singular arc, an extra quasi-optimality condition is fulfilled: the generalized Legendre condition. (See [8] for a formulation and [2] for a complete mathematical proof.) In the present situation this condition states that $\psi_-(\cdot)$ is positive along the singular arc and so along the entire trajectory. Thus we exclude that a trajectory has both a singular and a compactly contained bang arc corresponding to control -1. Finally the two "maximal" non–bang-bang concatenations are $-{+}\mathrm{S}\pm$ and $\pm\mathrm{S}{+}-$.

6. A result of infinite codimension. This last section presents a result which applies in a more general setting the techniques we used in the above proof. In particular, the result holds in any finite dimension.

PROPOSITION 2. *Take a finite-dimensional manifold M and two smooth vector fields f and g on M. Let X_+, $X_-, \dots$ be defined as usual. We will also write $X_{(m+)}$ for $X_{+\dots+}$, where $+$ is repeated m times. (and $X_{(m-)}$ for the natural counterpart). Assume that the system*

$$\dot{q} = f(q) + u\, g(q), \qquad |u| \le 1\,, \tag{27}$$

is such that for each $q_0 \in M$ and for $\star = +, -$, we have

$$T_{q_0} M = \operatorname{span}\{g(q_0), [f,g](q_0), X_\star(q_0), \dots, X_{(m\star)}(q_0), \dots\} \stackrel{\text{def}}{=} V_\star(q_0)\,.$$

Let $q : (T_1, T_2) \to M$ be an extremal trajectory of the system above such that the corresponding control function u verifies $|u| = 1$ on an open dense subset of (T_1, T_2) and let Σ be the set of discontinuities of u (not avoidable by changing u on a set of measure zero). Then either Σ is discrete or it contains a sequence of infinitely many subsequent isolated points.

Proof. Let O be the maximal open dense subset of (T_1, T_2) on which u, after modification on a set of measure zero, is smooth. Clearly u is constant on any arc, i.e., any connected component of O.

Assume by contradiction that Σ neither is discrete nor contains an infinite sequence of subsequent isolated points. With each (bang) arc (τ_1, τ_2) we can associate the smaller point of Σ which is larger than or equal to τ_2 and which is not isolated, unless there exists a finite sequence of subsequent isolated points including τ_2 and T_2. Denote by A the set of all points which can be associated as described with some arc. By our assumptions A is nonempty and preperfect, i.e., each point of A is a density point for A. We give a partition $A = A_+ \cup A_-$ by defining $A_\star$, $\star = +, -$, as the set of points a of A for which the control u on a left neighborhood of a has constant sign $\star 1$. We have that there exists $\star \in \{+, -\}$ and there exists a subset B of $A_\star$ which is nonempty and preperfect. (Indeed, if $a \in A_+$ is not a density point for A_+, then there exists a neighborhood U of a such that $(U \setminus \{a\}) \cap A$ is a preperfect nonempty subset of A_- and thus either A_+ is nonempty and preperfect or A_- has a preperfect nonempty subset.)

Let $p(\cdot)$ be a covector trajectory associated with $q(\cdot)$. To complete the argument we want to prove that at each point $\tau \in B$ the covector $p(\tau)$ annihilates $V_\star(q(\tau))$. It is clear that $\varphi(\tau) = \langle p(\tau), g(q(\tau))\rangle = 0$ for each $\tau \in \Sigma$. Since on the interval between two subsequent points of Σ there is at least one zero of $\varphi'(\cdot) = \langle p(\cdot), [f,g](q(\cdot))\rangle$, we have that, by continuity, $\langle p(\tau), [f,g](q(\tau))\rangle = 0$ for each $\tau \in A$. Reasoning similarly and using the preperfectness of B we prove by induction on m that $\langle p(\tau), X_{(m\star)}(q(\tau))\rangle = 0$ for each $\tau \in B$. □

Proposition 2 is an improvement of Proposition 2 in [1], where it was proved, under the same hypothesis, that Σ cannot be a perfect set, that is, a closed preperfect set. To express it consistently with Proposition 2, this older result says that either Σ is empty or it contains an isolated point.

REFERENCES

[1] A. A. AGRACHEV, *On regularity properties of extremals controls*, J. Dynam. Control Systems, 1 (1995), pp. 319–324.

[2] A. A. AGRACHEV AND R. V. GAMKRELIDZE, *A second order optimality principle for a time-optimal problem*, Math. USSR-Sb., 29 (1976), pp. 547–576.

[3] A. A. AGRACHEV AND R. V. GAMKRELIDZE, *Symplectic geometry for optimal control*, in Nonlinear Controllability and Optimal Control, Pure Appl. Math. 133, Hector J. Sussmann, ed., Marcel Dekker, New York, 1990, pp. 263–277.

[4] A. A. Agrachev, G. Stefani, and P. Zezza, *Strong optimality for a bang-bang trajectory*, SIAM J. Control Optim., 41 (2002), pp. 991–1014.

[5] B. Bonnard and J. de Morant, *Toward a geometric theory in the time-minimal control of chemical batch reactors*, SIAM J. Control Optim., 33 (1995), pp. 1279–1311.

[6] A. Bressan, *The generic local time-optimal local stabilizing controls in dimension* 3, SIAM J. Control Optim., 24 (1986), pp. 177–190.

[7] A. T. Fuller, *Study of an optimum nonlinear system*, J. Electronics Control, 15 (1963), pp. 63–71.

[8] H. J. Kelley, R. E. Kopp, and H. Gardner Moyer, *Singular extremals*, in Topics in Optimization, Academic Press, New York, 1967, pp. 63–101.

[9] A. J. Krener and H. Schättler, *The structure of small-time reachable sets in low dimensions*, SIAM J. Control Optim., 27 (1989), pp. 120–147.

[10] I. A. K. Kupka, *The ubiquity of Fuller's phenomenon*, in Nonlinear Controllability and Optimal Control, Pure Appl. Math. 133, Hector J. Sussmann, ed., Marcel Dekker, New York, 1990, pp. 313–350.

[11] B. Piccoli, *Regular time-optimal syntheses for smooth planar systems*, Rend. Sem. Mat. Univ. Padova, 95 (1996), pp. 59–79.

[12] H. Schättler, *On the local structure of time-optimal bang-bang trajectories in* $\mathbb{R}^3$, SIAM J. Control Optim., 26 (1988), pp. 186–204.

[13] H. Schättler, *The local structure of time-optimal trajectories in dimension three under generic conditions*, SIAM J. Control Optim., 26 (1988), pp. 899–918.

[14] H. Schättler, *Regularity properties of optimal trajectories: Recently developed techniques*, in Nonlinear Controllability and Optimal Control, Pure Appl. Math. 133, Hector J. Sussmann, ed., Marcel Dekker, New York, 1990, pp. 351–381.

[15] H. Schättler and M. Jankovic, *A synthesis of time-optimal controls in the presence of saturated singular arcs*, Forum Math., 5 (1993), pp. 203–241.

[16] H. J. Sussmann, *Lie brackets, real analyticity and geometric control*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 1–116.

[17] H. J. Sussmann, *Time-optimal control in the plane*, in Feedback Control of Linear and Nonlinear Systems, Lecture Notes in Control and Inform. Sci. 39, Springer-Verlag, Berlin, 1985, pp. 244–260.

[18] H. J. Sussmann, *Envelopes, conjugate points and optimal bang-bang extremals*, in Proceedings of the 1985 Paris Conference on Nonlinear Systems, M. Fliess and M. Hazewinkel, eds., D. Reidel Publishing Co., Dordrecht, The Netherlands, 1986, pp. 325–346.

[19] H. J. Sussmann, *A weak regularity for real analytic optimal control problems*, Rev. Mat. Iberoamericana, 2 (1986), pp. 307–317.

[20] H. J. Sussmann, *The structure of time-optimal trajectories for single-input systems in the plane: The* $\mathcal{C}^\infty$ *nonsingular case*, SIAM J. Control Optim., 25 (1987), pp. 433–465.

[21] H. J. Sussmann, *Envelopes, high order optimality conditions and Lie brackets*, in Proceedings of the 28th IEEE Conference on Decision and Control, Vol. 2, IEEE Press, Piscataway, NJ, 1989, pp. 1107–1112.

[22] H. J. Sussmann, *Synthesis, presynthesis, sufficient conditions for optimality and subanalytic sets*, in Nonlinear Controllability and Optimal Control, Pure Appl. Math. 133, Hector J. Sussmann, ed., Marcel Dekker, New York, 1990, pp. 1–19.

[23] M. I. Zelikin and V. F. Borisov, *Theory of Chattering Control. With Applications to Astronautics, Robotics, Economics, and Engineering*, Systems and Control: Foundations and Applications, Birkhäuser, Boston, 1994.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 532–558

AN ERGODIC CONTROL PROBLEM FOR CONSTRAINED DIFFUSION PROCESSES: EXISTENCE OF OPTIMAL MARKOV CONTROL*

AMARJIT BUDHIRAJA†

Abstract. An ergodic control problem for a class of constrained diffusion processes is considered. The goal is the almost sure minimization of long term cost per unit time. The main result of the paper is that there exists an optimal Markov control for the considered problem. It is shown that under the assumption of regularity of the Skorohod map and the assumption that the drift vector field takes values in a certain cone of stability, the class of controlled diffusion processes considered have strong, uniform in control, stability properties. The role of the boundary is critical in obtaining the stability and ergodic control results for the class of controlled constrained diffusion processes considered in this work since the domains are unbounded and the corresponding unconstrained diffusions are typically transient. These stability properties are key in obtaining appropriate tightness estimates. Once these estimates are available the remaining work lies in identifying weak limits of a certain family of occupation measures. In this regard an extension to the Echeverria–Weiss–Kurtz characterization of invariant measures of Markov processes, to the case of constrained-controlled processes considered in this paper, is proved. This characterization result is also crucially used in proving the compactness of the family of invariant measures of Markov processes corresponding to all possible Markov controls.

Key words. ergodic control, optimal Markov control, controlled reflected diffusions, constrained processes, control of queuing networks, patchwork martingale problem, controlled martingale problem, Echeverria's criterion

AMS subject classifications. 93E20, 60H30, 60J60

PII. S0363012901379073

1. Introduction.

Constrained diffusion processes arise in a natural fashion in the heavy traffic analysis of queuing networks coming from problems in computer, communications, and manufacturing systems. The problem of control of such queuing systems is of great current interest (cf. [31, 20, 23, 19, 32]). Except for a few special cases, the control problem for queuing networks is quite difficult to analyze directly and thus one tries to find more tractable approximations. In that respect, diffusion approximations obtained via appropriate scaling limits become very attractive since in the limit many fine details are eliminated and usually the only parameters remaining are the means, the variances of the various processes, and the mean routing structure. Because of the simple structure, the limit problem is considerably easier to solve. Once the optimal solution to the control problem for the constrained diffusions is obtained, one can then approximate the properties and suggest good policies for the actual physical system. Thus the study of constrained controlled diffusion processes is one of the central objectives in the optimal control of queuing networks.

In this work we consider a control problem for a class of diffusion processes which are constrained to lie in a polyhedral cone G with a vertex at origin. The domain $G \subset \mathbb{R}^k$ is given as an intersection of N half spaces G_i, $i = 1, \ldots, N$. Associated with each G_i are two vectors: the first vector, denoted as n_i, represents the inward normal to G_i while the second, denoted as d_i, gives the "direction of constraint." Roughly speaking, the constrained version of a given unrestricted trajectory in $\mathbb{R}^k$

*Received by the editors July 6, 2001; accepted for publication (in revised form) November 24, 2002; published electronically May 12, 2003.
http://www.siam.org/journals/sicon/42-2/37907.html
†Department of Statistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (budhiraj@email.unc.edu).

is obtained by pushing back the trajectory, whenever it is about to exit the domain, in a prespecified direction of constraint using the minimal force required to keep the trajectory within the domain. Precise definitions will be given in section 2. The constraining mechanism is described via the Skorohod map, denoted as $\Gamma(\cdot)$, which takes an unrestricted trajectory $\psi(\cdot)$ and maps it to a trajectory $\phi(\cdot) \doteq \Gamma(\psi)(\cdot)$ such that $\phi(t) \in G$ for all $t \in (0, \infty)$. Under appropriate conditions on $(d_i, n_i)_{i=1}^N$ it follows from the results in [13] that the Skorohod map is well defined and it enjoys a rather strong regularity property (see Theorem 2.3).

The controlled constrained diffusion process that we consider in this paper are obtained as a solution to the equation

$$X(t) = \Gamma\left(X(0) + \int_0^{\cdot} b(X(s), u(s))ds + \int_0^{\cdot} \sigma(X(s))dW(s)\right)(t), \quad t \in [0, \infty), \tag{1.1}$$

where $W(\cdot)$ is a standard Wiener process, $b : G \times U \to \mathbb{R}^k$, $\sigma : G \to \mathbb{R}^{k\times k}$ are suitable coefficients, U is a given control set, and $u(\cdot)$ is a U valued "admissible" control process. The control problem that we study is concerned with the ergodic cost criterion:

$$\limsup_{T\to\infty} \frac{1}{T} \int_0^T k(X(s), u(s))ds, \tag{1.2}$$

where the limit above is taken a.s. and $k : G \times U \to \mathbb{R}$ is a suitable map. The two key objectives of the controller are, first, to choose a control in a nonanticipative fashion, which minimizes the cost in (1.2) and, second, to obtain a control which is "easy" to implement. With regard to the second goal, one of the most desirable features of a good control is that the control depends only on the current value of the state and not on the whole history of the state and/or the control process. In other words, we are seeking controls $u(\cdot)$ such that there exists some measurable map $v : G \to U$ satisfying $u(t) = v(X(t))$, a.s. for all $t \in [0, \infty)$. Under such a control the solution to (1.1) becomes a Markov process and for this reason the map $v(\cdot)$ is referred to as a "Markov control." The objective of this work is to show that, under appropriate conditions on the model (cf. Conditions 2.2, 2.4, 2.5, 3.1) there is a Markov control which minimizes the cost in (1.2). The ergodic control problem for unconstrained diffusions is one of the classical problems in stochastic control. The problem has been studied extensively in [39, 30, 7, 6, 40, 22, 28]. The approach taken in the present paper has been inspired by the techniques and results in [7]. For ergodic control results on constrained jump-diffusion processes in bounded domains and *expected* long term cost per unit time criterion, we refer the reader to [4, 34, 31]. For the case of constrained diffusions in unbounded domains, we are not aware of any results which give the existence of optimal Markov control under the ergodic cost criterion considered in this paper.

As is classical in an ergodic control problem of the above form (cf. [6, 40, 31, 28]) the problem of existence of optimal Markov controls under the cost criterion in (1.2) is closely related to certain stability properties of the solutions to (1.1). In a recent work [2], which dealt with uncontrolled constrained diffusions, various stability properties of the solution to (1.1) with $b(x, u) \equiv \beta(x)$, where $\beta : \mathbb{R}^k \to \mathbb{R}^k$ is an appropriate drift vector, were obtained. In particular it was shown that if for all $x \in G$, $\beta(x)$ lies in a certain cone $\mathcal{C}$ (see (3.1)) and its distance from the boundary of the cone is uniformly bounded below by a positive constant, then the constrained diffusion is positive recurrent and has a unique invariant measure. The above results identify an

important, nontrivial class of ergodic constrained diffusions in unbounded domains, in the sense that the corresponding unconstrained version of these processes would typically be transient. To see this one needs only to consider the case where $b(x,u) \equiv b$, where b is some fixed vector in $\mathcal{C}^0$. For this latter case ($b(x,u) \equiv b$), the cone, in fact, provides a necessary and sufficient condition for positive recurrence of constrained diffusions, i.e., if $b \notin \mathcal{C}$, then the corresponding constrained diffusion is transient [9]. In the context of the constant drift case, this necessary and sufficient condition for positive recurrence, for constrained diffusions which correspond to single class networks, was first proved in [21].

The estimates used in the study of the stability properties of the uncontrolled form of (1.1) can be used for the controlled problem studied in this paper as well. Using these estimates, we will obtain rather strong (uniform in control) stability properties for the processes obtained as solutions to (1.1) (see section 6). These stability results are then used in various tightness arguments in this paper. As another consequence of results in [2] we have that under any Markov control $v(\cdot)$, under the assumption that the drift vector field $b(\cdot,\cdot)$ takes values in the set $\mathcal{C}(\delta)$ (see (3.2)) for some $\delta > 0$, the solution to (1.1) is positive recurrent and has a unique invariant measure, denoted as η_v. We abbreviate this statement by saying that "all Markov controls are stable." The above condition on the drift will be the "blanket stability" assumption made in most of the results in this work. In the classical theory of ergodic control (cf. [6]), there are generally two kinds of problems studied. The first is the so-called stable case, where a blanket stability condition on the model is assumed. The study of this case is the main goal of this paper. The second case does not assume stability conditions on the model but instead makes certain restrictive conditions on the cost function k which penalize unstable behavior of the controlled process. This is referred to as the "near monotone case" in [6]. The proof of the existence of a Markov control in the near monotone case, under our setup, is essentially the same as that for the classical model considered in [6]. We briefly sketch the argument in the appendix.

Using the "blanket stability" condition and an ergodic theorem of Khasminskii [24] it will follow that under a Markov control $v(\cdot)$ the limit in (1.2) is a.s. equal to

$$\int_G k(x, v(x))\eta_v(dx). \tag{1.3}$$

The next key step in the program is to show that for any admissible control $u(\cdot)$ the limit in (1.2) can be expressed in the form (1.3) for some measurable $v : G \to U$ which may depend on ω (the parameter of randomness), the control $u(\cdot)$, and other data. This is done in Proposition 7.3. As an immediate consequence of this step it follows that

$$\limsup_{T\to\infty} \frac{1}{T} \int_0^T k(X(s), u(s))ds \geq \inf_v \int_G k(x, v(x))\eta_v(dx), \tag{1.4}$$

a.s., where the infimum on the right side above is taken over all measurable maps $v : G \to U$. In order to prove the above step we need a characterization result for the invariant measures of solutions of (1.1) (with $u(\cdot) = v(X(\cdot))$ for some measurable $v(\cdot)$). The characterization results for invariant measures of solutions to martingale problems are classical (cf. [17, 5, 28]). For the class of constrained processes (see (2.2))

considered in this work, one has via an application of Ito's formula that for all $f \in C_0^\infty(G)$

$$f(X(t)) - \int_0^t \int_S (Lf)(X(s),\alpha)u(s)(d\alpha)ds - \sum_{i=1}^N \int_0^t (D_i f)(Z(s))dY_i(s)$$

is a martingale, where L is as in (5.1). Since the local time processes $\{Y_i(\cdot)\}$ are typically not absolutely continuous, this leads us to the study of controlled, *singular*, martingale problems which do not fall in the purview of the results in the above listed references. In section 5, we use the ideas of patchwork and constrained martingale problems introduced by Kurtz [26, 27] to prove a characterization result for invariant measures, suitable for our purposes. It was pointed out by the referee that Theorem 5.7 is a special case of a recently appeared result in [29].

As a final step we need to show that the infimum on the right side of (1.4) is attained for some Markov control $v : G \to U$. This step (Proposition 7.2), combined with the ergodic theorem in [24] yields our main theorem: Theorem 3.4. The fact that the infimum is attained is a consequence of the fact that the family $\{\eta_v : v \text{ is a Markov control}\}$ is compact. The proof of the compactness of this family once more uses the characterization result for the invariant measures studied in section 5 and various tightness estimates, which follow from the stability properties of our controlled diffusions. This is done in Lemma 7.1. Observe that since (1.4) holds a.s., we can replace the left side of the expression by

$$\text{ess inf} \limsup_{T\to\infty} \frac{1}{T} \int_0^T k(X(s),u(s))ds.$$

Thus our main result (Theorem (3.4)) says that there is a Markov control for which the cost, for almost every realization, is no worse than that for the (essentially) best possible realization corresponding to any other control.

In this paper we do not address the problem of construction of the optimal Markov control or the approximation of optimal control for physical queuing systems using the optimal Markov control for the limit diffusion. These questions will be addressed in a future work. One of the important tools in the study of the control problem is the existence and uniqueness of the HJB equation for the value function. However, neither is known for the ergodic control problem studied in this paper and its treatment will be undertaken in a future work.

The paper is organized as follows. In section 2 we present the basic definitions and properties of the Skorohod map. Section 3 introduces the ergodic cost problem that interests us in this work. We give our key condition on the drift vector which assures us that all Markov controls are stable. Finally in this section we state our main result: Theorem 3.4. Section 4 is an assortment of some background results used in the proof of our main theorem. In section 5 we present our extension of the Echeverria–Weiss–Kurtz criterion for the invariant measures, which is a special case of a recently published result by Kurtz and Stockbridge [29]. Section 6 is devoted to obtaining stability properties of our constrained diffusions. This section crucially uses some estimates derived in [2] (cf. Lemma 6.1). As a consequence of these results we prove strong, uniform in control, tightness properties of the solutions of (1.1). Finally, in section 7 we present the proof of Theorem 3.4. In section 8 we comment on the possibility of characterizing the value of the ergodic control problem via a suitable HJB equation.

2. Skorohod map and controlled constrained diffusions. Let $G \subset \mathbb{R}^k$ be a polyhedral cone in $\mathbb{R}^k$ with the vertex at the origin given as the intersection of half spaces G_i, $i = 1, \ldots, N$. Each half space G_i is associated with a unit vector n_i via the relation $G_i = \{x \in \mathbb{R}^k : \langle x, n_i \rangle \geq 0\}$, where $\langle \cdot, \cdot \rangle$ denotes the usual inner product in $\mathbb{R}^k$. Denote the boundary of a set $B \subset \mathbb{R}^k$ by ∂B. We will denote the set $\{x \in \partial G : \langle x, n_i \rangle = 0\}$ by F_i. For $x \in \partial G$, define the set, $n(x)$, of inward normals to G at x by $n(x) \doteq \{r : |r| = 1, \quad \langle r, x - y \rangle \leq 0 \ \ \forall y \in G\}$. With each face F_i we associate a unit vector d_i such that $\langle d_i, n_i \rangle > 0$. This vector defines the *direction of constraint* associated with the face F_i. For $x \in \partial G$ define $d(x) \doteq \{d \in \mathbb{R}^k : d = \sum_{i \in \mathrm{In}(x)} \alpha_i d_i; \alpha_i \geq 0; \ \ ||d|| = 1\}$, where $\mathrm{In}(x) \doteq \{i \in \{1, 2, \ldots, N\} : \langle x, n_i \rangle = 0\}$. We will denote the collection of all subsets of $\{1, \ldots, N\}$ by Λ. Also for $\lambda \in \Lambda$ we will define $F_\lambda \doteq \cap_{i \in \lambda} F_i$. As a convention we will take $F_\emptyset$ as G.

Let $D([0, \infty) : \mathbb{R}^k)$ denote the set of functions mapping $[0, \infty)$ to $\mathbb{R}^k$ that are right continuous and have limits from the left. We endow $D([0, \infty) : \mathbb{R}^k)$ with the usual Skorohod topology. Let $D_G([0, \infty) : \mathbb{R}^k) \doteq \{\psi \in D([0, \infty) : \mathbb{R}^k) : \psi(0) \in G\}$. For $\eta \in D([0, \infty) : \mathbb{R}^k)$ let $|\eta|(T)$ denote the total variation of η on $[0, T]$ with respect to the Euclidean norm on $\mathbb{R}^k$.

Definition 2.1. *Let $\psi \in D_G([0, \infty) : \mathbb{R}^k)$ be given. Then $(\phi, \eta) \in D([0, \infty) : \mathbb{R}^k) \times D([0, \infty) : \mathbb{R}^k)$ solves the Skorohod problem (SP) for ψ with respect to G and d if and only if $\phi(0) = \psi(0)$, and for all $t \in [0, \infty)$ (1) $\phi(t) = \psi(t) + \eta(t)$; (2) $\phi(t) \in G$; (3) $|\eta|(t) < \infty$; (4) $|\eta|(t) = \int_{[0,t]} I_{\{\phi(s) \in \partial G\}} d|\eta|(s)$; (5) there exists (Borel) measurable $\gamma : [0, \infty) \to \mathbb{R}^k$ such that $\gamma(t) \in d(\phi(t))$ ($d|\eta|$-almost everywhere) and $\eta(t) = \int_{[0,t]} \gamma(s) d|\eta|(s)$.*

On the domain $D \subset D_G([0, \infty) : \mathbb{R}^k)$, on which there is a unique solution to the SP, we define the Skorohod map (SM) Γ as $\Gamma(\psi) \doteq \phi$, if $(\phi, \psi - \phi)$ is the unique solution of the SP posed by ψ. We will make the following assumptions on the data defining the SP above.

Condition 2.2. (a) *There exists a compact, convex set $B \in \mathbb{R}^k$ with $0 \in B^0$, such that if $v(z)$ denotes the set of inward normals to B at $z \in \partial B$, then for $i = 1, 2, \ldots, N$, $z \in \partial B$ and $|\langle z, n_i \rangle| < 1$ implies that $\langle v, d_i \rangle = 0$ for all $v \in v(z)$.* (b) *There exists a map $\pi : \mathbb{R}^k \to G$ such that if $y \in G$, then $\pi(y) = y$, and if $y \notin G$, then $\pi(y) \in \partial G$, and $y - \pi(y) = \alpha\gamma$ for some $\alpha \leq 0$ and $\gamma \in d(\pi(y))$.* (c) *For every $x \in \partial G$, there is $n \in n(x)$ such that $\langle d, n \rangle > 0$ for all $d \in d(x)$.*

The above assumptions can be verified for a rich class of problems arising from queuing networks. For example, in the seminal work [18], it was shown that the above properties hold for SPs associated with open, single class queuing networks (cf. [15]). Other classes of network examples for which the above properties hold are in [13, 15, 16, 35, 36]. Condition (c) above is equivalent to the assumption that the $N \times N$ matrix with the (i, j)th entry $\langle d_i, n_j \rangle$ is completely-S (cf. [37, 13]).

The following result is taken from [13].

Theorem 2.3 (see [13]). *Under Condition 2.2 the SM is well defined on all of $D_G([0, \infty) : \mathbb{R}^k)$, i.e., $D = D_G([0, \infty) : \mathbb{R}^k)$ and the SM is Lipschitz continuous in the following sense. There exists a $K < \infty$ such that for all $\phi_1, \phi_2 \in D_G([0, \infty) : \mathbb{R}^k)$*

$$\sup_{0 \leq t < \infty} |\Gamma(\phi_1)(t) - \Gamma(\phi_2)(t)| < K \sup_{0 \leq t < \infty} |\phi_1(t) - \phi_2(t)|. \tag{2.1}$$

In rest of the paper Condition 2.2 will always be taken to hold. We will also assume without loss of generality that $K \geq 1$.

We now introduce the controlled constrained diffusion processes that will be studied in this paper. Throughout this paper we will assume the relaxed control frame-

work, i.e., there is a compact metric space S such that the control set is $U \doteq \mathcal{P}(S)$ (the space of all probability measures on S endowed with the weak convergence topology). All topological spaces in this paper will be endowed with their natural Borel σ-field. For a topological space $\mathcal{K}$, we will denote its Borel σ-field by $\mathcal{B}(\mathcal{K})$. The space of all real, continuous, and bounded functions defined on $\mathcal{K}$ will be denoted as $C_b(\mathcal{K})$ and the space of all probability measures on $(\mathcal{K}, \mathcal{B}(\mathcal{K}))$ by $\mathcal{P}(\mathcal{K})$. The space $\mathcal{P}(\mathcal{K})$ will be endowed with the weak convergence topology. For $A \in \mathcal{B}(\mathcal{K})$, $\mathcal{I}_A(\cdot)$ will denote the indicator function of the set A. Also, we will denote by $C_b^2(G)$ and $C_0^\infty(G)$ the space of real valued, bounded, and twice continuously differentiable functions on G and the space of real valued, infinitely differentiable, vanishing at infinity, functions on G, respectively. By a filtered probability space $(\Omega, \mathcal{F}, P, (\mathcal{F}_t))$ we will mean a probability space $(\Omega, \mathcal{F}, P)$ endowed by a filtration $(\mathcal{F}_t)_{t\geq 0}$ satisfying the usual hypothesis. A pair of stochastic processes $(u(\cdot), W(\cdot))$ defined on some filtered probability space $(\Omega, \mathcal{F}, P, (\mathcal{F}_t))$ is said to be an admissible pair if $W(\cdot)$ is an $\mathcal{F}_t$ standard Wiener process and $u(\cdot)$ is a U valued, measurable, $\{\mathcal{F}_t\}$ adapted process.

We will consider controlled constrained diffusion processes of the form defined in (1.1), where for $(x,u) \in G \times U$, $b(x,u) \doteq \int_S \bar{b}(x,\alpha)u(d\alpha)$ and the coefficients $\sigma : G \to \mathbb{R}^{k\times k}$ and $\bar{b} : G \times S \to \mathbb{R}^k$ are maps satisfying the following conditions.

CONDITION 2.4. *There exists $r \in (0,\infty)$ such that*

(i) *$\bar{b}$ is a continuous map and for all $x, y \in G$ and $\alpha \in S$*

$$||\bar{b}(x,\alpha) - \bar{b}(y,\alpha)|| + ||\sigma(x) - \sigma(y)|| \leq r||x-y||;$$

(ii) *for all $x \in G$ and $\alpha \in S$*

$$||\bar{b}(x,\alpha)|| + ||\sigma(x)|| \leq r.$$

We will also assume the following nondegeneracy assumption on σ.

CONDITION 2.5. *There exists $c_0 \in (0,\infty)$ such that for all $x \in G$ and $\alpha \in \mathbb{R}^k$ $\alpha'(\sigma(x)\sigma'(x))\alpha \geq c_0\alpha'\alpha$.*

In the rest of the paper, in addition to Condition 2.2, Conditions 2.4 and 2.5 will also be assumed to hold. The following result on the unique strong solution for (1.1) follows on using the Lipschitz property of the SM and the usual fixed point arguments.

THEOREM 2.6. *Let $(u(\cdot), W(\cdot))$ be an admissible pair on some filtered probability space $(\Omega, \mathcal{F}, P, (\mathcal{F}_t))$. Then there exists an $\{\mathcal{F}_t\}$ adapted process $X(\cdot)$ with continuous sample paths satisfying* (1.1) *for all t a.s. Furthermore, if $X_1(\cdot)$ and $X_2(\cdot)$ are two such processes, then*

$$P(X_1(t) = X_2(t);\ \ \forall t \in (0,\infty)) = 1.$$

Remark 2.7. If $X(\cdot)$ solves (1.1), then (cf. Theorem 3.5.1 in [31]) there exist continuous, increasing $\mathcal{F}_t$ adapted processes $\{Y_i(\cdot); 1 \leq i \leq N\}$ such that

$$X(t) = X(0) + \int_0^t b(X(s), u(s))ds + \int_0^t \sigma(X(s))dW(s) + \sum_{i=1}^N d_i Y_i(t) \tag{2.2}$$

for all t, a.s. Furthermore, $Y_i(0) = 0$ and for all $t > 0$ $\int_0^t \mathcal{I}_{F_i}(X(s))dY_i(s) = Y_i(t)$, a.s., $i = 1, \ldots, N$.

From the point of view of applications it is important to consider Markov controls, namely the case where $u(\cdot) = v(X(\cdot))$ for some measurable map $v : G \to U$. However,

in such a case (1.1) may not admit a strong solution. Thus we need to work with a weak solution of (1.1).

DEFINITION 2.8. *Let $v : G \to U$ be a measurable map. We say that the equation*

$$X(t) = \Gamma\left(X(0) + \int_0^{\cdot} b(X(s), v(X(s)))ds + \int_0^{\cdot} \sigma(X(s))dW(s)\right)(t), \quad X(0) \sim \mu, \tag{2.3}$$

admits a weak solution if there exists a filtered probability space $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\})$ on which is given a $\{\mathcal{F}_t\}$ Wiener process $W(\cdot)$ and an $\mathcal{F}_t$ adapted process $X(\cdot)$ with continuous paths such that $X(0)$ has the probability law μ and for all t the equality in (2.3) *holds a.s. We say that* (2.3) *admits a unique weak solution if whenever there are two sets of such spaces and processes denoted as $(\Omega^i, \mathcal{F}^i, P^i, \mathcal{F}^i_t)$, $(W^i(\cdot), X^i(\cdot))$, $i = 1, 2$, then the probability law of $X^1(\cdot)$ is same as that of $X^2(\cdot)$.*

With an abuse of terminology we will also call the map v above a "Markov control." Under the standing assumptions of this paper we have the following result. For the proof of this result we refer the reader to Theorem 4.2.2 of [31]. Although the proof there is for the case where G is a compact set, exactly the same arguments hold for the case of unbounded state space. The key idea in the proof, as in the case of unconstrained diffusions, is to use Condition 2.5 and Girsanov's theorem to get rid of the drift and then use the strong Feller property of the new process and estimates on the Radon–Nikodym derivative to conclude the strong Feller property of the original process.

THEOREM 2.9. *There is a unique weak solution for* (2.3). *Denoting the law of the solution process $X(\cdot)$, when $X(0) = x$ a.s., by P^v_x we have that $\{P^v_x\}_{x\in G}$ is a strongly Feller–Markov family. Furthermore, the transition probability law $P(t, x, dy)$ of this Markov process is mutually absolutely continuous with respect to the Lebesgue measure on G in the following uniform sense. Given $\delta > 0$ and $0 < t_0 < t_1 < \infty$ there exists an $\epsilon > 0$ such that for all $A \in \mathcal{B}(G)$ with $\lambda(A) \le \epsilon$, where λ denotes the Lebesgue measure on G, $P(t, x, A) \le \delta$, $x \in G$, $t \in [t_0, t_1]$. Finally for any $\epsilon > 0$, $0 < t_0 < t_1 < \infty$, and compact set $K_0 \subset G$ there is a $\delta > 0$ such that for all $x \in K_0$, $t \in [t_0, t_1]$, and $A \in \mathcal{B}(G)$ with $\lambda(A \cap K_0) \ge \epsilon$ we have that $P(t, x, A) \ge \delta$.*

3. The ergodic cost problem. In this work we are interested in a control problem with an ergodic cost criterion. Namely, we are interested in minimizing, over the class of all admissible controls, the cost defined in (1.2), where $X(\cdot)$ is given as a solution of (1.1) on some filtered probability space with an admissible pair $(u(\cdot), W(\cdot))$, the limit above is taken a.s. on the corresponding probability space, and $k : G \times U \to \mathbb{R}$ is a map defined as follows. For $(x, u) \in G \times U$, $k(x, u) \doteq \int_S \overline{k}(x, \alpha)u(d\alpha)$, where $\overline{k}$ is in $C_b(G \times S)$.

We will call a Markov control v a stable Markov control (SMC) if the corresponding controlled Markov process $\{P^v_x\}_{x\in G}$ is positive recurrent and has a unique invariant measure. We are interested in obtaining conditions under which there is an optimal SMC for the cost criterion (1.2). The following stability assumption on the underlying model will be assumed throughout this paper.

Define

$$\mathcal{C} \doteq \left\{ -\sum_{i=1}^{N} \alpha_i d_i : \alpha_i \ge 0;\ i \in \{1, \dots, N\} \right\}. \tag{3.1}$$

The cone $\mathcal{C}$ was used to characterize stability of a certain class of constrained diffusion processes in [9, 2].

Let $\delta \in (0, \infty)$ be fixed. Define the set

$$\mathcal{C}(\delta) \doteq \{v \in \mathcal{C} : \text{dist}(v, \partial C) \geq \delta\}. \tag{3.2}$$

Our next assumption, which also will be assumed throughout this paper, on the diffusion model stipulates the permissible drifts in the underlying diffusion.

CONDITION 3.1. *There exists a* $\delta \in (0, \infty)$ *such that for all* $(x, u) \in G \times U$, $b(x, u) \in \mathcal{C}(\delta)$.

Under the assumptions made above the results of [2] show that all Markov controls are SMC; more precisely, we have the following.

THEOREM 3.2. *The Markov family* $\{P_x^v\}_{x \in G}$ *of Theorem* 2.9 *is positive recurrent and admits a unique invariant measure, denoted as* η_v.

Remark 3.3. In [2] the proof of positive recurrence assumes that the drift coefficient in the constrained diffusion process satisfies a Lipschitz condition; however, as is pointed out in Remark 4.6 of that paper, the same proof continues to hold with the assumptions on the coefficients made in this paper.

Now we are able to state the main result of this paper.

THEOREM 3.4. *There exists a Markov control* $\overline{v}(\cdot)$ *such that if for some* $\mu \in \mathcal{P}(G)$ $\overline{X}(\cdot)$ *is the corresponding process solving* (2.3) *on some filtered probability space, with the probability law of* $\overline{X}(0)$ *being* μ, *then*

$$\limsup_{T \to \infty} \frac{1}{T} \int_0^T k(\overline{X}(s), \overline{v}(X(s)))ds = \inf \operatorname{ess\,inf} \limsup_{T \to \infty} \frac{1}{T} \int_0^T k(X(s), u(s))ds, \tag{3.3}$$

a.s., where the outside infimum on the right side above is taken over all controlled processes $X(\cdot)$ *with an arbitrary initial distribution and solving* (1.1) *over some filtered probability space with some admissible pair* $(W(\cdot), u(\cdot))$.

Remark 3.5. The referee has pointed out that the boundedness on the cost function can be relaxed to bounded below, and using Choquet's theorem and results in [1] one can show the existence of a regular (nonrelaxed) Markov control which is optimal.

The proof of the above theorem will be given in section 7.

4. Some background results. In this section we collect some background results which will be used in the proof of Theorem 3.4. We begin with an ergodic theorem of Khasminskii [24] which is applicable to the Markov family $\{P_x^v\}$ considered in this paper because of Theorem 2.9. As a consequence of this result the limiting time averages on the left side of (3.3) can be replaced with expectations with respect to the measure $\eta_{\overline{v}}$.

LEMMA 4.1 (see Khasminskii [24, Theorem 3.1]). *For a given* $\mu \in \mathcal{P}(G)$ *and a Markov control* v *let* $\overline{X}(\cdot)$ *be the process on some filtered probability space solving* (2.3) *with the distribution of* $\overline{X}(0)$ *being* μ. *Then for all* η_v *integrable functions* g *on* G, $\frac{1}{T}\int_0^T g(X(s))ds$ *converges a.s. to* $\int_G g(x)\eta_v(dx)$.

The following lemma has been proved in [12]; however, the domain G there is different from our problem. Thus for the sake of completeness we sketch the proof in the appendix. This lemma will be used several times in this paper in controlling the reflection term $Y(\cdot)$ in our constrained diffusion processes.

LEMMA 4.2. *There exists a* $g \in C_b^2(G)$ *such that*

$$\langle \nabla g(x), d_i \rangle \geq 1 \qquad \forall x \in F_i, \quad i \in \{1, \ldots, N\}. \tag{4.1}$$

The following lemma essentially says that in considering admissible controls, we can without loss of generality restrict ourselves to controls that are adapted with respect to the filtration generated by $(X(\cdot), Y(\cdot))$. The proof is similar to Theorem 1.2.2 (p. 18) of [6] and is therefore omitted.

LEMMA 4.3. *Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ be a filtered probability space on which is given an admissible pair $(u(\cdot), W(\cdot))$. Let $X(\cdot)$ be a solution to* (1.1) *with the corresponding boundary processes $\{Y_i(\cdot)\}_{i=1}^N$. Then there exists a enlargement $(\overline{\Omega}, \overline{\mathcal{F}}, \{\overline{\mathcal{F}}_t\}, \overline{P})$ of the above probability space on which is given a $\{\overline{\mathcal{F}}_t\}$ Wiener process $\tilde{W}(\cdot)$ and $\mathcal{P}(S)$ valued measurable stochastic process $\tilde{u}(\cdot)$ such that for a.e. $t \in [0,\infty)$, $\tilde{u}(t)$ is $\mathcal{F}_t^{X,Y}$ measurable, where $\mathcal{F}_t^{X,Y}$ denotes the P completion of $\sigma\{X(s); \{Y_i(s)\}_{i=1}^N; 0 \le s \le t\}$, and $X(\cdot)$ solves*

$$X(t) = X(0) + \int_0^t b(X(s), \tilde{u}(s))ds + \int_0^t \sigma(X(s))d\tilde{W}(s) + \sum_{i=1}^N d_i Y_i(t).$$

The following lemma will be used in some conditioning arguments in the proofs of Lemma 6.4 and Proposition 7.3. The proof is similar to Theorem 1.1.6 (p. 13) of [6]. Thus the proof is omitted. For a Polish space $\mathcal{K}$, denote by $C([0,\infty) : \mathcal{K})$ the space of continuous functions from $[0,\infty)$ to $\mathcal{K}$, endowed with the topology of uniform convergence on compacts.

LEMMA 4.4. *Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ be a filtered probability space. Let $(u(\cdot), W(\cdot))$ be an admissible pair on this probability space. Let $X(\cdot)$ be given as a solution of* (1.1) *and let τ be an a.s. finite $\{\mathcal{F}_t\}$ stopping time. Denote the conditional distribution of $X(\tau + \cdot)$ given $\mathcal{F}_\tau$ by $\pi(\omega)(\cdot)$, i.e., for $A \in \mathcal{B}(C([0,\infty) : G))$ and a.e. ω $P\left(X(\tau + \cdot) \in A \mid \mathcal{F}_\tau\right)(\omega) = \pi(\omega)(A)$. Then, for a.e. ω, $\pi(\omega)$ equals the probability law of $\overline{X}_\omega(\cdot)$, where $\overline{X}_\omega(\cdot)$ solves an equation of the form* (1.1) *with $(u(\cdot), W(\cdot))$ replaced by some other admissible pair, $(\overline{u}_\omega(\cdot), \overline{W}_\omega(\cdot))$, given on some filtered probability space $(\Omega^\omega, \mathcal{F}^\omega, \{\mathcal{F}_t^\omega\}, P^\omega)$ and $\overline{X}_\omega(0) = X(\tau(\omega))$.*

5. Characterization of the invariant measure. One of the key ingredients of the proof of Theorem 3.4 is an extension of Echeverria–Weiss characterization of invariant measures (cf. [17, 42]) to the class of constrained controlled Markov processes considered in this paper. The proof of this characterization uses a clever idea presented in the proof of a similar characterization result for constrained (uncontrolled) Markov processes in Kurtz [27]. It also uses ideas from [28]. We begin with the following definitions. For $f \in C_b^2(G)$ let $Lf : G \times U \to \mathbb{R}$ be defined as:

$$(Lf)(x,u) \doteq \frac{1}{2}\sum_{i,j=1}^k a_{i,j}(x)\frac{\partial^2 f}{\partial x_i \partial x_j}(x) + \sum_{i=1}^k b_i(x,u)\frac{\partial f}{\partial x_i}(x), \quad (x,u) \in G \times U, \tag{5.1}$$

where $a_{ij}(x) \doteq \sigma(x)\sigma^T(x)$. With an abuse of notation we will write for $\alpha \in S$, $(Lf)(x, \delta_{\{\alpha\}})$, merely as $(Lf)(x,\alpha)$, where $\delta_{\{\alpha\}}$ denotes the probability measure concentrated at the point α. Thus with this notation, for $(x,u) \in G \times U$, $(Lf)(x,u) = \int_S (Lf)(x,\alpha)u(d\alpha)$. For $i = 1, 2, \ldots, N$ and $f \in C_b^2(G)$ let $D_i f : G \to \mathbb{R}$ be defined as $(D_i f)(x) \doteq \langle d_i, \nabla f(x)\rangle$, $x \in G$.

DEFINITION 5.1 (constrained controlled martingale problem (CCMP)). *For $\mu \in \mathcal{P}(G)$ a solution to the $(\mu, L, G, (D_i, F_i)_{i=1}^N)$ CCMP is a pair of $\{\mathcal{F}_t\}$ adapted processes $(Z(\cdot), \Phi(\cdot))$ on some filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ such that the following*

hold. (i) $Z(\cdot)$ *is a* G *valued process with, a.s., continuous trajectories.* (ii) $Z(0)$ *has probability law* μ. (iii) $\Phi(\cdot)$ *is a* U *valued, measurable, and* $\{\mathcal{F}_t\}$ *adapted process.* (iv) *There is an* $\{\mathcal{F}_t\}$ *adapted,* N*-dimensional "boundary" process* $Y(\cdot) = (Y_1(\cdot), \ldots, Y_N(\cdot))$ *such that for each* $i \in \{1, 2, \ldots, N\}$, P*-a.s.,* $Y_i(0) = 0$, $Y_i(\cdot)$ *is continuous and nondecreasing, and for all* $t \in (0, \infty)$, $\int_0^t \mathcal{I}_{F_i}(Z(s))dY_i(s) = Y_i(t)$. (v) *For all* $f \in C_0^\infty(G)$, $f(Z(t)) - \int_0^t \int_S (Lf)(Z(s), \alpha)\Phi(s)(d\alpha)ds - \sum_{i=1}^N \int_0^t (D_i f)(Z(s))$ $dY_i(s)$ *is an* $\mathcal{F}_t$ *martingale.*

The proof of the following result is standard and thus is omitted (cf. Theorem 4.5.2 in [41]).

THEOREM 5.2. *Let* $(Z(\cdot), \Phi(\cdot))$ *be a solution of the* $(\mu, L, G, (D_i, F_i)_{i=1}^N)$ *CCMP on some filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$. *Then there exists an enlargement* $(\overline{\Omega}, \overline{\mathcal{F}}, \{\overline{\mathcal{F}}_t\}, \overline{P})$ *of the above space such that* (i) *there is a* $\overline{\mathcal{F}}_t$ *Wiener process* $\overline{W}(\cdot)$ *defined on the enlarged space,* (ii) *the processes* $(Z(\cdot), \Phi(\cdot))$ *are measurable and* $\overline{\mathcal{F}}_t$ *adapted, and* (iii) *for all* $t \geq 0$, *a.s.*

$$Z(t) = \Gamma\left(Z(0) + \int_0^t \int_S b(Z(s), \alpha)\Phi(s)(d\alpha)ds + \int_0^t \sigma(Z(s))d\overline{W}(s)\right). \tag{5.2}$$

Conversely, if there is a pair of processes $(Z(\cdot), \Phi(\cdot))$ *solving* (5.2) *on some filtered probability space* $(\overline{\Omega}, \overline{\mathcal{F}}, \{\overline{\mathcal{F}}_t\}, \overline{P})$ *satisfying* (i), (ii), *and* (iii) *above, then the pair is a solution of the* $(\mu, L, G, (D_i, F_i)_{i=1}^N)$ *CCMP where* μ *is the probability law of* $Z(0)$.

A solution to the CCMP is closely related to the following patchwork controlled martingale problem (PCMP) introduced in the context of uncontrolled constrained processes by Kurtz in [26].

DEFINITION 5.3. *For* $\mu \in \mathcal{P}(G)$ *a solution to the* $(\mu, L, G, (D_i, F_i)_{i=1}^N)$ *PCMP is an* $\{\mathcal{F}_t\}$ *adapted vector stochastic process* $(\xi(\cdot), \Lambda(\cdot), \lambda_0(\cdot), \ldots, \lambda_N(\cdot))$, *with values in* $G \times U \times \mathbb{R}_+^{N+1}$ *on some filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ *such that the following hold:* (i) $\xi(\cdot)$ *has continuous trajectories a.s.* (ii) $\xi(0)$ *has probability law* μ. (iii) $\Lambda(\cdot)$ *is a* U *valued, measurable,* $\{\mathcal{F}_t\}$ *adapted process.* (iv) *For all* $i \in \{0, 1, 2, \ldots, N\}$, P*-a.s.,* $\lambda_i(0) = 0$, $\lambda_i(\cdot)$ *is continuous and nondecreasing, and for all* $t \in [0, \infty)$, $\int_0^t \mathcal{I}_{F_i}(\xi(s))d\lambda_i(s) = \lambda_i(t)$, *where we define* $F_0 \doteq G$. (iv) *For all* $t \geq 0$, $\sum_{i=0}^m \lambda_i(t) = t$, *a.s.* (v) *For all* $f \in C_0^\infty(G)$, $f(\xi(t)) - \int_0^t \int_S (Lf)(\xi(s), \alpha)\Lambda(s)(d\alpha)d\lambda_0(s) - \sum_{i=1}^N \int_0^t (D_i f)(Z(s))d\lambda_i(s)$ *is a* $\mathcal{F}_t$ *martingale.*

The proof of the following result is similar to the proof of Lemma 3.1 of [12] except that instead of condition (S.a) and (S.b) of [12] we use Condition 2.2(c).

LEMMA 5.4. *Suppose that* $(\xi(\cdot), \Lambda(\cdot), \lambda_0(\cdot), \ldots, \lambda_N(\cdot))$ *is a solution of the* $(\mu, L, G,$ $(D_i, F_i)_{i=1}^N)$ *PCMP on some filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$. *Then, a.s.,* $\lambda_0(\cdot)$ *is a strictly increasing process such that* $\lambda_0(t) \to \infty$ *as* $t \to \infty$.

The following proposition establishes the connection between a solution of a CCMP and a solution of the PCMP. For the proof of the proposition we refer the reader to Theorem 3.4 of [12].

PROPOSITION 5.5. *Suppose that* $(\xi(\cdot), \Lambda(\cdot), \lambda_0(\cdot), \ldots, \lambda_N(\cdot))$ *is a solution of the* $(\mu, L, G, (D_i, F_i)_{i=1}^N)$ *PCMP on some filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$. *Then if for* $t \geq 0$ $\tau(t) \doteq \inf\{s \geq 0 : \lambda_0(s) \geq t\}$, $\mathcal{G}_t \doteq \mathcal{F}_{\tau(t)}$, $Z(t) \doteq \xi(\tau(t))$, $\Phi(t) \doteq \Lambda(\tau(t))$, *and for* $i = 1, \ldots, N$, $Y_i(t) \doteq \lambda_i(\tau(t))$, *then* $(Z(\cdot), \Phi(\cdot))$ *is the solution of the* $(\mu, L, G, (D_i, F_i)_{i=1}^N)$ *CCMP on* $(\Omega, \mathcal{F}, \{\mathcal{G}_t\}, P)$ *with the corresponding boundary processes* $\{Y_i(\cdot)\}_{i=1}^N$.

The following lemma is the first step in the characterization of the invariant measure for the family $\{P_x^v\}$ of Theorem 2.9.

For a measurable space $(\Omega, \mathcal{F})$ we denote by $\mathcal{M}_F(\Omega)$ the space of all finite, possibly identically zero, measures on $(\Omega, \mathcal{F})$. For a measurable map $v : G \to U$, $x \in G$, and $B \in \mathcal{B}(S)$ we will sometimes write $v(x)(B)$ as $v(x, B)$.

LEMMA 5.6. *Let $v : G \to U$ be a measurable map. Let η_v be as in Theorem 3.2. Then there exist measures $\mu_i \in \mathcal{M}_F(F_i)$ such that for all $f \in C_0^\infty(G)$*

$$\int_{G\times S} (Lf)(x,\alpha)\mu_0(dx, d\alpha) + \sum_{i=1}^N \int_{F_i} (D_i f)(x)\mu_i(dx) = 0, \tag{5.3}$$

where $\mu_0 \in \mathcal{P}(G\times S)$ is given as $\mu_0(A\times B) \doteq \int_A v(x, B)\eta_v(dx)$, $A \in \mathcal{B}(G)$, $B \in \mathcal{B}(S)$.

Proof. Let $X(\cdot)$ be a solution of (2.3) with $X(0) \sim \eta_v$ on some filtered probability space. Then $X(\cdot)$ is a stationary process. From Remark 2.7 there exist continuous increasing adapted processes $Y_i(\cdot)$, $i = 1, \ldots, N$, such that (2.2) holds with $u(\cdot)$ replaced by $v(X(\cdot))$. Let $g \in C_b^2(G)$ be as in Lemma 4.2. Then via an application of Ito's formula we have that

$$\begin{aligned} g(X(t)) = g(X(0)) &+ \int_0^t (Lg)(X(s), v(X(s)))ds + \sum_{i=1}^N \int_0^t (D_i g)(X(s))dY_i(s) \\ &+ \int_0^t \langle \nabla g(X(s)), \sigma(X(s))dW(s)\rangle. \end{aligned}$$

Taking expectations in the above equality, using the stationarity of $X(\cdot)$, and recalling the properties of the function $g(\cdot)$, we have that for all $t \geq 0$

$$\begin{aligned} \sum_{i=1}^N E(Y_i(t)) &\leq \sum_{i=1}^N E\left(\int_0^t (D_i g)(X(s))dY_i(s)\right) \\ &\leq \int_0^t E|Lg(X(s), v(X(s)))|ds \\ &\leq \overline{C}t, \end{aligned}$$

where

$$\overline{C} \doteq \sup_{x\in G, u\in U} |Lg(x, u)|. \tag{5.4}$$

Thus if we define for $A \in \mathcal{B}(F_i)$, $i = 1, \ldots, N$, $\mu_i(A) \doteq E(\int_0^1 \mathcal{I}_A(X(s))dY_i(s))$, then $\mu_i \in \mathcal{M}_F(F_i)$ since

$$E\left(\int_0^1 \mathcal{I}_{F_i}(X(s))dY_i(s)\right) = E(Y_i(1)) \leq \overline{C}. \tag{5.5}$$

Now let $f \in C_0^\infty(G)$ be arbitrary. Then another application of Ito's formula gives

$$\begin{aligned} f(X(1)) = f(X(0)) &+ \int_0^1 (Lf)(X(s), v(X(s)))ds + \sum_{i=1}^N \int_0^1 (D_i f)(X(s))dY_i(s) \\ &+ \int_0^1 \langle \nabla f(X(s)), \sigma(X(s))dW(s)\rangle. \end{aligned}$$

Taking expectations and using the stationarity of $X(\cdot)$ we have that

$$\int_G (Lf)(x, v(x))\eta_v(dx) + \sum_{i=1}^{N} \int_{F_i} (D_i f)(x)\mu_i(dx) = 0.$$

The proof now follows on recalling the definition of μ_0 and observing that for $(x, u) \in G \times U$, $(Lf)(x, u) = \int_S (Lf)(x, \alpha)u(d\alpha)$. □

The following extension of Echeverria–Weiss–Kurtz criterion (cf. [42, 27]) is an essential step in our proof of Theorem 3.4.

THEOREM 5.7. *Suppose that there exist measures $\mu_0 \in \mathcal{M}(G \times S)$, $\mu_i \in \mathcal{M}_F(F_i)$, $i = 1, \ldots, N$, such that for all $f \in C_0^\infty(G)$* (5.3) *holds. Decompose $\hat{\mu}_0$ as*

$$\hat{\mu}_0(dx, d\alpha) = v(x, d\alpha)\eta(dx), \tag{5.6}$$

where $\eta \in \mathcal{P}(G)$ is given as $\eta(A) = \hat{\mu}_0(A \times S)$, $A \in \mathcal{B}(G)$, and $v(x, d\alpha)$ is the appropriate regular conditional distribution. Then there exists a solution $(Z(\cdot), \Phi(\cdot))$ to the CCMP $(\eta, L, G, (D_i, F_i)_{i=1}^N)$ on some filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ such that (i) *$Z(\cdot)$ is a stationary process with the invariant measure η;* (ii) *$\Phi(s)(\cdot) = v(Z(s), \cdot)$ for all $s \in [0, \infty)$, a.s.;* (iii) *$Z(\cdot)$ is a positive recurrent, strongly Feller–Markov process with transition probability family $\{P_x^v\}_{x \in G}$ and $\eta = \eta_v$ is its unique invariant measure (cf. Theorem* 3.2*).*

Proof. Assume without loss of generality that $S \cap \{1, \ldots, N\} = \emptyset$. Define a new control set $\tilde{S} \doteq S \cup \{1, \ldots, N\}$. Let $\tilde{d}(\cdot, \cdot)$ be a distance on it defined as follows. For $x, y \in \tilde{S}$, $\tilde{d}(x, y) \doteq d(x, y)$ if $x \in S$ and $y \in S$; $\tilde{d}(x, x) = 0$ and we set $\tilde{d}(x, y) \doteq 1$ otherwise, where $d(\cdot, \cdot)$ is the given metric on S. Clearly $(\tilde{S}, \tilde{d}(\cdot, \cdot))$ is a compact metric space. For $n \in \mathbb{N}$ (where $\mathbb{N}$ is the set of all positive integers), define the linear operator $C_n : C_0^\infty(G) \to C_b(G \times \tilde{S})$ as follows. For $f \in C_0^\infty(G)$ and $(x, \tilde{\alpha}) \in G \times \tilde{S}$

$$\begin{aligned}(C_n f)(x, \tilde{\alpha}) &\doteq (Lf)(x, \tilde{\alpha}) \text{ if } \tilde{\alpha} \in S \\ &\doteq n(D_i f)(x) \text{ if } \tilde{\alpha} \in \{1, \ldots, N\}.\end{aligned}$$

Define $\tilde{\nu}_n \in \mathcal{P}(G \times \tilde{S})$ as follows. For $h \in C_b(G \times \tilde{S})$

$$\begin{aligned}&\int_{G \times \tilde{S}} h(x, \tilde{\alpha})\tilde{\nu}_n(dx, d\tilde{\alpha}) \\ &\quad \doteq \frac{1}{K_n}\left(\int_{G \times S} h(x, \alpha)\mu_0(dx, d\alpha) + \frac{1}{n}\sum_{i=1}^{N} \int_{F_i} h(x, i)\mu_i(dx)\right),\end{aligned} \tag{5.7}$$

where $K_n \doteq \mu_0(G \times S) + \frac{1}{n}\sum_{i=1}^{N} \mu_i(F_i)$.

From the assumption that (5.3) holds it follows now that for all $f \in C_0^\infty(G)$

$$\int_{G \times \tilde{S}} (C_n f)(x, \tilde{\alpha})\tilde{\nu}_n(dx, d\tilde{\alpha}) = 0. \tag{5.8}$$

Disintegrate $\tilde{\nu}_n$ as follows. For $A \in \mathcal{B}(G)$ and $B \in \mathcal{B}(\tilde{S})$,

$$\tilde{\nu}_n(A \times B) = \int_A \tilde{v}_n(x, B \cap S)\eta^n(dx) + \sum_{i=1}^{N} \int_A \tilde{v}_{i,n}(x)\delta_{\{i\}}(B)\eta^n(dx),$$

where for $B \in \mathcal{B}(S)$ the maps $\tilde{v}_n(\cdot, B)$ and $\tilde{v}_{i,n}(\cdot)$ are measurable; for all $x \in G$ $\tilde{v}_n(x,\cdot) \in \mathcal{M}_F(S)$, $\tilde{v}_{i,n}(x) \geq 0$, $i = 1, \ldots, N$; $\tilde{v}_n(x, S) + \sum_{i=1}^N \tilde{v}_{i,n}(x) = 1$ and $\eta^n \in \mathcal{P}(G)$ is given as follows. For $A \in \mathcal{B}(G)$

$$\eta^n(A) \doteq \frac{1}{K_n}\left(\mu_0(A \times S) + \frac{1}{n}\sum_{i=1}^N \mu_i(A \cap F_i)\right). \tag{5.9}$$

Also define $\tilde{\nu} \in \mathcal{P}(G \times S)$ as the normalization of μ_0, i.e., $\tilde{\nu} \doteq \hat{\mu}_0$. Recall that from (5.6), $\tilde{\nu}(dx, d\alpha) = v(x, d\alpha)\eta(dx)$. For fixed $x \in G$ define a probability measure $v_n(x, d\tilde{\alpha})$ on $\tilde{S}$ as follows. For $A \in \mathcal{B}(\tilde{S})$

$$\begin{aligned} v_n(x, A) &\doteq v(x, A \cap S) \text{ if } x \in G^0 \\ &\doteq \tilde{v}_n(x, A \cap S) + \sum_{i=1}^N \tilde{v}_{i,n}(x)\delta_{\{i\}}(A) \text{ otherwise.} \end{aligned} \tag{5.10}$$

It is easy to check that for all $h \in C_b(G \times \tilde{S})$

$$\int_{G\times\tilde{S}} h(x,\tilde{\alpha})\tilde{\nu}_n(dx, d\tilde{\alpha}) = \int_{G\times\tilde{S}} h(x,\tilde{\alpha}) v_n(x, d\tilde{\alpha})\eta^n(dx). \tag{5.11}$$

Using (5.8), (5.11), and Theorem 2.4 of [28] we now have that there exists a filtered probability space (for the sake of simplicity we suppress the dependence of the filtered probability space on n in our notation) $(\Omega, \mathcal{F}, P, (\mathcal{F}_t))$ on which is given an adapted G valued stationary process $X_n(\cdot)$ with continuous paths such that the probability law of $X_n(0)$ is η^n and for all $f \in C_0^\infty(G)$,

$$f(X_n(t)) - \int_0^t \left(\int_{\tilde{S}} (C_n f)(X_n(s), \tilde{\alpha}) v_n(X_n(s), d\tilde{\alpha})\right) ds \tag{5.12}$$

is an $\mathcal{F}_t$ martingale.

For $x \in G$ let $v_n^0(x, d\alpha) \in \mathcal{M}_F(S)$ be defined as follows. For $A \in \mathcal{B}(S)$

$$\begin{aligned} v_n^0(x, A) &\doteq v(x, A) \text{ if } x \in G^0 \\ &\doteq \tilde{v}_n(x, A) \text{ if } x \in \partial G. \end{aligned} \tag{5.13}$$

Also, define for $i \in \{1, \ldots, N\}$ $v_{i,n}^0 : G \to [0, 1]$ as

$$v_{i,n}^0(x) \doteq \tilde{v}_{i,n}(x)\mathcal{I}_{\partial G}(x). \tag{5.14}$$

Note that for all $x \in G$

$$\begin{aligned} v_n(x, A) &= v_n^0(x, A) \text{ if } A \in \mathcal{B}(S) \\ &= \sum_{i=1}^N v_{i,n}^0(x)\mathcal{I}_{\{i\}}(A) \text{ if } A \subset \{1, \ldots, N\}. \end{aligned} \tag{5.15}$$

Rewriting (5.12) using (5.15) we have that for all $f \in C_0^\infty(G)$

$$\begin{aligned} f(X_n(t)) - \int_0^t \left(\int_S (Lf)(X_n(s), \alpha) v_n^0(X_n(s), d\alpha)\right) ds \\ - \sum_{i=1}^N n \int_0^t (D_i f)(X_n(s)) v_{i,n}^0(X_n(s)) ds \end{aligned}$$

is an $\mathcal{F}_t$ martingale. Now define for $t \in [0,\infty)$, $i \in \{1,\dots,N\}$,

$$\lambda_0^n(t) \doteq \int_0^t v_n^0(X_n(s),S)ds,$$
$$\lambda_i^n(t) \doteq \int_0^t v_{i,n}^0(X_n(s))ds.$$

Also, for $x \in G$ define $\Lambda_n(x,\cdot) \in \mathcal{P}(S)$ as follows. For $A \in \mathcal{B}(S)$

$$\Lambda_n(x,A) \doteq \frac{v_n^0(x,A)}{v_n^0(x,S)} \text{ if } v_n^0(x,S) \neq 0 \tag{5.16}$$
$$\doteq \pi(A) \text{ otherwise,}$$

where π is an arbitrary probability measure on S. Then in this new notation we have that for all $f \in C_0^\infty(G)$

$$f(X_n(t)) - \int_0^t \left(\int_S (Lf)(X_n(s),\alpha)\Lambda_n(X_n(s),d\alpha) \right) d\lambda_0^n(s)$$
$$- \sum_{i=1}^N n \int_0^t (D_i f)(X_n(s))d\lambda_i^n(s)$$

is an $\mathcal{F}_t$ martingale. Clearly, for all $t \geq 0$, $\sum_{i=0}^N \lambda_i^n(t) = t$. Furthermore, for $i = 1,\dots,N$,

$$\lambda_i^n(t) = \int_0^t \mathcal{I}_{F_i}(X_n(s))d\lambda_i^n(s) \qquad \forall t \geq 0. \tag{5.17}$$

To prove (5.17) note that, it suffices to show that for all $t \geq 0$, $E(v_{i,n}^0(X_n(t))) = E(\mathcal{I}_{F_i}(X_n(t))v_{i,n}^0(X_n(t)))$. Also,

$$\begin{aligned} E(v_{i,n}^0(X_n(t))) &= \int_G v_{i,n}^0(x)\eta^n(dx) \\ &= \tilde{\nu}_n(G \times \{i\}) \\ &= \tilde{\nu}_n(F_i \times \{i\}) \\ &= E(\mathcal{I}_{F_i}(X_n(t))v_{i,n}^0(X_n(t))), \end{aligned}$$

where the next to last equality follows from (5.7). This proves (5.17). Thus it follows that $(X_n(\cdot),\Lambda_n(X_n(\cdot)),\lambda_0^n(\cdot),\dots,\lambda_N^n(\cdot))$ solves the $(\eta^n,L,G,(nD_i,F_i)_{i=1}^N)$ PCMP on $(\Omega,\mathcal{F},P,(\mathcal{F}_t))$. From Lemma 5.4 it follows that λ_0^n is a.s. strictly increasing. Now define $\tau_n : [0,\infty) \to [0,\infty)$ as $\tau_n(t) \doteq \inf\{s \geq 0 : \lambda_0^n(s) \geq t\}$, $t \in [0,\infty)$. Also for $t \geq 0$, let $\mathcal{G}_t^n \doteq \mathcal{F}_{\tau_n(t)}$, $Z_n(t) \doteq X_n(\tau_n(t))$, $\Phi_n(t) \doteq \Lambda_n(Z_n(t))$, and for $i = 1,\dots,N$ $Y_{i,n}(t) \doteq \lambda_{i,n}(\tau_n(t))$. Then from Proposition 5.5, $(Z_n(\cdot),\Phi_n(\cdot))$ solve the CCMP on $(\Omega,\mathcal{F},P,(\mathcal{G}_t^n))$ with the corresponding boundary processes $\{Y_{i,n}(\cdot)\}_{i=1}^N$. Next note that since $\lambda_0^n(0) = 0$ and for $0 \leq s \leq t < \infty$

$$\begin{aligned} |\lambda_0^n(t) - \lambda_0^n(s)| &= \left| \int_s^t v_n^0(X_n(r),S)dr \right| \\ &\leq |t-s|, \quad a.s., \end{aligned}$$

we have that the family $\{\lambda_0^n(\cdot)\}$ is tight in $C([0,\infty);[0,\infty))$. Next observe that from Theorem 5.2 there exists an enlargement $(\overline{\Omega},\overline{\mathcal{F}},\{\overline{\mathcal{F}}_t\},\overline{P})$ of the above space such that there is a $\overline{\mathcal{F}}_t$ Wiener process $\overline{W}(\cdot)$ defined on the enlarged space and

$$Z_n(t) = \Gamma\left(Z_n(0) + \int_0^\cdot \int_S b(Z_n(s),\alpha)\Phi_n(s)(d\alpha)ds + \int_0^\cdot \sigma(Z_n(s))d\overline{W}(s)\right)(t), \tag{5.18}$$

where the dependence of the Wiener process and the space on n is again suppressed in the notation. Since the probability law of $Z_n(0)$ is same as that of $X_n(0)$, i.e., η^n and from (5.9) $\eta^n(A) \to \eta(A)$ as $n \to \infty$ for all $A \in \mathcal{B}(G)$, we have that the family $\{Z_n(0)\}$ is tight. Furthermore, using the Lipschitz property of the Skorohod map we have that for $0 \le s \le t < \infty$

$$|Z_n(t) - Z_n(s)| \le Kr|t-s| + K\left|\int_s^t \sigma(Z_n(s))dW(s)\right|.$$

Recalling that $\sigma(\cdot)$ is bounded we have as a result of the above observations that the family $\{Z_n(\cdot)\}$ is tight in $C([0,\infty):G)$. Let $(Z(\cdot),\lambda_0(\cdot))$ be a weak limit point of the sequence $(Z_n(\cdot),\lambda_0^n(\cdot))$ and relabel the convergent subsequence as $(Z_n(\cdot),\lambda_0^n(\cdot))$. Observing that $\lambda_0^n(t) \le t$, a.s., for all $t \ge 0$ and $n \in \mathbb{N}$ and $E(\lambda_0^n(t)) = \frac{t}{K_n}\mu_0(G\times S) \to t$, as $n \to \infty$, we have that $\lambda_0(t) = t$ for all $t \ge 0$, a.s. Next observe that from the weak convergence of $Z_n(\cdot)$ and $\lambda_0^n(\cdot)$ we have that as $n \to \infty$, $X_n(\cdot) = Z_n(\lambda_0^n(\cdot))$ converges weakly to $Z(\lambda_0(\cdot)) \equiv Z(\cdot)$. Since for each $n \in \mathbb{N}$, $X_n(\cdot)$ is stationary, we must have that the limit $Z(\cdot)$ is a stationary process too. Also, since the law of $Z(0)$ is η we have that the stationary distribution is η. Next note that, from (5.13) and (5.16), for all $x \in G^0$, $\Lambda_n(x,d\alpha) = v(x,d\alpha)$. Also from Theorem 4.2.1 in [31] we have that for all $n \in \mathbb{N}$, $E\left(\int_0^\infty \mathcal{I}_{\partial G}(Z_n(s))ds\right) = 0$. From these two observations, (5.18), and the Lipschitz property of the Skorohod map, it follows that

$$Z_n(t) = \Gamma\left(Z_n(0) + \int_0^\cdot \int_S b(Z_n(s),\alpha)v(Z_n(s),d\alpha)ds + \int_0^\cdot \sigma(Z_n(s))d\overline{W}(s)\right)(t)$$

for all $t \ge 0$, a.s. The Feller property of the family $\{P_x^v\}$ (see Theorem 2.9) now gives that $Z_n(\cdot)$ converges weakly to the solution of

$$\tilde{Z}(t) = \Gamma\left(\tilde{Z}(0) + \int_0^\cdot \int_S b(\tilde{Z}(s),\alpha)v(\tilde{Z}(s),d\alpha)ds + \int_0^\cdot \sigma(\tilde{Z}(s))d\overline{W}(s)\right)(t).$$

Since $Z_n(\cdot)$ also converges weakly to $Z(\cdot)$ we must have that $Z(\cdot)$ and $\tilde{Z}(\cdot)$ have the same distribution, in particular $Z(\cdot)$ is a stationary Markov process with the stationary distribution η.

This proves (i) and (ii) of the theorem. Finally, part (iii) follows from Theorem 3.2. □

6. Stability properties of the constrained controlled diffusions. We will now like to obtain some stability properties of the class of processes obtained as a solution of an equation of the form (1.1). We begin with the following lemma, the proof of which is contained in the proof of Theorem 4.4 of [2]. For $x \in G$ let $X^x(\cdot)$ be the solution of (1.1) with $X^x(0) = x$ on some filtered probability space $(\Omega,\mathcal{F},\{\mathcal{F}_t\},P)$ on which is given an admissible pair $(u(\cdot),W(\cdot))$.

LEMMA 6.1 (cf. Theorem 4.4 of [2]). *There exists $\alpha, r_1, C_1, \Delta, \theta \in (0,\infty)$ such that for all $r_2 \in (0,\infty)$ and $t \in (0,\infty)$,*

$$\sup_{x\in G:|x|=r_2} \sup P(\tau_1(x) > t) \le \frac{e^{\alpha C_1 r_2}}{e^{(\alpha-\theta)\Delta}} e^{-\theta t},$$

where $\tau_1 \doteq \inf\{t > 0 : |X^x(t)| = r_1\}$ *and the inner supremum on the left side above is taken over all possible solutions* $X(\cdot)$ *of* (1.1) *with* $X(0) = x$ *given on some filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ *with some admissible pair* $(u(\cdot), W(\cdot))$.

Remark 6.2. Theorem 4.4 of [2] is stated for uncontrolled constrained diffusion processes; however, the result (and most of the proof) holds in the generality considered here. The only place where the proof in [2] needs to be modified is as follows. The proof of Theorem 4.4 of [2] relies on Lemma 4.3 of the same paper. However, the proof of Lemma 4.3, presented in [2], at one place uses the Markov property of $X^x(\cdot)$. Thus in the appendix of this work we provide an alternate proof of this lemma which does not appeal to the Markov property and holds for the class of processes $X^x(\cdot)$ considered here.

LEMMA 6.3. *Let* r_1 *be as in Lemma* 6.1 *and let* $r_2 \in (r_1, \infty)$ *be arbitrary. For* $x \in G$ *let* $X^{x,(u,W)}(\cdot)$ *denote the solution of* (1.1), *with* $X^{x,(u,W)}(0) = x$, *given on some filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ *and with an admissible pair* $(u(\cdot), W(\cdot))$. *Let*

$$\tau(x) \doteq \inf\{t \geq 0 : |X^{x,(u,W)}(t)| = r_1 \ \text{ and } \ |X^{x,(u,W)}(s)| = r_2 \ \text{ for some } \ \mathrm{s} \in [0, \mathrm{t}]\},$$

where we have suppressed the dependence of $\tau(x)$ *on* $(u(\cdot), W(\cdot))$ *in the notation.*

Then there exists a $\delta_0 \in (0, \infty)$ *such that*

$$\inf_{x \in G : |x| = r_1} \inf_{u(\cdot), W(\cdot)} E(\tau(x)) > \delta_0 \tag{6.1}$$

and there exist $\kappa_0, \theta_0 \in (0, \infty)$ *such that for all* $t \in [0, \infty)$

$$\sup_{x \in G : |x| = r_1} \sup_{u(\cdot), W(\cdot)} P(\tau(x) > t) < \kappa_0 e^{-\theta_0 t}. \tag{6.2}$$

Proof. For notational simplicity we denote $X^{x,(u,W)}(\cdot)$ by $X^x(\cdot)$. We first prove (6.2). Given $X^x(\cdot)$ as in the statement of the lemma, define $\tau_0(x) \doteq \inf\{t \geq 0 : |X^x(t)| = r_2\}$. In view of Lemma 6.1 and Lemma 4.4 it suffices to show that there exist $\kappa', \theta_0' \in (0, \infty)$ such that for all $t \geq 0$, $\sup_{x \in G : |x| = r_1} \sup_{u(\cdot), W(\cdot)} P(\tau_0(x) > t) < \kappa' e^{-\theta_0' t}$. This will follow, if we show that for all $k \in \mathbb{N}$

$$\sup_{u(\cdot), W(\cdot)} \sup_{x \in G : |x| = r_1} P(\tau_0(x) > k) < e^{-\theta_0' k}. \tag{6.3}$$

Noting that $P(\tau_0(x) > k) = E\left(E\left(\mathcal{I}_{[\tau_0(x) > k]} \mid \mathcal{F}_{k-1}\right) \mathcal{I}_{[\tau_0(x) > k-1]}\right)$, we have from Lemma 4.4 that in order to show (6.3) it suffices to show that there exists an $\epsilon_0 \in (0, 1)$ such that

$$\sup_{x \in G : |x| \leq r_2} \sup_{u(\cdot), W(\cdot)} P(\tau_0(x) > 1) < \epsilon_0. \tag{6.4}$$

We will prove this by the method of contradiction. Suppose that (6.4) does not hold for any $\epsilon_0 \in (0, 1)$. Then there exist $\{x^n, u^n(\cdot), W^n(\cdot), X^n(\cdot), Y^n(\cdot)\}_{n \geq 1}$ such that for each $n \in \mathbb{N}$, $x^n \in \{x \in G : |x| \leq r_2\}$, $(u^n(\cdot), W^n(\cdot))$ is an admissible pair on some filtered probability space, $X^n(\cdot)$ and $Y^n(\cdot) \doteq (Y_1^n(\cdot), \ldots, Y_N^n(\cdot))$ are obtained as a solution of (1.1) with $(u(\cdot), W(\cdot))$ replaced by $(u^n(\cdot), W^n(\cdot))$, $X^n(0) = x^n$, and $\lim_{n \to \infty} P(\tau_{0,n} > 1) = 1$, where $\tau_{0,n} \doteq \inf\{t \geq 0 : |X^n(t)| = r_2\}$.

Let $\{f_i\}_{i=1}^{\infty}$ be a countable dense set in the unit ball of $C(S)$. Define, for $t \geq 0$ and $j \geq 1$, $\beta_j^n(t) \doteq \int_S f_j(\alpha) u^n(t, d\alpha)$. Let B denote the closed unit ball of

$L_\infty[0,\infty)$ endowed with the metric $\overline{d}(\cdot,\cdot)$ defined as follows. For $x, y \in B$, $\overline{d}(x,y) \doteq \sum_{M=1}^{\infty}\sum_{j=1}^{\infty}\frac{\langle x-y, e_j^M\rangle_M}{2^M 2^j}$, where for each $M \in \mathbb{N}$, $\{e_j^M(\cdot)\}_{j=1}^{\infty}$ is a complete orthonormal system in $L^2[0,M]$ and $\langle\cdot,\cdot\rangle_M$ denotes the usual inner product in $L^2[0,M]$. Clearly $(B,\overline{d}(\cdot,\cdot))$ is a compact metric space. Let E denote the countable product of B endowed with the product topology. Then E is a compact Polish space and $\beta^n(\cdot) \doteq (\beta_1^n(\cdot),\ldots)$ is an E valued random variable.

Recalling that $X^n(0) = x^n$ and $|x^n| \le r_2$ we see that the family $X^n(0)$ is tight. Furthermore, using the Lipschitz property of the Skorohod map and Condition 2.4(ii) we see that there exists $\tilde{C} < \infty$ such that for all $0 \le s \le t < \infty$

$$|X^n(t) - X^n(s)| \le \tilde{C}\left[|t-s| + \left|\int_s^t \sigma(X^n(q))dW^n(q)\right|\right].$$

Using the boundedness of $\sigma(\cdot)$ we now have that $\{X^n(\cdot)\}$ is tight in $C([0,\infty):G)$. Next choosing $g \in C_b^2(G)$ as in Lemma 4.2 we see that for $0 \le s \le t < \infty$

$$\sum_{i=1}^{N}|Y_i^n(t) - Y_i^n(s)| \le \overline{C}|t-s| + \left|\int_s^t \langle \nabla g(X^n(q)), \sigma(X^n(q))dW^n(q)\rangle\right|,$$

where $\overline{C}$ is as defined in (5.4). Combining this with the fact that $Y^n(0) = 0$ we have that $\{Y^n(\cdot)\}$ is tight in $C([0,\infty):[0,\infty)^N)$. Thus $(\beta^n(\cdot), X^n(\cdot), Y^n(\cdot))$ is a tight family of random variables with values in $\overline{E} \doteq E \times C([0,\infty):G) \times C([0,\infty):[0,\infty)^N)$. Pick a weakly convergent subsequence of the above sequence and relabel it as the original sequence. By going to the Skorohod representation space $(\overline{\Omega}, \overline{\mathcal{F}}, P)$—however, keeping the same notation for random variables for convenience—we have that there exists an $\overline{E}$ valued random element $(\beta(\cdot), X(\cdot), Y(\cdot))$ such that $(\beta^n(\cdot), X^n(\cdot), Y^n(\cdot))$ converges a.s. to $(\beta(\cdot), X(\cdot), Y(\cdot))$ as $n \to \infty$.

Next note that for $f \in C_0^\infty(G)$ and $0 \le t_1 \le t_2 < \infty$

(6.5)

$$E\left(\left(f(X^n(t_2)) - f(X^n(t_1)) - \int_{t_1}^{t_2}\left(\int_S (Lf)(X^n(s),\alpha)u^n(s,d\alpha)\right)ds \right.\right.$$
$$\left.\left. - \sum_{i=1}^{N}\int_{t_1}^{t_2}(D_i f)(X^n(s))dY_i^n(s)\right)\psi(X^n(s_1), Y^n(s_1),\ldots,X^n(s_m),Y^n(s_m))\right) = 0,$$

where $m \in \mathbb{N}$, $0 \le s_1 \le \cdots \le s_m \le t_1$, and ψ is an arbitrary continuous and bounded function defined on the obvious domain. From Lemma 2.4 of [12] we have that $\int_{t_1}^{t_2}(D_i f)(X^n(s))dY_i^n(s) \to \int_{t_1}^{t_2}(D_i f)(X(s))dY_i(s)$ a.s. as $n \to \infty$. Also from the Lipschitz property of the coefficients $b(\cdot)$ and σ (cf. Condition 2.4(i), (iii)) and Lemma II.1.3 of [6] we have that

$$\int_{t_1}^{t_2}\left(\int_S (Lf)(X^n(s),\alpha)u^n(s,d\alpha)\right)ds \to \int_{t_1}^{t_2}\left(\int_S (Lf)(X(s),\alpha)u(s,d\alpha)\right)ds$$

a.s., as $n \to \infty$, where $u(\cdot)$ is a U valued measurable process satisfying $\int_S f_i(\alpha)u(t,d\alpha) = \alpha_i(t)$ for all $i \in \mathbb{N}$.

Thus taking limit as $n \to \infty$ in (6.5) we have that

$$E\left(\left(f(X(t_2)) - f(X(t_1)) - \int_{t_1}^{t_2}\left(\int_S (Lf)(X(s),\alpha)u(s,d\alpha)\right)ds \right.\right.$$
$$\left.\left. - \sum_{i=1}^{N}\int_{t_1}^{t_2}(D_i f)(X(s))dY_i(s)\right)\psi(X(s_1),Y(s_1),\ldots,X(s_m),Y(s_m))\right) = 0.$$

Furthermore, without loss of generality, we can take $u(\cdot)$ to be $\mathcal{F}_t^{X,Y}$ adapted. Also noting that for any $f \in C_b(G)$ such that $f = 0$ on F_i we have that $\int_0^\infty f(X^n(s))dY_i^n(s) = 0$, it follows that for such an f, $\int_0^\infty f(X(s))dY_i(s) = 0$. Thus for all $i \in \{1,\ldots,N\}$ and $t \geq 0$, $Y_i(t) = \int_0^t \mathcal{I}_{F_i}(X(s))dY_i(s)$, a.s. Thus $(X(\cdot),u(\cdot))$ solves the CCMP for $(\delta_{\{x\}}, L, G, (D_i,F_i)_{i=1}^N)$ on the filtered probability space $(\overline{\Omega},\overline{\mathcal{F}},\mathcal{F}_t^{X,Y})$. Finally, defining $\tau \doteq \inf\{t : |X(t)| = r_2\}$ we have that $\tau_{0,n} \to \tau$ a.s. and thus $P(\tau \geq 1) = 1$. But this is clearly impossible in view of Condition 2.5. Thus we have arrived at a contradiction. This proves (6.2). The proof of (6.1) follows via a similar argument via contradiction. This proves the lemma. □

As an immediate consequence of above lemmas we have the following result.

LEMMA 6.4. *For $\pi \in \mathcal{P}(G)$ with support contained in S_0 and admissible pair $(u(\cdot),W(\cdot))$ given on some filtered probability space $(\Omega,\mathcal{F},\{\mathcal{F}_t\},P)$, let $X^{\pi,(u,W)}(\cdot)$ denote the solution of* (1.1), *with $X^{\pi,(u,W)}(0)$ having the probability law π. Let*

$$\tau(\pi) \doteq \inf\{t \geq 0 : |X^{\pi,(u,W)}(t)| = r_1 \ \text{ and } \ |X^{\pi,(u,W)}(s)| = r_2 \ \text{ for some } \ s \in [0,t]\}.$$

Define $\eta^{\pi,(u,W)} \in \mathcal{P}(G)$ as follows. For $f \in C_b(G)$

$$\int_G f(y)\eta^{\pi,(u,W)}(dy) \doteq \frac{E\left(\int_0^{\tau(\pi)} f(X^{\pi,(u,W)}(t))dt\right)}{E(\tau(\pi))}.$$

Then the family $\{\eta^{\pi,(u,W)} : \pi \in \mathcal{P}(G);\ \mathrm{supp}(\pi) \subset S_0;\ (u(\cdot),W(\cdot))$ admissible$\}$ is tight.

Proof. Let $\epsilon > 0$ be arbitrary. Let $\theta_0,\kappa_0,\delta_0$ be as in Lemma 6.3. It follows from Lemma 4.5 of [2] that for all $M \in (0,\infty)$ the family $\{X^{x,(u,W)}(t), t \geq 0; |x| \leq M; (u(\cdot),W(\cdot))$ admissible$\}$ is tight. Using this observation and Lemma 4.4 we have that there exists a compact set K^ϵ in G such that

$$P(X^{\pi,(u,W)}(t) \notin K^\epsilon) \leq \frac{(\epsilon\theta_0\delta_0)^2}{4\kappa_0}$$

for all $t \in (0,\infty)$, $\pi \in \mathcal{P}(G)$, $\mathrm{supp}(\pi) \subset S_0$, and $(u(\cdot),W(\cdot))$ admissible. Hence

$$\begin{aligned}
\eta^{\pi,(u,W)}((K^\epsilon)^c) &= \frac{\int_0^\infty E\left(\mathcal{I}_{\{\tau(\pi)>t\}}\mathcal{I}_{\{X^{\pi,(u,W)}(t)\notin K^\epsilon\}}\right)dt}{E(\tau(\pi))} \\
&\leq \frac{\epsilon\theta_0\delta_0}{2\sqrt{\kappa_0}}\frac{\int_0^\infty \sqrt{P(\tau(\pi)>t)}dt}{E(\tau(\pi))} \\
&\leq \frac{\epsilon\theta_0}{2}\int_0^\infty e^{-\frac{\theta_0 t}{2}}dt \\
&\leq \epsilon,
\end{aligned}$$

where the next to last inequality follows from Lemmas 4.4 and 6.3. This proves the lemma. □

7. Proof of Theorem 3.4. Let $X^x(\cdot)$ solve (1.1) with $X^x(0) = x$ on some filtered probability space with an admissible pair $(u(\cdot), W(\cdot))$. Define stopping times τ_i, ξ_i, $i \in \mathbb{N}$, as follows:

$$\tau_1 \doteq \inf\{t \geq 0 : |X^x(t)| = r_1\}. \tag{7.1}$$

For $n \geq 1$

$$\xi_n \doteq \inf\{t \geq \tau_n : |X^x(t)| = r_2\} \tag{7.2}$$

and

$$\tau_{n+1} \doteq \inf\{t \geq \xi_n : |X^x(t)| = r_1\}. \tag{7.3}$$

From Lemma 6.1 and Lemma 6.3 it follows that for all $i \in \mathbb{N}$, $E(\xi_i) < \infty$ and $E(\tau_i) < \infty$.

Now suppose that there is a $v : G \to U$ such that for all $t \geq 0$, $v(X^x(t)) = u(t)$, a.s. From the strong Markov property of the solution of (2.3) it follows that $\{X^x(\tau_i)\}_{i\geq 1}$ is a $S_0 \doteq \{x \in G : |x| = r_1\}$ valued Markov chain. Furthermore, from Condition 2.5 it follows that for $i \in \mathbb{N}$ the probability law of $X(\tau_i)$ has a density, with respect to the surface measure on S_0, which is bounded away from 0. Using the above property of the Markov chain $\{X(\tau_i)\}_{i\geq 1}$, it follows along the lines of Lemma IV.4.1 of [25] that there exists a unique invariant measure, ρ, for this chain. Extend the measure ρ to $\tilde{\rho} \in \mathcal{P}(G)$ by setting $\tilde{\rho}(A) \doteq \tilde{\rho}(A \cap S_0)$ for $A \in \mathcal{B}(G)$. Let $X(\cdot)$ be given as a solution of (2.3) with $X(0)$ having the probability law $\tilde{\rho}$. Define $\eta \in \mathcal{P}(G)$ as follows: For $f \in C_b(G)$ $\int_G f(x)\eta(dx) \doteq \frac{E\left(\int_0^{\tau_2} f(X(t))dt\right)}{E(\tau_2)}$. Then it follows as in Theorem IV.4.1 of [25] that η is the unique invariant measure for the Markov process $X(\cdot)$, i.e., in the notation of Theorem 3.2, $\eta = \eta_v$.

The following compactness result is a crucial step in the proof.

LEMMA 7.1. *The family* $\{\eta_v | v : G \to U;\ \ v \ \text{is measurable}\}$ *is a compact set in* $\mathcal{P}(G)$.

Proof. We begin by observing that as an immediate consequence of the above representation of η_v and Lemma 6.4, we have that the above family is tight. Now let $\{v_n\}$ be a sequence of measurable maps from G to U. Suppose that η_{v_n} converges to $\eta \in \mathcal{P}(G)$. We will like to show that there exists a measurable $v : G \to U$ such that $\eta = \eta_v$. Define the sequence $\{\nu_n\}$ of elements of $\mathcal{P}(G \times S)$ as follows. For $A \in \mathcal{B}(G)$, $B \in \mathcal{B}(S)$ $\nu_n(A \times B) \doteq \int_A v_n(x, B)\eta_{v_n}(dx)$. From Lemma 5.6 there exist measures $\mu_i^n \in \mathcal{M}_F(F_i)$, $i = 1, \ldots, N$, $n \in \mathbb{N}$, such that for all $f \in C_0^\infty(G)$

$$\int_{G\times S} (Lf)(x, \alpha)\nu_n(dx, d\alpha) + \sum_{i=1}^{N} \int_{F_i} (D_i f)(x)\mu_i^n(dx) = 0. \tag{7.4}$$

From the compactness of S and Lemma 7.1 we have that $\{\nu_n\}_{n\geq 1}$ is a tight family. Denote by $\overline{F}_i$ the one point compactification of F_i. Extend, for each $i = 1, \ldots, N$, μ_i^n to an element of $\mathcal{M}_F(\overline{F}_i)$ in a natural way, denoting the extension as $\overline{\mu}_i^n$. Also note that from (5.5) we have that the measures μ_i^n can be chosen such that $\mu_i^n(F_i) = \overline{\mu}_i^n(\overline{F}_i) \leq \overline{C}$, where $\overline{C}$ is the constant defined in (5.4). Thus, by going to a subsequence if necessary, we have that there exist $\nu \in \mathcal{P}(G \times S)$ and $\overline{\mu}_i \in \mathcal{M}_F(\overline{F}_i)$ such that for all $h \in C_b(G \times S)$ and $h_i \in C_b(\overline{F}_i)$, $i = 1, \ldots, N$, $\int_{G\times S} h(x, \alpha)\nu_n(dxd\alpha)$ converges to

$\int_{G\times S} h(x,\alpha)\nu(dxd\alpha)$ and $\int_{\overline{F}_i} h_i(x)\overline{\mu}^i_n(dx)$ converges to $\int_{\overline{F}_i} h_i(x)\overline{\mu}^i(dx)$, as $n\to\infty$. Also note that $\nu(dx\times S)=\eta(dx)$. Let $v: G\to U$ be a measurable map such that $\nu(dxd\alpha)=v(x,d\alpha)\eta(dx)$. For $i=1,\dots,N$, let μ_i be the restriction of $\overline{\mu}_i$ to F_i. Then from (7.4) we have that $\int_G (Lf)(x,\alpha)\nu(dx,d\alpha)+\sum_{i=1}^N \int_{F_i}(D_if)(x)\mu_i(dx)=0$. From Theorem 5.7 it now follows that $\eta=\eta_v$. This proves the lemma. □

Let

$$\beta^* \doteq \inf_v \int_G k(x,v(x))\eta_v(dx), \tag{7.5}$$

where the infimum on the right side is taken over all Markov controls v. In rest of the section we will show that the infimum above is attained by some Markov control v and furthermore for any admissible pair $(u(\cdot),W(\cdot))$ given on some filtered probability space $\limsup_{T\to\infty}\frac{1}{T}\int_0^T k(X(s),u(s))ds\ge\beta^*$, a.s., where $X(\cdot)$ solves (1.1) with an arbitrary initial distribution. This fact along with Lemma 4.1 will prove Theorem 3.4.

PROPOSITION 7.2. *Let β^* be as defined in* (7.5). *Then, there exists a measurable map $v: G\to U$ such that $\int_G k(x,v(x))\eta_v(dx)=\beta^*$.*

Proof. Let $v_n: G\to U$ be a sequence of maps such that $\int_G k(x,v_n(x))\eta_{v_n}(dx)$ converges to β^*, as $n\to\infty$. For $n\in\mathbb{N}$, define $\nu_n\in\mathcal{P}(G\times S)$ as follows. For $f\in C_b(G\times S)$

$$\int_{G\times S} f(x,\alpha)\nu_n(dxd\alpha)\doteq\int_{G\times S} f(x,\alpha)v_n(x,d\alpha)\eta_{v_n}(dx).$$

From Lemma 7.1 we have that $\{\nu_n\}_{n\ge1}$ is a tight sequence of probability measures. By going to a subsequence if necessary, we have that there exists a $\nu\in\mathcal{P}(G\times S)$ and $\eta\in\mathcal{P}(G)$ such that $\nu_n\to\nu$ and $\eta_{v_n}\to\eta$ as $n\to\infty$. Thus recalling that $\int_G k(x,v_n(x))\eta_{v_n}(dx)=\int_{G\times S}\overline{k}(x,\alpha)\nu_n(dxd\alpha)$ we have that $\int_{G\times S}\overline{k}(x,\alpha)\nu(dxd\alpha)=\beta^*$. Clearly, $\eta(dx)=\nu(dx\times S)$. Furthermore, as in the proof of Lemma 7.1 we have via an application of Theorem 5.7 that if $v: G\to U$ is a measurable map such that $\nu(dxd\alpha)=v(x,d\alpha)\eta(dx)$, then $\eta=\eta_v$. Hence

$$\begin{aligned}\beta^* &= \int_{G\times S}\overline{k}(x,\alpha)\nu(dxd\alpha)\\ &= \int_G\left(\int_S \overline{k}(x,\alpha)v(x,d\alpha)\right)\eta_v(dx)\\ &= \int_G k(x,v(x))\eta_v(dx).\end{aligned}$$ □

PROPOSITION 7.3. *Let $X(\cdot)$ be the solution of* (1.1) *on some filtered probability space $(\Omega,\mathcal{F},\{\mathcal{F}_t\},P)$ on which is given an admissible pair $(u(\cdot),W(\cdot))$. Then $\limsup_{t\to\infty}\frac{1}{t}\int_0^t k(X(s),u(s))ds\ge\beta^*$, a.s., where β^* is as defined in* (7.5).

Proof. In view of Proposition 7.2 it suffices to show that for a.e. ω and for every sequence $t_k\to\infty$ (as $k\to\infty$), there exists a further subsequence (denoted again as the original sequence) such that

$$\lim_{k\to\infty}\frac{1}{t_k}\int_0^{t_k} k(X(s),u(s))ds=\int_G k(x,v(x))\eta_v(dx) \tag{7.6}$$

for some measurable $v: G\to U$, possibly depending on ω and the subsequence.

Define a family of probability measures $\{\nu_t\}_{t\geq 0}$ on $G \times S$ as follows. For $f \in C_b(G \times S)$,

$$\nu_t(f) \doteq \frac{1}{t}\int_0^t \int_S f(X(s),\alpha)u(s,d\alpha)ds. \tag{7.7}$$

We first claim that the family $\{\nu_{t_k}\}$ is a tight family for any sequence $t_k \to \infty$. Since S is compact, in order to prove the claim, it suffices to show that the family $\{\tilde{\nu}_{t_k}\}_{k\geq 1}$ of probability measures on G defined as

$$\tilde{\nu}_{t_k}(f) \doteq \frac{1}{t_k}\int_0^{t_k} f(X(s))ds, \qquad f \in C_b(G),$$

is tight. For $n \in \mathbb{N}$, let f_n be a nonnegative smooth map defined on G such that $f_n(x) = 0$ for $|x| < n$ and $f_n(x) = 1$ for $|x| > n+1$. Using estimates obtained in Lemmas 6.3 (cf. 6.1), 6.4, 4.4, we have as in [6, Chapter 6, pp. 153–154] that for all $\epsilon > 0$, there exists $N(\epsilon) \in \mathbb{N}$ such that for all $n \geq N(\epsilon)$, $\limsup_{t\to\infty} \frac{1}{t}\int_0^t f_n(X(s))ds \leq \epsilon$. This immediately implies the a.s. tightness of $\{\tilde{\nu}_{t_k}\}_{k\geq 1}$ and hence the claim. Now define measures $\mu_{i,k} \in \mathcal{M}_F(F_i)$ as

$$\mu_{i,k}(f) \doteq \frac{1}{t_k}\int_0^{t_k} f(X(s))dY_i(s), \qquad f \in C_b(F_i), \tag{7.8}$$

$i = 1,\dots,N$. Observe that if $g \in C_b^2(G)$ is defined as in the proof of Lemma 5.4, then

$$\begin{aligned}\mu_{i,k}(F_i) &= \frac{Y_i(t_k)}{t_k}\\ &\leq \overline{C} + \frac{1}{t_k}\left|\int_0^{t_k}\langle \nabla g(X(s)), \sigma(X(s))dW(s)\rangle\right| + o(1),\end{aligned}$$

where $\overline{C}$ is defined via (5.4). Noting that the second term on the right side of the above display converges to 0, a.s., as $k \to \infty$, we have that a.s. $\sup_k \mu_{i,k}(F_i) < \infty$, $i = 1,\dots$. Denote the extension of $\mu_{i,k}$ to $\overline{F}_i$, the one point compactification of F_i, by $\overline{\mu}_{i,k}$. Now fix an ω outside a suitable null set. Then, by going to a subsequence if necessary, we have that, for the given ω, there exist $\nu \in \mathcal{P}(G \times S)$, $\overline{\mu}_i \in \mathcal{M}_F(\overline{F}_i)$ such that for all $f \in C_b(G \times S)$ and $f_i \in C_b(\overline{F}_i)$, $i = 1,\dots,N$, $\int_{G\times S} f(x,\alpha)\nu_{t_k}(dxd\alpha)$ converges to $\int_{G\times S} f(x,\alpha)\nu(dxd\alpha)$ and $\int_{\overline{F}_i} f_i(x)\overline{\mu}_{i,k}(dx)$ converges to $\int_{\overline{F}_i} f_i(x)\overline{\mu}_i(dx)$, as $k \to \infty$. Let $f \in C_0^\infty(G)$. Applying Ito's formula to $f(X(t_k))$, dividing by t_k, and taking the limit as $k \to \infty$, we have that

$$\int_{G\times S}(Lf)(x,\alpha)\nu(dxd\alpha) + \sum_{i=1}^N \int_{F_i}(D_i f)(x)\mu_i(dx) = 0.$$

Hence from Theorem 5.7 we have that if $\eta(dx) \doteq \nu(dx \times S)$ and ν is disintegrated as $\nu(dxd\alpha) = v(x,d\alpha)\eta(dx)$, then $\eta = \eta_v$. This immediately yields that

$$\lim_{k\to\infty}\frac{1}{t_k}\int_0^{t_k} k(X(s),u(s))ds = \int_G \overline{k}(x,v)\eta_v(dx).$$

This proves the proposition. □

We are now ready to prove our main result.

Proof of Theorem 3.4. From Proposition 7.2 there exists a measurable map $\overline{v}: G \to U$ such that

$$\beta^* = \int_G k(x, \overline{v}(x))\eta_{\overline{v}}(dx). \tag{7.9}$$

Let $\overline{X}(\cdot)$ solve (2.3), with v replaced by $\overline{v}$ on some filtered probability space with probability law of $\overline{X}(0)$ equal to μ. Then, from Lemma 4.1

$$\limsup_{T\to\infty} \frac{1}{T}\int_0^T k(\overline{X}(s), \overline{v}(\overline{X}(s)))ds = \int_G k(x, \overline{v}(x))\eta_{\overline{v}}(dx), \tag{7.10}$$

a.s. Combining (7.9) and (7.10) we have that $\beta^* = \limsup_{T\to\infty} \frac{1}{T}\int_0^T k(\overline{X}(s), v(\overline{X}(s)))$ ds. The result now follows from Proposition 7.3. □

8. Characterizing the value via the HJB equation. One of the important goals in optimal control theory is to derive the HJB equation for the value function and characterize the value function as the unique solution (in an appropriate class) of the PDE. The detailed treatment of this aspect of the ergodic control problem studied in this paper will be undertaken elsewhere; however, we briefly outline below the key steps in such a program.

The classical approach to the HJB equation for the ergodic control is by the "vanishing discount method" (cf. [10, 8, 6, 3]). In this approach the first step is to characterize the value function $V_\alpha(x)$ of the discounted control problem,

$$V_\alpha(x) = \inf_u E\left(\int_0^\infty e^{-\alpha t}k(X^x(s), u(s))ds\right), \tag{8.1}$$

where $\alpha \in (0, \infty)$, the infimum is taken over all admissible controls u and $X^x(\cdot)$ is the solution of (1.1) with $X(0) \equiv x$. The natural HJB equation associated with the control problem (8.1) will be

$$\begin{aligned} \inf_{u\in U}(L\psi(x,u) + k(x,u) - \alpha\psi(x)) = 0, &\qquad x \in G, \\ \langle \nabla\psi(x), d_i\rangle = 0, &\qquad x \in \partial G,\ \ i \in In(x). \end{aligned} \tag{8.2}$$

Theory of classical solutions for such PDEs in domains with corners is not available; however, using the seminal ideas of Crandall and Lions [11, 33], Dupuis and Ishii [14] have proved the existence of a viscosity solution to (8.2). We remark that [14] considers the case of bounded domains; however, the key ideas there can be adapted to cover the current setting. By standard techniques (cf. Theorem III.2.1 in [6]) it can be shown that $V_\alpha(\cdot) \in C_b(G)$. Next, in order to establish that V_α is the unique viscosity solution of (8.2), we will need to adapt the proofs of Theorems I.1 and II.1 of [33] to oblique derivative problems in domains with corners.

The next key step in the program is to show that the family $\{V_\alpha(x) - V_\alpha(0); \alpha \in (0,1)\}$ is precompact in $C(G)$. The proof of this statement is currently the biggest obstacle since the classical derivation (see Theorem VI.3.1 of [6]) makes use of certain gradient estimates on $V_\alpha(x)$, uniform in α, which are currently unavailable. Another approach based on viscosity solutions, taken in [3], avoids this difficulty by making some strong stability assumptions on the model (a restoring force towards bounded sets that grows without bound as $|x| \to \infty$), which are not satisfied in the current setup because of the radially homogeneous nature of the problem.

Once the above compactness issue is resolved one can, by usual limiting arguments, as $\alpha \to 0$, (cf. [3]) and the stability properties of the viscosity solution (cf. Proposition I.3 [33]), exhibit a solution (V^*, ρ^*) for the HJB equation for the ergodic control problem:

$$\begin{aligned} \inf_{u\in U} (LV^*(x,u) + k(x,u) - \rho^*) = 0, \qquad & x \in G, \\ \langle \nabla\psi(x), d_i\rangle = 0, \qquad & x \in \partial G, \ \ i \in In(x). \end{aligned} \tag{8.3}$$

Finally one would like to establish that $\rho^* = \text{ess inf}_u \limsup_{T\to\infty} \frac{1}{T}\int_0^T k(X(s),u(s))ds$, where the infimum is taken over all admissible controls. The proof of this statement will proceed along the lines of Theorem 4.2 of [3].

Appendix.

Sketch of the proof for the near monotone case. We assume that all the conditions on the model imposed in the earlier sections, except Condition 3.1, hold. We will replace this "blanket stability" condition by the following assumption on the function k. Following [6], we will call a Markov control v a stable stationary Markov control if the corresponding solution to (2.3) is a positive recurrent Markov process and has a unique invariant measure, denoted as η_v. We denote the collection of all such controls by $\mathcal{F}^*$. The near monotone condition on the cost function k is as follows.

CONDITION 9.1. (a) $\beta \doteq \inf_{v\in\mathcal{F}^*} \int_G k(x,v(x))\eta_v(dx) < \infty$; (b) $\liminf_{|x|\to\infty} \inf_{u\in U} k(x,u) > \beta$.

Part (a) of the above condition is satisfied if, for example, there exist a $\delta, M \in (0,\infty)$, and some measurable map $v : G \to U$ such that $b(x,v(x)) \in \mathcal{C}(\delta)$ for all $x \in G$ such that $|x| > M$. The (b) part of the condition is obviously satisfied if $k(x,u) = \psi(|x|)$ for a monotone increasing map $\psi : \mathbb{R}_+ \to \mathbb{R}$.

The key steps in the proof of Theorem 3.4 for this setup are, once more, Propositions 7.2 and 7.3 (with β^* there replaced by β). Since in the near monotone case the proof of Proposition 7.2 is very similar to that of Proposition 7.3, we sketch only the proof of the latter case.

Sketch of Proposition 7.3 in the near monotone case. Let $\{\nu_t\}_{t\geq 0}$ be as defined in (7.7) and let $\overline{G}$ be the one point compactification of G and denote the point at ∞ by p_∞. Since S is compact, we have that $\{\nu_t\}_{t\geq 0}$ is a tight family of probability measures on $\overline{G}\times S$. Define $\mu_{i,k} \in \mathcal{M}_F(F_i)$, via (7.8). Then there exist $\mu \in \mathcal{P}(\overline{G}\times S)$, $\overline{\mu}_i \in \mathcal{M}_F(F_i)$, such that for all $f \in C_b(\overline{G}\times S)$ and $f_i \in C_b(\overline{F}_i)$, $i = 1,\dots,N$, $\int_{\overline{G}\times S} f(x,\alpha)\nu_{t_k}(dxd\alpha)$ converges to $\int_{\overline{G}\times S} f(x,\alpha)\nu(dxd\alpha)$ and $\int_{\overline{F}_i} f_i(x)\overline{\mu}_{i,k}(dx)$ converges to $\int_{\overline{F}_i} f_i(x)\overline{\mu}_i(dx)$, as $k\to\infty$. Let ν be decomposed as $\nu = (1-\rho)\nu_1 + \rho\nu_\infty$, where $\nu_1 \in \mathcal{P}(G\times S)$ and $\nu_\infty \in \mathcal{P}(\{p_\infty\}\times S)$. Now using the near monotonicity of k, one shows, exactly as in [6, pp. 146–147] that $\rho = 0$. Thus we have that $\nu = \nu_1 \in \mathcal{P}(G\times S)$. The rest of the proof is exactly the same as in the stable case.

Proof of Lemma 4.2. We begin by observing that the geometry of the space G implies that there exists $C \in (1,\infty)$ such that for all $x \in G$ and $\lambda \in \Lambda$, $d(x,F_\lambda) \leq C\max_{i\in\lambda}\langle x, n_i\rangle$. Next, from Condition 2.2(c), it follows that for all $\lambda \in \Lambda$, there exist positive constants $\{c_i^\lambda\}_{i\in\lambda}$ such that $\eta^\lambda \doteq \sum_{i\in\lambda} c_i^\lambda n_i$ satisfies $\langle \eta^\lambda, d_i\rangle > 0$ for all $i \in \lambda$. Define $\tilde{c} \doteq \inf_{i\in\lambda;\ \lambda\in\Lambda}\langle \eta^\lambda, d_i\rangle$. Furthermore as a convenient normalization we take $\sum_{i\in\lambda} c_i^\lambda = \frac{1}{2}$. Next define constants (γ_k, β_k), $k = 0,1,\dots,N$, inductively as follows:

$\gamma_N \doteq \frac{1}{2(C+1)}$, $\beta_N \doteq \tilde{c}\gamma_N$, and for $k = 1, \ldots, N$, $\gamma_{N-k} \doteq \frac{\beta_{N-k+1}}{C}$, $\beta_{N-k} \doteq \tilde{c}\gamma_{N-k}$. Let ϕ, ψ be maps from $\mathbb{R}_+$ to $\mathbb{R}_+$ defined as follows:

$$\begin{aligned}\phi(x) &\doteq x - 1, \qquad x \in \left[0, \frac{1}{2}\right] \\ &\doteq 0, \qquad x \geq 1\end{aligned}$$

and

$$\begin{aligned}\psi(x) &\doteq 0, \qquad x \in \left[0, \frac{1}{2}\right] \\ &\doteq 1, \qquad x \geq 1.\end{aligned}$$

Now define for $\lambda \in \Lambda$, $f_\lambda : G \to \mathbb{R}$ as follows:

$$f_\lambda(x) \doteq a_\lambda \phi\left(\frac{\langle \eta^\lambda, x\rangle}{\beta_{|\lambda|}}\right) \prod_{j \notin \lambda} \psi\left(\frac{\langle n_j, x\rangle}{\gamma_{|\lambda|}}\right),$$

where $|\lambda|$ denotes the cardinality of the set λ and a_λ are suitable positive chosen inductively as follows. For all λ with $|\lambda| = 1$ we choose a_λ so that $a_\lambda \langle \eta_\lambda, d_i\rangle \geq \beta_1$, $i \in \lambda$. Having chosen a_λ for all λ with $1 \leq |\lambda| < k$, we choose a_λ, for a $\lambda \in \Lambda$ with $|\lambda| = k$, such that $a_\lambda \langle \eta_\lambda, d_i\rangle \geq \beta_k(M_{k-1} + 1)$, $i \in \lambda$, where

$$M_{k-1} \doteq \sum_{j=1}^{k-1} \sum_{\lambda : |\lambda| = j} \left(- \inf_{x \in F_i ; i \in \lambda} \langle \nabla f_\lambda(x), d_i\rangle \right)^+.$$

Finally define $g : G \to \mathbb{R}$ as $g(x) \doteq \sum_{\lambda \in \Lambda} f_\lambda(x)$. It can be verified as in [12] that g defined as above satisfies (4.1). □

Alternate proof of Lemma 4.3 of [2]. We give a proof which, unlike the proof in [2], does not appeal to the Markov property of $X^x(\cdot)$. This lemma is needed for the proof of Theorem 4.4 of [2] and therefore the proof of Lemma 6.1 of the present paper.

LEMMA 9.2 (Lemma 4.3 of [2]). *For $x \in G$ and an admissible pair $(u(\cdot), W(\cdot))$ given on some filtered probability space, $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$. Let $X^x(\cdot)$ denote the solution of (1.1) with $X^x(0) = x$. Let $\Delta > 0$ be fixed. For $n \in \mathbb{N}$ let ν_n be defined as follows:*

$$\nu_n \doteq \sup_{(n-1)\Delta \leq s \leq n\Delta} \left| \int_{(n-1)\Delta}^{s} \sigma(X^x(s)) dW(s) \right|.$$

Then for any $\rho \in (0, \infty)$ and $m, n \in \mathbb{N}$, $m \leq n$,

$$E\left(e^{\rho \sum_{i=m}^n \nu_i}\right) \leq \left(2\sqrt{2} e^{k^2 \rho^2 r^2 \Delta}\right)^{(n-m+1)}.$$

Proof. Let $\int_0^\cdot \sigma(X^x(s)) dW(s) \equiv (M_1(\cdot), \ldots, M_k(\cdot))$. Then for $i = 1, \ldots, k$, $M_i(\cdot)$ is a square integrable $\{\mathcal{F}_t\}$ martingale. An application of Cauchy–Schwarz inequality gives that

$$E\left(e^{\rho \sum_{i=m}^n \nu_i}\right) \leq \left(\prod_{j=1}^{k} E\left(e^{k\rho \sum_{i=m}^n M_j^i}\right)\right)^{\frac{1}{k}},$$

where for $j = 1, \dots, k$, $i = m, m+1, \dots, n$, $M_j^i \doteq \sup_{(i-1)\Delta \le s \le i\Delta} |M_j(s) - M_j((i-1)\Delta)|$. Now fix a $j \in \{1, \dots, k\}$. In view of Condition 2.5, $|\langle M_j \rangle(t)| \to \infty$, a.s. as $t \to \infty$. Thus from Theorem V.1.6 in [38], $M_j(\cdot)$ has the same probability law as $B(\tau(\cdot))$, where $\tau(\cdot) \doteq \langle M_j \rangle(\cdot)$, $B(t)$ is a $\mathcal{G}_t$ standard Brownian motion, $\mathcal{G}_t \doteq \mathcal{F}_{S(t)}$, and $S(t) \doteq \inf\{s \ge 0 : \langle M_j \rangle(s) > t\}$. Next observe that

$$\begin{aligned} E\left(e^{k\rho \sum_{i=m}^n M_j^i}\right) &= E\left(e^{k\rho \sum_{i=m}^n \sup_{0\le s\le\Delta} |B(\tau((i-1)\Delta+s)) - B(\tau((i-1)\Delta))|}\right) \\ &= E\left(H_{n-1} e^{k\rho \sup_{0\le s\le\Delta} |B(\tau((n-1)\Delta+s)) - B(\tau((n-1)\Delta))|}\right), \end{aligned}$$

where $H_{n-1} \doteq e^{k\rho \sum_{i=m}^{n-1} \sup_{0\le s\le\Delta} |B(\tau((i-1)\Delta+s)) - B(\tau((i-1)\Delta))|}$. Note that H_{n-1} is $\mathcal{G}_{\tau((n-1)\Delta)}$ measurable. Furthermore, from Condition 2.4(iv) we have that for $0 < s < t < \infty$, $|\tau(t) - \tau(s)| \le r^2 |t - s|$ and therefore

$$\begin{aligned} &\sup_{0\le s\le\Delta} |B(\tau((n-1)\Delta + s)) - B(\tau((n-1)\Delta))| \\ &\quad \le \sup_{0\le t\le r^2\Delta} |B(\tau((n-1)\Delta) + t) - B(\tau((n-1)\Delta))|. \end{aligned}$$

Finally note that

$$\begin{aligned} &E\left(H_{n-1} e^{k\rho \sup_{0\le s\le\Delta} |B(\tau((n-1)\Delta+s)) - B(\tau((n-1)\Delta))|}\right) \\ &\le E\left(H_{n-1} E\left(e^{k\rho \sup_{0\le t\le r^2\Delta} |B(\tau((n-1)\Delta)+t) - B(\tau((n-1)\Delta))|} \mid \mathcal{G}_{\tau((n-1)\Delta)}\right)\right) \\ &\le E\left(H_{n-1} 2\left(E(e^{2k\rho |B(r^2\Delta)|})\right)^{\frac{1}{2}}\right) \\ &\le 2\sqrt{2} e^{k^2\rho^2 r^2 \Delta} E\left(H_{n-1}\right), \end{aligned}$$

where the second inequality above follows from Doob's inequality for submartingales. Iterating the above inequalities we have the result. □

Acknowledgments. I would like to thank the referees and an associate editor for suggesting various improvements to the manuscript.

REFERENCES

[1] M. K. Ghosh, A. Arapostathis, and S. I. Marcus, *Ergodic control of switching diffusions*, SIAM J. Control Optim., 35 (1997), pp. 1952–1988.

[2] R. Atar, A. Budhiraja, and P. Dupuis, *On positive recurrence of constrained diffusion processes*, Ann. Probab., 29 (2001), pp. 979–1000.

[3] G. K. Basak, V. S. Borkar, and M. K. Ghosh, *Ergodic control of degenerate diffusions*, Stochastic Anal. Appl., 1 (1997), pp. 1–17.

[4] A. Bensoussan, *Perturbation Methods in Optimal Control*, John Wiley, New York, 1988.

[5] A. G. Bhatt and V. S. Borkar, *Occupation measures for controlled Markov processes: Characterization and optimality*, Ann. Probab., 24 (1996), pp. 1531–1562.

[6] V. Borkar, *Optimal Control of Diffusion Processes*, Longman Scientific and Technical, Harlow, UK, 1989.

[7] V. S. Borkar and M. K. Ghosh, *Ergodic control of multidimensional diffusions* I: *The existence results*, SIAM J. Control Optim., 26 (1988), pp. 112–126.

[8] V. S. Borkar and M. K. Ghosh, *Ergodic control of multidimensional diffusions.* II. *Adaptive control*, Appl. Math. Optim., 21 (1990), pp. 191–220.

[9] A. Budhiraja and P. Dupuis, *Simple necessary and sufficient conditions for the stability of constrained processes*, SIAM J. Appl. Math., 59 (1999), pp. 1686–1700.

[10] R. M. Cox, *Stationary and Discounted Control of Diffusion Processes*, Ph.D. thesis, Columbia University, New York, 1984.
[11] M. G. Crandall and P. L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.
[12] J. G. Dai and R. J. Williams, *Existence and uniqueness of semimartingale reflecting Brownian motions in convex polyhedrons*, Theory Probab. Appl., 40 (1995), pp. 1–40.
[13] P. Dupuis and H. Ishii, *On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications*, Stochastics, 35 (1991), pp. 31–62.
[14] P. Dupuis and H. Ishii, *On oblique derivative problems for fully nonlinear second-order elliptic PDEs on domains with corners*, Hokkaido Math. J., 20 (1991), pp. 135–164.
[15] P. Dupuis and K. Ramanan, *Convex duality and the Skorokhod Problem.* I, II, Probab. Theory Related Fields, 2 (1999), pp. 153–195, 197–236.
[16] P. Dupuis and K. Ramanan, *A multiclass feedback queueing network with a regular Skorokhod problem*, Queuing Syst. Theory Appl., 36 (2000), pp. 327–349.
[17] P. E. Echeverria, *A criterion for invariant measures of Markov processes*, Z. Warsch. Verw. Gebiete., 61 (1982), pp. 1–16.
[18] J. M. Harrison and M. I. Reiman, *Reflected Brownian motion on an orthant*, Ann. Probab., 9 (1981), pp. 302–308.
[19] J. M. Harrison and J. A. Van Miegham, *Dynamic control of Brownian networks: State space collapse and equivalent workload formulation*, Ann. Appl. Probab., 7 (1997), pp. 747–771.
[20] J. M. Harrison and L. Wein, *Scheduling networks of queues: Heavy traffic analysis of a simple open network*, Queuing Systems Theory Appl., 5 (1989), pp. 265–280.
[21] J. M. Harrison and R. J. Williams, *Brownian models of open queueing networks with homogeneous customer populations*, Stochastics, 22 (1987), pp. 77–115.
[22] U. G. Haussmann and J. P. Lepeltier, *On the existence of optimal controls*, SIAM J. Control Optim., 28 (1990), pp. 851–902.
[23] F. P. Kelly and C. N. Laws, *Dynamic routing in open queuing networks: Brownian models, cut constraints and resource pooling*, Queuing Systems Theory Appl., 13 (1993), pp. 47–86.
[24] R. Z. Khasminskii, *Ergodic properties of recurrent diffusion processes and stabilization of the solution to the Cauchy problem for parabolic equations*, Theory Probab. Appl., 5 (1960), pp. 179–196.
[25] R. Z. Khasminskii, *Stochastic Stability of Differential Equations*, Sijthoff & Noordhoff, Groningen, The Netherlands, 1980.
[26] T. G. Kurtz, *Martingale problems for constrained Markov processes*, in Advances in Stochastic Calculus, J. S. Baras and V. Mirelli, eds., Springer-Verlag, New York, 1990, pp. 151–168.
[27] T. G. Kurtz, *A control formulation for constrained Markov processes*, in Mathematics of Random Media, Lectures in Appl. Math., 27, AMS, Providence, RI, 1991, pp. 139–150.
[28] T. G. Kurtz and R. H. Stockbridge, *Existence of Markov controls and characterization of optimal Markov controls*, SIAM J. Control Optim., 36 (1998), pp. 609–653.
[29] T. G. Kurtz and R. H. Stockbridge, *Stationary solutions and forward equations for controlled and singular martingale problems*, Electron. J. Probab., 6 (2001), 52 pp.
[30] H. J. Kushner, *Optimality conditions for the average cost per unit time problem with a diffusion model*, SIAM J. Control Optim., 16 (1978), pp. 330–346.
[31] H. J. Kushner, *Heavy Traffic Analysis of Controlled Queueing and Communication Networks*, Springer-Verlag, New York, 2001.
[32] L. F. Martins and H. J. Kushner, *Routing and singular control for queuing networks in heavy traffic*, SIAM J. Control Optim, 28 (1990), pp. 1209–1233.
[33] P.-L. Lions, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations.* II*, Viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 8 (1983), pp. 1229–1276.
[34] J. L. Menaldi and M. Robin, *Ergodic control of reflected diffusions with jumps*, Appl. Math. Optim., 35 (1997), pp. 117–138.
[35] V. Nguyen, *Processing networks with parallel and sequential tasks: Heavy traffic analysis and Brownian limits*, Ann. Appl. Probab., 3 (1993), pp. 28–55.
[36] W. Peterson, *A heavy traffic limit theorem for networks of queues with multiple customer types*, Math. Oper. Res., 16 (1991), pp. 90–118.
[37] M. Reiman and R. J. Williams, *A boundary property for semimartingale reflecting Brownian motions*, Probab. Theory Related Fields, 77 (1988), pp. 87–97.
[38] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*, Springer-Verlag, Berlin, 1991.
[39] M. Robin, *Long-term average cost control problems for continuous time markov processes—A survey*, Acta Appl. Math., 1 (1983), pp. 281–300.

[40] R. H. Stockbridge, *Time Average Control of Martingale Problems*, Ph.D. thesis, University of Wisconsin, Madison, WI, 1987.
[41] D. Stroock and S. Varadhan, *Multidimensional Diffusion Processes*, Springer-Verlag, Berlin-New York, 1979.
[42] A. Weiss, *Invariant Measures of Diffusions in Bounded Domains*, Ph.D. thesis, New York University, New York, 1981.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 559–575

ON REFLECTING BOUNDARY PROBLEM FOR OPTIMAL CONTROL*

OANA-SILVIA SEREA†

Abstract. This paper deals with Mayer's problem for controlled systems with reflection on the boundary of a closed subset K. The main result is the characterization of the *possibly discontinuous* value function in terms of a unique solution *in a suitable sense* to a partial differential equation of Hamilton–Jacobi–Bellman type.

Key words. control of variational inequality, boundary reflection, viscosity solutions

AMS subject classifications. 34K35, 49J24, 49L20, 49L25, 93C15

PII. S0363012901395935

1. Introduction.

We investigate the Mayer control problem:

$$\text{Minimize } g(x(T)) \tag{1}$$

for a given $T > 0$ over all absolutely continuous solutions of the following differential variational inequality:

$$\begin{cases} \text{(i) } x'(t) \in f(x(t), u(t)) - N_K(x(t)) \text{ for almost all } t \geq t_0, \\ \text{(ii) } x(t) \in K \text{ for all } t \geq t_0,\ x(t_0) = x_0, \text{ and} \\ u(\cdot) : [0, \infty) \to U \text{ is a measurable function,} \end{cases} \tag{2}$$

where $N_K(x)$ is the normal cone to K at $x \in K$ (see Definition 1).

Here K is a nonempty closed subset of $\mathbb{R}^N$, $g : K \to \mathbb{R}$ and f is a function from $K \times U$ into $\mathbb{R}^N$.

If $\mathcal{U}(t_0)$ is the set of measurable controls on $[t_0, \infty)$ with values in U, the value function corresponding to the optimal control problem (1), (2) is given by

$$V(t_0, x_0) = \inf_{u(\cdot) \in \mathcal{U}(t_0)} g(x(T; t_0, x_0, u(\cdot)) \text{ for all } (t_0, x_0) \in [0, T] \times K, \tag{3}$$

where $x(\cdot; t_0, x_0, u(\cdot))$ denotes the solution of (2) starting from (t_0, x_0).

By the very definition it is easy to see that the value function is finite on its domain $[0, T] \times K$, if and only if (2) has solutions. This explains the choice of the form of the right-hand side of the differential inclusion (2). We notice that $N_K(x) = \{0\}$ whenever $x \in \text{int}K$; f is modified only on the boundary of K, so (2) is a problem with reflection at the boundary. We shall show that this reflection allows us to obtain the existence of solutions to (2) (see section 1).

Our main purpose in this paper is to characterize the value function (3) by an equation of Hamilton–Jacobi type.

Of course, the characterization is based on a suitable definition for the notion of viscosity solutions of a Hamilton–Jacobi–Bellman inequality (HJBI) that we will introduce below.

*Received by the editors December 10, 2001; accepted for publication (in revised form) December 2, 2002; published electronically May 12, 2003.
http://www.siam.org/journals/sicon/42-2/39593.html

†Laboratoire de Mathématiques, Unité CNRS, FRE-2218, Université de Bretagne Occidentale, 6 avenue Victor Le Gorgeu, BP 809, 29285 Brest cedex, France (Oana-Silvia.Serea@univ-brest.fr).

More precisely, we will prove that the value function V is the unique solution to

$$\text{(HJBI)} \qquad \begin{cases} \frac{\partial V}{\partial t}(t,x) + H(x, \frac{\partial V}{\partial x}(t,x)) - \left\langle \frac{\partial V}{\partial x}(t,x),\ N_K(x) \right\rangle \ni 0 \\ \qquad \text{if } (t,x) \in [0,T) \times K, \\ \text{with the condition } V(T,x) = g(x) \text{ if } x \in K, \end{cases}$$

where $H(x,p) := \min_{u \in U} \langle f(x,u), p \rangle$.

If the boundary of K, $\partial K \in C^1$, and K is the closure of an open set, we will show that V is a viscosity solution of the following Hamilton–Jacobi equation with Neumann-type boundary condition in the sense of [13]:

$$(4) \qquad \begin{cases} \frac{\partial V}{\partial t}(t,x) + H(x, \frac{\partial V}{\partial x}(t,x)) = 0 \text{ if } (t,x) \in [0,T) \times K, \\ \frac{\partial V}{\partial n}(t,x) = 0 \text{ if } (t,x) \in [0,T) \times \partial K, \\ \text{with the condition } V(T,x) = g(x),\ x \in K, \end{cases}$$

where $n(x)$ is the unit outward normal to K at $x \in \partial K$.

It is well known that the value function for the Skorokhod control problem (see [3], [13]) with a smooth K is a viscosity solution of (4). The Skorokhod problem for a smooth K has been considered and solved by Lions [13] and Lions and Snitzman [14]. Another study was made by Tanaka in [18] when K is convex with normal reflection. By a different approach, this problem was considered like a viability problem for a differential inclusion by Frankowska in [8]. Note that the notion of solutions of the Skorokhod problem is not the same as the notion of solutions to (2) that we use in this paper, but for the smooth case the two control problems lead to the same Hamilton–Jacobi equation (4).

Our second interest is to establish that the two following systems,

$$(5) \qquad \begin{cases} \text{(i) } x'(t) \in F(x(t)) - N_K(x(t)) \text{ for almost all } t \geq t_0, \\ \text{(ii) } x(t) \in K \text{ for all } t \in [t_0, \infty),\ t_0 \geq 0,\ x(t_0) = x_0 \in K \end{cases}$$

and

$$(6) \qquad \begin{cases} \text{(i) } x'(t) \in \Pi_{\overline{co} T_K(x(t))} F(x(t)) \text{ for almost all } t \geq t_0, \\ \text{(ii) } x(t) \in K \text{ for all } t \in [t_0, \infty),\ t_0 \geq 0,\ x(t_0) = x_0 \in K, \end{cases}$$

have the same set of solutions.

Here K is compact, $F : K \to \mathbb{R}^N$ is a set valued map and $\overline{co} A$ is the closed convex hull of a set A.

In general, the map $x \to N_K(x)$ has no easy continuity properties and so the right side of the differential inclusion is (2). For this reason the set of solutions to (5) or (6) may be empty. So it is necessary to find regularity hypotheses for K in order to provide existence and eventually uniqueness results for (5) or (6).

These kind of results for a general map F can be applied, in particular, when $F(\cdot) = f(\cdot, U)$, allowing us to obtain properties of the set of solutions to (2).

Our main contribution here is the fact that by introducing the projection on the closed convex hull of $T_K(x)$ in (6) we succeed in treating the case where the set K is only compact, improving the already known equivalence and existence results of [2] where K is supposed to be sleek.

Existence and equivalence results for (5) and (6) are established by Henry [11] for a convex set. The convexity assumption on the set K, has been relaxed by Cornet in [6], who merely required the tangential regularity. We also refer to Thibault [19] for

the case of a closed set K for an existence result of viable solution, but the reflection is made using the Clarke normal cone. Note that in [19] the set K may depend on t.

We note that the boundary reflection control problem was not yet well studied for nonsmooth K. We also succeed in generalizing some existence and equivalence results of [2] for the systems (5) and (6).

Let us explain how this paper is organized.

In the first section we introduce some preliminaries and we study the systems (5), (6).

In the second section we prove that the value function is a viscosity solution of (HJBI) in the sense of Definition 3, and a uniqueness result for the solutions of this partial differential inequality is also established.

In the third section we study the case of discontinuous and only bounded value functions for our control problem. Our main result says that V is the unique generalized solution to the corresponding (HJBI) for arbitrary discontinuous terminal cost g.

The fourth section concerns existence and uniqueness results of l.s.c. solutions to (HJBI) in the sense of Definition 16.

The last section is an appendix with technical proofs of our claims.

2. Preliminaries.

2.1. Definitions, assumptions, and notations. We assume that $f : K \times U \to \mathbb{R}$ is continuous and satisfies

$$(\mathrm{H}_f) \qquad \begin{cases} ||f(x,u)|| \le a(1+\|x\|), \\ ||f(x,u)-f(y,u)|| \le c_1 \|x-y\| \text{ for all } x,y \in K, u \in U, \\ \text{the set } f(x,U) \text{ is convex,} \end{cases}$$

where $c_1, a > 0$ are constants; U is a compact metric space.

We recall the notions of tangent and normal cones.

Definition 1. *For $x \in K$, we define by*

$$T_K(x) = \left\{ v \in \mathbb{R}^N | \lim_{h \to 0^+} \inf d_K(x+hv)/h = 0 \right\}$$

the tangent cone to K at x and by

$$N_K(x) = T_K(x)^- = \{ p \in \mathbb{R}^N | \langle p,\ v \rangle \le 0 \text{ for all } v \in T_K(x) \}$$

the normal cone to K at x.

Recall that $T_K(x)$ is a closed cone and $N_K(x)$ is a closed convex cone.

Let us describe some classes of sets which will be used in the following sections.

Definition 2. *A closed set $K \subset \mathbb{R}^N$ is called proximal retract if there exists a neighborhood I of K such that the projection $\Pi_K(\cdot)$ is single-valued in I, with $\Pi_K(x) := \{ z \in K \mid ||x-z|| = \inf_{y \in K} ||x-y|| \}$ for all $x \in \mathbb{R}$.*

We will describe some of the properties of such sets. This will be the key for the proof of the existence and uniqueness results concerning (5) and (HJBI). The class of proximal retracts includes closed, convex subsets of $\mathbb{R}^N$ and submanifolds of $\mathbb{R}^N$ of class $\mathrm{C}^{1,1}$. Another class of proximal retracts is the class of weakly convex sets (see [8] for the definition and the geometrical interpretation). A complete characterization of proximal retract sets is made in [17] (see Theorem 4.1, p. 5245). In particular, such sets have the property that there exists $\rho > 0$ such that every nonzero normal "can be

realized" by a ball with a radius equal to ρ. This characterization says, in particular, that only "exterior" corners are allowed.

So, if K is proximal retract, then from Theorem 4.1 in [17], Lemma 4.2 and Theorem 2.2 in [6] we have the following:

- There exist $r, c > 0$ such that the application $x \to N_K(x) \cap B(0,r) + cx$ is monotone[1] on K. This monotonicity property, which is equivalent to Definition 2, is very important because it allows us to establish the uniqueness of solutions to (2).

- The set K is sleek, i.e., the map $x \to T_K(x)$ is l.s.c.

- For all $x \in K$, $T_K(x) = C_K(x)$, where $C_K(x)$ denotes Clarke's tangent cone.[2] Note that the class of sleek sets is larger then the class of proximal retracts.

2.2. Viscosity solutions. To describe the value function as a unique solution to the corresponding HJBI, we introduce the following definition of solutions to (HJBI).

DEFINITION 3. *A viscosity supersolution of* (HJBI) *is an l.s.c. function* $\psi : (0,T) \times K \to \mathbb{R}$ *such that*

$$\begin{gathered}
\text{for any } \phi \in C^1 \text{ and } (t_0,x_0) \in \arg\min(\psi - \phi), \\
\text{if } x_0 \in \operatorname{int} K,\ \frac{\partial \phi}{\partial t}(t_0,x_0) + H\left(x_0, \frac{\partial \phi}{\partial x}(t_0,x_0)\right) \le 0 \\
\text{and if } x_0 \in \partial K,\ \text{there exists } y_0 \in N_K(x_0) \text{ such that} \\
\frac{\partial \phi}{\partial t}(t_0,x_0) + H\left(x_0, \frac{\partial \phi}{\partial x}(t_0,x_0)\right) - \left\langle y_0,\ \frac{\partial \phi}{\partial x}(t_0,x_0)\right\rangle \le 0
\end{gathered}$$

and a viscosity subsolution of (HJBI) *is a u.s.c. function* $\varphi : (0,T) \times K \to \mathbb{R}$ *such that*

$$\begin{gathered}
\text{for any } \phi \in C^1 \text{ and } (t_0,x_0) \in \arg\max(\varphi - \phi), \\
\text{if } x_0 \in \operatorname{int} K,\ \frac{\partial \phi}{\partial t}(t_0,x_0) + H\left(x_0, \frac{\partial \phi}{\partial x}(t_0,x_0)\right) \ge 0 \\
\text{and if } x_0 \in \partial K,\ \text{there exists } z_0 \in N_K(x_0) \text{ such that} \\
\frac{\partial \phi}{\partial t}(t_0,x_0) + H\left(x_0, \frac{\partial \phi}{\partial x}(t_0,x_0)\right) - \left\langle z_0,\ \frac{\partial \phi}{\partial x}(t_0,x_0)\right\rangle \ge 0.
\end{gathered}$$

A viscosity solution of (HJBI) *is a function which is both subsolution and supersolution.*

It is clear that a viscosity solution is a continuous function because it is simultaneously u.s.c. and l.s.c.

Remark 4. A motivation for our definition of (HJBI) is the fact that, when (t_0,x_0) is a differentiability point of V, we have in the usual sense

$$\frac{\partial V}{\partial t}(t_0,x_0) + H\left(x_0, \frac{\partial V}{\partial x}(t_0,x_0)\right) - \left\langle N_K(x_0),\ \frac{\partial V}{\partial x}(t_0,x_0)\right\rangle \ni 0. \tag{7}$$

[1] Recall that a set valued map $G : K \to \mathbb{R}^N$ is monotone if $\langle y_1 - y_2, x_1 - x_2\rangle \ge 0$ for all $y_i \in G(x_i)$, $i \in \{1,2\}$.

[2] $C_K(x) = \{v \mid \lim_{h \to 0^+, K \ni x' \to x} d_K(x' + hv)/h = 0\}$. This tangent cone is always convex.

Indeed, there exists $\lambda \in [0,1]$ such that

$$0 = \lambda \left(\frac{\partial V}{\partial t}(t_0, x_0) + H\left(x_0, \frac{\partial V}{\partial x}(t_0, x_0)\right) - \left\langle y_0, \ \frac{\partial V}{\partial x}(t_0, x_0) \right\rangle \right) + (1-\lambda) \left(\frac{\partial V}{\partial t}(t_0, x_0) + H\left(x_0, \frac{\partial V}{\partial x}(t_0, x_0)\right) - \left\langle z_0, \ \frac{\partial V}{\partial x}(t_0, x_0) \right\rangle \right),$$

and because $N_K(x_0)$ is convex, (7) is verified.

It is quite natural to obtain an equation of the form (7), namely a partial differential inequality. The motivation lies in the fact that for a smooth set, the reflection is channeled in a fixed direction, given by the outward normal. For nonsmooth sets the outward normal will be replaced with the normal cone which, in general, contains many directions.

Note that this definition contains those given by Lions in [13], when the boundary of K, $\partial K \in C^1$.

2.3. Control systems with reflection on the boundary of a constraint set. In this section we study the differential inequalities (5) and (6) by explaining the method which we use in order to get a boundary reflection for closed sets K. This allows us to give some applications to the properties of solutions to the controlled system (2).

We consider a closed set K, a set valued map $F : K \to \mathbb{R}^N$, and the following differential inclusion:

$$\begin{cases} \text{(i) } x'(t) \in F(x(t)) \text{ for almost all } t \geq t_0, \\ \text{(ii) } x(t_0) = x_0 \in K, \ t_0 \geq 0. \end{cases} \tag{8}$$

The equation (6) appears naturally if we want a given closed set to become a viability[3] domain of a new system which is "as close as possible" to the original dynamic system (8).

Indeed, when the necessary and sufficient condition for the existence of viable solutions

$$F(x) \cap T_K(x) \neq \emptyset \text{ for all } x \in K$$

is not satisfied, the natural way to solve the above problem is to introduce the projected problem (6).

We note that $\Pi_{\overline{co}T_K(x)}F(x) = F(x)$ whenever $x \in \text{int}K$; F is modified only on the boundary of K, so (6) is a problem of reflection at the boundary. Moreover, the application $x \to \Pi_{\overline{co}T_K(x)}F(x)$ has no easy continuity properties, but, thanks to the properties of the projection on a convex cone, it is possible to prove that the solutions to (5) and (6) coincide. We do not make any assumption on the regularity of the set K, improving already known results of [2] where the set K is sleek.

It is easier to find sufficient conditions for the set K in order to obtain continuity properties of the right-hand side of (5). So, for the study of existence and uniqueness of solutions we consider (5). We have the following proposition.

PROPOSITION 5. (i) *Suppose that K is closed and F is a set valued map. Then the sets of absolutely continuous solutions to* (5) *and* (6) *are equal.*

[3]Recall that a solution $x(\cdot)$ to (8) is called viable in K if $x(t) \in K$ for all $t \geq 0$. The set K is a viability domain for (8) if for all $x_0 \in K$ there exists a solution to (8) which is viable in K.

Moreover if F is a Marchaud map[4] *and K is bounded and sleek, then*

(ii) *for every $(t_0, x_0) \in [0, \infty) \times K$ there exists a solution of* (5) *or equivalently of* (6).

(iii) *the restriction of the map $(t_0, x_0) \in [0, T] \times K \to S_F(t_0, x_0)$ to a compact set C is compact into $[0, \infty) \times K \times W^{1,1}(0, \infty\ ; K)e^{-bt}$ for all b with $b > a$. Here $S_F(t_0, x_0)$ denotes the set of solutions to* (5) *starting from (t_0, x_0).*

Before giving the proof, we note that the first part of the above proposition is a generalization of Theorem 10.1.1 in [2] where the set K is supposed to be sleek; here K is only bounded. The second part recalls well-known existence and compactness results (see [1] and [2]).

Proof. (i) Using Proposition 0.6.4 from [1] we deduce that

$$\Pi_{\overline{co}T_K(x)} F(x) \subset F(x) - N_K(x) \text{ for all } x \in K,$$

and, consequently, a solution to (6) is also a solution to (5).

Conversely, if $x(t) \in K$ for all $t \geq t_0$, we have

$$\lim_{h \to 0} \frac{x(t+h) - x(t)}{h} \in T_K(x(t)) \text{ and } \lim_{h \to 0} \frac{x(t) - x(t-h)}{h} \in -T_K(x(t)) \text{ for a.e. } t \geq t_0,$$

so $x'(t) \in T_K(x(t)) \cap -T_K(x(t)) \subset N_K(x(t))^{\perp}$ a.e. $t \geq t_0$.

Let $t \geq t_0$ be a derivability point of $x(\cdot)$, and let $x'(t) = f(t) - p(t)$ with $f(t) \in F(x(t))$ and $p(t) \in N_K(x(t))$.

The above arguments say that $\langle x'(t) - f(t), x'(t) \rangle = 0$.

Thus, $x'(t) \in \Pi_{\overline{co}T_K(x(t))} f(t) \subset \Pi_{\overline{co}T_K(x(t))} F(x(t))$ for a.e. $t \geq t_0$.

(ii) We prove now the existence of a solution to the differential inequality (5). If K is sleek, then the map $x \to N_K(x)$ has a closed graph.

For all $x \in K$, we set $H(x) = F(x) - a(1 + ||x||)B \cap N_K(x)$, where B denotes the unit ball of $\mathbb{R}^N$.

Because the map $x \to a(1 + ||x||)B \cap N_K(x)$ is Marchaud, H is also Marchaud. Hence by Theorem 2.1.3 in [1] the existence of solutions of (5) follows.

Let us prove that the closed subset K is a viability domain for the differential inclusion (5).

Indeed, using the equality $I - \Pi_{N_K(x)} = \Pi_{T_K(x)}$, we have that for any $x \in K$ and $f \in F(x)$, $f - \Pi_{N_K(x)} f \in (F(x) - N_K(x)) \cap T_K(x)$.

Using the estimation $||\Pi_{N_K(x)} f|| \leq ||f|| \leq a(1 + ||x||)$, we get that $\Pi_{N_K(x)} f \in a(1 + ||x||)B \cap N_K(x)$ and consequently $f - \Pi_{N_K(x)} f \in H(x) \cap T_K(x)$.

The above arguments say that H satisfies the hypotheses of viability theorem 4.2.1 of [1], and since $H(x) \subset F(x) - N_K(x)$, the second part ensues.

(iii) See Theorem 2.2.1 in [1]. □

Now, let us begin a short study of the optimal control problem with reflected trajectories. From now on, we consider that the set valued map F is given by the equality

$$F(x) = f(x, U) = \{f(x, u), u \in U\} \text{ for all } x \in K.$$

[4] A set valued map F from $\mathbb{R}^N$ onto $\mathbb{R}^N$ is called Marchaud map if F is u.s.c. with nonempty compact convex values and has a linear growth.

We denote by $S_f(t_0, x_0)$ the set of absolutely continuous solutions to

$$\begin{cases} \text{(i) } x'(t) \in f(x(t), u(t)) - N_K(x(t)) \text{ for almost all } t \geq t_0, \\ \text{(ii) } x(t) \in K \text{ for all } t \geq t_0,\ x(t_0) = x_0 \in K \text{ for all } u(\cdot) \in \mathcal{U}(t_0), \end{cases} \tag{9}$$

and by $S_F(t_0, x_0)$ the set of absolutely continuous solutions to

$$\begin{cases} \text{(i) } x'(t) \in F(x(t)) - N_K(x(t)) \text{ for almost all } t \geq t_0, \\ \text{(ii) } x(t) \in K \text{ for all } t \geq t_0,\ x(t_0) = x_0 \in K. \end{cases} \tag{10}$$

We will prove now that (9) and (10) are equivalent. In this paper, we will use one of these systems to simplify our proofs.

PROPOSITION 6. *Suppose that K is a compact sleek set and (H_f) holds.*

(i) *If $x(\cdot)$ is a solution to* (10) *starting from $(t_0, x_0) \in [0, T] \times K$, then there exists $u(\cdot) \in \mathcal{U}(t_0)$ such that $x(\cdot)$ is equal to $x(\cdot; t_0, x_0, u(\cdot))$, the solution of* (9).

(ii) *As a direct consequence of* (i),

$$S_F(t_0, x_0) = S_f(t_0, x_0) \text{ for all } (t_0, x_0) \in [0, \infty) \times K.$$

Proof. We essentially use the fact that K is sleek (which implies that the application $x \to N_K(x)$ has a closed graph) and Theorem 1.14.1 from [1]. Consider $\Phi(t) := \{v \in U |\ x'(t) \in f(x(t), v) - N_K(x(t))\}$ for a.e. $t \geq t_0$. We can prove that the multivalued function Φ has a measurable selection which gives our measurable control $u(\cdot) \in \mathcal{U}(t_0)$. □

Moreover, with an easy computation, using the fact that K is proximal retract and Gronwall's inequality, we obtain the following estimation.

LEMMA 7. *Assume that (H_f) holds true and K is a bounded proximal retract. Then for $x_0(\cdot) \in S_f(t_0, x_0)$, $x_1(\cdot) \in S_f(t_1, x_1)$ with fixed $u(\cdot) \in \mathcal{U}(t_0)$ and for $t \geq t_1 \geq t_0$, there exists $C > 0$, a constant depending on t, such that*

$$\|x_0(t; t_0, x_0, u(\cdot)) - x_1(t; t_1, x_1, u(\cdot))\| \leq C(\|x_0 - x_1\| + |t_0 - t_1|).$$

We omit the proof of Lemma 7 because it is an easy adaptation of Lemma 4.4, p. 143, proved in [6]. As a direct consequence of the above estimation we obtain the following.

COROLLARY 8. *Assume that (H_f) holds true and K is a bounded proximal retract. Then for fixed $u(\cdot) \in \mathcal{U}(t_0)$ there exists an unique solution of* (2).

2.4. The optimal control problem. First, we give some standard results concerning the regularity of V without proof. Later we shall prove the existence and the uniqueness of viscosity solutions of (HJBI).

LEMMA 9. *Suppose that (H_f) holds true and K is a compact proximal retract. Then we have the following:*

(i) *(Existence of an optimal control.) If g is l.s.c., then V is l.s.c. and there exists an optimal trajectory starting from each point $(t_0, x_0) \in [0, T] \times K$, i.e., there exists $\bar{x}(\cdot) \in S_F(t_0, x_0)$ such that*

$$V(t_0, x_0) = g(\bar{x}(T; t_0, x_0, \bar{u}(\cdot))) \text{ for all } (t_0, x_0) \in [0, T] \times K.$$

(ii) *If g is a Lipschitz function, then V is locally Lipschitz and bounded.*

Next we give the Bellman dynamic programming.

PROPOSITION 10 (dynamic programming principle). *Let $g : K \to \mathbb{R}$ be a bounded function, K a compact proximal retract, and suppose that (H_f) holds. Then, for all $(t_0, x_0) \in [0, T] \times K$ we have*

$$V(t_0, x_0) = \inf_{x \in S_F(t_0, x_0)} V(t_0 + h, x(t_0 + h)) \text{ with } h > 0 \text{ small enough.} \tag{11}$$

3. The Hamilton–Jacobi partial differential variational inequality.

3.1. Existence result. The aim of this section is to provide an existence and a comparison result for viscosity solutions to a partial differential inequality with a kind of boundary conditions for nonsmooth sets, which generalizes first order Hamilton–Jacobi equations with Neumann conditions for smooth sets.

Using the dynamic programming principle we prove that the value function for the control problem (1), (2) is the viscosity solution of (HJBI) in the sense of Definition 3.

PROPOSITION 11. *If K is a compact proximal retract, g a Lipschitz function, and (H_f) holds true, then V is a locally Lipschitz viscosity solution of* (HJBI) *with the final condition $V(T,x) = g(x)$ for all $x \in K$.*

This theorem can be considered as an existence result of solutions to (HJBI).

Proof. First we prove that V is a supersolution.

We consider $(t_0, x_0) \in \arg\min(V - \psi), \psi \in C^1$, with

$$V(t_0, x_0) = \psi(t_0, x_0) \text{ and } V(t, x) \geq \psi(t, x)$$

in a neighborhood of (t_0, x_0).

For all $h > 0$ small enough, there exists $x_h(\cdot) \in S_F(t_0, x_0)$ such that

$$\psi(t_0, x_0) + h^2 = V(t_0, x_0) + h^2 \geq V(t_0 + h, x_h(t_0 + h)) \geq \psi(t_0 + h, x_h(t_0 + h)).$$

For a subset A of $\mathbb{R}^N$ we denote by $B(A, \varepsilon) = \{x \in \mathbb{R}^N | \inf_{y \in A} ||y - x|| \leq \varepsilon\}$. $B(A, \varepsilon)$ denotes the neighborhood of the set A with a radius equal to $\varepsilon > 0$.

Let M be a bound of F on K. Using the Lipschitz property of $F(\cdot)$ and the upper semicontinuity of $N_K(\cdot) \cap B(0, M)$, we have that, for all $\varepsilon > 0$, there exists an $h > 0$ small enough such that the following inclusions hold:

$$\begin{aligned}
\frac{1}{h}(x_h(t_0 + h) - x_0) &\in \frac{1}{h}\int_{t_0}^{t_0+h} (F(x_h(s)) - N_K(x_h(s)) \cap B(0, M))ds \\
&\subset \frac{1}{h}\int_{t_0}^{t_0+h} (F(x_0))ds + B(0,1)\frac{1}{h}\int_{t_0}^{t_0+h} L||x_h(s) - x_0||ds \\
&\quad - \frac{1}{h}\int_{t_0}^{t_0+h} B(N_K(x_0) \cap B(0, M), \varepsilon)ds \\
&= F(x_0) + B(0,1)\frac{1}{h}\int_{t_0}^{t_0+h} L||x_h(s) - x_0||ds - B(N_K(x_0) \cap B(0, M), \varepsilon).
\end{aligned}$$

Hence for all $\varepsilon > 0$, there exists a sequence h_n such that $\lim_{n\to\infty} h_n = 0$ and

$$\lim_n \frac{1}{h_n}(x_{h_n}(t_0 + h_n) - x_0) \in F(x_0) - B(N_K(x_0) \cap B(0, M), \varepsilon).$$

Letting $\varepsilon \to 0$ we obtain that

$$\lim_n \frac{1}{h_n}(x_{h_n}(t_0 + h_n) - x_0) \in F(x_0) - N_K(x_0) \cap B(0, M). \tag{12}$$

Moreover,

$$\begin{aligned}
&\lim_n \left(\frac{1}{h_n}[\psi(t_0 + h_n, x_{h_n}(t_0 + h_n; t_0, x_0, u(\cdot))) - \psi(t_0, x_0)] - h_n\right) \\
&= \lim_n \left(\frac{1}{h_n}\left[\psi\left(t_0 + h_n, x_0 + h_n\left(\frac{1}{h_n}(x_{h_n}(t_0 + h_n) - x_0)\right) - \psi(t_0, x_0)\right] - h_n\right).
\end{aligned} \tag{13}$$

Using (12) and (13) we have the following.

First case ($x_0 \in \text{int}K$). Then $N_K(x_0) = \{0\}$ and there exists $u \in U$ such that

$$\frac{\partial \psi}{\partial t}(t_0, x_0) + \left\langle \frac{\partial \psi}{\partial x}(t_0, x_0),\ f(x_0, u) \right\rangle \leq 0,$$

and consequently

$$\frac{\partial \psi}{\partial t}(t_0, x_0) + \inf_{u \in U} \left\langle \frac{\partial \psi}{\partial x}(t_0, x_0), f(x_0, u) \right\rangle \leq 0.$$

Second case ($x_0 \in \partial K$). Then $\{0\} \subset N_K(x_0)$ and there exist $u \in U$, $y_u \in N_K(x_0)$ such that:

$$\frac{\partial \psi}{\partial t}(t_0, x_0) + \left\langle \frac{\partial \psi}{\partial x}(t_0, x_0),\ f(x_0, u) - y_u \right\rangle \leq 0.$$

So, there exists $w_0 = y_u \in N_K(x_0) \cap B(0, M)$ such that

$$\frac{\partial \psi}{\partial t}(t_0, x_0) + \inf_{u \in U} \left\langle \frac{\partial \psi}{\partial x}(t_0, x_0),\ f(x_0, u) \right\rangle - \left\langle w_0,\ \frac{\partial \psi}{\partial x}(t_0, x_0) \right\rangle \leq 0$$

and V is a supersolution.

The proof of the fact that V is subsolution is similar and we omit it. □

3.2. Uniqueness result. This section concerns the uniqueness of the viscosity solutions of (HJBI). The importance of this result leads us to treat it separately. Moreover, the characterization of the value function as the unique solution of (HJBI) ensues.

THEOREM 12 (uniqueness result in the Lipschitz case). *Assume that* (H_f) *holds true. Let K be a compact proximal retract and g be a Lipschitz function. Then there exists at most one uniformly continuous viscosity solution of* (HJBI) *which satisfies the final condition $V(T, x) = g(x)$ for all $x \in K$.*

The proof can be adapted from Evans [7]. We only underline that the difference to Evans' proof is due to the monotonicity of the multivalued function $x \to N_K(x) \cap B(0, M) + cx$.

4. The discontinuous case. In this section we investigate the value function V when $g : K \to \mathbb{R}$ is supposed to be bounded. In this case the value is only a bounded function. A natural question is how to use the viscosity theory to describe V. Here we establish a relation between the value and the viscosity sub or supersolutions of (HJBI). This kind of problem has been studied for the Bolza problem in [15], [16].

The main point of this section is to prove the following.

THEOREM 13. *Suppose that K is a proximal retract and* (H_f) *holds.*

(i) *If g is bounded, then for every $(t, x) \in [0, T] \times K$*

$$V(t, x) = \inf\{\psi(t, x) |\ \psi \text{ l.s.c. supersolution of (HJBI)};\ \psi(T, \cdot) \geq g(\cdot)\} \text{ and}$$
$$V(t, x) = \sup\{\varphi(t, x) |\ \varphi \text{ u.s.c. subsolution of (HJBI)};\ \varphi(T, \cdot) \leq g(\cdot)\}.$$

(ii) *If g is l.s.c., then*

$$V = \min\{\psi \mid \psi \text{ l.s.c. supersolution of (HJBI)};\ \psi(T, \cdot) \geq g(\cdot)\}.$$

(iii) *If g is u.s.c., then*

$$V = \max\{\varphi \mid \varphi \text{ u.s.c. subsolution of (HJBI) } ;\ \varphi(T,\cdot) \le g(\cdot)\}.$$

Before giving the proof, we note that the above theorem allows us to get, in particular, a stronger uniqueness result. More precisely, if ψ is an l.s.c. supersolution and φ is a u.s.c. subsolution of (HJBI) satisfying $\psi(T,\cdot) \ge \varphi(T,\cdot)$ on K, then $\psi \ge \varphi$ on $[0,T] \times K$.

Proof. (i) Let ψ be an l.s.c. supersolution of (HJBI) with $\psi(T,\cdot) \ge g(\cdot)$. We want to prove that $V \le \psi$ on $[0,T] \times K$. To do this we use the following lemma proved in the appendix.

LEMMA 14. *Assume that (H_f) holds true, K is a compact proximal retract, and $\psi : (0,T) \times K \to \mathbb{R}$ is an l.s.c. viscosity supersolution of* (HJBI). *Then for every $(t_0,x_0) \in (0,T) \times K$ there exists a solution $x(\cdot;t_0,x_0,u(\cdot))$ of* (2) *such that*

$$\psi(t,x(t)) \le \psi(t_0,x_0) \text{ for all } t \in [t_0,T]. \tag{14}$$

So we obtain that there exists an $x(\cdot) \in S_F(t_0,x_0)$ satisfying (14). Hence we have $V(t_0,x_0) \le g(x(T)) \le \psi(T,x(T)) \le \psi(t_0,x_0)$.

Using the very definition of the value function, for all $\varepsilon > 0$ there exists $u_\varepsilon(\cdot) \in \mathcal{U}(t_0)$ such that $g\left(x\left(T;t_0,x_0,u_\varepsilon(\cdot)\right)\right) < V(t_0,x_0) + \varepsilon$.

For $M_1 > \sup_{x\in K} g(x)$ we define $l_\epsilon : \mathbb{R}^N \to \mathbb{R}$ by the following formula:

$$l_\epsilon(x) = \begin{cases} g\left(x\left(T;t_0,x_0,u_\varepsilon(\cdot)\right)\right) & \text{if } x = x\left(T;t_0,x_0,u_\varepsilon(\cdot)\right), \\ M_1, & \text{if } x \ne x\left(T;t_0,x_0,u_\varepsilon(\cdot)\right). \end{cases}$$

Obviously l_ε is l.s.c. so V_{l_ε}, the value function of the control problem with g replaced by l_ε, is a l.s.c. supersolution of (HJBI) and $V_{l_\epsilon}(T,\cdot) = l_\epsilon(\cdot) \ge g(\cdot)$.

We also have $V_{l_\epsilon}(t_0,x_0) = g\left(x\left(T;t_0,x_0,u_\varepsilon(\cdot)\right)\right) \le V(t_0,x_0) + \varepsilon$. By the definition of the infimum we obtain

$$V(t_0,x_0) = \inf\{\psi(t_0,x_0) \mid \psi \text{ l.s.c. supersolution of (HJBI); } \psi(T,\cdot) \ge g(\cdot)\}.$$

Now let us prove the second relation. Let $(t_0,x_0) \in (0,T) \times K$. We denote by $A(t_0,x_0) := \{x(T) |\ x(\cdot) \in S_F(t_0,x_0)\}$. By Proposition 5 $A(t_0,x_0)$ is a compact set. We define $h : \mathbb{R}^N \to \mathbb{R}$ by

$$h(y) = \begin{cases} V(t_0,x_0) \text{ if } y \in A(t_0,x_0), \\ m \text{ if } y \in \mathbb{R}^N \setminus A(t_0,x_0), \end{cases}$$

where $m = \inf_{x\in K} g(x)$. So, h is u.s.c. because $A(t_0,x_0)$ is closed.

Obviously we have that $V_h(t_0,x_0) = V(t_0,x_0)$ and $V_h(T,\cdot) \le h(\cdot) \le g(\cdot)$.

Moreover, V_h is (see Lemma 21 in the appendix) a u.s.c. subsolution for (HJBI).

Now, to complete the proof of (i) we use the definition of the supremum and the following lemma proved in the appendix.

LEMMA 15. *Assume that (H_f) holds true, K is a compact proximal retract, and $\varphi : (0,T) \times K \to \mathbb{R}$ is a u.s.c. viscosity subsolution of* (HJBI) *such that $\varphi(T,x) \le g(x)$ for all $x \in K$. Then $V(t,x) \ge \varphi(t,x)$ for every $(t,x) \in (0,T) \times K$.*

The proofs of (ii) and (iii) are direct consequences of Lemma 14, Lemma 15, and Lemma 21. ∎

5. On l.s.c. solutions of HJBI with reflection on smooth sets. If $g : \mathbb{R}^N \to \mathbb{R}$ is an l.s.c. function, then V is also l.s.c. In [5], [8] a modification of the concept of viscosity solutions for semicontinuous functions was proposed. This approach is based on a construction of "touching from one side" functions, which is usual for viscosity solutions theory.

We suppose that K is a $\mathrm{C}^{1,1}$ submanifold with boundary. If we denote by $n(x)$ the unit outward normal to K at $x \in \partial K$, the normal cone $N_K(x)$ is generated by $n(x)$, i.e., $N_K(x) = [0,\infty)n(x)$ for $x \in \partial K$ and $N_K(x) = \{0\}$ for $x \in \mathrm{int}K$.

We propose a definition for l.s.c. solutions to the HJBI of Barron–Jensen–Frankowska type.

DEFINITION 16. *A viscosity l.s.c. solution of* (HJBI) *is a function* $\psi : [0,T]\times K \to \mathbb{R}$ *such that*

$$\text{for any } \phi \in C^1 \text{ and } (t_0,x_0) \in \arg\min(\psi-\phi),$$

$$\text{if } (t_0,x_0) \in [0,T)\times \mathrm{int}K, \text{ we have } \frac{\partial\phi}{\partial t}(t_0,x_0) + H\left(x_0, \frac{\partial\phi}{\partial x}(t_0,x_0)\right) \le 0;$$

$$\text{if } (t_0,x_0) \in (0,T]\times \mathrm{int}K, \text{ we have } \frac{\partial\phi}{\partial t}(t_0,x_0) + H\left(x_0, \frac{\partial\phi}{\partial x}(t_0,x_0)\right) \ge 0;$$

$$\text{if } (t_0,x_0) \in [0,T)\times \partial K, \text{ then there exists } u \in U \text{ such that}$$

$$\frac{\partial\phi}{\partial t}(t_0,x_0) + \left\langle (f(x_0,u) - \Pi_{N_K(x)} f(x_0,u)), \frac{\partial\phi}{\partial x}(t_0,x_0)\right\rangle \le 0;$$

$$\text{if } (t_0,x_0) \in (0,T]\times \partial K \text{ and } \min_{u\in U}\langle f(x_0,u), n(x_0)\rangle > 0, \text{ then for all } u \in U,$$

$$\frac{\partial\phi}{\partial t}(t_0,x_0) + \left\langle (f(x_0,u) - \Pi_{N_K(x)} f(x_0,u)), \frac{\partial\phi}{\partial x}(t_0,x_0)\right\rangle \ge 0.$$

Note that in $\mathrm{int}K$ the equation is satisfied in the Barron–Jensen–Frankowska sense (see [5], [8]).

We obtain the following uniqueness results.

PROPOSITION 17. *Suppose that K is a $C^{1,1}$ submanifold with boundary and for any $u \in U$ and for all $x_0 \in K$ we have $\langle f(x_0,u), n(x_0)\rangle < 0$. If g is l.s.c. and (H_f) holds true, then the value function V is the unique l.s.c. viscosity solution of* (HJBI) *which verifies the final condition $V(T,x) = g(x)$ for all $x \in K$, and for all $(t,x) \in (0,T]\times\partial K$ we have*

$$\lim \inf_{\substack{(t',x')\to(t,x)\\ x\in intK}} V(t',x') = V(t,x).$$

The proof is similar to the proof of Theorem 2.3 in [8].

PROPOSITION 18. *Suppose that K is a $C^{1,1}$ submanifold with boundary and for any $u \in U$ and for all $x_0 \in K$ we have $\langle f(x_0,u), n(x_0)\rangle > 0$. If g is l.s.c. and (H_f) holds true, then the value function V is the unique l.s.c. viscosity solution of* (HJBI) *which verifies the final condition $V(T,x) = g(x)$ for all $x \in K$, and for all $(t,x) \in (0,T]\times\partial K$ we have*

$$\lim \inf_{\substack{(t',x')\to(t,x)\\ x\in intK}} V(t',x') = V(t,x).$$

Proof. Step 1. V satisfies Definition 16. The proof of the first inequality is similar to the proof of the fact that V is an l.s.c. supersolution of (HJBI).

For proving the second inequality we observe that for all $x_0 \in \partial K$ and $u \in U$,

$$\Phi_u(x_0) := f(x_0,u) - \Pi_{N_K(x)} f(x_0,u) \in \Pi_{T_K(x)} f(x_0,u) = \Pi_{\partial T_K(x)} f(x_0,u).$$

Consequently, $-f(x_0,u) + \Pi_{N_K(x)} f(x_0,u) \in -\Pi_{\partial T_K(x)} f(x_0,u)$.

As K is a $C^{1,1}$ submanifold and for all $x_0 \in K$, $\min_{u\in U}\langle f(x_0,u), n(x_0)\rangle > 0$, $\Phi_u(\cdot)$ is a Lipschitz application on ∂K. Moreover, ∂K is locally invariant (see [2, viability theorem 3.2.4] by $\Phi_u(\cdot)$ and by $-\Phi_u(\cdot)$ (because $\partial T_K(x) = -\partial T_K(x)$).

Now let $(t_0,x_0) \in \arg\min (V-\phi)$, $\phi \in C^1$. We have two cases.

First case ($x_0 \in \partial K$). For a fixed constant control $u \in U$, we consider the solution of

$$\begin{cases} x'(t) = -f(x(t),u) + \Pi_{N_K(x(t))} f(x(t),u), \\ x(t_0) = x_0, \end{cases}$$

which stays in ∂K because of the invariance properties of $\Phi_u(\cdot)$. Using the dynamic programming principle we get $V(t_0,x_0) \geq V(t_0-h, x(t_0-h))$ with $h>0$ small enough. So, $\phi(t_0,x_0) \geq \phi(t_0-h,x(t_0-h))$ with $h>0$ small enough. Recall that $\phi \in C^1$ and consequently

$$\frac{\partial \phi}{\partial t}(t_0,x_0) + \left\langle (f(x_0,u) - \Pi_{N_K(x)} f(x_0,u)), \frac{\partial \phi}{\partial x}(t_0,x_0) \right\rangle \geq 0.$$

Second case ($x_0 \in \mathrm{int}K$). $N_K(x_0) = \{0\}$ and for all $u \in U$, because f is a Lipschitz application, there exists $B(x_0;r_u)$, $r_u > 0$, such that the solution to

$$\begin{cases} x'(t) = -f(x(t),u), \\ x(t_0) = x_0 \end{cases}$$

stays in $B(x_0;r_u)$. Using the dynamic programming principle, for $h>0$ small enough $V(t_0,x_0) \geq V(t_0-h,x(t_0-h))$ so $\phi(t_0,x_0) \geq \phi(t_0-h,x(t_0-h))$. Because $\phi \in C^1$ we obtain $\frac{\partial \phi}{\partial t}(t_0,x_0) + \langle (f(x_0,u), \frac{\partial \phi}{\partial x}(t_0,x_0)\rangle \geq 0$. This allows us to say that V is an l.s.c. solution of (HJBI).

Step 2 (uniqueness). Now let us prove that V is the unique l.s.c. solution of (HJBI). Let W be an l.s.c. solution of (HJBI) with $W(T,x) = g(x)$ for all $x \in K$. We have already proved (see Theorem 13) that $W \geq V$.

For the reverse inequality we consider $(t_0,x_0) \in (0,T) \times K$ and $x(\cdot) \in S_F(t_0,x_0)$. There exists $u(\cdot) \in \mathcal{U}(t_0)$ such that $x(\cdot) = x(\cdot;t_0,x_0,u(\cdot))$. We have two cases.

First case ($x(T) \in \mathrm{int}K$). For a fixed $u(\cdot) \in \mathcal{U}(t_0)$, ∂K is invariant by $\Phi_{u(.)}(\cdot)$ and $\Phi_{u(.)}(\cdot)$ is Lipschitz in the second variable, so we have that $x([t_0,T]) \subset \mathrm{int}K$.

By the measurable viability theorem (see Theorem 4.7 in [10], [2]) $Epi(W)$ is viable for the dynamics given by $(t,x,y) \to (-1, -f(x(t),u(t)), 0)$. For the solution starting from $(T,x(T),W(T,x(T)))$, we have for all $t \in [t_0,T]$, $W(T-t,x(T-t)) \leq W(T,x(T))$, so $W(t_0,x_0) \leq W(T,x(T)) = g(x(T))$.

Second case ($x(T) \in \partial K$). Denote by τ the first time with the property $x(\tau) \in \partial K$. Using invariance properties of $\Phi_{u(.)}(\cdot)$ and because $\Phi_{u(.)}(\cdot)$ is Lipschitz in the second variable, we obtain that $x([t_0,\tau)) \subset \mathrm{int}K$ and $x([\tau,T]) \subset \partial K$.

As in the above case, we apply the measurable viability theorem (see Theorem 4.7 in [10], [2]) to $Epi(W)$, on the one hand, to $[\tau,T]$ for the dynamics given by $(t,x,y) \to (-1, -f(x(t),u(t)) + \Pi_{N_K(x(t))} f(x(t),u(t)), 0)$ for the solution starting from $(T,x(T),W(T,x(T)))$ and, on the other hand, to $[t_0,\tau_n)$ for the dynamics given by

$(t,x,y) \to (-1, -f(x(t),u(t)), 0)$ for the solution starting from $(\tau, x(\tau), W(\tau, x(\tau)) = \lim_n (\tau_n, x_n, W(\tau, x(\tau)), x_n \in \text{int}K$.

We have $W(t_0,x_0) \le W(\tau, x(\tau))$ and $W(\tau, x(\tau)) \le W(T, x(T)) = g(x(T))$. Consequently, by definition of the value function $W(t_0,x_0) \le V(t_0,x_0)$. □

We note that here we can obtain uniqueness for l.s.c. solutions only in two (extremal) cases, where the vector field $f(x,u)$ is pointing only outside of the domain or only inside. For the intermediate situation it seems that we cannot obtain uniqueness (see the counterexample given below). The lack of uniqueness can be a consequence of the fact that in the intermediate situation we lose the Lipschitz regularity of $\Phi_u(\cdot)$ in ∂K and the idea of the above proof will fail.

Counterexample. Now we will show that a uniqueness result is not possible using our definition without imposing boundary properties on our dynamics as we did in the above propositions. We do this by giving a counterexample.

Let $K = [0,1] \subset \mathbb{R}$. For a dynamics given by $f = 0$ and $g = 1$ the value function is $V(t,x) = 1$ for all $(t,x) \in [0,1] \times [0,1]$. Moreover V is an l.s.c. solution of HJBI in the sense of our definition. Define

$$u(t,x) = \begin{cases} 1 \text{ if } (t,x) \in [0,1] \times (0,1], \\ 0 \text{ if (t,x)} \in [0,1] \times \{0\}. \end{cases}$$

It is easy to verify that u is also an l.s.c. solution for the HJBI and we do not have uniqueness because $\langle f(x_0), n(x_0) \rangle = 0$ for all $x_0 \in \partial K$.

For another definition of the discontinuous solution Ley [12] obtained a counterexample proving that there is no uniqueness to HJB with a notion of the solution in the Ishii–Barles–Perthame sense.

6. Appendix. Let us give the proof of Lemma 14 and Lemma 15. We shall use the following classical viability theorem and the fact that the definition of super and subsolutions to (HJBI) can be written equivalently in terms of subdifferentials. (See [15] to get formulations of viscosity solutions in terms of subdifferentials of the PDE associated to the Mayer control problem with $K = R^N$.)

THEOREM 19 (see [2, viability theorem 3.2.4]). *Assume that G is a Marchaud map and let $D \subset \mathbb{R}^N$ be closed. If for every $z \in D$ we have*

$$\text{for all } p \in N_D(z), \quad \min_{y \in G(z)} \langle y, p \rangle \le 0, \tag{15}$$

then for every $x_0 \in D$, $t_0 < T$, there exists a solution $x(\cdot)$ to the Cauchy problem $x'(s) \in G(x(s))$, $x(t_0) = x_0$ such that $x(t) \in D$ for all $t \in [t_0, T]$.

Now we give the proof of Lemma 14 and Lemma 15.

Proof of Lemma 14. Fix $t_0 \in (0,T)$. We set

$$D_\psi = \text{cl}(\{(t,x,r) \ : \ t \in (0,T], \ x \in K, \ r \ge \psi(t,x)\}) \cup [T, \infty) \times K \times \mathbb{R},$$

$$\tilde{F}(t,x,r) = \begin{cases} 0 \text{ if } t < 0, \\ \frac{t}{t_0}(1, F(x) - N_K(x) \cap B(0,M), 0) \text{ if } t \in [0, t_0], \\ (1, F(x) - N_K(x) \cap B(0,M), 0) \text{ if } t \in [t_0, T], \\ (1, F(x) - N_K(x) \cap B(0,M), 0) \text{ if } t > T, \end{cases}$$

where cl denote the closure and M is a bound of F on K. We show that (15) holds true for $\tilde{F}$ and D_ψ.

First case ($x_0 \in \text{int}K$). Let $z_0 = (s_0, x_0, r_0 := \psi(t_0,x_0)) \in D_\psi$. If $s_0 = 0$, then $\tilde{F} = 0$. Obviously (15) holds.

If $s_0 \geq T$ and $(p_s, p_x, p_r) \in N_{D_\psi}(s_0, x_0, r_0)$, then $p_s \leq 0$, $p_x = 0$, $p_r = 0$. Hence (15) holds.

It remains to consider the case $s_0 \in (0, T)$. We have $N_{D_\psi}(s_0, x_0, r_0) \subset N_{D_\psi}(s_0, x_0, \psi(s_0, x_0))$. Let $(p_s, p_x, p_r) \in N_{D_\psi}(s_0, x_0, \psi(s_0, x_0))$.

If $p_r < 0$, then $(p_s/-p_r,\, p_x/-p_r) \in \partial_- \psi(s_0, x_0)$ (see Proposition 4.1 in [9]).

Since ψ is a supersolution of (HJBI) there exists $y_0 \in N_K(x_0)$ such that

$$\frac{p_s}{-p_r} + \min_{z \in F(x_0)} \left\langle z,\ \frac{p_x}{-p_r} \right\rangle - \left\langle y_0,\ \frac{p_x}{-p_r} \right\rangle \leq 0 \quad \text{and}$$
$$\frac{p_s}{-p_r} + \min_{z \in \{F(x_0) - N_K(x_0) \cap B(0,M)\}} \left\langle z,\ \frac{p_x}{-p_r} \right\rangle \leq 0.$$

Hence $\min_{\tilde{y} \in \tilde{F}(s_0, x_0, r_0)} \langle \tilde{y}, (p_s, p_x, p_r) \rangle \leq 0$.

Now we consider the case $p_r = 0$. By a Rockafellar's lemma (see, for instance, Lemma 4.2 in [9]) there exists a sequence $s_n \to s_0$, $x_n \to x_0$, and $p_{s_n} \to p_s$, $p_{x_n} \to p_x$, $p_{r_n} \to 0$, $p_{r_n} < 0$ such that $(p_{s_n}, p_{x_n}, p_{r_n}) \in N_{D_\psi}(p_{s_n}, p_{x_n}, p_{r_n})$. Since $p_{r_n} < 0$ we obtain from the previous case that

$$\min_{\tilde{y}_n \in \tilde{F}(s_n, x_n, r_n)} \langle \tilde{y}, (p_{s_n}, p_{x_n}, p_{r_n}) \rangle \leq 0.$$

We get $\min_{\tilde{y} \in \tilde{F}(s_0, x_0, r_0)} \langle \tilde{y}, (p_s, p_x, p_r) \rangle \leq 0$, because $\tilde{F}$ is Marchaud.

Second case ($x_0 \in \partial K$). Let $z_0 = (s_0, x_0, r_0 := \psi(t_0, x_0)) \in D_\psi$. If $s_0 = 0$, then $\tilde{F} = 0$. Obviously (15) holds true.

If $s_0 \geq T$ and $(p_s, p_x, p_r) \in N_{D_\psi}(s_0, x_0, r_0)$, then $p_s \leq 0$, $p_x \in N_K(x_0)$, $p_r = 0$. Hence $[F(x_0) - N_K(x_0)] \cap T_K(x_0) \neq \emptyset$ and (15) holds.

It remains to consider $s_0 \in (0, T)$, which is similar to the first case.

Finally we obtain $\min_{\tilde{y} \in \tilde{F}(s_0, x_0, r_0)} \langle \tilde{y}, (p_s, p_x, p_r) \rangle \leq 0$.

In view of the above theorem we have a solution $z(\cdot)$ to the Cauchy problem $z'(s) \in \tilde{F}(z(s))$, $z(t_0) = z_0$. Let $z(s) = (t(s), x(s), r(s))$.

By the definition of $\tilde{F}$ we have $t(s) = s$, $r(s) = r_0 = \psi(t_0, x_0)$. Hence, (14) holds true and the proof is completed. □

Proof of Lemma 15. The proof is divided into several steps.

First step. We fix $t_0 \in (0, T)$, $u(\cdot) \in \mathcal{U}(t_0)$ such that $u(\cdot)$ is a continuous function. We set

$$D_\varphi = \mathrm{cl}(\{(t, x, r)\ :\ t \in (0, T],\ x \in K,\ r \leq \varphi(t, x)\}) \cup [T, \infty) \times K \times \mathbb{R},$$
$$G(t, x, r) = \begin{cases} 0 \text{ if } t < 0, \\ \frac{t}{t_0}(1, F(x, u(t)) - N_K(x) \cap B(0, M), 0) \text{ if } t \in [0, t_0], \\ (1, F(x, u(t)) - N_K(x) \cap B(0, M), 0) \text{ if } t \in [t_0, T], \\ (1, F(x, u(t)) - N_K(x) \cap B(0, M), 0) \text{ if } t > T, \end{cases}$$

and we want to prove that for every $(t_0, x_0, \varphi(t_0, x_0)) \in (0, T) \times K \times \mathbb{R}$ the solution $x(\cdot; t_0, x_0, u(\cdot))$ to (2) satisfies

$$\varphi(t, x(t)) \geq \varphi(t_0, x_0) \text{ for all } t \in [t_0, T]. \tag{16}$$

Since φ is a u.s.c. viscosity subsolution of (HJBI) we have

$$\text{for any } \phi \in C^1 \text{ and } (t_0, x_0) \in \arg\max(\varphi - \phi),$$
$$\text{there exists } z_0 \in N_K(x_0) \text{ such that}$$
$$\frac{\partial \phi}{\partial t}(t_0, x_0) + H\left(x_0, \frac{\partial \phi}{\partial x}(t_0, x_0)\right) - \left\langle z_0, \frac{\partial \phi}{\partial x}(t_0, x_0)\right\rangle \geq 0,$$
$$\text{so } -\frac{\partial \phi}{\partial t}(t_0, x_0) + \min_{y_0 \in N_K(x_0) \cap B(0,M)} \left\{ \left\langle (f(x_0, u(t_0)) - y_0), -\frac{\partial \phi}{\partial x}(t_0, x_0)\right\rangle \right\} \leq 0.$$

Using the same arguments as we used in the proof of the above lemma and because G is a Marchaud map, we obtain

$$\min_{y_0 \in N_K(x_0) \cap B(0,M)} \langle (f(x_0, u(t_0)) - y_0), (p_s, p_x, p_r)\rangle \leq 0.$$

So, for every $(p_t, p_x, p_r) \in N_{D_\varphi}(t_0, x_0, \varphi(t_0, x_0))$,

$$\min_{\tilde{y} \in G(t_0, x_0, \varphi(t_0, x_0))} \langle \tilde{y}, (p_s, p_x, p_r)\rangle \leq 0.$$

Using Theorem 19 we obtain that for every $(t_0, x_0, \varphi(t_0, x_0)) \in (0, T) \times K \times \mathbb{R}$ the solution $x(\cdot; t_0, x_0, u(\cdot))$ to (2) satisfies (16).

Second step. Fix $u(\cdot) \in \mathcal{U}(t_0)$. Then there exists a sequence $u_n(\cdot) \subset \mathcal{U}(t_0)$ of continuous functions such that $u_n(\cdot) \to u(\cdot)$ in $L^\infty(0, T; U)$.

For all $n \in \mathbb{N}$, $x_n(\cdot; t_0, x_0, u_n(\cdot))$ satisfies

$$x_n'(t) \in f(x_n(s), u_n(s))ds - \int_{t_0}^t N_K(x_n(s)) \cap B(0, M)ds, \text{ or equivalently}$$
$$x_n(t) \in x_0 + \int_{t_0}^t f(x_n(s), u_n(s))ds - \int_{t_0}^t N_K(x_n(s)) \cap B(0, M)ds. \tag{17}$$

As $B(0, M)$ is a compact set and the application $N_K(\cdot) \cap B(0, M)$ is u.s.c., there exists a measurable selection $y_n(\cdot) \in N_K(x_n(s)) \cap B(0, M)$. Moreover, for all $s \in [t_0, T]$ there exists

$$\lim_{n \to \infty} y_n(s) = y(s) \in N_K(x(s)) \cap B(0, M).$$

Hence, by the Lebesgue theorem we obtain that

$$\int_{t_0}^t y_n(s)ds \to \int_{t_0}^t y(s)ds \in \int_{t_0}^t N_K(x(s)) \cap B(0, M)ds \text{ a.e. } t \in [t_0, T].$$

Recall that the restriction of the application $(t_0, x_0) \in [0, T] \times K \to S_F(t_0, x_0)$ to a compact set C is compact into $[0, \infty) \times K \times W^{1,1}(0, \infty; K)e^{-bt}$ for all b with $b > a$. Since $x_n(\cdot; t_0, x_0, u_n(\cdot)) \in S_F(t_0, x_0)$ there exists $x(\cdot) \in S_F(t_0, x_0)$ such that $\lim_{n \to \infty} x_n(\cdot) = x(\cdot)$ in $W^{1,1}(0, T; K)$.

Passing to the limit in (17) we obtain that for almost all $t \in [t_0, T]$ $x(t) \in x_0 + \int_{t_0}^t f(x(s), u(s))ds - \int_{t_0}^t N_K(x(s)) \cap B(0, M))ds$ and consequently, $x(\cdot) = x(\cdot; t_0, x_0, u(\cdot))$.

Third step. Assume that $\bar{x}(\cdot; t_0, x_0, \bar{u}(\cdot))$ is an optimal trajectory for V, starting from $(t_0, x_0) \in [0, T] \times K$, i.e., $V(t_0, x_0) = g(\bar{x}(T; t_0, x_0, \bar{u}(\cdot))$.

Then there exists a sequence of continuous functions $u_n(\cdot)$, such that $u_n(\cdot) \to \bar{u}(\cdot)$ in $L^\infty(0, T; U)$ and consequently $x_n(\cdot; t_0, x_0, u_n(\cdot)) \to \bar{x}(\cdot; t_0, x_0, \bar{u}(\cdot))$ in $W^{1,1}(0, T; K)$.

Using the above arguments for every $n \in \mathbb{N}$, $\varphi(t, x_n(t; t_0, x_0, u_n(\cdot)) \geq \varphi(t_0, x_0)$ for all $t \in [t_0, T]$ and consequently, $\varphi(T, x_n(T; t_0, x_0, u_n(\cdot)) \geq \varphi(t_0, x_0)$.

As φ is u.s.c., we obtain by letting $n \to \infty$

$$\begin{aligned} V(t_0, x_0) = g(\bar{x}(T; t_0, x_0, \bar{u}(\cdot))) &\geq \varphi(T, (\bar{x}(T; t_0, x_0, \bar{u}(\cdot))) \\ &\geq \limsup_{n \to \infty} \varphi(T, x_n(T; t_0, x_0, u_n(\cdot)) \geq \varphi(t_0, x_0). \end{aligned}$$

Then $V \geq \varphi$ and the proof is complete. □

For the proof of Lemma 21 let us establish a stability result for (HJBI).

LEMMA 20. *Assume that $H : K \times \mathbb{R}^N \to \mathbb{R}$ is a continuous Hamiltonian and let K be a compact proximal retract. If $w_n : (0, T) \times K \to \mathbb{R}$ is an increasing (decreasing) sequence of uniformly locally bounded l.s.c. (u.s.c.) supersolutions (subsolutions) of* (HJBI) *and w is a pointwise limit of w_n, then w is an l.s.c. (u.s.c.) supersolution (subsolution) of* (HJBI).

The proof of this lemma is adapted from [3]. The main difficulties and changes with Barles's proof are given by the regularity of the application $N_K(\cdot)$.

LEMMA 21. *Assume that (H_f) holds. Let K be a compact proximal retract and g be a u.s.c. (respectively, l.s.c.) function. Then V is a u.s.c. subsolution (respectively, l.s.c. supersolution) of* (HJBI).

Proof. We define a sequence $g_n : K \to \mathbb{R}$ by

$$g_n(x) = \sup_{y \in K} (g(y) - n||x - y||).$$

The supconvolutions g_n are Lipschitz, $g_n(x) \geq g_{n+1}(x)$, and $\lim g_n(x) = g(x)$ for every $x \in K$. Using Proposition 11, V_{g_n} is a Lipschitz solution to (HJBI) with $V_{g_n}(T, \cdot) = g_n(\cdot)$ and $V_{g_n} \geq V$.

Denote $U(t, x) = \lim V_{g_n}(t, x)$. Then, using the above result, U is a u.s.c. subsolution of (HJBI) and $U(T, x) = g(x)$. Obviously $U \geq V$ and by Lemma 15 we have that $U = V$. So V is a u.s.c. subsolution for (HJBI).

The proof in the l.s.c. case is similar to the u.s.c. case. □

REFERENCES

[1] J.-P. AUBIN AND A. CELLINA (1984), *Differential Inclusions*, Springer–Verlag, Berlin.

[2] J.-P. AUBIN (1991), *Viability Theory*, Birkhäuser, Berlin.

[3] G. BARLES (1994), *Solutions de viscosité des équations de Hamilton-Jacobi*, Springer–Verlag, Paris.

[4] G. BARLES (1999), *Fully Neumann boundary conditions for quasilinear degenerate elliptic equations and applications,* J. Differential Equations, 154, pp. 191–224.

[5] E. N. BARRON AND R. JENSEN (1990), *Semicontinuous viscosity solutions of Hamilton-Jacobi equations with convex Hamiltonian*, Comm. Partial Differential Equations, 15, pp. 1713–1742.

[6] B. CORNET (1983), *Existence of slow solutions for a class of differential inclusions*, J. Math. Anal. Appl., 96, pp. 130–147.

[7] L. C. EVANS (1998), *Partial Differential Equations*, Grad. Stud. Math., 19, A.M.S., Providence, RI.

[8] H. FRANKOWSKA (1985), *A viability approach to the Skorohod problem*, Stochastics, 4, pp. 227–244.

[9] H. FRANKOWSKA (1993), *Lower semicontinuous solutions of the Hamilton–Jacobi–Bellman equations,* SIAM J. Control Optim., 31, pp. 257–272.

[10] H. FRANKOWSKA, S. PLASKACZ, AND T. RZEZUCHOWSKI (1995), *Measurable viability theorems and the Hamilton-Jacobi-Bellman equation*, J. Differential Equations, 116, pp. 265–305.

[11] C. HENRY (1973), *An existence theorem for a class of differential equations with multivalued right-hand side*, J. Math. Anal. Appl., 41, pp. 179–186.

[12] O. Ley (2002), *A counter-example to the characterization of the discontinuous value function of control problems with reflection*, C. R. Acad. Sci. Paris, Optimal Control, Ser. I, 335, pp. 469–473.

[13] P. L. Lions (1985), *Neumann type boundary conditions for Hamilton-Jacobi equations*, Duke Math. J., 52, pp. 793–820.

[14] P. L. Lions and A. S. Sznitman (1984), *Stochastic differential equations with reflecting boundary conditions*, Comm. Pure Appl. Math., 37, pp. 511–537.

[15] S. Plaskacz and M. Quincampoix (2000), *Discontinuous Mayer problem under state-constraints*, Topol. Methods Nonlinear Anal., 15, pp. 91–100.

[16] S. Plaskacz and M. Quincampoix (2001), *Value-functions for differential games and control systems with discontinuous terminal cost*, SIAM J. Control Optim., 39, pp. 1485–1498.

[17] R. A. Poliquin, R. T. Rockafellar, and L. Thibault (2000), *Local differentiability of distance functions*, Trans. Amer. Math. Soc., 352, pp. 5231–5249.

[18] H. Tanaka (1979), *Stochastic differential equations with reflecting boundary condition in convex regions*, Hiroshima Math. J., 9, pp. 163–177.

[19] L. Thibault, *Sweeping process with regular and nonregular sets*, J. Differential Equations, to appear.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 576–595

VARIATIONAL ANALYSIS APPLIED TO THE PROBLEM OF OPTICAL PHASE RETRIEVAL*

JAMES V. BURKE† AND D. RUSSELL LUKE‡

Abstract. We apply nonsmooth analysis to a well-known optical inverse problem, phase retrieval. The phase retrieval problem arises in many different modalities of electromagnetic imaging and has been studied in the optics literature for over forty years. The state of the art for this problem in two dimensions involves iterated projections for solving a nonconvex feasibility problem. Despite widespread use of these algorithms, current mathematical theory cannot explain their success. At the heart of projection algorithms is a nonconvex, nonsmooth optimization problem. We obtain some insight into these algorithms by applying techniques from nonsmooth analysis. In particular, we show that the weak closure of the set of directions toward the projection generate the subdifferential of the corresponding squared set distance function. Following a pattern of proof described in [F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998], this result is generalized to provide conditions under which the subdifferential of an integral function equals the integral of the subdifferential.

Key words. phase retrieval, least squares, nonsmooth analysis, variational analysis

AMS subject classifications. 78A45, 93E24, 49J52, 49J53

PII. S0363012902406436

1. Introduction.

The phase retrieval problem arises frequently in a number of different optical imaging modalities including diffraction imaging and interferometry. While the imaging models differ slightly, the feature common to these techniques is the problem of recovering the phase of a complex-valued function from measurements of the amplitude of that function, as well as other a priori constraints. There are many unsolved mathematical problems surrounding wavefront reconstruction and phase retrieval in general. Nevertheless, engineers and physicists have been solving this problem in some sense for over thirty years. The most famous application of phase retrieval came with NASA's Hubble Space Telescope (HST). Optical wavefront reconstruction played a central role in the effort to identify gross manufacturing errors in the HST and to design, in effect, a pair of glasses for the near-sighted telescope. We refer the reader to [16] for a review and tutorial of wavefront reconstruction. Here we present only the abstract setting.

The *forward imaging* model is formulated on the space $L^2[\mathbb{R}^2, \mathbb{R}^2]$ of square integrable functions mapping $\mathbb{R}^2$ to $\mathbb{R}^2$. The model input $u : \mathbb{R}^2 \to \mathbb{R}^2$ is an optical field generated by the object we are trying to observe. The optical device is characterized by a unitary bounded linear operator $\mathcal{F}_m : L^2[\mathbb{R}^2, \mathbb{R}^2] \to L^2[\mathbb{R}^2, \mathbb{R}^2]$. The subscript m indicates certain parameter settings in the optical device that constitute a particular known "tuning" such as focus. Let $\mathbb{R}_+$ denote the nonnegative orthant. The model output, or data, corresponding to the mth tuning of the device, $\psi_m : \mathbb{R}^2 \to \mathbb{R}_+$, is

*Received by the editors April 26, 2002; accepted for publication (in revised form) November 22, 2002; published electronically May 12, 2003.

http://www.siam.org/journals/sicon/42-2/40643.html

†Department of Mathematics, University of Washington, Seattle, WA 98195 (burke@math.washington.edu). This author's work was partially supported by NSF grant DMS-0203175.

‡Corresponding author. Institut für Numerische und Angewandte Mathematik, Universität Göttingen, Germany (luke@math.uni-goettingen.de). This author's work was partially supported by NASA grant NGT 5-66 and a Postdoctoral Fellowship from the Pacific Institute for the Mathematical Sciences through Simon Fraser University.

amplitude measurements. The imaging model is given by

$$|\mathcal{F}_m(u(\cdot))| = \psi_m(\cdot), \qquad m = 0, 1, \ldots, M, \tag{1}$$

where the modulus $|\cdot|$ is the *pointwise Euclidean* magnitude. Our discussion switches frequently between the finite- and infinite-dimensional settings. Whenever there is chance for confusion, we indicate a mapping F on the function space explicitly as $F(u(\cdot))$.

Wavefront reconstruction is an *inverse problem*: given $\mathcal{F}_m$ and ψ_m, $m = 0, 1, \ldots, M$, determine u satisfying (1). For a more detailed review of the existence and uniqueness theory behind this problem we refer to [16] and references therein. For our purposes it suffices to note that there is no known closed-form solution to this inverse problem. Moreover, in the presence of noise it is likely that a solution does not exist, thus solution techniques involve minimizing a performance measure. Even though the performance measure that we consider is smooth, the modulus in (1) leads to a nonsmooth objective (see Theorem 3.1 in section 3). At first glance, it would seem that one could easily handle nonsmoothness by squaring both sides of (1). It turns out, however, that objectives based on the *modulus* function, or a nearby smooth approximation, perform better than objectives built upon the modulus squared [16]. Therefore, it can be advantageous to exploit nonsmoothness rather than to avoid it.

Since noise in the data is most often modeled as additive white noise, the least squares error metric is used to find the best fit to (1). For $m = 0, 1, \ldots, M$ and ψ_m not equal to zero a.e., define

$$\mathbb{Q}_m := \left\{ u \in L^2[\mathbb{R}^2, \mathbb{R}^2] \;\middle|\; |\mathcal{F}_m(u)| = \psi_m \text{ a.e.} \right\}. \tag{2}$$

The phase retrieval problem is given by

$$\begin{aligned} &\text{minimize } J(u) \\ &\quad \text{over} \quad u \in L^2[\mathbb{R}^2, \mathbb{R}^2], \end{aligned} \tag{3}$$

where

$$J(u) = \sum_{m=0}^{M} \frac{\beta_m}{2} \text{dist}^{\,2}(u; \mathbb{Q}_m) \tag{4}$$

is the weighted ($\beta_m > 0$ for $m = 0, \ldots, M$) squared set distance error for the phase retrieval problem and

$$\text{dist}\,(u; \mathbb{Q}_m) := \inf_{w \in \mathbb{Q}_m} \|u - w\|. \tag{5}$$

The error metric (4) has a long tradition in the optics literature [9, 10]. It has also been studied in the convex setting where each of the sets $\mathbb{Q}_m$ is assumed to be convex (e.g., see [2, 7]).

Problem (3) is often reformulated as a feasibility problem: the function u must lie in the intersection of the sets $\mathbb{Q}_0 \cap \mathbb{Q}_1 \cap \cdots \mathbb{Q}_m$, assuming that this intersection is nonempty. Projection algorithms are often used to find a point in the intersection of such a collection of sets. Independent of the mathematical literature on projections (and in some cases *before* these algorithms appeared in the mathematical literature) optical scientists developed image processing algorithms for recovering the phase from amplitude measurements known in the optics literature as *iterative transform methods*.

Here one adjusts the phase of the current estimate, $u^{(\nu)}$, at iteration ν by replacing the magnitude of the image $\mathcal{F}_m(u^{(\nu)}(\cdot))$ with the known pointwise magnitude $\psi_m(\cdot)$ and then inverse transforming the result, $\mathcal{F}_m^*(\psi_m(\cdot)\exp(\sqrt{-1}\ \arg(\mathcal{F}_m(u^{(\nu)}(\cdot)))))$. It is straightforward to show that this operation is a projection [16]. The Gerchberg–Saxton algorithm [10] is a classical example of this type of algorithm. When the sets $\mathbb{Q}_m$ are convex and the intersection is nonempty, then this approach is perfectly reasonable since cyclic projections onto such a finite collection of convex sets converges to the intersection (e.g., see [3] and the references therein). In the setting of phase retrieval, however, the sets $\mathbb{Q}_m$ are not even weakly closed, let alone convex [16, Property 4.1]. This poses serious challenges to any convergence theory for algorithms based on projections. Not surprisingly, many have noted that iterative transform algorithms often stagnate. There are some well-known strategies for dealing with these problems [9], but it has recently been observed that these too are applications of convex operator splitting strategies in nonconvex, nonlinear settings [4], so convergence is still problematic.

To overcome some of the problems inherent in treating the leading algorithms as nonconvex instances of projection algorithms, we approach the problem in its variational form (3) using the tools of nonsmooth analysis. We show that, for the squared set distance error metric (4), some projection algorithms can be viewed as *subgradient descent algorithms*. Thus, the critical object for our analysis is the *subdifferential*, or generalized derivative of the squared set distance error metric $J(u)$. In this analysis, the space to which the data $\{\psi_m : m = 0, 1, \dots M\}$ belongs is of critical importance. We require these functions to be nonnegative and finite-valued with their value tending to zero as their argument diverges to infinity in norm. Specifically, we assume that the data belongs to the set $\mathbb{U}$ where

(6)
$$\mathbb{U} = \left\{ v \in L^1 \cap L^2 \cap L^\infty[\mathbb{R}^2, \mathbb{R}] \text{ such that } v(x) \geq 0 \text{ a.e. and } |v(x)| \to 0 \text{ as } |x| \to \infty \right\}.$$

In section 2 we review the theory of projections applied to this problem. The most common projection algorithms, stated in general form in section 2.3, are central to current numerical techniques for this problem. In section 3, we look at the problem from the perspective of nonsmooth least squares, beginning first with finite-dimensional nonsmooth analysis in section 3.2 and building toward the infinite-dimensional analysis in section 3.5. We then apply these results to the problem of wavefront reconstruction in section 3.6. In the final section of the paper we present a result on the exchange of subdifferentiation and integration. Such results have a long history, beginning with Rockafellar's result [20] for convex normal integrands. Our result is in the spirit of [6, Theorem 3.5.18]. Indeed, our method of proof parallels that given by Clarke, Ledyaev, Stern, and Wolenski. The key difference between our result and [6, Theorem 3.5.18] is that our domain of integration is all of $\mathbb{R}^2$ as opposed to an interval in $\mathbb{R}$.

2. Geometric approaches.

2.1. Projections. In general, it may be difficult to prove that the projection of a given point onto a given set exists, much less to identify it with a formula. Much of the general theory of projections [24] does not apply since the sets in question are neither weakly closed nor convex [16, Property 4.1]. However, in the application to phase retrieval there is a very simple characterization in terms of pointwise, finite-dimensional projections.

Our focus is on sets of the form

$$\mathbb{Q}(b) := \left\{ u \in L^2[\mathbb{R}^2, \mathbb{R}^2] \;\middle|\; |u| = b \text{ a.e.} \right\}. \tag{7}$$

Here the set $\mathbb{Q}(b)$ is parameterized by the *function* $b : \mathbb{R}^2 \to \mathbb{R}_+$. Alternatively, one can think of this set as being parameterized *pointwise* by $x \in \mathbb{R}^2$, that is, at each point x, the set $\mathbb{Q}(b(x)) \subset \mathbb{R}^2$ is simply the sphere of radius $b(x)$, denoted $b(x)\mathbb{S}$, where $\mathbb{S}$ is the unit sphere in $\mathbb{R}^2$. For the closed set $\mathbb{Q}$ in the Hilbert space $\mathbb{X}$, we define the projection operator $\Pi_{\mathbb{Q}}(v)$ as the multivalued mapping, or multifunction, given as the set of all solutions to the minimum distance problem for the set $\mathbb{Q}$:

$$\Pi_{\mathbb{Q}}(v) := \arg\min_{u \in \mathbb{Q}} \|v - u\| = \{\bar{u} \in \mathbb{Q} : \|v - \bar{u}\| = \inf_{u \in \mathbb{Q}} \|v - u\|\}. \tag{8}$$

It is a simple matter to characterize the pointwise projection $\Pi_{b(x)\mathbb{S}} : \mathbb{R}^2 \rightrightarrows \mathbb{R}^2$:

$$\Pi_{b(x)\mathbb{S}}(v) = b(x)\Pi_{\mathbb{S}}(v) = b(x) \times \begin{cases} \frac{v}{|v|} & \text{for } v \neq 0, \\ \mathbb{S} & \text{for } v = 0, \end{cases} \qquad v \in \mathbb{R}^2. \tag{9}$$

Note that the projection is multivalued at $v = 0$. In the following sections we construct the infinite-dimensional projection $\Pi_{\mathbb{Q}(b)} : L^2[\mathbb{R}^2, \mathbb{R}^2] \rightrightarrows L^2[\mathbb{R}^2, \mathbb{R}^2]$ onto $\mathbb{Q}(b)$ from the corresponding pointwise projection at the point x, $\Pi_{b(x)\mathbb{S}} : \mathbb{R}^2 \rightrightarrows \mathbb{R}^2$ onto $b(x)\mathbb{S}$.

2.2. Measurable multifunctions. We now review some of the properties of measurable multifunctions used in this study [1, 6, 11, 21]. In section 3.3 we extend this review to include the integration theory of measurable multivalued mappings. For more information on this and related topics, we refer the interested reader to [21, chapter 14].

Let $\Omega \neq \emptyset$ and let $\mathcal{A}$ be a σ-field of subsets of Ω, called the *measurable* subsets of Ω or the $\mathcal{A}$-*measurable* subsets. The corresponding measure space is denoted $(\Omega, \mathcal{A})$. Our discussion is limited to *complete nonatomic* measure spaces.

The multifunction $F : \Omega \rightrightarrows \mathbb{R}^n$ is said to be $\mathcal{A}$-measurable, or simply measurable, if for all open sets $\mathbb{V}$ the set $\{x \mid \mathbb{V} \cap F(x) \neq \emptyset\}$ is in $\mathcal{A}$. The multifunction F is said to be $\mathcal{A} \otimes \mathcal{B}^n$-*measurable* if $\text{gph}(F) = \{(x, v) \mid v \in F(x)\} \in \mathcal{A} \otimes \mathcal{B}^n$. Here $\mathcal{B}^n$ denotes the Borel σ-field on $\mathbb{R}^n$ and $\mathcal{A} \otimes \mathcal{B}^n$ is the σ-field on $\Omega \times \mathbb{R}^n$ generated by all sets $A \times D$ with $A \in \mathcal{A}$ and $D \in \mathcal{B}^n$. If $F(x)$ is closed for each x, then F is *closed*. Similarly, F is said to be *convex* if $F(x)$ is convex for each x. Finally, we note that the completeness of the measure space guarantees the measurability of subsets of Ω obtained as the projections of measurable subsets $\mathbb{G}$ of $\Omega \times \mathbb{R}^n$:

$$\mathbb{G} \in \mathcal{A} \otimes \mathcal{B}^n \quad \Longrightarrow \quad \{\omega \in \Omega \mid \exists\, x \in \mathbb{R}^n \text{ with } (\omega, x) \in \mathbb{G}\} \in \mathcal{A},$$

and thus F is $\mathcal{A}$-measurable if and only if F is $\mathcal{A} \otimes \mathcal{B}^n$-measurable [21, Theorem 14.8].

Let $F : \Omega \rightrightarrows \mathbb{R}^n$. Denote by $\mathcal{S}(F)$ the set of μ-measurable functions $f : \Omega \to \mathbb{R}^n$ that satisfy $f(x) \in F(x)$ a.e. in Ω $(x \in \Omega)$. We call $\mathcal{S}(F)$ the *set of measurable selections* of F.

THEOREM 2.1 (measurable selections [21, Corollary 14.6]). *A closed-valued measurable map $F : \Omega \rightrightarrows \mathbb{R}^n$ always admits a measurable selection.*

For a measurable function $f = (f_1, \ldots, f_n)$, $f_i : \Omega \to \mathbb{R}$, for $i = 1, \ldots, n$, the integral $\int f d\mu$ is defined to be the vector

$$\left(\int f_1, d\mu, \ldots, \int f_n d\mu \right).$$

The set

$$\left\{ \int f d\mu \mid f \in \mathcal{S}(F) \right\}$$

is the *integral* of the multivalued mapping $F : \Omega \rightrightarrows \mathbb{R}^n$ and is denoted by $\int F d\mu$ or $\int F$. We say that $F : \Omega \rightrightarrows \mathbb{R}^n$ is *integrably bounded*, or for emphasis μ-integrably bounded, if there is a μ-integrable $a : \Omega \to \mathbb{R}^n_+$ such that

$$(|v_1|, \ldots, |v_n|) \leq a(x)$$

for all pairs $(x, v) \in (\Omega, \mathbb{R}^n)$ satisfying $v \in F(x)$. Here and elsewhere we interpret vector inequalities as elementwise inequalities. If $a(x)$ in the above inequality is *square-integrable* with respect to the measure μ on the measure space $(\Omega, \mathcal{A}, \mu)$, then the multifunction F is said to be L^2-bounded. When $\Omega = \mathbb{R}^n$, we let $L^2_m(\mathbb{R}^n, \mathcal{A}, \mu)$ denote the Hilbert space of functions mapping $\mathbb{R}^n$ to $\mathbb{R}^m$ with inner product on the measure space $(\mathbb{R}^n, \mathcal{A}, \mu)$ given by

$$\langle f, g \rangle = \int_{\mathbb{R}^n} (f(x), g(x)) \mu(dx), \tag{10}$$

where $(\cdot, \cdot)$ denotes the usual finite-dimensional vector inner product.

The next property is a generalization of [6, Exercise 3.5.14].

PROPOSITION 2.2 (weak compactness of measurable selections). *Let the multifunction $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ be closed, convex-valued, and L^2-bounded on $L^2_m(\mathbb{R}^n, \mathcal{M}^n, \nu_n)$, where $\mathcal{M}^n$ is the Lebesgue field on $\mathbb{R}^n$ and ν_n is the n-dimensional Lebesgue measure. Then the set of measurable selections $\mathcal{S}(F)$ is a weakly compact, convex set in $L^2_m(\mathbb{R}^n, \mathcal{M}^n, \nu_n)$.*

Proof. This set is clearly convex since F is pointwise convex-valued. Thus, by [8, Theorem 1, p. 58] we need only show that $\mathcal{S}(F)$ is weakly sequentially compact. Consider any sequence $\{f_i\} \subset \mathcal{S}(F)$. We must show that $\{f_i\}$ has a weakly convergent subsequence with limit $f_* \in \mathcal{S}(F)$. Since the sequence is L^2-bounded, reflexivity, separability, and Alaoglu's theorem [23, Exercise 18(b), p. 269] imply that there exists a weakly convergent subsequence whose limit belongs to the weak closure of $\mathcal{S}(F)$. Since $\mathcal{S}(F)$ is convex, the strong and weak closures of $\mathcal{S}(F)$ coincide. Hence the result follows if $\mathcal{S}(F)$ is strongly closed. Since strong convergence implies the existence of a subsequence that is almost everywhere pointwise convergent [23, Theorem 3.12], and $F(x)$ is pointwise closed, we have that $\mathcal{S}(F)$ is strongly closed. □

2.3. Application to wavefront reconstruction: Projection algorithms. We now characterize the projections associated with the problem of phase retrieval in terms of the corresponding pointwise projections. This allows us to describe a general algorithmic framework that includes many of the currently used phase retrieval algorithms. Let $b \in L^2[\mathbb{R}^2, \mathbb{R}]$ with $b(x) \geq 0$ a.e., let the pointwise projection $b(x)\Pi_{\mathbb{S}}$ be defined by (9), and let $\mathbb{Q}(b)$ be defined by (7). For $u, v \in L^2[\mathbb{R}^2, \mathbb{R}^2]$, it is shown in [16, Theorem 4.2] that the projection $\Pi_{\mathbb{Q}(b)} : L^2[\mathbb{R}^2, \mathbb{R}^2] \rightrightarrows L^2[\mathbb{R}^2, \mathbb{R}^2]$ onto $\mathbb{Q}(b)$ is characterized as the collection of measurable selections from the pointwise projection mapping (9):

$$\Pi_{\mathbb{Q}(b)}(u) = \mathcal{S}\left(b(\cdot)\Pi_{\mathbb{S}}(u(\cdot))\right) \quad \text{and} \quad \operatorname{dist}(u; \mathbb{Q}(b)) = \| \, |u| - b \, \|. \tag{11}$$

One can characterize the projection onto the sets $\mathbb{Q}_m$ defined in (2) in a similar fashion. The $\mathcal{F}_m$-transform of $\Pi_{\mathbb{Q}(b)}(u)$ is the $\mathcal{F}_m$-transform of all $v \in \Pi_{\mathbb{Q}(b)}(u)$ and is

written $\mathcal{F}_m\left(\Pi_{\mathbb{Q}(b)}(u)\right)$. For each of the unitary operators $\mathcal{F}_m$ and all $u \in L^2[\mathbb{R}^2, \mathbb{R}^2]$, we know from [16, Corollary 4.3] that

$$\Pi_{\mathbb{Q}_m}(u) = \mathcal{F}_m^* \left(\Pi_{\mathbb{Q}(\psi_m)}(\mathcal{F}_m(u))\right) \quad \text{and} \quad \operatorname{dist}(u; \mathbb{Q}_m) = \| \, |\mathcal{F}_m(u)| - \psi_m \| \, . \tag{12}$$

A general framework for projection algorithms can be found in [3], which considers sequences of weighted relaxed projections of the form

$$u^{(\nu+1)} \in \left(\sum_{m=0}^{M} \gamma_m^{(\nu)} \left[(1 - \alpha_m^{(\nu)}) \mathcal{I} + \alpha_m^{(\nu)} \Pi_{\mathbb{Q}_m} \right] \right) (u^{(\nu)}). \tag{13}$$

Here $\mathcal{I}$ is the identity mapping, $\alpha_m^{(\nu)}$ is a relaxation parameter usually in the interval $[0, 2]$, and the weights $\gamma_m^{(\nu)}$ are nonnegative scalars summing to one. General results for these types of algorithms apply only to convex sets. In the convex setting the inclusion in (13) is an equality since projections onto convex sets are single-valued. In the nonconvex setting this is not the case.

It is shown in [16] that the Gerchberg–Saxton algorithm [10] and its variants can be viewed as an instance of (13). As in [16] we use the change of variables $\lambda^{(\nu)} \beta_m^{(\nu)} = \gamma_m^{(\nu)} \alpha_m^{(\nu)}$ to rewrite (13) as

$$u^{(\nu+1)} \in \left(\mathcal{I} - \lambda^{(\nu)} \mathcal{G}^{(\nu)} \right) \left(u^{(\nu)} \right), \tag{14}$$

where for all ν the operators $\mathcal{G}^{(\nu)} : L^2 \to L^2$ are given by

$$\mathcal{G}^{(\nu)} := \sum_{m=0}^{M} \mathcal{G}_m^{(\nu)} \qquad \text{with} \qquad \mathcal{G}_m^{(\nu)} := \beta_m^{(\nu)} \left(\mathcal{I} - \Pi_{\mathbb{Q}_m} \right). \tag{15}$$

In (14) the nonnegative weights $\beta_m^{(\nu)}$ do not necessarily sum to 1, and the parameters $\lambda^{(\nu)}$ are to be interpreted as *step lengths*. This formulation of the projection algorithm is shown in the next section to correspond to a steepest descent algorithm for a weighted squared distance function.

3. Nonsmooth analysis. Convergence results for projection methods applied to the phase retrieval problem are not possible in general due to the nonconvexity of the constraint sets. The nonconvexity of the constraint sets is associated with the nonsmoothness of the square of the set distance error $\operatorname{dist}(u; \mathbb{Q}_m)$ defined in (5). This is fundamentally different from the convex setting in a Hilbert space where the squared distance function is smooth.

3.1. Least squares. In general the optimal value of the weighted squared set distance error $J(u)$ defined by (4) is nonzero. Classical techniques for solving the problem numerically are based on satisfying a first-order necessary condition for optimality. For smooth functions this condition simply states that the gradient takes the value zero at any local solution to the optimization problem. However, the functions $\operatorname{dist}^2(u; \mathbb{Q}_m)$ are not differentiable. The easiest way to see this is to consider the one-dimensional function $a(x) = \, ||x| - b|^2$, where $b > 0$. This function is not differentiable at $x = 0$. (Indeed, it is not even subdifferentiably regular at $x = 0$—see (19)). It is precisely at these points that the finite-dimensional projection operator $\Pi_{b\mathbb{S}}$ is multivalued. Similarly, $\operatorname{dist}^2(u; \mathbb{Q}_m)$ is not differentiable at functions u for which there exists a set $\Omega \subset \operatorname{supp}(\psi_m)$ of positive measure on which u vanishes.

In the nonsmooth setting the usual first-order necessary condition for optimality is replaced by a first-order variational principle of the form $0 \in \partial J(u_*)$, where ∂ denotes a *subdifferential* operator such as those studied in [5, 6, 12, 13, 15, 18]. In this paper, the phrase *the subdifferential* refers to the nonconvex subdifferential introduced by Kruger and Mordukhovich [15]. This subdifferential is precisely described in Definition 3.12, and its calculus is extensively developed in [18]. The main result of this paper is the characterization of the subdifferential of the distance functions $\text{dist}^{\,2}(\cdot;\mathbb{Q}_m)$ and the objective function J (equation (4)). We do this by following the pattern of proof used by Clarke, Ledyaev, Stern, and Wolenski in [6, Theorem 3.5.18]. A consequence of this approach is that we also establish the subdifferential regularity of the functions $\text{dist}^{\,2}(\cdot;\mathbb{Q}_m)$ and J. This in turn implies that for these functions the Clarke subdifferential [5, 6] and the nonconvex subdifferential [15] are equivalent. The statement of the main result now follows.

THEOREM 3.1 (projections and subdifferentials). *Let $\psi_m : \mathbb{R}^2 \to \mathbb{R}_+$ belong to $\mathbb{U}$ where the set $\mathbb{U}$ is defined in* (6)*, and let $\Pi_{\mathbb{Q}_m} : L^2 \rightrightarrows \mathbb{Q}_m$ be defined by* (8)*. Then the functions $\text{dist}^{\,2}(\cdot;\mathbb{Q}_m)$ and J are everywhere subdifferentially regular and for $u \in L^2[\mathbb{R}^2,\mathbb{R}^2]$ we have*

$$\partial\left(\text{dist}^{\,2}(u;\mathbb{Q}_m)\right) = 2\text{cl}^*\left(\mathcal{I} - \Pi_{\mathbb{Q}_m}(u)\right) \tag{16}$$

and

$$\partial J(u) = \sum_{m=0}^{M} \text{cl}^*\left(\mathcal{G}_m(u)\right), \tag{17}$$

where $\mathcal{G}_m$ is defined by (15)*, J is defined in* (4)*, and $\text{cl}^*(\,\cdot\,)$ denotes the weak-star closure.*

Note that in a Hilbert-space setting $\text{cl}^*(\,\cdot\,) = w\text{–cl}(\,\cdot\,)$, where $w\text{–cl}(\,\cdot\,)$ denotes the weak closure. The proof is given at the end of this section. In passing, we note that in the convex case Theorem 3.1 is an elementary consequence of a much more general result for convex functions given in [19, Theorem 20]. For further results along these lines we refer the reader to [5, Proposition 2.5.4] and [21, Example 8.53].

3.2. Finite-dimensional nonsmooth analysis. In [16, Theorem 4.2] it is shown that the squared set distance error $\text{dist}^{\,2}(u;\mathbb{Q}(b))$ defined in (7) is given as the integral of the *pointwise* distance function defined by (11). In Theorem 3.1 we extend this correspondence to the subdifferentials of the associated infinite- and finite-dimensional functions. We begin this analysis by introducing the necessary tools from finite-dimensional variational analysis.

Recall that

$$\text{dist}^{\,2}(u;\mathbb{Q}(b)) \;=\; \int_{\mathbb{R}^2} r^2(u(x);b(x))dx \;=\; \|u\|^2 + \|b\|^2 + 2h(u;b),$$

where the pointwise residual $r : \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}$ and the mapping $h : L^2[\mathbb{R}^2,\mathbb{R}^2] \to \mathbb{R}$ are given by

$$r(u(x);b(x)) = |u(x)| - b(x) \qquad \text{and} \qquad h(u;b) := \int_{\mathbb{R}^2} -|u(x)|\,b(x)dx, \tag{18}$$

respectively. While $\text{dist}^{\,2}(u;\mathbb{Q}(b))$ is not smooth, it is straightforward to show that it is Lipschitz continuous on bounded subsets of $L^2[\mathbb{R}^2,\mathbb{R}^2]$.

A function $f : \mathbb{X} \to \mathbb{R}$ is *locally Lipschitz near* x if there exists a constant $K \geq 0$ and a neighborhood $\mathbb{V}(x) \subset \mathbb{X}$ of x such that

$$|f(z) - f(y)| \leq K\|z - y\| \qquad \forall\, z, y \in \mathbb{V}(x).$$

For any set $\mathbb{V} \subset X$ over which f is finite-valued, f is said to be *locally Lipschitz on* $\mathbb{V}$ if it is locally Lipschitz at every $x \in \mathbb{V}$. The function is said to be *(globally) Lipschitz on* $\mathbb{V}$ if

$$|f(x) - f(y)| \leq K\|x - y\| \qquad \forall\, x, y \in \mathbb{V}.$$

PROPOSITION 3.2 (Lipschitz constants). *If* $b \in L^2[\mathbb{R}^2, \mathbb{R}]$ *with* $b(x) \geq 0$ *a.e., then the mapping* $\operatorname{dist}^2(\cdot; \mathbb{Q}(b)) : L^2[\mathbb{R}^2, \mathbb{R}^2] \to \mathbb{R}_+$ *is finite-valued and Lipschitz on any bounded subset* $\mathbb{V} \subset L^2[\mathbb{R}^2, \mathbb{R}^2]$ *with Lipschitz constant*

$$K = K_{\|\cdot\|^2} + K_{2h(\cdot;b)},$$

where $K_{\|\cdot\|^2} = 2\sup_{u \in \mathbb{V}} \|u\|$ *is a Lipschitz constant for* $\|u\|^2$ *on* $\mathbb{V}$ *and* $K_{2h(\cdot;b)} = 2\|b\|$ *is a Lipschitz constant for* $h(\cdot; b)$*, independent of* $\mathbb{V}$*.*

Proof. This follows from the proof of [16, Lemma B.2]. ∎

Lipschitz continuity of the squared set distance error J is a straightforward consequence of Proposition 3.2 and the fact the mappings $\mathcal{F}_m$ are unitary.

We now introduce some basic definitions from nonsmooth analysis. In our discussion we allow mappings to have infinite values; thus it is convenient to define the extended reals $\overline{\mathbb{R}}$, where $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$. The *effective domain* of $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, denoted $\operatorname{dom} f \subset \mathbb{R}^n$, is the set on which f is finite. To avoid certain pathological mappings the discussion is restricted to *proper*, i.e., not everywhere infinite, *lower semicontinuous* (l.s.c.) functions.

DEFINITION 3.3 (subderivatives [21]). *For a Lipschitz function* $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ *and a point* $u_* \in \mathbb{R}^m$ *with* $f(u_*)$ *finite,*

(i) *the* subderivative *function* $df(u_*) : \mathbb{R}^m \to \overline{\mathbb{R}}$ *is defined by*

$$df(u_*)(w) := \liminf_{\tau \searrow 0} \frac{f(u_* + \tau w) - f(u_*)}{\tau};$$

(ii) *the* regular subderivative *function (or the Clarke generalized directional derivative when* f *is Lipschitz)* $\widehat{d}f(u_*) : \mathbb{R}^m \to \overline{\mathbb{R}}$ *is defined by*

$$\widehat{d}f(u_*)(w) := \limsup_{u \to u_*,\, \tau \searrow 0} \frac{f(u + \tau w) - f(u)}{\tau}.$$

DEFINITION 3.4 (subgradients: finite-dimensions [21]). *Consider a function* $f : \mathbb{R}^m \to \overline{\mathbb{R}}$*, a point* $v \in \mathbb{R}^m$*, and a point* $u_* \in \mathbb{R}^m$ *with* $f(u_*)$ *finite.*

(i) v *is a* regular subgradient *of* f *at* u_* *if*

$$\liminf_{\substack{u \to u_* \\ u \neq u_*}} \frac{f(u) - f(u_*) - \langle v,\ u - u_* \rangle}{|u - u_*|} \geq 0.$$

We call the set of regular subgradients v *the* regular subdifferential *of* f *at* u_* *and denote this set by* $\widehat{\partial} f(u_*)$*.*

(ii) *v is a* (general) subgradient *of f at u_* if there are sequences $u^{(\nu)} \to u_*$ and $v^{(\nu)} \in \widehat{\partial} f(u^{(\nu)})$ with $f(u^{(\nu)}) \to f(u_*)$ and $v^{(\nu)} \to v$. We call the set of (general) subgradients v the* (general) subdifferential *of f at u_* and denote this set by $\partial f(u_*)$.*

(iii) *v is a* Clarke subgradient *of f at u_* if f is l.s.c. on a neighborhood of u_* and v satisfies*

$$\langle v,\ w\rangle \leq \widehat{d} f(u_*)(w)\ \forall w \in \mathbb{R}^m.$$

We call the set of Clarke subgradients v the Clarke subdifferential *of f at u_* and denote this set by $\overline{\partial} f(u_*)$.*

(iv) *A Lipschitz function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to be* (subdifferentially) regular *at $u_* \in \operatorname{dom} f$ with $\partial f(u_*) \neq \emptyset$ if*

$$\partial f(u_*) = \widehat{\partial} f(u_*). \tag{19}$$

Remark 3.5 (subdifferentials with closed graphs). From the definitions it can be shown that if $f : \mathbb{R}^n \to \mathbb{R}$ is continuous, then the subgradients ∂f and $\widehat{\partial} f$ are closed with $\widehat{\partial} f$ convex and $\widehat{\partial} f \subset \partial f$. Moreover, the mapping ∂f is outer semicontinuous [21, Definition 5.4]. Therefore, by [21, Theorem 5.7] the graph of ∂f is closed.

Remark 3.6 (subdifferentials of compositions). If $g : \mathbb{X} \to \overline{\mathbb{R}}$ is given as the composition of two functions $f : \mathbb{Y} \to \overline{\mathbb{R}}$ and $h : \mathbb{X} \to \mathbb{Y}$, i.e., $g(x) = (f \circ h)(x) = f(h(x))$, then we write $\partial g(x) = \partial(f \circ h)(x)$. On the other hand, we write $\partial f(h(x))$ to denote the subdifferential of f evaluated at $h(x)$.

The subdifferential definitions are illustrated with the following important example.

Example 3.7 (subdifferential of the modulus). Let $b \in (0, \infty)$. Since the function $b|u|$ is convex it is subdifferentially regular for all u, and

$$\partial\left(b|u|\right) = b\partial\left(|u|\right) = \begin{cases} b\frac{u}{|u|} & \text{if } u \neq 0, \\ b\mathbb{B} & \text{if } u = 0, \end{cases}$$

where $b\mathbb{B}$ is the ball of radius b: $\mathbb{B} = \{u : |u| \leq 1\}$.

In contrast, the function $-b|u|$ for $b \in (0, \infty)$ is not regular at 0. Nevertheless for all u

$$\partial\left(-b|u|\right) = b\partial\left(-|u|\right) = \begin{cases} -b\frac{u}{|u|} & \text{if } u \neq 0, \\ b\mathbb{S} & \text{if } u = 0, \end{cases}$$

where $b\mathbb{S}$ is the sphere of radius b: $\mathbb{S} = \{u : |u| = 1\}$. The Clarke subdifferential of $-b|u|$ is the convex hull, denoted $\operatorname{conv}(\cdot)$, of the generalized subdifferential:

$$\overline{\partial}\left(-b|u|\right) \ = \ \operatorname{conv} \partial\left(-b|u|\right) \ = \ -\partial(b|u|\,).$$

Proof. The first part of the statement is a trivial modification of [21, Exercise 8.27]. The last statement follows from [21, Theorem 8.49]. ∎

This example yields the following correspondence between finite-dimensional projections $\Pi_{b\mathbb{S}}$ and the subdifferential $\partial(-b|u|\,)$.

Proposition 3.8 (pointwise projections and subdifferentials). *Let $\Pi_{b\mathbb{S}}(u)$ be the projection defined in* (9)*. For $u \in \mathbb{R}^2$, $b \in \mathbb{R}_+$, and $r^2 : \mathbb{R}^2 \to \mathbb{R}_+$ defined in* (18) *we have*

$$\partial(-b|u|\,) = -\Pi_{b\mathbb{S}}(u), \quad \overline{\partial}(-b|u|\,) = -\operatorname{conv}\left(\Pi_{b\mathbb{S}}(u)\right), \quad \textit{and} \quad \partial r^2(u; b) = 2(I - \Pi_{b\mathbb{S}}(u)),$$

where I is the finite-dimensional identity operator. Moreover,

$$\overline{\partial} r^2(u;b) = \operatorname{conv}\,[2(I - \Pi_{b\mathbb{S}}(u))].$$

As with the finite-dimensional projection $\Pi_{b\mathbb{S}}$ and the infinite-dimensional projection $\Pi_{\mathbb{Q}(b)} : L^2[\mathbb{R}^2, \mathbb{R}^2] \rightrightarrows L^2[\mathbb{R}^2, \mathbb{R}^2]$ defined in (11), there is a relationship between the finite-dimensional Clarke subdifferential $\overline{\partial} r^2(u(x); b(x))$ (x fixed) and the "subdifferential" of the square distance function, $\partial(\text{dist}^{\,2}(u; \mathbb{Q}(b)))$. In infinite-dimensional spaces there are several possible definitions for the subdifferential depending on the underlying geometry and topology of the space. Fortunately, in the separable Hilbert-space setting of phase retrieval many of these definitions coincide [18, Theorem 9.2]. Thus we can choose the characterization that is most convenient. The following development parallels that of Clarke, Ledyaev, Stern, and Wolenski in [6, chapter 3, section 5]. We begin by recalling the definitions and theorems necessary for the analysis.

3.3. Integrals of multivalued functions. We now develop some properties of integrals of multivalued mappings. The next theorem, due to Hildenbrand [11], is a restatement of Theorems 3 and 4 of Aumann [1] for multifunctions on the nonatomic measure space $(\Omega, \mathcal{A}, \mu)$. These results are central to the theory of integrals of multivalued functions.

THEOREM 3.9 (integrals of multifunctions [11, Theorem 4 and Proposition 7]). *The following properties hold for integrably bounded multifunctions $F : \Omega \rightrightarrows \mathbb{R}^n$ on nonatomic measure spaces $(\Omega, \mathcal{A}, \mu)$:*

(i) *if F is $\mathcal{A} \otimes \mathcal{B}^n$-measurable, then $\int F = \int \operatorname{conv} F$;*

(ii) *if F is closed (not necessarily $\mathcal{A} \otimes \mathcal{B}^n$-measurable), then $\int F$ is compact.*

The following result is instrumental in the proof of our main result. It is a generalization of [6, Exercise 3.5.17].

PROPOSITION 3.10 (weak closure of nonconvex multivalued integrands). *Let v be chosen from the set of selections $\mathcal{S}(\operatorname{conv} F)$, where $F : \mathbb{R}^2 \rightrightarrows \mathbb{R}^2$ is a nonempty, closed, $\mathcal{M}^2 \otimes \mathcal{B}^2$-measurable, L^2-bounded multifunction on $L_2^2(\mathbb{R}^2, \mathcal{M}^2, P)$ for the probability measure $P(dx) = b(x)dx$ defined by the density $b : \mathbb{R}^2 \to \mathbb{R}_+$. Then there exists a sequence $\{f_i\}$ of measurable selections of F which converges weakly to v. Consequently,*

$$\mathcal{S}(\operatorname{conv} F) \subset \operatorname{cl}^*(\mathcal{S}(F)). \tag{20}$$

Proof. Consider the box $\mathbb{I}_n = [-n, n] \times [-n, n]$ for $n = 1, 2, 3, \ldots$. Suppose each box $\mathbb{I}_n$ is partitioned into $(2n^2)^2$ pixels of width $1/n$. Set

$$t_k^n = \frac{k}{n} - n \quad \text{for } k = 0, 1, \ldots, 2n^2,$$

and for each $t \in [-n, n]$ define

$$\underline{(t)}_n = \max\{t_k^n : t_k^n \le t, k = 0, \ldots, 2n^2\} \quad \text{and} \quad \overline{(t)}_n = \min\{t_k^n : t_k^n \ge t, k = 0, \ldots, 2n^2\}.$$

Note that $0 < \max\{t - \underline{(t)}_n, \overline{(t)}_n - t\} \le 1/n$ whenever $t \in [-n, n]$. By Theorem 3.9 there exists a selection $f_n \in F$ on $(\mathbb{R}^2, \mathcal{M}^2, P)$ corresponding to the partition of the box $\mathbb{I}_n$ such that

$$\int_{\mathbb{R}^2} f_n(x) b(x) dx = \int_{\mathbb{R}^2} v(x) b(x) dx$$

with

$$\int_{t_j^n}^{t_{j+1}^n}\int_{t_k^n}^{t_{k+1}^n} f_n(x)b(x)dx = \int_{t_j^n}^{t_{j+1}^n}\int_{t_k^n}^{t_{k+1}^n} v(x)b(x)dx, \quad n = 1,2,3,\ldots,\ j,k = 0,\ldots,2n^2.$$

We show that the sequence f_n converges weakly to v. Let $g \in C^\infty[\mathbb{R}^2,\mathbb{R}^2]$ and $\mathcal{X}_{\mathbb{M}}$ be the indicator of the box $\mathbb{M} = [\alpha,\beta]\times[\gamma,\eta]$. Given $\epsilon > 0$ we will show that there exists n' such that $|\langle g\mathcal{X}_{\mathbb{M}},\ f_n - v\rangle| \le \epsilon$ for all $n \ge n'$, i.e., $\langle g\mathcal{X}_{\mathbb{M}},\ f_n - v\rangle \to 0$.

Let n_1 be such that $\mathbb{M} \subset \mathbb{I}_{n_1}$ for all $n \ge n_1$. Choose $n \ge n_1$. Integration by parts yields

$$\langle g\mathcal{X}_{\mathbb{M}},\ f_n - v\rangle = \left(g(\beta,\eta),\ \int_\gamma^\eta\int_\alpha^\beta [f_n(s,t)-v(s,t)]b(s,t)ds\,dt\right) \tag{21}$$

$$-\int_\gamma^\eta \left(g_y(\beta,y),\ \int_\gamma^y\int_\alpha^\beta [f_n(s,t)-v(s,t)]b(s,t)ds\,dt\right)dy \tag{22}$$

$$-\int_\alpha^\beta \left(g_x(x,\eta),\ \int_\gamma^\eta\int_\alpha^x [f_n(s,t)-v(s,t)]b(s,t)ds\,dt\right)dx \tag{23}$$

$$+\int_\gamma^\eta\int_\alpha^\beta \left(g_{xy}(x,y),\int_\gamma^y\int_\alpha^x [f_n(s,t)-v(s,t)]b(s,t)ds\,dt\right)dx\,dy. \tag{24}$$

Note that each of these terms contains an expression of the form

$$\begin{aligned}\int_{\hat\gamma}^{\hat\eta}\int_{\hat\alpha}^{\hat\beta}(f_n(s,t)-v(s,t))b(s,t)ds\,dt &= \int_{\underline{(\hat\eta)}_n}^{\hat\eta}\int_{\hat\alpha}^{\hat\beta}(f_n(s,t)-v(s,t))b(s,t)ds\,dt\\ &+\int_{\hat\gamma}^{\overline{(\hat\gamma)}_n}\int_{\hat\alpha}^{\hat\beta}(f_n(s,t)-v(s,t))b(s,t)ds\,dt\\ &+\int_{\overline{(\hat\gamma)}_n}^{\underline{(\hat\eta)}_n}\int_{\hat\alpha}^{\overline{(\hat\alpha)}_n}(f_n(s,t)-v(s,t))b(s,t)ds\,dt\\ &+\int_{\overline{(\hat\gamma)}_n}^{\underline{(\hat\eta)}_n}\int_{\underline{(\hat\beta)}_n}^{\hat\beta}(f_n(s,t)-v(s,t))b(s,t)ds\,dt,\end{aligned} \tag{25}$$

where $[\hat\gamma,\hat\eta]\times[\hat\alpha,\hat\beta] \subset [\gamma,\eta]\times[\alpha,\beta] \subset [-n,n]\times[-n,n]$. Let $a \in L_2^2(\mathbb{R}^2,\mathcal{M}^2,P)$ be an L^2-bound for conv F. For any box of the form $[\alpha',\beta']\times[\gamma',\eta']$, we have the bound

$$\begin{aligned}\left|\int_{\gamma'}^{\eta'}\int_{\alpha'}^{\beta'}(f_n(s,t)-v(s,t))b(s,t)ds\,dt\right| &\le \int_{\gamma'}^{\eta'}\int_{\alpha'}^{\beta'}|f_n(s,t)-v(s,t)|b(s,t)ds\,dt\\ &\le \int_{\gamma'}^{\eta'}\int_{\alpha'}^{\beta'}2|a(s,t)|b(s,t)ds\,dt\\ &= 2\int_{\mathbb{R}^2}|a(x)|\mathcal{X}_{[\alpha',\beta']\times[\gamma',\eta']}(x)b(x)dx\\ &\le 2\|a\|\int_{\mathbb{R}^2}\mathcal{X}_{[\alpha',\beta']\times[\gamma',\eta']}(x)b(x)dx\\ &= 2\|a\|\int_{[\alpha',\beta']\times[\gamma',\eta']}b(x)dx.\end{aligned}$$

Next note that the Lebesgue measure of each of the sets $[\underline{(\hat{\eta})}_n, \hat{\eta}] \times [\hat{\alpha}, \hat{\beta}]$, $[\hat{\gamma}, \overline{(\hat{\gamma})}_n] \times [\hat{\alpha}, \hat{\beta}]$, $[\overline{(\hat{\gamma})}_n, \hat{\eta}] \times [\hat{\alpha}, \overline{(\hat{\alpha})}_n]$, and $[\overline{(\hat{\gamma})}_n, \underline{(\hat{\eta})}_n] \times [\underline{(\hat{\beta})}_n, \hat{\beta}]$ appearing in (25) is bounded by

$$\frac{1}{n} \max\{(\eta - \gamma), (\beta - \alpha)\},$$

which can be made arbitrarily small. By [23, Exercise 12, p. 33], for every $\bar{\epsilon} > 0$ there is an $\delta(\bar{\epsilon}) > 0$ such that

$$\int_{\mathbb{E}} b(x)dx \leq \bar{\epsilon} \quad \text{whenever } \mathcal{M}(\mathbb{E}) \leq \delta(\bar{\epsilon}),$$

where $\mathcal{M}(\mathbb{E})$ is the Lebesgue measure of the set $\mathbb{E}$. Therefore, given $\bar{\epsilon} > 0$, we can choose n so that $\frac{1}{n} \max\{(\eta - \gamma), (\beta - \alpha)\} < \delta(\bar{\epsilon})$. By combining this with (25), we obtain the bound

$$\left| \int_{\hat{\gamma}}^{\hat{\eta}} \int_{\hat{\alpha}}^{\hat{\beta}} (f_n(s,t) - v(s,t)) b(s,t) ds\, dt \right| \leq 8\|a\|\bar{\epsilon}. \tag{26}$$

If we set

$$\Gamma = \max\left\{ |g(s,t)|,\ |g_y(s,t)|,\ |g_x(s,t)|,\ |g_{xy}(s,t)| \ :\ (s,t) \in [\alpha, \beta] \times [\gamma, \eta] \right\},$$

the bound (26) yields the following bound for the sum of the four integrands (21)–(24):

$$|\langle g\mathcal{X}_{\mathbb{M}},\ f_n - v \rangle| \leq \Gamma[1 + (\eta - \gamma) + (\beta - \alpha) + (\eta - \gamma)(\beta - \alpha)]\,[8\|a\|\bar{\epsilon}]\,.$$

Given any $\epsilon > 0$ there exists an $\bar{\epsilon} > 0$ such that the left-hand side, and so also the right-hand side, of this inequality is less than ϵ; moreover, for this $\bar{\epsilon}$ there is an n' such that

$$\frac{1}{n} \max\{(\eta - \gamma), (\beta - \alpha)\} < \delta(\bar{\epsilon}) \qquad \forall\, n \geq n'.$$

Therefore, for all $n \geq n'$ we have $|\langle g\mathcal{X}_{\mathbb{M}},\ f_n - v \rangle| \leq \epsilon$, which is what we set out to show. Since functions of the form $g\mathcal{X}_{\mathbb{M}}$, where $g \in C^\infty[\mathbb{R}^2, \mathbb{R}^2]$ and $\mathbb{M} \subset \mathbb{R}^2$ is a box, are dense in $L_2^2(\mathbb{R}^2, \mathcal{M}^2, P)$ we have that the sequence f_n converges weakly to v. ∎

3.4. Application to wavefront reconstruction. We now apply the above results to the weighted negative modulus mapping $-b(\cdot)|u(\cdot)|$.

PROPOSITION 3.11 (integrals of projections and subgradients). *Let $b \in \mathbb{U}$ be a density function for the probability measure $P(dx) = b(x)dx$ on $(\mathbb{R}^2, \mathcal{M}^2)$ and let $u \in L^2[\mathbb{R}^2, \mathbb{R}^2]$. The negative modulus function $-|u(x)|$ has the following properties:*

(i) *$\mathcal{S}\left(b(\cdot)\overline{\partial}(-|u(\cdot)|)\right)$ is a weakly compact, convex set in $L_2^2(\mathbb{R}^2, \mathcal{M}^2, \nu_2)$;*

(ii) *$\int \partial(-|u(x)|)b(x)dx = \int -\operatorname{conv}\left(\Pi_{\mathbb{S}}(u(x))\right) b(x)dx$, and $\int \partial(-|u(x)|)b(x)dx$ is a compact subset of $\mathbb{R}^2$;*

(iii) *$\mathcal{S}\left(b(\cdot)\overline{\partial}(-|u(\cdot)|)\right) \subset -\mathrm{cl}^*\left(\Pi_{\mathbb{Q}(b)}(u)\right)$ for all $u \in L^2[\mathbb{R}^2, \mathbb{R}^2]$, where $\mathbb{Q}(b)$ is defined by (7) and $\Pi_{\mathbb{Q}(b)}(u)$ by (8).*

Proof. (i) At each x, $b(x)\overline{\partial}(-|u(x)|)$ is closed and convex-valued. In addition, by Example 3.7 every element of the set $\overline{\partial}(-|u(x)|)$ has magnitude less than or equal to 1 and so the multifunction $b(\cdot)\overline{\partial}(-|u(\cdot)|)$ is L^2-bounded in $(\mathbb{R}^2, \mathcal{M}^2, \nu_2)$. Hence, by Proposition 2.2, the multifunction $\mathcal{S}\left(b(x)\overline{\partial}(-|u(x)|)\right)$ is weakly compact in $L_2^2(\mathbb{R}^2, \mathcal{M}^2, \nu_2)$.

(ii) We wish to apply Theorem 3.9, so we must show that the multifunction F written as the composition of a multifunction with a measurable function

$$F(x) = [\partial(-|\cdot|) \circ u](x) = \partial(-|u(x)|)$$

is P-integrably bounded and $\mathcal{M}^2 \otimes \mathcal{B}^2$-measurable. By Example 3.7, the multifunction $F : \mathbb{R}^2 \rightrightarrows \mathbb{R}^2$ is P-integrably bounded with bound equal to 1. By Remark 3.5 $\partial(-|\cdot|) : \mathbb{R}^2 \rightrightarrows \mathbb{R}^2$ has closed graph and is therefore $\mathcal{M}^2 \otimes \mathcal{B}^2$-measurable. By hypothesis, the function u is a Lebesgue measurable mapping from $(\mathbb{R}^2, \mathcal{M}^2)$ into $(\mathbb{R}^2, \mathcal{M}^2)$. Thus, by [11, Proposition 1.b, p. 59] the composite multifunction F defined above is $\mathcal{M}^2 \otimes \mathcal{B}^2$-measurable. Therefore Theorem 3.9 applies to give the result.

(iii) By Proposition 3.10 every $v(\cdot) \in \mathcal{S}\left(b(\cdot)\overline{\partial}(-|u(\cdot)|)\right)$ is the weak limit of a sequence of functions in $\mathcal{S}\left(b(\cdot)\partial(-|u(\cdot)|)\right)$, since $\operatorname{conv}\left(\partial(-|u(\cdot)|)\right) = \overline{\partial}(-|u(\cdot)|)$ (see Example 3.7). If $v \in \mathcal{S}\left(b(\cdot)\partial(-|u(\cdot)|)\right)$, then by [16, Theorem 4.2] and Proposition 3.8 $-v \in \mathcal{S}(b(\cdot)\Pi_{\mathbb{S}}(u(\cdot)))$. Hence, by (11),

$$\mathcal{S}\left(b(\cdot)\partial(-|u(\cdot)|)\right) \subset -\Pi_{\mathbb{Q}(b)}(u),$$

from which the result follows. □

3.5. Infinite-dimensional nonsmooth analysis. The next step is to relate the subdifferential of the integral to the integral of the subdifferential. We begin with a brief review of infinite-dimensional nonsmooth analysis. For a complete discussion see [5, 6, 12, 13, 14, 15, 17, 18] and the references therein. To begin with, let $df(u)$ and $\widehat{d}f(u)$ be defined in exactly the same way that they were defined in the finite-dimensional setting in Definition 3.3.

DEFINITION 3.12 (subgradients: infinite-dimensions). *Let $\mathbb{X}$ be a separable Hilbert space, let $f : \mathbb{X} \to \overline{\mathbb{R}}$ be locally Lipschitz continuous, and let $u_* \in \operatorname{dom} f$.*

(i) *A vector $v \in \mathbb{X}^*$ is a* Dini ϵ-subgradient *of f at u_* if*

$$\langle v, w\rangle \le df(u_*)(w) + \epsilon\|w\| \qquad \forall\, w \in \mathbb{X},$$

where $df(u_*)(w)$ is the infinite-dimensional version of the subderivative defined in Definition 3.3(i). *We call the set of Dini ϵ-subgradients v the* Dini ϵ-subdifferential *of f at u_* and denote this set by $\partial_\epsilon^- f(u_*)$. When $\epsilon = 0$, we write $\partial^- f(u_*)$ instead of $\partial_0^- f(u_*)$. By the definition of the subderivative function Definition* 3.3(i) *and the regular subgradient Definition* 3.4(i) *it can be shown that for Lipschitz f the Dini 0-subdifferential is simply the infinite-dimensional version of the regular subgradient, $\partial^- f(u_*) = \widehat{\partial} f(u_*)$.*

(ii) *A vector $v \in \mathbb{X}^*$ is a* subgradient *of f at u_* if there are sequences $\epsilon^{(\nu)} \searrow 0$, $u^{(\nu)} \to u_*$, and $v^{(\nu)} \in \partial_\epsilon^- f(u^{(\nu)})$ with $v^{(\nu)} \overset{w^*}{\to} v$, where $\overset{w^*}{\to}$ denotes weak-star convergence. We call the set of subgradients v the* subdifferential *of f at u_* and denote this set by $\partial f(u_*)$.*

(iii) *We define the Clarke generalized subdifferential, $\overline{\partial} f(u_*)$ of f at u_*, as in the finite-dimensional case, Definition* 3.4(iii).

(iv) *The function f is said to be* subdifferentially regular *at u_* if $\partial f(u_*) \neq \emptyset$ and*

$$\partial f(u_*) = \widehat{\partial} f(u_*).$$

Remark 3.13. This construction of the subdifferential comes from [14] where it is used the to construct the A-subdifferential, or approximate subdifferential. However,

due to the equivalence theorem of Mordukhovich and Shao [18, Theorem 9.2] it can also be used in the separable Hilbert space setting to define the subdifferential given in [15]. From Mordukhovich and Shao [18, Theorem 8.11], we also obtain the relation

$$\overline{\partial} f(u_*) = \mathrm{cl}^*\,(\mathrm{conv}\ \partial f(u_*)). \tag{27}$$

In particular, this implies that f is subdifferentiably regular at u_* if and only if

$$\overline{\partial} f(u_*) = \partial f(u_*).$$

In addition, when f is strictly differentiable, then $\partial f(u)$ coincides with the Fréchet derivative. Finally, we note that the sets $\partial f(u)$ are weakly closed.

Until now we have been concerned with the issue of when a subset of $\mathbb{R}^n$ depends measurably on the parameter $x \in \Omega$. It is equally important for us to consider the properties of measurable real-valued functions on $\mathbb{R}^n$. For this we make use of *normal* integrands as defined in [21, Definition 14.27]. A function $f : \Omega \times \mathbb{R}^n \to \overline{\mathbb{R}}$ is called a normal integrand if its epigraphical mapping $\mathrm{epi}\, f(x, \cdot\,)$, $x \in \Omega$, is closed-valued and measurable. Any autonomous, Lipschitz continuous mapping, i.e., $f(x, u) := g(u)$, where $g : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz, is a normal integrand [21, Example 14.30]. For example, the mapping $|u|$ is a normal integrand. We use normal integrands to prove the measurability of the following important mappings.

Lemma 3.14 (measurability of exposed faces). *Consider a closed-valued Lebesgue measurable multifunction $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$. For $x \in \mathbb{R}^m$ and $w \in \mathbb{R}^n$ define $F_* : \mathbb{R}^m \times \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ by*

$$F_*(x, w) = \mathrm{argmax}\,\{\langle v\,,\, w\rangle \mid v \in F(x)\}\,.$$

Then F_* is closed-valued and Lebesgue measurable.

Remark 3.15. Whenever the set $F_*(x, w)$ is nonempty it is called an exposed face of the convex set $F(x)$ [22, section 18]. It is easily shown that these sets are indeed *faces* of $F(x)$ in the sense of [22, section 18]. Here we have focused on Lebesgue measure, but other σ-finite complete measures are possible.

Proof. Since F is closed-valued and measurable, [21, Example 14.32] implies that the function $f : (\mathbb{R}^m \times \mathbb{R}^n) \times \mathbb{R}^n \to \overline{\mathbb{R}}$ given by

$$f(x, w, v) = \langle v\,,\, -w\rangle + \delta_{F(x)}(v)$$

is a normal integrand. Hence the result follows from [21, Theorem 14.37] since

$$F_*(x, w) = \mathrm{argmin}\, f(x, w, v). \qquad \square$$

We remark that if, in addition, F is compact-valued, then so is F_*.

Lemma 3.16 (subgradients of normal integrands [21, Theorem 14.56]). *Let $(\Omega, \mathcal{A}, \mu)$ be a complete measure space. For the proper normal integrand $f : \Omega \times \mathbb{R}^n \to \mathbb{R}$, and any $u(x) \in \mathrm{dom}\, f(x, \cdot\,)$ depending measurably on $x \in \Omega$, the subderivative functions*

$$(x, w) \mapsto \widehat{d} f(x, u(x))(w), \qquad (x, w) \mapsto df(x, u(x))(w)$$

are normal integrands and the subdifferential mappings

$$x \mapsto \widehat{\partial} f(x, u(x)), \qquad x \mapsto \partial f(x, u(x))$$

are closed-valued and measurable.

In the remainder of this section, whenever we speak of measure we will be referring to Lebesgue measure.

LEMMA 3.17 (measurable selections for the regular subderivative). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz and let $u : \mathbb{R}^m \to \mathbb{R}^n$ and $w : \mathbb{R}^m \to \mathbb{R}^n$ be measurable mappings. Then the subdifferential mapping $\overline{\partial} f(u(\cdot))$ is measurable and possesses a measurable selection $v : \mathbb{R}^m \to \mathbb{R}^n$ such that*

$$\langle v(x)\,,\, w(x)\rangle = \widehat{d}f(u(x))(w(x)) \qquad \text{a.e. } x \in \mathbb{R}^m. \tag{28}$$

Proof. By [21, Theorem 14.56] the mapping ∂f is measurable. Since $\overline{\partial} f(u)$ is simply the convex hull of $\partial f(u)$ for all $u \in \mathbb{R}^n$, [21, Exercise 14.12] implies that $\overline{\partial} f$ is compact convex-valued and measurable. Hence, by [21, Theorem 14.13], the mapping $\overline{\partial} f(u(\cdot))$ is also compact convex-valued and measurable. It remains to establish the existence of a measurable selection satisfying (28).

By [21, Theorem 8.49], we have $\widehat{d}f(u)(w) = \sup\langle \overline{\partial} f(u)\,,\, w\rangle$ for all $w \in \mathbb{R}^n$, and we have shown that the mapping $\overline{\partial} f$ is compact convex-valued and measurable. Therefore, by Lemma 3.14, the mapping

$$F_*(u, w) = \operatorname{argmax}\left\{\langle v\,,\, w\rangle \;\middle|\; v \in \overline{\partial} f(u)\right\}$$

is also compact convex-valued and measurable with

$$\operatorname{dom}(F_*) = \{(u, w) \mid F_*(u, w) \neq \emptyset\} = \mathbb{R}^n \times \mathbb{R}^n.$$

Again, by [21, Theorem 14.13], the mapping $F_*(u(\cdot), w(\cdot))$ is also compact convex-valued and measurable. The measurable selection theorem Theorem 2.1 now implies the existence of a measurable function $v(\cdot)$ such that $v(x) \in F_*(u(x), w(x))$ a.e., which proves the lemma. ∎

We now have our first general result on the interchange of integration and subdifferentiation.

LEMMA 3.18 (interchange of subdifferentiation and integration. I). *Let $\mathcal{H} = L^2_m(\mathbb{R}^n, \mathcal{M}^n, \nu_n)$ be the Hilbert space of square integrable functions mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$ defined in section 2.2, where $\mathcal{M}^n$ is the σ-field of Lebesgue measurable sets on $\mathbb{R}^n$ and ν_n is Lebesgue measure. For simplicity, we write $dx = \nu_n(dx)$. Let $f : \mathbb{R}^m \to \mathbb{R}$ be globally Lipschitz continuous with Lipschitz constant K, and suppose there exists $\hat{u} \in \mathcal{H}$ such that $f \circ \hat{u}$ is an L^2-bounded function on the space $(\mathbb{R}^n, \mathcal{M}^n, \mu)$ where $\mu = b\nu_n$, where $b : \mathbb{R}^n \to \mathbb{R}_+$ with $b \in L^1 \cap L^2 \cap L^\infty[\mathbb{R}^n, \mathbb{R}]$. Define the integral functional $J : \mathcal{H} \to \overline{\mathbb{R}}$ by*

$$J(u) = \int f(u(x))b(x)dx.$$

Then J is globally Lipschitz with Lipschitz constant $K\|b\|_2$, and for every $u \in \mathcal{H}$ the mapping $f \circ u$ is L^2-bounded and

$$\overline{\partial} J(u) \subset \mathcal{S}(b(\cdot)\overline{\partial} f(u(\cdot))). \tag{29}$$

Proof. Let $u \in \mathcal{H}$. The fact that $f \circ u$ is L^2-bounded follows immediately from the inequality

$$|f(u(x))| \leq |f(\hat{u}(x))| + K|u(x) - \hat{u}(x)|.$$

The global Lipschitz continuity of J is a consequence of the following derivation:

$$\begin{aligned}|J(u) - J(v)| &\leq \int K|u(x) - v(x)|b(x)dx \\ &= K\langle |u - v|,\, b\rangle \\ &\leq K\|b\|_2\|u - v\|_2.\end{aligned}$$

Remark 3.13 tells us that $\overline{\partial}J(u)$ is a weakly compact convex subset of $\mathcal{H}$ for all $u \in \mathcal{H}$. We also have from Proposition 2.2 that the set $\mathcal{S}(b(\cdot)\overline{\partial}f(u(\cdot)))$ is also a weakly compact convex subset of $\mathcal{H}$ for all $u \in \mathcal{H}$. Hence the inclusion (29) follows if it can be shown that

$$\sup\left\{\langle v\,,\, w\rangle \;\middle|\; v \in \mathcal{S}(b(\cdot)\overline{\partial}f(u(\cdot)))\right\} \geq \widehat{d}J(u)(w)$$

for all $w \in \mathcal{H}$.

Let $w \in \mathcal{H}$ and let $\{u_i\} \subset \mathcal{H}$ and $\{\tau_i\} \subset \mathbb{R}_+$ be such that $\{u_i\}$ strongly converges to u and $\tau_i \downarrow 0$ with

$$\widehat{d}J(u)(w) = \lim_{i\to\infty} \frac{J(u_i + \tau_i w) - J(u_i)}{\tau_i}.$$

Then, by Fatou's lemma,

$$\begin{aligned}\widehat{d}J(u)(w) &= \lim_{i\to\infty} \int \frac{f(u_i(x) + \tau_i w(x)) - f(u_i(x))}{\tau_i} b(x)dx \\ &\leq \int \limsup_{i\to\infty} \frac{f(u_i(x) + \tau_i w(x)) - f(u_i(x))}{\tau_i} b(x)dx \\ &\leq \int \widehat{d}f(u(x))(w(x))b(x)dx.\end{aligned} \tag{30}$$

By Lemma 3.17, the multifunction $\overline{\partial}f(u(\cdot))$ possesses a measurable selection v such that $\widehat{d}f(u(x))(w(x)) = \langle v(x)\,,\, w(x)\rangle$ a.e. on $\mathbb{R}^n$. Therefore, by (30) we have

$$\begin{aligned}\widehat{d}J(u)(w) &\leq \int \langle v(x)\,,\, w(x)\rangle\, b(x)dx \\ &\leq \sup\left\{\langle v\,,\, w\rangle \;\middle|\; v \in \mathcal{S}(b(\cdot)\overline{\partial}f(u(\cdot)))\right\},\end{aligned}$$

proving the result. □

3.6. Application to wavefront reconstruction. In the next proposition we establish the connection between the projection $\Pi_{\mathbb{Q}(b)}$ defined by (8) and the subdifferential of $h : L^2[\mathbb{R}^2, \mathbb{R}^2] \to \mathbb{R}$ defined by (18), where $b \in \mathbb{U}$ with $\mathbb{U}$ defined in (6). Proposition 3.19 is a special case of a more general result to be proved in the final section (Theorem 4.2). However, here we provide a separate and fundamentally different proof which provides the motivation for the perturbation methods studied in [16].

PROPOSITION 3.19 (projection-subdifferential equivalence). *Let $b \in \mathbb{U}$ and let $\Pi_{\mathbb{Q}(b)} : L^2 \rightrightarrows \mathbb{Q}(b)$ be as defined in* (8) *with $\mathbb{Q}(b)$ defined by* (7)*, and $h : L^2[\mathbb{R}^2, \mathbb{R}^2] \to \overline{\mathbb{R}}$ be as defined by* (18)*. Then for all $u \in L^2[\mathbb{R}^2, \mathbb{R}^2]$*

$$\overline{\partial}\left(h(u; b)\right) = \mathcal{S}\left(b(\cdot)\overline{\partial}\left(-|u(\cdot)|\right)\right) = \mathrm{cl}^*\left(-\Pi_{\mathbb{Q}(b)}(u)\right) = \partial\left(h(u; b)\right). \tag{31}$$

Thus, in particular, $h(\cdot; b)$ is everywhere subdifferentiably regular.

Proof. Note that the equivalences in (31) are scale invariant in the sense that if they are shown to be true for a given function b, then they must be true with b replaced by αb for any choice of $\alpha > 0$ since

$$\alpha\partial(h(u;b)) = \partial(h(u;\alpha b)), \qquad \alpha\mathcal{S}(b(\cdot)\bar{\partial}(-|u(\cdot)|)) = \mathcal{S}(\alpha b(\cdot)\bar{\partial}(-|u(\cdot)|)),$$

and

$$\alpha \operatorname{cl}^*(-\Pi_{Q(b)}(u)) = \operatorname{cl}^*(-\Pi_{Q(\alpha b)}(u)).$$

Since b is nonnegative and integrable, we may therefore assume with no loss in generality that b is a probability density function for some probability measure $P(dx) = b(x)dx$.

If (31) holds, then the subdifferential regularity of $h(\cdot;b)$ follows immediately from Proposition 3.11(i) and (27). By Lemma 3.18 and part (iii) of Proposition 3.11,

$$\overline{\partial}h(u;b) \subset \mathcal{S}\left(b(\cdot)\overline{\partial}(-|u(\cdot)|)\right) \subset \operatorname{cl}^*\left(-\Pi_{\mathbb{Q}(b)}(u)\right).$$

Since $\partial h(u;b) \subset \overline{\partial}h(u;b)$, the result follows once it is shown that

$$\operatorname{cl}^*\left(-\Pi_{\mathbb{Q}(b)}(u)\right) \subset \partial h(u;b). \tag{32}$$

By Proposition 3.2 the mapping h is globally Lipschitz continuous with Lipschitz constant $K = \|b\|$, and by Remark 3.13 $\partial h(u;b)$ is weakly closed. Therefore, if $-\Pi_{\mathbb{Q}(b)}(u) \subset \partial h(u;b)$, then $\operatorname{cl}^*\left(-\Pi_{\mathbb{Q}(b)}(u)\right) \subset \partial h(u;b)$. We now show that $-\Pi_{\mathbb{Q}(b)}(u) \subset \partial h(u;b)$.

Let $v \in -\Pi_{\mathbb{Q}(b)}(u)$ and for all $\epsilon > 0$ define $\tilde{u}_\epsilon := u\mathcal{X}_{\operatorname{supp}(u)} + \epsilon v(1 - \mathcal{X}_{\operatorname{supp}(u)})$. Then, by [16, Theorem 4.1],

$$\|u - \tilde{u}_\epsilon\| = \epsilon\|v(1 - \mathcal{X}_{\operatorname{supp}(u)})\| \leq \epsilon\|b\|,$$

and $|\cdot|$ is differentiable at $\tilde{u}_\epsilon(x)$ for every $x \in \operatorname{supp}(b)$ with

$$v(x) = -\nabla|\tilde{u}_\epsilon(x)|b(x) \qquad \forall\, \epsilon > 0.$$

For every $w \in L^2[\mathbb{R}^2, \mathbb{R}^2]$ and $x \in \operatorname{supp}(b)$, we have

$$\frac{|\tilde{u}_\epsilon(x) + tw(x)| - |\tilde{u}_\epsilon(x)|}{t} \to (\nabla|\tilde{u}_\epsilon(x)|, w(x)),$$

and, since $|\cdot|$ is Lipschitz with Lipschitz constant 1,

$$\left|\frac{|\tilde{u}_\epsilon(x) + tw(x)| - |\tilde{u}_\epsilon(x)|}{t}\right| \leq |w(x)| \qquad \forall\, x \in \operatorname{supp}(b).$$

Therefore, by the Lebesgue dominated convergence theorem, the function $h(\cdot;b)$ is Gâteaux differentiable at $\tilde{u}_\epsilon$ with Gâteaux derivative $-\nabla|\tilde{u}_\epsilon|b = v$. Hence, since $|\cdot|$ is Lipschitz continuous the lim inf in Definition 3.3(ii) is attained as a limit yielding $dh(\tilde{u}_\epsilon;b)(w) = \langle v, w\rangle$. Consequently

$$v \in \partial^- h(\tilde{u}_\epsilon;b) \qquad \forall\, \epsilon > 0.$$

Taking the limit as $\epsilon \downarrow 0$, we find that $v \in \partial h(u;b)$. Therefore, $-\Pi_{\mathbb{Q}(b)}(u) \subset \partial h(u;b)$. ∎

The proof of Theorem 3.1 now follows easily from the calculus of subdifferentials.

Proof of Theorem 3.1. [16, Corollary 4.3] gives the representation

$$\begin{aligned}\operatorname{dist}^2(u, \mathbb{Q}_m) &= \operatorname{dist}^2(\mathcal{F}_m(u), \mathbb{Q}(\psi_m)) \\ &= \|\mathcal{F}_m(u)\|^2 + \|\psi_m\|^2 + 2h[\mathcal{F}_m(u); \psi_m].\end{aligned}$$

By applying [18, Theorem 6.7] together with Proposition 3.19 and [16, Corollary 4.3], we obtain

$$\begin{aligned}\partial \operatorname{dist}^2(\mathcal{F}_m(u), \mathbb{Q}(\psi_m)) &= 2\partial\left(\left(\frac{1}{2}\|\cdot\|^2 + h(\cdot; \psi_m)\right) \circ \mathcal{F}_m\right)(u) \\ &= 2\mathcal{F}_m^*\left[\mathcal{F}_m(u) + \operatorname{cl}^*\left(-\Pi_{\mathbb{Q}_m}(\mathcal{F}_m(u))\right)\right] \\ &= 2\operatorname{cl}^*\left(\mathcal{I} - \Pi_{\mathbb{Q}_m}(u)\right).\end{aligned}$$

Hence the subdifferential regularity of all the functions involved in conjunction with [18, Theorem 4.1] yields the result. ∎

4. Concluding remarks. We conclude with a generalization of Theorem 3.19. Theorem 4.2 establishes the equivalence of the infinite-dimensional subdifferential objects in the setting relevant to phase retrieval and establishes their relation to the finite-dimensional Clarke subdifferential. The result, and its proof, closely parallels that given in [6, Theorem 3.5.18].

LEMMA 4.1 (interchange of subdifferentiation and integration. II). *Let the hypotheses of Lemma* 3.18 *hold. Then*

$$\mathcal{S}(b(\cdot)\partial f(u(\cdot))) \subset \partial J(u). \tag{33}$$

Proof. Let $z \in \mathcal{S}(b(\cdot)\partial f(u(\cdot)))$. Since $\partial f(u(\cdot))$ is closed-valued and measurable, there exists $v \in \mathcal{S}(\partial f(u(\cdot)))$ for which $z = bv$. We show that $z \in \partial J(u)$. For this purpose, let C be a countably dense subset of $\operatorname{gph} \widehat{\partial} f$. Observe that

$$\partial f(u) = \left\{ \lim_{j\to\infty} v^j \;\middle|\; \{(u^j, v^j)\} \subset C,\, u^j \to u \right\}.$$

Let $\{(u^k, v^k)\}$ be an enumeration of C. Then for each $x \in \mathbb{R}^n$ and each integer $i \in \{1, 2, \dots\}$, define $k_i(x)$ be the first integer k for which

$$|u^k - u(x)| \le \frac{1}{i} \quad \text{and} \quad |v^k - v(x)| \le \frac{1}{i}.$$

For each $i = 1, 2, \dots$, define $u^i : \mathbb{R}^n \to \mathbb{R}^m$ and $v^i : \mathbb{R}^n \to \mathbb{R}^m$ by

$$u^i(x) = u^{k_i(x)} \quad \text{and} \quad v^i(x) = v^{k_i(x)}.$$

We claim that the functions u^i and v^i are measurable with

$$v^i(x) \in \widehat{\partial} f(u^i(x)) \text{ a.e.} \tag{34}$$

for $i = 1, 2, \dots$. Indeed, the range of both u^i and v^i is contained in the set C and so is countable. Moreover, for a given integer k,

$$\begin{aligned}&\left\{x \;\middle|\; (u^i(x), v^i(x)) = (u^k, v^k)\right\} \\ &= \left[\bigcap_{j=1}^{k-1} \left\{x \;\middle|\; \max\{|u^j - u(x)|, |v^j - v(x)|\} > \frac{1}{i}\right\}\right] \\ &\quad \cap \left\{x \;\middle|\; \max\{|u^k - u(x)|, |v^k - v(x)|\} \le \frac{1}{i}\right\},\end{aligned}$$

where each of the sets on the left-hand side is measurable.

Next observe that for all $w \in \mathcal{H}$, we have from Fatou's lemma that

$$dJ(u^i)(w) = \liminf_{\tau \searrow 0} \frac{J(u^i + \tau w) - J(u^i)}{\tau} \geq \int_{\mathbb{R}^2} df(u^i(x))(w(x))b(x)dx \geq \left\langle bv^i \, , \, w \right\rangle,$$

where the last inequality follows from (34). Hence $bv^i \in \widehat{\partial} J(u^i)$ for $i = 1, 2, \ldots$. Finally, since $u^i \to u$ and $v^i \to v$ by construction, we have $bv \in \partial J(u)$. ∎

THEOREM 4.2 (interchange of subdifferentiation and integration). *Let the hypotheses of Lemma* 3.18 *hold with* $n = m = 2$. *Then, for all* $u \in \mathcal{H} = L^2[\mathbb{R}^2, \mathbb{R}^2]$,

$$\partial J(u) = \mathrm{cl}^* \mathcal{S}(b(\cdot)\partial f(\cdot)) = \mathcal{S}(b(\cdot)\overline{\partial} f(\cdot)) = \overline{\partial} J(u).$$

In particular, this implies that J *is everywhere subdifferentially regular.*

Proof. By Proposition 3.10 we have

$$\mathcal{S}(b(\cdot)\overline{\partial} f(u(\cdot))) \subset \mathrm{cl}^* \, \mathcal{S}(b(\cdot)\partial f(u(\cdot))).$$

Since the set $\partial J(u)$ is weakly closed, Lemma 4.1 implies that

$$\mathrm{cl}^* \, \mathcal{S}(b(\cdot)\partial f(u(\cdot))) \subset \partial J(u).$$

Combining these facts with Lemma 3.18 yields

$$\begin{aligned} \overline{\partial} J(u) &\subset \mathcal{S}(b(\cdot)\overline{\partial} f(u(\cdot))) \\ &\subset \mathrm{cl}^* \, \mathcal{S}(b(\cdot)\partial f(u(\cdot))) \\ &\subset \partial J(u) \\ &\subset \overline{\partial} J(u), \end{aligned}$$

which proves the result. ∎

The restriction in Theorem 4.2 to the case $n = m = 2$ follows from the use of this hypothesis in Proposition 3.10. However, we believe that it is possible to extend this proposition to the general case, which would allow us to remove the restriction $n = m = 2$ from Theorem 4.2.

REFERENCES

[1] R. J. AUMANN, *Integrals of set-valued functions*, J. Math. Anal. Appl., 12 (1965), pp. 1–12.

[2] H. BAUSCHKE AND J. BORWEIN, *Dykstra's alternating projection algorithm for two sets*, J. Approx. Theory, 79 (1994), pp. 418–443.

[3] H. H. BAUSCHKE AND J. M. BORWEIN, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.

[4] H. H. BAUSCHKE, P. L. COMBETTES, AND D. R. LUKE, *Phase retrieval, error reduction algorithm, and Fienup variants: A view from convex optimization*, J. Opt. Soc. Amer. A, 19 (2002), pp. 1334–1345.

[5] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5, SIAM, Philadelphia, 1990.

[6] F. H. CLARKE, YU. S. LEDYAEV, R. J. STERN, AND P. R WOLENSKI, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.

[7] P. L. COMBETTES, *Inconsistent signal feasibility problems: Least-squares solutions in a product space*, IEEE Trans. Signal Process., 42 (1994), pp. 2955–2966.

[8] M. M. DAY, *Normed Linear Spaces*, 3rd ed., Springer-Verlag, New York, 1973.

[9] J. FIENUP, *Phase retrieval algorithms: A comparison.* Appl. Optim., 21 (1982), pp. 2758–2769.

[10] R. GERCHBERG AND W. SAXTON, *A practical algorithm for the determination of phase from image and diffraction plane pictures*, Optik, 35 (1972), pp. 237–246.

[11] W. Hildenbrand, *Core and Equilibria of a Large Economy*, Princeton University Press, Princeton, NJ, 1974.
[12] A. D. Ioffe, *Approximate subdifferentials and applications:* II, Mathematika, 33 (1986), 111–128.
[13] A. D. Ioffe, *Approximate subdifferentials and applications:* III, Mathematika, 36 (1989), pp. 1–38.
[14] A. D. Ioffe, *Proximal analysis and approximate subdifferentials*, J. London Math. Soc., 41 (1990), pp. 175–192.
[15] A. Y. Kruger and B. S. Mordukhovich, *Extremal points and the Euler equation in nonsmooth optimization*, Dokl. Akad. Nauk BSSR, 24 (1980), pp. 684–687.
[16] D. R. Luke, J. V. Burke, and R. G. Lyon, *Optical wavefront reconstruction: Theory and numerical methods*, SIAM Rev., 44 (2002), pp. 169–224.
[17] B. S. Mordukhovich, *Maximum principle in problems of time optimal control with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.
[18] B. S. Mordukhovich and Y. Shao, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., 328 (1996), pp. 1235–1280.
[19] R. T. Rockafellar, *Conjugate Duality and Optimization*, SIAM, Philadelphia, 1974.
[20] R. T. Rockafellar, *Integral functionals, normal integrands and measurable selections*, in Nonlinear Operators in the Calculus of Variations, Lecture Notes in Math. 543, Springer-Verlag, New York, 1976, pp. 157–207.
[21] R. T. Rockafellar and R. J. Wets, *Variational Analysis*, Springer-Verlag, New York, 1998.
[22] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[23] W. Rudin, *Real and Complex Analysis*, 2nd ed., McGraw–Hill, San Francisco, CA, 1974.
[24] L. P. Vlasov, *Approximative properties of sets in normed linear spaces*, Russian Math. Surveys, 28 (1973), pp. 1–66.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 596–636

BREGMAN MONOTONE OPTIMIZATION ALGORITHMS*

HEINZ H. BAUSCHKE†, JONATHAN M. BORWEIN‡, AND PATRICK L. COMBETTES§

Abstract. A broad class of optimization algorithms based on Bregman distances in Banach spaces is unified around the notion of Bregman monotonicity. A systematic investigation of this notion leads to a simplified analysis of numerous algorithms and to the development of a new class of parallel block-iterative surrogate Bregman projection schemes. Another key contribution is the introduction of a class of operators that is shown to be intrinsically tied to the notion of Bregman monotonicity and to include the operators commonly found in Bregman optimization methods. Special emphasis is placed on the viability of the algorithms and the importance of Legendre functions in this regard. Various applications are discussed.

Key words. Banach space, block-iterative method, Bregman distance, Bregman monotone, Bregman projection, $\mathfrak{B}$-class operator, convex feasibility problem, essentially smooth function, essentially strict convex function, Fejér monotone, Legendre function, monotone operator, proximal mapping, proximal point algorithm, resolvent, subgradient projection

AMS subject classifications. 90C25, 90C48, 47H05

PII. S0363012902407120

1. Introduction.

A sequence $(x_n)_{n\in\mathbb{N}}$ in a Banach space $\mathcal{X}$ is *Fejér monotone* with respect to a set $S \subset \mathcal{X}$ if

$$(\forall x \in S)(\forall n \in \mathbb{N}) \quad \|x_{n+1} - x\| \leq \|x_n - x\|. \tag{1.1}$$

In Hilbert spaces, this notion has proven to be remarkably useful and successful in attempts to unify and harmonize the convergence proofs of a large number of optimization algorithms; see, e.g., [5, 6, 9, 40, 41, 49, 60]. A classical example is the method of cyclic projections for finding a point in the intersection $S \neq \emptyset$ of a finite family of closed convex sets $(S_i)_{1\leq i\leq m}$. In 1965, Bregman [14, Thm. 1] showed that for every initial point $x_0 \in \mathcal{X}$ the sequence $(x_n)_{n\in\mathbb{N}}$ generated by the cyclic projections algorithm

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = P_{n \,(\mathrm{mod}\, m)+1} x_n, \tag{1.2}$$

where P_i denotes the metric projector onto S_i and where the mod m function takes values in $\{0, \ldots, m-1\}$, is Fejér monotone with respect to S and converges weakly to a point in that set. Two years later [15], the same author investigated the convergence of this method in a general topological vector space $\mathcal{X}$. To this end, he introduced a distance-like function $D\colon E \times E \to \mathbb{R}$, where E is a convex subset of $\mathcal{X}$ such that $S = E \cap \bigcap_{i=1}^m S_i \neq \emptyset$. The conditions defining D require, in particular, that for

*Received by the editors May 6, 2002; accepted for publication (in revised form) January 8, 2003; published electronically May 29, 2003.
http://www.siam.org/journals/sicon/42-2/40712.html

†Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario N1G 2W1, Canada (hbauschk@uoguelph.ca). This author's research was supported by the Natural Sciences and Engineering Research Council of Canada.

‡Centre for Experimental & Constructive Mathematics, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada (jborwein@cecm.sfu.ca). This author's research was supported by the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chair Programme.

§Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie – Paris 6, 75005 Paris, France (plc@math.jussieu.fr).

every $i \in \{1, \dots, m\}$ and every $y \in E$, there exists a point $P_i y \in E \cap S_i$ such that $D(P_i y, y) = \min D(E \cap S_i, y)$. In this broader context, Bregman showed that for every initial point $x_0 \in E$ the cyclic projections algorithm (1.2) produces a sequence that satisfies the monotonicity property

$$(\forall x \in S)(\forall n \in \mathbb{N}) \quad D(x, x_{n+1}) \leq D(x, x_n) \tag{1.3}$$

and whose cluster points are in S [15, eq. (1.2) and Thm. 1]. If $\mathcal{X}$ is a Hilbert space, an example of a D-function satisfying the required conditions relative to the weak topology is $D\colon \mathcal{X}^2 \to \mathbb{R}\colon (x, y) \mapsto \|x - y\|^2/2$. In this case, we recover the previous convergence result [15, Example 1] and observe that (1.3) reduces to (1.1). If $\mathcal{X}$ is the Euclidean space $\mathbb{R}^N$, another example of a suitable D-function is

$$D\colon E \times E \to \mathbb{R}\colon (x, y) \mapsto f(x) - f(y) - \langle x - y, \nabla f(y) \rangle, \tag{1.4}$$

where $f\colon E \subset \mathbb{R}^N \to \mathbb{R}$ is a convex function which is differentiable on E and satisfies a set of auxiliary properties [15, Example 2]. Due to its importance in applications, this particular type of D-function was further studied in [30] and has since been known as a *Bregman distance* (see [33] for an historical account). In $\mathbb{R}^N$, various investigations have focused on the use of Bregman distances in projection, proximal point, and fixed point algorithms; see [7, 31, 32, 33, 46, 47, 83]. (See also [58, 59], where extensions of (1.4) to nondifferentiable functions were studied.) Extensions to Hilbert [18, 20, 61] and Banach [1, 8, 21, 23, 24, 25, 26, 27, 55, 56, 75] spaces have also been considered more recently. In the present paper, we adopt the following definition for Bregman distances.

DEFINITION 1.1. *Let $\mathcal{X}$ be a real Banach space and let $f\colon \mathcal{X} \to]-\infty, +\infty]$ be a lower semicontinuous convex function which is Gâteaux-differentiable on* $\operatorname{int}\operatorname{dom} f \neq \emptyset$. *The* Bregman distance *(for brevity D-*distance*) associated with f is the function*

$$\begin{aligned} &D\colon \mathcal{X} \times \mathcal{X} \to [0, +\infty], \\ &(x, y) \mapsto \begin{cases} f(x) - f(y) - \langle x - y, \nabla f(y) \rangle & \text{if } y \in \operatorname{int}\operatorname{dom} f, \\ +\infty & \text{otherwise.} \end{cases} \end{aligned} \tag{1.5}$$

In addition, the Bregman distance to a set $C \subset \mathcal{X}$ is the function

$$\begin{aligned} D_C\colon \mathcal{X} &\to [0, +\infty], \\ y &\mapsto \inf D(C, y). \end{aligned} \tag{1.6}$$

In Hilbert spaces, one recovers $D\colon (x, y) \mapsto \|x - y\|^2/2$ by setting $f = \|\cdot\|^2/2$. This observation suggests that the following natural variant of the notion of Fejér monotonicity suits the environment described in Definition 1.1.

DEFINITION 1.2. *A sequence $(x_n)_{n\in\mathbb{N}}$ in $\mathcal{X}$ is* Bregman monotone *(for brevity D-*monotone*) with respect to a set $S \subset \mathcal{X}$ if the following conditions hold:*

(i) $S \cap \operatorname{dom} f \neq \emptyset$.
(ii) $(x_n)_{n\in\mathbb{N}}$ *lies in* $\operatorname{int}\operatorname{dom} f$.
(iii) $(\forall x \in S \cap \operatorname{dom} f)(\forall n \in \mathbb{N})$ $D(x, x_{n+1}) \leq D(x, x_n)$.

Let us note that item (ii) is stated only for the sake of clarity and that it could be replaced by $x_0 \in \operatorname{int}\operatorname{dom} f$ since, in view of (1.5), (iii) then forces the whole sequence $(x_n)_{n\in\mathbb{N}}$ to lie in $\operatorname{int}\operatorname{dom} f$.

The importance of the notion of Bregman monotonicity is implicit in [15]. In the Euclidean space setting of [32] (see also [33, page 55]), Bregman monotone sequences were called "D_f Fejér monotone" by analogy with (1.1).

The goal of this paper is to provide a broad framework for the design and the analysis of algorithms based on Bregman distances around the notion of D-monotonicity. This framework not only will lead to a unified convergence analysis for existing algorithms, but also will serve as a basis for the development of a new class of parallel, block-iterative, surrogate Bregman projection methods for solving convex feasibility problems involving variational inequalities, convex inequalities, equilibrium constraints, and fixed point constraints. The tools developed in this paper also provide the main building blocks for the algorithms proposed in [10] to find best Bregman approximations from intersections of closed convex sets in reflexive Banach spaces.

Guide to the paper. We proceed towards our goal of constructing a broad framework for Bregman distance-based algorithms in several steps.

We collect assumptions, notation, and basic results in section 2. The standing assumptions on the underlying space $\mathcal{X}$ and the function f that generates the Bregman distance are stated in section 2.1. In sections 2.2–2.6, we introduce basic notation and terminology, including D-viable operators and Legendre functions. Useful identities for the Bregman distance are provided in section 2.7.

A general and powerful class of operators based on Bregman distances is introduced and analyzed in section 3. This so-called $\mathfrak{B}$-class includes types of operators fundamental in Bregman optimization such as D-firm operators, D-resolvents, D-prox operators, and (subgradient) D-projections, which correspond to their classical counterparts when $\mathcal{X}$ is a Hilbert space and $f = \|\cdot\|^2/2$. For example, it is shown that if $\mathcal{X}$ is reflexive and f is Legendre, then D-prox operators belong to $\mathfrak{B}$ (Corollary 3.25). This result underscores the importance of Legendreness. Moreover, $\mathfrak{B}$-class operators are stable under a certain type of parallel combination, which will be crucial in the formulation of a new block-iterative algorithmic framework in section 5.

Section 4 is devoted to D-monotonicity. This is a central notion in the analysis of Bregman optimization methods because it describes the behavior of a wide class of algorithms based on Bregman distances. Assumptions are given under which simple characterizations can be established for the weak and strong convergence of D-monotone sequences. In conjunction with the results of section 3, D-monotonicity provides a global framework for the development and analysis of algorithms. Indeed, we show that D-monotone sequences can be generated systematically via the iterative scheme

$$x_0 \in \operatorname{int}\operatorname{dom} f \;\text{ and }\; (\forall n \in \mathbb{N})\;\; x_{n+1} \in T_n x_n, \;\text{ where }\; T_n \in \mathfrak{B}. \tag{1.7}$$

A detailed convergence analysis of this unifying model is carried out which, in turn, covers and extends known convergence results.

Finally, in section 5, we are in a position to construct a new block-iterative algorithmic framework. Results obtained in sections 3 and 4 are combined to construct and investigate new classes of parallel, block-iterative methods for solving convex feasibility problems. The main result, Theorem 5.7, provides conditions sufficient for the weak and strong convergence of sequences generated by the new algorithm. Section 5.4 presents several scenarios in which these sufficient conditions are satisfied, including the frequently encountered situation when f is a separable Legendre function on $\mathbb{R}^N$ such that $\operatorname{dom} f^*$ is open (Example 5.14). The concluding sections, sections 5.5 and 5.6, discuss how the main result can be applied to specific optimization problems such as solving convex inequalities, finding common zeros of maximal monotone operators, finding common minimizers of convex function, and finding common fixed points of D-firm operators.

2. Notation, assumptions, and basic facts.

2.1. Standing assumptions. We assume throughout the paper that $\mathcal{X}$ is a real Banach space and that $f\colon \mathcal{X} \to]-\infty,+\infty]$ is a lower semicontinuous convex function which is Gâteaux-differentiable on $\operatorname{int}\operatorname{dom} f \neq \emptyset$.

2.2. Basic notation. Throughout, $\mathbb{N}$ is the set of nonnegative integers. The norm of $\mathcal{X}$ and that of its topological dual $\mathcal{X}^*$ is denoted by $\|\cdot\|$, the associated metric distance by d, and the canonical bilinear form on $\mathcal{X} \times \mathcal{X}^*$ by $\langle\cdot,\cdot\rangle$. (If $\mathcal{X}$ is a Hilbert space, $\langle\cdot,\cdot\rangle$ denotes also its scalar (or inner) product.) The metric distance function to a set $C \subset \mathcal{X}$ is $d_C\colon \mathcal{X} \to [0,+\infty] : y \mapsto \inf_{x\in C} \|x-y\|$ where, by convention, $\inf \emptyset = +\infty$. For every $y \in \operatorname{int}\operatorname{dom} f$, we set $f_y = f - \nabla f(y)$. The symbols $\rightharpoonup$, $\overset{*}{\rightharpoonup}$, and $\to$ denote, respectively, weak, weak$*$, and strong convergence. $\mathfrak{S}(x_n)_{n\in\mathbb{N}}$ and $\mathfrak{W}(x_n)_{n\in\mathbb{N}}$ are, respectively, the sets of strong and weak cluster points of a sequence $(x_n)_{n\in\mathbb{N}}$ in $\mathcal{X}$. $\operatorname{bdry} C$ denotes the boundary of a set $C \subset \mathcal{X}$, $\operatorname{int} C$ its interior, and $\overline{C}$ its closure. The closed ball of center x and radius ρ is denoted by $B(x;\rho)$. The normalized duality mapping J of $\mathcal{X}$ is defined by

$$(\forall x \in \mathcal{X}) \quad J(x) = \big\{x^* \in \mathcal{X}^* \mid \|x\|^2 = \langle x, x^*\rangle = \|x^*\|^2\big\}. \tag{2.1}$$

$\mathbb{R}^N$ is the standard N-dimensional Euclidean space.

2.3. Set-valued operators. Let $\mathcal{Y}$ be a Banach space and $2^{\mathcal{Y}}$ the family of all subsets of $\mathcal{Y}$. A set-valued operator from $\mathcal{X}$ to $\mathcal{Y}$ is an operator $A\colon \mathcal{X} \to 2^{\mathcal{Y}}$. It is characterized by its graph $\operatorname{gr} A = \{(x,u) \in \mathcal{X} \times \mathcal{Y} \mid u \in Ax\}$; its domain is $\operatorname{dom} A = \{x \in \mathcal{X} \mid Ax \neq \emptyset\}$ (with closure $\overline{\operatorname{dom}} A$); its range is $\operatorname{ran} A = \bigcup_{x\in\mathcal{X}} Ax$ (with closure $\overline{\operatorname{ran}} A$); and, if $\mathcal{Y} = \mathcal{X}$, its fixed point set is $\operatorname{Fix} A = \{x \in \mathcal{X} \mid x \in Ax\}$ (with closure $\overline{\operatorname{Fix}} A$). The graph of the inverse A^{-1} of A is $\{(u,x) \in \mathcal{Y} \times \mathcal{X} \mid (x,u) \in \operatorname{gr} A\}$. If $B\colon \mathcal{X} \to 2^{\mathcal{Y}}$ and $\alpha \in \mathbb{R}$, then $\operatorname{gr}(\alpha A + B) = \{(x, \alpha u + v) \in \mathcal{X} \times \mathcal{Y} \mid (x,u) \in \operatorname{gr} A, (x,v) \in \operatorname{gr} B\}$. As is customary, if $x \in \operatorname{dom} A$ and A is single-valued on $\operatorname{dom} A$, we shall denote the unique element in Ax by Ax. Finally, A is locally bounded at $x \in \mathcal{X}$ if there exists $\rho \in]0,+\infty[$ such that $A\big(B(x;\rho)\big)$ is bounded. (We adopt the same definition as in [79, section 17]; it differs slightly from Phelps' definition [71, Chap. 2] which requires $x \in \operatorname{dom} A$.)

2.4. Orbits and suborbits of algorithms. In section 4 and subsequent sections, we shall discuss various algorithms. Sequences generated by algorithms are called orbits, and their subsequences are referred to as suborbits.

2.5. Functions. The domain of a function $g\colon \mathcal{X} \to]-\infty,+\infty]$ is $\operatorname{dom} g = \{x \in \mathcal{X} \mid g(x) < +\infty\}$ (with closure $\overline{\operatorname{dom}} g$), and g is proper if $\operatorname{dom} g \neq \emptyset$. Moreover, g is subdifferentiable at $x \in \operatorname{dom} g$ if its subdifferential at this point,

$$\partial g(x) = \big\{x^* \in \mathcal{X}^* \mid (\forall y \in \mathcal{X}) \ \ \langle y - x, x^*\rangle + g(x) \leq g(y)\big\}, \tag{2.2}$$

is not empty; a subgradient of g at x is an element of $\partial g(x)$. The domain of continuity of g is

$$\operatorname{cont} g = \big\{x \in \mathcal{X} \mid \ |g(x)| < +\infty \ \text{ and } g \text{ is continuous at } x\big\}, \tag{2.3}$$

and its lower level set at height $\eta \in \mathbb{R}$ is $\operatorname{lev}_{\leq\eta} g = \{x \in \mathcal{X} \mid g(x) \leq \eta\}$. Recall that the value of g^*, the conjugate of g, at point $x^* \in \mathcal{X}^*$ is defined by

$$g^*(x^*) = \sup_{x\in\mathcal{X}} \langle x, x^*\rangle - g(x); \tag{2.4}$$

g is cofinite if $\operatorname{dom} g^* = \mathcal{X}^*$. Furthermore, g is coercive if $\lim_{\|x\|\to+\infty} g(x) = +\infty$, supercoercive if $\lim_{\|x\|\to+\infty} g(x)/\|x\| = +\infty$, (weak) lower semicontinuous if its lower level sets $\big(\operatorname{lev}_{\leq\eta} g\big)_{\eta\in\mathbb{R}}$ are (weakly) closed, and (weak) inf-compact if they are (weakly) compact. If $\mathcal{X}$ is reflexive, the notions of weak inf-compactness and coercivity coincide for weak lower semicontinuous functions. The set of minimizing sequences of g is denoted by

$$\mathcal{M}(g) = \big\{(x_n)_{n\in\mathbb{N}} \text{ in } \operatorname{dom} g \mid g(x_n) \to \inf g(\mathcal{X})\big\} \tag{2.5}$$

and the set of global minimizers of g by $\operatorname{Argmin} g$. (If it is a singleton, its unique element is denoted by $\operatorname{argmin} g$.) The inf-convolution of two functions $g_1, g_2 \colon \mathcal{X} \to]-\infty,+\infty]$ is $g_1 \,\square\, g_2 \colon \mathcal{X} \to [-\infty,+\infty] : x \mapsto \inf_{y\in\mathcal{X}} g_1(y) + g_2(x-y)$.

The indicator function of a set $C \subset \mathcal{X}$ is the function $\iota_C \colon \mathcal{X} \to \{0,+\infty\}$ that takes value 0 on C and $+\infty$ on its complement, and its normal cone is

$$N_C = \partial\iota_C \colon \mathcal{X} \to 2^{\mathcal{X}^*} : x \mapsto \begin{cases} \big\{x^* \in \mathcal{X}^* \mid (\forall y \in C) \;\; \langle y-x, x^*\rangle \leq 0\big\} & \text{if } x \in C,\\ \varnothing & \text{otherwise.}\end{cases} \tag{2.6}$$

2.6. *D*-viability and Legendre functions. Operators based on Bregman distances are not defined outside of $\operatorname{int}\operatorname{dom} f$. Thus, using the terminology of [3], for an algorithm such as (1.7) to be viable in the sense that its iterates remain in $\operatorname{int}\operatorname{dom} f$, the operators involved must satisfy the following viability condition.

DEFINITION 2.1. *An operator $T \colon \mathcal{X} \to 2^{\mathcal{X}}$ is* D-viable *if* $\operatorname{ran} T \subset \operatorname{dom} T = \operatorname{int}\operatorname{dom} f$.

It was shown in [7] that a sufficient condition for Bregman projection operators onto closed convex sets in Euclidean spaces to be D-viable is that f be a Legendre function. (In this context, "D-viability" was called "zone consistency" after [30].) The classical finite-dimensional definition of a Legendre function, as introduced by Rockafellar in [77, section 26], is of limited use in general Banach spaces since the resulting class of functions loses some of its remarkable finite-dimensional properties. In the context of Banach spaces, we introduced in [8] the following notion a Legendre function. It not only generalizes Rockafellar's classical definition but also preserves its salient properties in reflexive spaces. (For results on Legendre functions in nonreflexive spaces, see [13].)

DEFINITION 2.2 ([8, Def. 5.2]). *The function f is*

(i) essentially smooth *if ∂f is both locally bounded and single-valued on its domain;*

(ii) essentially strictly convex *if $(\partial f)^{-1}$ is locally bounded on its domain and f is strictly convex on every convex subset of* $\operatorname{dom}\partial f$*;*

(iii) Legendre *if it is both essentially smooth and essentially strictly convex.*

Such functions will be of prime importance in our analysis as they will be shown to provide a simple and convenient sufficient condition for the D-viability of the operators commonly encountered in Bregman optimization methods in Banach spaces.

2.7. Basic properties of Bregman distances. The following properties follow directly from (1.5).

PROPOSITION 2.3. *Let $\{x,y\} \subset \mathcal{X}$ and $\{u,v\} \subset \operatorname{int}\operatorname{dom} f$. Then*

(i) $D(u,v) + D(v,u) = \langle u-v, \nabla f(u) - \nabla f(v)\rangle$;

(ii) $D(x,u) = D(x,v) + D(v,u) + \langle x-v, \nabla f(v) - \nabla f(u)\rangle$;

(iii) $D(x,v) + D(y,u) = D(x,u) + D(y,v) + \langle x-y, \nabla f(u) - \nabla f(v)\rangle$.

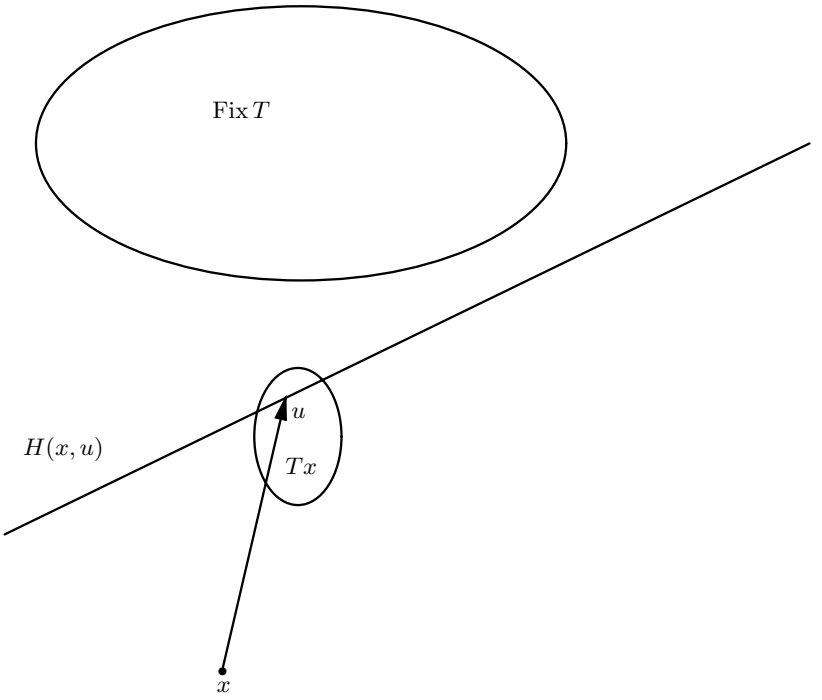


FIG. 1. *If $T \in \mathfrak{B}$, $x \in \operatorname{int}\operatorname{dom} f$, and $u \in Tx$, the half-space $H(x,u)$ contains* $\operatorname{Fix} T$.

3. Operators associated with Bregman distances. In Hilbert spaces, various nonlinear operators are involved in the design of algorithms, including projection operators, proximal operators, resolvents, subgradient projection operators, firmly nonexpansive operators, and combinations of these. Such operators arise in convex feasibility problems, in equilibrium theory, in systems of convex inequalities, in variational inequalities, as well as in numerous fixed point problems [5, 6, 9, 17, 35, 40, 41, 60, 72, 78]. Intrinsically tied to the very definition of these operators is the use of the standard notion of metric distance to measure the proximity between two points. In the context of Bregman distances, it is therefore natural to attempt to define variants of these operators. This effort has been undertaken by several authors at various levels of generality. In this section, we systematically study nonlinear operators associated with Bregman distances in order to bring together and extend a collection of results disseminated in the literature. Specifically, we investigate when D-firm operators, D-resolvents, D-prox operators, D-projectors, and subgradient D-projectors belong to class $\mathfrak{B}$. (For relationships among these operators in the classical case, i.e., when $\mathcal{X}$ is a Hilbert space and $f = \|\cdot\|^2/2$, see [9, Prop. 2.3].) Moreover, the class $\mathfrak{B}$ is shown to be closed under a certain type of relaxed parallel combination. The discussion is not limited to convex problems as nonconvex extensions of standard algorithms have been found to be quite useful in a number of applications; see [12, 28, 43, 52, 62].

3.1. The class $\mathfrak{B}$. Ultimately, our goal is to define a class of operators for which (1.7) systematically generates D-monotone sequences. In this perspective, the operators employed in (1.7) must be D-viable (see Definition 2.1) and induce a certain monotonicity property (see Definition 1.2). These requirements lead to the following class of operators (see Figure 1).

DEFINITION 3.1. *For every x and u in $\operatorname{int}\operatorname{dom} f$, set*

$$H(x,u) = \big\{y \in \mathcal{X} \mid \langle y-u, \nabla f(x) - \nabla f(u)\rangle \leq 0\big\}. \tag{3.1}$$

Then

$$\mathfrak{B} = \big\{T\colon \mathcal{X} \to 2^{\mathcal{X}} \mid \operatorname{ran} T \subset \operatorname{dom} T = \operatorname{int}\operatorname{dom} f,\ (\forall (x,u) \in \operatorname{gr} T)\ \operatorname{Fix} T \subset H(x,u)\big\}.$$

If $\mathcal{X}$ is Hilbertian, $f = \|\cdot\|^2/2$, and only single-valued operators are considered, then $\mathfrak{B}$ reverts to the class $\mathfrak{T}$ of operators introduced in [9] and further investigated in this context in [41, 42]. In these studies, $\mathfrak{T}$ was shown to play a central role in the analysis of Fejér monotone algorithms. Because of Proposition 3.3(i) below, there is some overlap between the "paracontractions" introduced in [31, 75] (see also [24, 26]) and operators in $\mathfrak{B}$. Furthermore, if f satisfies certain conditions and $T \in \mathfrak{B}$ is single-valued with $\operatorname{Fix} T \neq \emptyset$, then T is "totally nonexpansive" in the sense of [24].

LEMMA 3.2. *Let C_1 and C_2 be two convex subsets of $\mathcal{X}$ such that C_1 is closed and $C_1 \cap \operatorname{int} C_2 \neq \emptyset$. Then $\overline{C_1 \cap \operatorname{int} C_2} = C_1 \cap \overline{C_2}$.*

Proof. Since C_2 is convex with nonempty interior, $\overline{C_1 \cap \operatorname{int} C_2} \subset \overline{C_1} \cap \overline{\operatorname{int} C_2} = C_1 \cap \overline{C_2}$. To show the reverse inclusion, fix $x_0 \in C_1 \cap \operatorname{int} C_2$ and $x_1 \in C_1 \cap \overline{C_2}$. By convexity, $[x_0, x_1] \subset C_1$ and $[x_0, x_1[\subset \operatorname{int} C_2$. Therefore, $(\forall \alpha \in [0,1[)\ x_\alpha = (1-\alpha)x_0 + \alpha x_1 \in C_1 \cap \operatorname{int} C_2$. Consequently $x_1 = \lim_{\alpha \uparrow 1^-} x_\alpha \in \overline{C_1 \cap \operatorname{int} C_2}$, and we conclude $C_1 \cap \overline{C_2} \subset \overline{C_1 \cap \operatorname{int} C_2}$. ∎

PROPOSITION 3.3. *Let T be an operator in $\mathfrak{B}$ and let $F = \bigcap_{(x,u)\in \operatorname{gr} T} H(x,u)$. Then*

(i) $(\forall (x,u) \in \operatorname{gr} T)(\forall y \in \operatorname{Fix} T)\ \ D(y,u) \leq D(y,x) - D(u,x)$;
(ii) $(\forall (x,u) \in \operatorname{gr} T)\ \ D(u,x) \leq D_{\operatorname{Fix} T}(x)$;
(iii) $(\forall (x,u) \in \operatorname{gr} T)(\forall y \in \overline{\operatorname{Fix}} T)\ \ D(x,u) + D(u,x) \leq \langle y - x, \nabla f(u) - \nabla f(x)\rangle$.

Now suppose that $f|_{\operatorname{int}\operatorname{dom} f}$ is strictly convex; then

(iv) $\operatorname{Fix} T = F \cap \operatorname{int}\operatorname{dom} f$;
(v) $\operatorname{Fix} T$ *is convex;*
(vi) T *is single-valued on* $\operatorname{Fix} T$.

If, in addition, $\operatorname{Fix} T \neq \emptyset$, then

(vii) $\overline{\operatorname{Fix}} T = F \cap \overline{\operatorname{dom}} f$;
(viii) $(\forall (x,u) \in \operatorname{gr} T)(\forall y \in \overline{\operatorname{Fix}} T)\ \ D(y,u) \leq D(y,x) - D(u,x)$.

Proof. (i) Take $(x,u) \in \operatorname{gr} T$ and $y \in \operatorname{Fix} T$. Then Proposition 2.3(ii) and the inclusion $y \in H(x,u)$ yield $D(y,u) = D(y,x) - D(u,x) + \langle y-u, \nabla f(x) - \nabla f(u)\rangle \leq D(y,x) - D(u,x)$. (ii) By (i), $(\forall (x,u) \in \operatorname{gr} T)(\forall y \in \operatorname{Fix} T)\ D(u,x) \leq D(y,x)$. (iii) Take $(x,u) \in \operatorname{gr} T$ and $y \in \overline{\operatorname{Fix}} T$, and suppose $y_n \to y$ for some sequence $(y_n)_{n\in\mathbb{N}}$ in $\operatorname{Fix} T$. Then it follows from Proposition 2.3(i) that

$$\begin{aligned}(\forall n \in \mathbb{N})\ \ D(x,u) + D(u,x) &= \langle x-u, \nabla f(x) - \nabla f(u)\rangle \\ &= \langle x - y_n, \nabla f(x) - \nabla f(u)\rangle + \langle y_n - u, \nabla f(x) - \nabla f(u)\rangle \\ &\leq \langle x - y_n, \nabla f(x) - \nabla f(u)\rangle\,. \end{aligned} \tag{3.2}$$

Since $\langle x - y_n, \nabla f(x) - \nabla f(u)\rangle \to \langle x - y, \nabla f(x) - \nabla f(u)\rangle$, the proof is complete. (iv) Take $y \in F \cap \operatorname{int}\operatorname{dom} f$. Then $y \in \bigcap_{u\in Ty} H(y,u)$ and, in turn,

$$(\forall u \in Ty)\ \ \langle y-u, \nabla f(y) - \nabla f(u)\rangle \leq 0. \tag{3.3}$$

However, $\{y\} \cup Ty \subset \operatorname{int}\operatorname{dom} f$ and, since $f|_{\operatorname{int}\operatorname{dom} f}$ is strictly convex, ∇f is strictly monotone on $\operatorname{int}\operatorname{dom} f$. Therefore $Ty = \{y\}$ and $y \in \operatorname{Fix} T$. Thus, $F \cap \operatorname{int}\operatorname{dom} f \subset$

$\operatorname{Fix} T$. Since $T \in \mathfrak{B}$, the reverse inclusion is clear. (iv) $\Rightarrow$ (v) Since the sets $\big(H(x,u)\big)_{(x,u)\in \operatorname{gr} T}$ and $\operatorname{int}\operatorname{dom} f$ are convex, so is their intersection $\operatorname{Fix} T$. (vi) was proved in the proof of (iv). (iv) $\Rightarrow$ (vii) Observe that F is closed and apply Lemma 3.2. (viii) Take $(x,u) \in \operatorname{gr} T$, $y_0 \in \operatorname{Fix} T$, and $y \in \overline{\operatorname{Fix}} T$. By (iv) and (vii), $\operatorname{Fix} T = F \cap \operatorname{int}\operatorname{dom} f$ and $\overline{\operatorname{Fix}} T = F \cap \overline{\operatorname{dom}} f$. Since F and $\operatorname{dom} f$ are convex, $[y_0, y] \subset F$ and $[y_0, y[\subset \operatorname{int}\operatorname{dom} f$. Therefore,

$$(\forall \alpha \in [0,1[)\ \ y_\alpha = (1-\alpha)y_0 + \alpha y \in \operatorname{Fix} T. \tag{3.4}$$

Invoking the lower semicontinuity and convexity of f, we get

$$f(y) \leq \underline{\lim}_{\alpha\uparrow 1^-} f(y_\alpha) \leq \overline{\lim}_{\alpha\uparrow 1^-} f(y_\alpha) \leq \overline{\lim}_{\alpha\uparrow 1^-} (1-\alpha)f(y_0) + \alpha f(y) = f(y). \tag{3.5}$$

Hence $\lim_{\alpha\uparrow 1^-} f(y_\alpha) = f(y)$ and, in turn,

$$(\forall z \in \operatorname{int}\operatorname{dom} f)\ \ \lim_{\alpha\uparrow 1^-} D(y_\alpha, z) = D(y,z). \tag{3.6}$$

On the other hand, since $u \in Tx$ and $T \in \mathfrak{B}$, (3.4) and (i) yield

$$(\forall \alpha \in [0,1[)\ \ D(y_\alpha, u) \leq D(y_\alpha, x) - D(u,x). \tag{3.7}$$

Consequently, $D(y,u) \leq D(y,x) - D(u,x)$. □

3.2. *D*-firm operators. An operator $T\colon \mathcal{X} \to \mathcal{X}$ is said to be firmly nonexpansive if for all x and y in $\operatorname{dom} T$ one has [51]

$$(\forall \alpha \in]0,+\infty[)\ \ \|Tx - Ty\| \leq \|\alpha(x-y) + (1-\alpha)(Tx - Ty)\|. \tag{3.8}$$

For the sake of notational simplicity, let us now suppose that $\mathcal{X}$ is smooth. Then its normalized duality map J is single-valued and, upon invoking the equivalence $(\forall \alpha \in]0,+\infty[)\ \|u\| \leq \|u + \alpha v\| \Leftrightarrow 0 \leq \langle v, Ju\rangle$ [51], we observe that (3.8) is equivalent to

$$\langle Tx - Ty, J(Tx - Ty)\rangle \leq \langle x - y, J(Tx - Ty)\rangle. \tag{3.9}$$

If $\mathcal{X}$ is not a Hilbert space, then J is not linear and this type of inequality may be difficult to manipulate. In Hilbert spaces, $J = \operatorname{Id} = \nabla f$ for $f = \|\cdot\|^2/2$, and (3.9) can therefore be written

$$\langle Tx - Ty, \nabla f(Tx) - \nabla f(Ty)\rangle \leq \langle Tx - Ty, \nabla f(x) - \nabla f(y)\rangle. \tag{3.10}$$

In the framework of Bregman distances, this inequality suggests the following definition.

DEFINITION 3.4. *An operator* $T\colon \mathcal{X} \to 2^{\mathcal{X}}$ *with* $\operatorname{dom} T \cup \operatorname{ran} T \subset \operatorname{int}\operatorname{dom} f$ *is* D-firm *if*

$$(\forall (x,u) \in \operatorname{gr} T)(\forall (y,v) \in \operatorname{gr} T)\ \ \langle u - v, \nabla f(u) - \nabla f(v)\rangle \leq \langle u - v, \nabla f(x) - \nabla f(y)\rangle. \tag{3.11}$$

PROPOSITION 3.5. *Let* $T\colon \mathcal{X} \to 2^{\mathcal{X}}$ *be a D-firm operator. Then*

(i) $(\forall (x,u) \in \operatorname{gr} T)\ \ \operatorname{Fix} T \subset H(x,u)$;

(ii) $T \in \mathfrak{B}$ *if* $\operatorname{int}\operatorname{dom} f = \operatorname{dom} T$;

(iii) T *is single-valued on its domain if* $f|_{\operatorname{int}\operatorname{dom} f}$ *is strictly convex;*

(iv) $(\forall (x,u) \in \operatorname{gr} T)(\forall (y,v) \in \operatorname{gr} T) \quad D(u,v) + D(v,u) \leq D(u,y) + D(v,x) - D(u,x) - D(v,y)$.

Proof. (i) Suppose $y \in Ty$. Then (3.11) implies that

$$(\forall (x,u) \in \operatorname{gr} T) \quad \langle y - u, \nabla f(x) - \nabla f(u) \rangle \leq 0. \tag{3.12}$$

(i) $\Rightarrow$ (ii) is clear. (iii) Fix $x \in \operatorname{dom} T$ and $\{u,v\} \subset Tx$. Then (3.11) implies that

$$\langle u - v, \nabla f(u) - \nabla f(v) \rangle \leq 0. \tag{3.13}$$

Since ∇f is strictly monotone on $\operatorname{int}\operatorname{dom} f \supset \{u,v\}$, we obtain $u = v$. (iv) follows from Proposition 2.3(i), (3.11), and Proposition 2.3(iii). □

Remark 3.6. For single-valued operators in Hilbert spaces and f strongly convex (i.e., $f - \beta\|\cdot\|^2/2$ is convex for some $\beta \in]0,+\infty[$), item (iv) above was used to define D-firmness in [18].

3.3. *D*-resolvents. The resolvent of an operator $A\colon \mathcal{X} \to 2^{\mathcal{X}}$ is $(\operatorname{Id} + A)^{-1}$. It is known that an operator $T\colon \mathcal{X} \to \mathcal{X}$ is firmly nonexpansive if and only if it is the resolvent of an accretive operator $A\colon \mathcal{X} \to 2^{\mathcal{X}}$ [19].

Now let $A\colon \mathcal{X} \to 2^{\mathcal{X}^*}$ be a nontrivial operator, i.e., $\operatorname{gr} A \neq \emptyset$. Then, in the context of Bregman distances, it is reasonable to introduce the following variant of the notion of a resolvent to obtain an operator from $\mathcal{X}$ to $\mathcal{X}$ (this definition appears to have first been proposed in $\mathbb{R}^N$ in [46]).

DEFINITION 3.7. *The* D-resolvent *associated with* $A\colon \mathcal{X} \to 2^{\mathcal{X}^*}$ *is the operator*

$$R_A = (\nabla f + A)^{-1} \circ \nabla f\colon \mathcal{X} \to 2^{\mathcal{X}}. \tag{3.14}$$

An a posteriori motivation for (3.14) is that it preserves the usual fixed point characterization of the zeros of A, namely,

$$(\forall x \in \mathcal{X})(\forall \gamma \in]0,+\infty[) \quad 0 \in Ax \quad \Leftrightarrow \quad x \in \operatorname{Fix} R_{\gamma A}, \tag{3.15}$$

as $0 \in Ax \Leftrightarrow \nabla f(x) \in \nabla f(x) + \gamma A(x) = (\nabla f + \gamma A)(x) \Leftrightarrow x \in (\nabla f + \gamma A)^{-1}\big(\nabla f(x)\big)$. It is also consistent with previous attempts to define resolvents for monotone operators:

- Let $\mathcal{X}$ be smooth and set $f = \|\cdot\|^2/2$. Then $\nabla f = J$ and $R_A = (J + A)^{-1} \circ J$. This type of resolvent was used in [57].
- If $\mathcal{X}$ is Hilbertian and $f\colon x \mapsto \|\Pi x\|^2/2$, where Π is the metric projector onto a closed vector subspace of $\mathcal{X}$, then $\nabla f = \Pi$ and $R_A = (\Pi + A)^{-1} \circ \Pi$. This generalized resolvent was used in [54].

PROPOSITION 3.8. *R_A satisfies the following properties:*

(i) $\operatorname{dom} R_A \subset \operatorname{int}\operatorname{dom} f$.
(ii) $\operatorname{ran} R_A \subset \operatorname{int}\operatorname{dom} f$.
(iii) $\operatorname{Fix} R_A = (\operatorname{int}\operatorname{dom} f) \cap A^{-1}0$.
(iv) *Suppose A is monotone. Then the following conditions hold:*
  (a) *R_A is D-firm.*
  (b) *R_A is single-valued on its domain if $f|_{\operatorname{int}\operatorname{dom} f}$ is strictly convex.*
  (c) *Suppose* $\operatorname{ran}\nabla f \subset \operatorname{ran}(\nabla f + A)$. *Then* $R_A \in \mathfrak{B}$. *If, in addition, $f|_{\operatorname{int}\operatorname{dom} f}$ is strictly convex, then* $\operatorname{Fix} R_A$ *is convex.*

Proof. (i) is clear. (ii) We have

$$\begin{aligned} \operatorname{ran} R_A \subset \operatorname{ran}(\nabla f + A)^{-1} &= \operatorname{dom}(\nabla f + A) = \operatorname{dom}\nabla f \cap \operatorname{dom} A \subset \operatorname{dom}\nabla f \\ &= \operatorname{int}\operatorname{dom} f. \end{aligned} \tag{3.16}$$

(iii) $\operatorname{Fix} R_A \subset \operatorname{int}\operatorname{dom} f$ by (i) and $(\forall x \in \operatorname{int}\operatorname{dom} f)\ 0 \in Ax \Leftrightarrow x \in R_A x$ by (3.15). Hence, $A^{-1}0 \cap \operatorname{int}\operatorname{dom} f = \operatorname{Fix} R_A \cap \operatorname{int}\operatorname{dom} f = \operatorname{Fix} R_A$. (iv) Suppose that A is monotone. (a) In view of (i) and (ii), let us show that (3.11) is satisfied. Fix (x,u) and (y,v) in $\operatorname{gr} R_A$. Then $\nabla f(x) - \nabla f(u) \in Au$ and $\nabla f(y) - \nabla f(v) \in Av$. Consequently, since A is monotone, we get $\langle u-v, \nabla f(x) - \nabla f(u) - (\nabla f(y) - \nabla f(v))\rangle \geq 0$. (b) follows from (a) and Proposition 3.5(iii). (c) $\operatorname{ran}\nabla f \subset \operatorname{ran}(\nabla f + A) \Leftrightarrow \operatorname{ran}\nabla f \subset \operatorname{dom}(\nabla f + A)^{-1} \Leftrightarrow \operatorname{dom} R_A = \operatorname{dom}\nabla f = \operatorname{int}\operatorname{dom} f$. In view of (a) and Proposition 3.5(ii), $R_A \in \mathfrak{B}$. Proposition 3.3(v) implies the convexity of $\operatorname{Fix} R_A$. □

DEFINITION 3.9 (see [86, sections 32.14 and 32.21]). *A is*

(i) weakly coercive *if* $\lim_{\|x\|\to+\infty} \inf \|Ax\| = +\infty$;

(ii) strongly coercive *if*

$$(\forall x \in \operatorname{dom} A) \quad \lim_{\|y\|\to+\infty} \inf \frac{\langle y - x, Ax\rangle}{\|y\|} = +\infty;$$

(iii) 3-monotone *if*

$$\big(\forall\big((x,x^*),(y,y^*),(z,z^*)\big) \in (\operatorname{gr} A)^3\big)\ \ \langle x-y, x^*\rangle + \langle y-z, y^*\rangle + \langle z-x, z^*\rangle \geq 0;$$

(iv) 3^*-monotone *if it is monotone and*

$$(\forall (x,x^*) \in \operatorname{dom} A \times \operatorname{ran} A)\ \ \sup\big\{\langle x-y, y^*-x^*\rangle \mid (y,y^*) \in \operatorname{gr} A\big\} < +\infty.$$

LEMMA 3.10 (see [86, section 32.21], [16]). *Suppose that $\mathcal{X}$ is reflexive and that A is monotone and satisfies one of the following properties:*

(i) *A is 3-monotone.*

(ii) *A is strongly coercive.*

(iii) *$\operatorname{ran} A$ is bounded.*

(iv) *$A = \partial\varphi$, where $\varphi\colon \mathcal{X} \to]-\infty, +\infty]$ is a proper function.*

Then A is 3^*-monotone.

The following lemma is Reich's extension to a reflexive Banach space setting of the Brézis–Haraux theorem [16] on the range of the sum of two monotone operators.

LEMMA 3.11 (see [74, Thm. 2.2]). *Suppose that $\mathcal{X}$ is reflexive and let $A_1, A_2\colon \mathcal{X} \to 2^{\mathcal{X}^*}$ be two monotone operators such that $A_1 + A_2$ is maximal monotone and A_1 is 3^*-monotone. In addition, suppose that $\operatorname{dom} A_2 \subset \operatorname{dom} A_1$ or A_2 is 3^*-monotone. Then $\operatorname{int}\operatorname{ran}(A_1 + A_2) = \operatorname{int}(\operatorname{ran} A_1 + \operatorname{ran} A_2)$ and $\overline{\operatorname{ran}}\,(A_1 + A_2) = \overline{\operatorname{ran} A_1 + \operatorname{ran} A_2}$.*

PROPOSITION 3.12. *Let $\gamma \in]0, +\infty[$. Suppose that $\mathcal{X}$ is reflexive and that A is maximal monotone with $(\operatorname{int}\operatorname{dom} f) \cap \operatorname{dom} A = \operatorname{dom}\partial f \cap \operatorname{dom} A \neq \emptyset$. Then $\nabla f + \gamma A$ is maximal monotone. Moreover, the inclusions*

$$\begin{cases} \operatorname{int}(\operatorname{ran}\nabla f + \gamma \operatorname{ran} A) \subset \operatorname{ran}(\nabla f + \gamma A) \\ \operatorname{ran}\nabla f + \gamma \operatorname{ran} A \subset \overline{\operatorname{ran}}\,(\nabla f + \gamma A) \end{cases} \tag{3.17}$$

are satisfied if one of the following conditions holds:

(i) *$\operatorname{dom} A \subset \operatorname{int}\operatorname{dom} f$.*

(ii) *A is 3^*-monotone.*

Proof. Since f is proper, lower semicontinuous, and convex, ∂f is maximal monotone [79, Thm. 30.3] and $\operatorname{int}\operatorname{dom} f = \operatorname{cont} f \subset \operatorname{dom}\partial f \subset \operatorname{dom} f$ [48, Chap. I]. Since $(\operatorname{int}\operatorname{dom} f) \cap \operatorname{dom} A = \operatorname{dom}\partial f \cap \operatorname{dom} A \neq \emptyset$, we have $(\operatorname{int}\operatorname{dom}\partial f) \cap \operatorname{dom}\gamma A = (\operatorname{int}\operatorname{dom} f) \cap \operatorname{dom} A \neq \emptyset$, and it follows from Rockafellar's sum theorem [79, section 23] that $\partial f + \gamma A$ is maximal monotone. However, the above assumption implies

that $\operatorname{dom}(\nabla f + \gamma A) = \operatorname{dom}(\partial f + \gamma A)$ and, in turn, that $\nabla f + \gamma A = \partial f + \gamma A$ since $\{\nabla f\} = \partial f|_{\operatorname{int}\operatorname{dom} f}$. Thus, $\nabla f + \gamma A$ is maximal monotone. The second assertion is an application of Lemma 3.11 with $A_1 = \nabla f$ and $A_2 = \gamma A$. Indeed, $\operatorname{dom}\nabla f = \operatorname{int}\operatorname{dom} f$ and, by Lemma 3.10(iv), ∂f is 3^*-monotone and so is, therefore, ∇f since $\operatorname{gr}\nabla f \subset \operatorname{gr}\partial f$. ∎

THEOREM 3.13. *Let $\gamma \in]0,+\infty[$. Suppose that $\mathcal{X}$ is reflexive, that A is maximal monotone with $(\operatorname{int}\operatorname{dom} f) \cap \operatorname{dom} A = \operatorname{dom}\partial f \cap \operatorname{dom} A \neq \varnothing$, and that one of the following conditions holds:*

(i) *$\mathcal{X}$ is smooth and $f = \|\cdot\|^2/2$.*
(ii) *$(\nabla f + \gamma A)^{-1}$ is locally bounded at every point in $\mathcal{X}^*$.*
(iii) *$\nabla f + \gamma A$ is weakly coercive.*
(iv) *$\operatorname{dom} A \subset \operatorname{int}\operatorname{dom} f$ or A is 3^*-monotone, and one of the following conditions holds:*
  (a) *$\operatorname{ran}\nabla f + \gamma \operatorname{ran} A = \mathcal{X}^*$.*
  (b) *f is Legendre and cofinite.*
  (c) *$\operatorname{ran}(\nabla f + \gamma A)$ is closed and $0 \in \operatorname{ran} A$.*
  (d) *$\operatorname{ran}\nabla f$ is open and $0 \in \operatorname{ran} A$.*

Then $R_{\gamma A} \in \mathfrak{B}$.

Proof. In view of Proposition 3.8(iv)(c), it suffices to show that $\operatorname{ran}\nabla f \subset \operatorname{ran}(\nabla f + \gamma A)$. (i) Since $\mathcal{X}$ is smooth, $\nabla f = J$ [34, Corollary I.4.5] and Rockafellar's surjectivity theorem [79, Thm. 10.7] yields $\operatorname{ran}(\nabla f + \gamma A) = \mathcal{X}^*$. (ii) Proposition 3.12 asserts that $\nabla f + \gamma A$ is maximal monotone. It therefore follows from the Brézis–Browder surjectivity theorem (see [34, Thm. V.3.8] or [86, Thm. 32.G]) that $\operatorname{ran}(\nabla f + \gamma A) = \mathcal{X}^*$. (iii) $\Rightarrow$ (ii) follows from [86, Cor. 32.35] since $\nabla f + \gamma A$ is maximal monotone. (iv) By Proposition 3.12, (3.17) holds. (a) By (3.17), $\mathcal{X}^* = \operatorname{int}(\operatorname{ran}\nabla f + \gamma \operatorname{ran} A) \subset \operatorname{ran}(\nabla f + \gamma A)$. (b) $\Rightarrow$ (a) By [8, Thm. 5.10], Legendreness guarantees $\operatorname{ran}\nabla f = \operatorname{int}\operatorname{dom} f^*$ while cofiniteness gives $\operatorname{int}\operatorname{dom} f^* = \mathcal{X}^*$. Consequently, $\operatorname{ran}\nabla f + \gamma \operatorname{ran} A = \mathcal{X}^*$. (c) By (3.17), $\operatorname{ran}\nabla f = \operatorname{ran}\nabla f + \{0\} \subset \operatorname{ran}\nabla f + \gamma \operatorname{ran} A \subset \overline{\operatorname{ran}}\,(\nabla f + \gamma A) = \operatorname{ran}(\nabla f + \gamma A)$. (d) By (3.17), $\operatorname{ran}\nabla f = \operatorname{int}(\operatorname{ran}\nabla f + \{0\}) \subset \operatorname{int}(\operatorname{ran}\nabla f + \gamma \operatorname{ran} A) \subset \operatorname{ran}(\nabla f + \gamma A)$. ∎

In connection with the problem of finding zeros of maximal monotone operators, the following corollary is particularly useful.

COROLLARY 3.14. *Let $\gamma \in]0,+\infty[$. Suppose that $\mathcal{X}$ is reflexive, that A is maximal monotone with $0 \in \operatorname{ran} A$, and that one of the following conditions holds:*

(i) *$\operatorname{ran}\nabla f$ is open and $\operatorname{dom} A \subset \operatorname{int}\operatorname{dom} f$.*
(ii) *f is Legendre and $\operatorname{dom} A \subset \operatorname{int}\operatorname{dom} f$.*
(iii) *f is Legendre, A is 3^*-monotone, and $\operatorname{dom} A \cap \operatorname{int}\operatorname{dom} f \neq \varnothing$.*

Then $R_{\gamma A} \in \mathfrak{B}$.

Proof. The assertions follow from Theorem 3.13(iv)(d). Indeed, in (i), $\operatorname{dom} A \subset \operatorname{int}\operatorname{dom} f = \operatorname{cont} f \subset \operatorname{dom}\partial f \Rightarrow (\operatorname{int}\operatorname{dom} f) \cap \operatorname{dom} A = \operatorname{dom}\partial f \cap \operatorname{dom} A = \operatorname{dom} A \neq \varnothing$. On the other hand, in (ii) and (iii), $\operatorname{ran}\nabla f$ is open since Legendreness yields $\operatorname{ran}\nabla f = \operatorname{int}\operatorname{dom} f^*$ [8, Thm. 5.10]. Consequently, if $\operatorname{dom} A \subset \operatorname{int}\operatorname{dom} f$, then (ii) is a consequence of (i). Otherwise, if A is 3^*-monotone and $(\operatorname{int}\operatorname{dom} f) \cap \operatorname{dom} A \neq \varnothing$, then it suffices to note that essential smoothness yields $\operatorname{dom}\partial f = \operatorname{int}\operatorname{dom} f$ [8, Thm. 5.6], whence $(\operatorname{int}\operatorname{dom} f) \cap \operatorname{dom} A = \operatorname{dom}\partial f \cap \operatorname{dom} A \neq \varnothing$. ∎

Remark 3.15. In $\mathbb{R}^N$, Corollary 3.14(i) corresponds to [46, Thm. 4].

3.4. *D*-prox operators. The classical notion of a proximal operator was introduced by Moreau [64, 65, 67] in Hilbert spaces. The proximal operator associated with a function $\varphi\colon \mathcal{X} \to]-\infty,+\infty]$ is $\operatorname{prox}^{\varphi}\colon y \mapsto \operatorname{argmin}\varphi + \|\cdot - y\|^2/2$. Outside of

Hilbert spaces, this notion is of less interest since Fermat's rule for the minimization of $\varphi + \|\cdot - y\|^2/2$ becomes a nonseparable inclusion, namely, $0 \in \partial\varphi(x) + J(x-y)$.

In $\mathbb{R}^N$, the idea of defining proximal operators based on D-distance—rather than quadratic—penalizations was introduced in [32]. In our setting, they will be defined as follows.

DEFINITION 3.16. *Let* $\varphi\colon \mathcal{X} \to]-\infty,+\infty]$. *The* D-prox operator *of index* $\gamma \in]0,+\infty[$ *associated with* φ *is the operator*

$$\operatorname{prox}_\gamma^\varphi\colon \mathcal{X} \to 2^{\mathcal{X}},$$
$$y \mapsto \left\{ x \in \operatorname{dom} f \cap \operatorname{dom}\varphi \mid \varphi(x) + \frac{1}{\gamma}D(x,y) = \min\left(\varphi + \frac{1}{\gamma}D(\cdot,y)\right)(\mathcal{X}) < +\infty \right\}.$$

It follows from this definition that

$$\operatorname{dom}\operatorname{prox}_\gamma^\varphi \subset \operatorname{int}\operatorname{dom} f \quad \text{and} \quad \operatorname{ran}\operatorname{prox}_\gamma^\varphi \subset \operatorname{dom} f \cap \operatorname{dom}\varphi. \tag{3.18}$$

Recall (see section 2.5) that a function is weak inf-compact if all its lower level sets are weakly compact.

LEMMA 3.17. *Suppose that* $g_1\colon \mathcal{X} \to]-\infty,+\infty]$ *is weak lower semicontinuous and bounded from below and that* $g_2\colon \mathcal{X} \to]-\infty,+\infty]$ *is weak inf-compact. Then* $g_1 + g_2$ *is weak inf-compact.*

Proof. Set $\beta = \inf g_1(\mathcal{X})$ and let $\eta \in \mathbb{R}$. Since g_1 and g_2 are weak lower semicontinuous, so is their sum, and therefore $\operatorname{lev}_{\le\eta}(g_1+g_2)$ is weakly closed. On the other hand, $\operatorname{lev}_{\le\eta}(g_1+g_2)$ is contained in the weakly compact set $\operatorname{lev}_{\le\eta-\beta} g_2$. We conclude that $\operatorname{lev}_{\le\eta}(g_1+g_2)$ is weakly compact. ∎

The following result concerns the domain requirement for the D-viability of D-prox operators. Recall (see sections 2.5 and 2.2) that $\mathcal{M}$ denotes the set of minimizing sequences of a function and that $\mathfrak{W}$ is the set of weak cluster points of a sequence.

THEOREM 3.18. *Let* $\gamma \in]0,+\infty[$, *let* $\varphi\colon \mathcal{X} \to]-\infty,+\infty]$ *be such that* $\operatorname{dom} f \cap \operatorname{dom}\varphi \neq \emptyset$, *and assume that one of the following conditions holds:*

(i) $(\forall y \in \operatorname{int}\operatorname{dom} f)(\exists\,(x_n)_{n\in\mathbb{N}} \in \mathcal{M}(f_y + \gamma\varphi))(\exists\, x \in \mathfrak{W}(x_n)_{n\in\mathbb{N}})$ $f + \gamma\varphi$ *is weak lower semicontinuous at* x.

(ii) $(\forall y \in \operatorname{int}\operatorname{dom} f)$ $f_y + \gamma\varphi$ *is weak inf-compact.*

(iii) φ *is weak lower semicontinuous and bounded from below, and, for every* $y \in \operatorname{int}\operatorname{dom} f$, f_y *is weak inf-compact.*

(iv) φ *is weak inf-compact.*

Then $\operatorname{dom}\operatorname{prox}_\gamma^\varphi = \operatorname{int}\operatorname{dom} f$.

Proof. Fix $y \in \operatorname{int}\operatorname{dom} f$ and set $g = f_y + \gamma\varphi$. (i) Pick $(x_n)_{n\in\mathbb{N}} \in \mathcal{M}(g)$ such that $x_{k_n} \rightharpoonup x$ and g is weak lower semicontinuous at x. It follows that $g(x) \le \underline{\lim}\, g(x_{k_n}) = \inf g(\mathcal{X})$ and hence $g(x) = \inf g(\mathcal{X})$. Therefore, g achieves its infimum and the result holds since $\operatorname{prox}_\gamma^\varphi y = \operatorname{Argmin}(f_y+\gamma\varphi) = \operatorname{Argmin}(g)$. (ii) ⇒ (i) Take $(x_n)_{n\in\mathbb{N}} \in \mathcal{M}(g)$. Then it follows from weak inf-compactness of g that $(x_n)_{n\in\mathbb{N}}$ lies in a weakly compact set and therefore that $\mathfrak{W}(x_n)_{n\in\mathbb{N}} \neq \emptyset$. On the other hand, as g is weak inf-compact, it is weak lower semicontinuous and so is $f + \gamma\varphi = f_y + \gamma\varphi + \nabla f(y) = g + \nabla f(y)$. (iii) ⇒ (ii) follows from Lemma 3.17. (iv) ⇒ (ii) It is clear that f_y is weak lower semicontinuous. On the other hand, it follows from the convexity of f that, for every $x \in \mathcal{X}$, $\langle x - y, \nabla f(y)\rangle + f(y) \le f(x)$ and, therefore, $f_y(x) \ge f_y(y)$. Hence $\inf f_y(\mathcal{X}) \ge f_y(y) > -\infty$ and, by Lemma 3.17, g is weak inf-compact. ∎

The following fundamental result is due to Moreau [66] and Rockafellar [76].

LEMMA 3.19. *Let* $y^* \in \mathcal{X}^*$. *Then* $f - y^*$ *is coercive if and only if* $y^* \in \operatorname{int}\operatorname{dom} f^*$.

LEMMA 3.20. *Let* $g_1, g_2\colon \mathcal{X} \to]-\infty,+\infty]$ *be two convex functions. Then*

(i) *if* g_1 *and* g_2 *are lower semicontinuous and* $0 \in \operatorname{int}\big(\operatorname{dom} g_1 - \operatorname{dom} g_2\big)$, *then* $(g_1 + g_2)^* = g_1^* \,\square\, g_2^*$ [2];
(ii) *if* $\operatorname{cont} g_1 \cap \operatorname{dom} g_2 \neq \varnothing$, *then* $\partial(g_1 + g_2) = \partial g_1 + \partial g_2$ [79, Thm. 28.2].

PROPOSITION 3.21. *Let* $\varphi\colon \mathcal{X} \to \left]-\infty, +\infty\right]$ *be a lower semicontinuous convex function such that* $\operatorname{dom} f \cap \operatorname{dom} \varphi \neq \varnothing$ *and let* $\gamma \in \left]0, +\infty\right[$. *Suppose that* $\mathcal{X}$ *is reflexive and that one of the following conditions holds:*

(i) $(\forall y \in \operatorname{int}\operatorname{dom} f)(\exists\, (x_n)_{n\in\mathbb{N}} \in \mathcal{M}(f_y + \gamma\varphi))\ \ \sup_{n\in\mathbb{N}} \|x_n\| < +\infty$.
(ii) $(\forall y \in \operatorname{int}\operatorname{dom} f)\ \ f_y + \gamma\varphi$ *is coercive.*
(iii) $\operatorname{ran} \nabla f \subset \operatorname{int}\operatorname{dom} (f + \gamma\varphi)^*$.
(iv) $f + \gamma\varphi$ *is cofinite.*
(v) $0 \in \operatorname{int}(\operatorname{dom} f - \operatorname{dom}\varphi)$ *and* $\operatorname{dom} f^* + \gamma \operatorname{dom}\varphi^* = \mathcal{X}^*$.
(vi) φ *is bounded from below and* f *is essentially strictly convex.*
(vii) $f + \gamma\varphi$ *is supercoercive.*
(viii) φ *is bounded from below and* f *is supercoercive.*
(ix) φ *is coercive.*

Then $\operatorname{dom} \operatorname{prox}_\gamma^\varphi = \operatorname{int}\operatorname{dom} f$.

Proof. Let y be an arbitrary point in $\operatorname{int}\operatorname{dom} f$. Note that, since φ is weak lower semicontinuous, so are $f + \gamma\varphi$ and $f_y + \gamma\varphi$ and that, since $\mathcal{X}$ is reflexive, coercive weak lower semicontinuous functions are weak inf-compact. (i) is a consequence of Theorem 3.18(i). Indeed, take a bounded sequence $(x_n)_{n\in\mathbb{N}} \in \mathcal{M}(f_y + \gamma\varphi)$. Then it follows from the reflexivity of $\mathcal{X}$ that $\mathfrak{W}(x_n)_{n\in\mathbb{N}} \neq \varnothing$. (ii) follows at once from Theorem 3.18(ii). (iii) $\Leftrightarrow$ (ii) $\nabla f(y) \in \operatorname{int}\operatorname{dom} (f+\gamma\varphi)^* \Leftrightarrow f+\gamma\varphi-\nabla f(y)$ is coercive by Lemma 3.19. (iv) $\Rightarrow$ (iii) is clear. (v) $\Rightarrow$ (iv) Lemma 3.20(i) yields

$$\begin{aligned} \operatorname{dom} f^* + \gamma \operatorname{dom}\varphi^* &= \operatorname{dom} f^* + \operatorname{dom}\gamma\varphi^*(\cdot/\gamma) = \operatorname{dom} f^* + \operatorname{dom}(\gamma\varphi)^* \\ &= \operatorname{dom}\big(f^* \,\square\, (\gamma\varphi)^*\big) \end{aligned} \tag{3.19}$$

and

$$0 \in \operatorname{int}(\operatorname{dom} f - \operatorname{dom}\varphi) \ \ \Rightarrow \ \ f^* \,\square\, (\gamma\varphi)^* = (f + \gamma\varphi)^*. \tag{3.20}$$

Hence $\operatorname{dom} f^* + \gamma \operatorname{dom}\varphi^* = \mathcal{X}^* \Rightarrow \operatorname{dom}(f + \gamma\varphi)^* = \mathcal{X}^*$. (vi) is a consequence of Theorem 3.18(iii): indeed, by [8, Thm. 5.9(ii)], $\nabla f(y) \in \operatorname{int}\operatorname{dom} f^*$ and f_y is therefore coercive by Lemma 3.19. (vii) $\Rightarrow$ (iv) See [8, Thm. 3.4]. (viii) $\Rightarrow$ (vii) is clear. (ix) is a consequence of Theorem 3.18(iv). $\square$

The next result gathers some facts concerning D-prox operators for convex functions.

PROPOSITION 3.22. *Let* $\varphi\colon \mathcal{X} \to \left]-\infty, +\infty\right]$ *be convex and let* $\gamma \in \left]0, +\infty\right[$. *Then the following hold:*

(i) $\operatorname{prox}_\gamma^\varphi = \big(\partial(f + \gamma\varphi)\big)^{-1} \circ \nabla f$.
(ii) *If, in addition,* $\operatorname{ran}\operatorname{prox}_\gamma^\varphi \subset \operatorname{int}\operatorname{dom} f$, *then*
  (a) $\operatorname{prox}_\gamma^\varphi = R_{\gamma\partial\varphi}$;
  (b) $\operatorname{Fix}\operatorname{prox}_\gamma^\varphi = (\operatorname{int}\operatorname{dom} f) \cap \operatorname{Argmin}\,\varphi$;
  (c) $\operatorname{prox}_\gamma^\varphi$ *is D-firm;*
  (d) $\operatorname{prox}_\gamma^\varphi$ *is single-valued on its domain if* $f|_{\operatorname{int}\operatorname{dom} f}$ *is strictly convex.*

Proof. Fix $y \in \operatorname{int}\operatorname{dom} f$. (i) By (3.18), $\operatorname{ran}\operatorname{prox}_\gamma^\varphi \subset \operatorname{dom} f \cap \operatorname{dom}\varphi$. If $\operatorname{dom} f \cap \operatorname{dom}\varphi = \varnothing$, both sides of the desired identity reduce to the trivial operator $z \mapsto \varnothing$. If not, take $x \in \operatorname{dom} f \cap \operatorname{dom}\varphi$. Since $\operatorname{cont}\nabla f(y) = \mathcal{X}$, Lemma 3.20(ii) yields

$\partial(f_y + \gamma\varphi)(x) = \partial(f + \gamma\varphi)(x) - \nabla f(y)$. Consequently,

$$\begin{aligned} x \in \operatorname{prox}_\gamma^\varphi y &\Leftrightarrow 0 \in \partial(f_y + \gamma\varphi)(x) \\ &\Leftrightarrow \nabla f(y) \in \partial(f + \gamma\varphi)(x) \\ &\Leftrightarrow x \in \big(\partial(f + \gamma\varphi)\big)^{-1}\big(\nabla f(y)\big). \end{aligned} \tag{3.21}$$

(ii) Suppose $\operatorname{ran}\operatorname{prox}_\gamma^\varphi \subset \operatorname{int}\operatorname{dom} f$. (a) On the one hand, it follows from (3.18) that $\operatorname{ran}\operatorname{prox}_\gamma^\varphi \subset (\operatorname{int}\operatorname{dom} f) \cap \operatorname{dom}\varphi$. On the other hand, $\operatorname{ran} R_{\gamma\partial\varphi} \subset \operatorname{dom}(\nabla f + \gamma\partial\varphi) \subset (\operatorname{int}\operatorname{dom} f) \cap \operatorname{dom}\varphi$. Therefore, if $(\operatorname{int}\operatorname{dom} f) \cap \operatorname{dom}\varphi = \varnothing$, both sides of the desired identity reduce to the trivial operator $z \mapsto \varnothing$. If not, take $x \in (\operatorname{int}\operatorname{dom} f) \cap \operatorname{dom}\varphi = \operatorname{cont} f \cap \operatorname{dom}\varphi$. Lemma 3.20(ii) now yields $\partial(f+\gamma\varphi)(x) = \nabla f(x) + \gamma\partial\varphi(x)$ and (3.21) becomes

$$x \in \operatorname{prox}_\gamma^\varphi y \Leftrightarrow \nabla f(y) \in \nabla f(x) + \gamma\partial\varphi(x) \Leftrightarrow x \in R_{\gamma\partial\varphi} y. \tag{3.22}$$

(a) $\Rightarrow$ (b) follows from Proposition 3.8(iii). (a) $\Rightarrow$ (c) Since $\partial\varphi$ is monotone, $R_{\gamma\partial\varphi}$ is D-firm by Proposition 3.8(iv)(a). (a) $\Rightarrow$ (d) follows from Proposition 3.8(iv)(b). ∎

We now turn our attention to the range requirement for the D-viability of D-prox operators.

PROPOSITION 3.23. *Let $\varphi\colon \mathcal{X} \to]-\infty, +\infty]$ be convex such that $\operatorname{dom} f \cap \operatorname{dom}\varphi \neq \varnothing$, and let $\gamma \in]0, +\infty[$. Assume that one of the following conditions holds:*

(i) $\operatorname{dom}\partial(f + \gamma\varphi) \subset \operatorname{int}\operatorname{dom} f$.
(ii) $\operatorname{dom} f \cap \operatorname{dom}\varphi \subset \operatorname{int}\operatorname{dom} f$.
(iii) $\operatorname{dom} f$ *is open.*
(iv) $\operatorname{dom}\varphi \subset \operatorname{int}\operatorname{dom} f$.
(v) $(\operatorname{int}\operatorname{dom} f) \cap \operatorname{dom}\varphi \neq \varnothing$ *and one of the following conditions holds:*
  (a) $\operatorname{dom}\partial f \cap \operatorname{dom}\partial\varphi \subset \operatorname{int}\operatorname{dom} f$.
  (b) f *is essentially smooth.*
  (c) $\operatorname{dom}\partial\varphi \subset \operatorname{int}\operatorname{dom} f$.

Then $\operatorname{ran}\operatorname{prox}_\gamma^\varphi \subset \operatorname{int}\operatorname{dom} f$.

Proof. (i) By Proposition 3.22(i),

$$\operatorname{ran}\operatorname{prox}_\gamma^\varphi \subset \operatorname{ran}\big(\partial(f + \gamma\varphi)\big)^{-1} = \operatorname{dom}\partial(f + \gamma\varphi) \subset \operatorname{int}\operatorname{dom} f. \tag{3.23}$$

(ii) $\Rightarrow$ (i) $\operatorname{dom}\partial(f + \gamma\varphi) \subset \operatorname{dom}(f + \gamma\varphi) = \operatorname{dom} f \cap \operatorname{dom}\varphi \subset \operatorname{int}\operatorname{dom} f$. (iii) $\Rightarrow$ (ii) and (iv) $\Rightarrow$ (ii) are clear. (v) $\Rightarrow$ (i) It results from Lemma 3.20(ii) that $\partial(f + \gamma\varphi) = \partial f + \gamma\partial\varphi$. Whence, (a) $\Rightarrow$ (i). (b) $\Rightarrow$ (a) Essential smoothness $\Rightarrow \operatorname{dom}\partial f = \operatorname{int}\operatorname{dom} f$ [8, Thm. 5.6(iii)]. (c) $\Rightarrow$ (a) is clear. ∎

Upon combining Propositions 3.23, 3.22(ii)(c), 3.21, and 3.5(ii), we obtain the following theorem.

THEOREM 3.24. *Let $\varphi\colon \mathcal{X} \to]-\infty, +\infty]$ be a lower semicontinuous convex function such that $\operatorname{dom} f \cap \operatorname{dom}\varphi \neq \varnothing$, and let $\gamma \in]0, +\infty[$. Suppose that $\mathcal{X}$ is reflexive and that one of conditions* (i)–(ix) *in Proposition* 3.21 *holds together with one of conditions* (i)–(v) *in Proposition* 3.23. *Then* $\operatorname{prox}_\gamma^\varphi \in \mathfrak{B}$.

The following special case underscores the importance of the notion of Legendreness.

COROLLARY 3.25. *Let $\varphi\colon \mathcal{X} \to]-\infty, +\infty]$ be a lower semicontinuous convex function such that $(\operatorname{int}\operatorname{dom} f) \cap \operatorname{dom}\varphi \neq \varnothing$, and let $\gamma \in]0, +\infty[$. Suppose that $\mathcal{X}$ is reflexive, that f is Legendre, and that φ is bounded below. Then*

(i) $\operatorname{prox}_\gamma^\varphi$ *is single-valued on its domain and* $\operatorname{prox}_\gamma^\varphi \in \mathfrak{B}$;

(ii) *for every* x *and* y *in* $\operatorname{int}\operatorname{dom} f$,

$$x = \operatorname{prox}_\gamma^\varphi y \;\Leftrightarrow\; (\forall z \in \operatorname{dom}\varphi)\;\; \langle z - x, \nabla f(y) - \nabla f(x)\rangle / \gamma + \varphi(x) \le \varphi(z).$$

Proof. (i) Combine Propositions 3.23(v)(b), 3.22(ii)(c) and (d), 3.21(vi), and 3.5(ii). (ii) By (3.22), $x = \operatorname{prox}_\gamma^\varphi y \Leftrightarrow \nabla f(y) - \nabla f(x) \in \gamma\partial\varphi(x)$. ∎

Remark 3.26. A special case of Theorem 3.18(iii) in $\mathbb{R}^N$ can be found in [32, Prop. 3.1]. In $\mathbb{R}^N$, assertions (iv) and (v)(b) of Proposition 3.23 appear in [58, Lemma 3.3]. In the case when $\mathcal{X}$ is Hilbertian and $f = \|\cdot\|^2/2$, the characterization supplied by Corollary 3.25(ii) is well known; see, e.g., [48, section II.2].

3.5. *D*-projections. The following concept goes back to Bregman's original paper [15].

DEFINITION 3.27. *The* D-projector *onto a set* $C \subset \mathcal{X}$ *is the operator*

$$\begin{aligned} P_C\colon \mathcal{X} &\to 2^{\mathcal{X}},\\ y &\mapsto \big\{x \in C \cap \operatorname{dom} f \mid D(x,y) = D_C(y) < +\infty\big\}. \end{aligned} \tag{3.24}$$

It is clear that, for any $\gamma \in\,]0,+\infty[$, $P_C = \operatorname{prox}_\gamma^{\iota_C}$. Hence, the results of section 3.4 will automatically yield results on D-projections when specialized to $\varphi = \iota_C$. Before we proceed in this direction, let us introduce a couple of definitions, which are natural adaptations of standard ones in metric approximation theory [81].

DEFINITION 3.28. *A set* $C \subset \mathcal{X}$ *is* D-proximinal *if* $\operatorname{dom} P_C = \operatorname{int}\operatorname{dom} f$ *and* D-semi-Chebyshev *if* P_C *is single-valued on its domain.* C *is* D-Chebyshev *if it is* D*-proximinal and* D*-semi-Chebyshev.*

DEFINITION 3.29. *A set* $C \subset \mathcal{X}$ *is* D-approximately weakly compact *if*

$$(\forall y \in \operatorname{int}\operatorname{dom} f)(\forall (x_n)_{n\in\mathbb{N}} \text{ in } C \cap \operatorname{dom} f)\;\; D(x_n, y) \to D_C(y) \;\Rightarrow\; \mathfrak{W}(x_n)_{n\in\mathbb{N}} \cap C \neq \varnothing.$$

THEOREM 3.30. *Let* C *be a subset of* $\mathcal{X}$ *such that* $C \cap \operatorname{dom} f \neq \varnothing$ *and assume that one of the following conditions holds:*

(i) C *is* D*-approximately weakly compact.*
(ii) $(\forall y \in \operatorname{int}\operatorname{dom} f)(\exists\, \eta \in \mathbb{R})\;\; C \cap \operatorname{lev}_{\le\eta} f_y$ *is nonempty and weakly compact.*
(iii) C *is weakly closed and, for every* $y \in \operatorname{int}\operatorname{dom} f$, f_y *is weak inf-compact.*
(iv) C *is weakly compact.*

Then C *is* D*-proximinal.*

Proof. (i) Since f is weak lower semicontinuous, $f + \iota_C$ is weak lower semicontinuous at every point in C. Now fix $y \in \operatorname{int}\operatorname{dom} f$ and $(x_n)_{n\in\mathbb{N}} \in \mathcal{M}(f_y + \iota_C)$. Then $D(x_n, y) \to D_C(y)$ and Definition 3.29 yields $\mathfrak{W}(x_n)_{n\in\mathbb{N}} \cap C \neq \varnothing$. Now take $x \in \mathfrak{W}(x_n)_{n\in\mathbb{N}} \cap C$. Since $f + \iota_C$ is weak lower semicontinuous at x, the claims follow from Theorem 3.18(i) with $\varphi = \iota_C$. (ii) Fix $y \in \operatorname{int}\operatorname{dom} f$. As minimizing $D(\cdot, y)$ over C is equivalent to minimizing the weak lower semicontinuous function f_y over the weakly compact set $C \cap \operatorname{lev}_{\le\eta} f_y$, the result follows. Assertions (iii) and (iv) follow, respectively, from assertions (iii) and (iv) in Theorem 3.18 with $\varphi = \iota_C$. ∎

Upon setting $\varphi = \iota_C$, Proposition 3.21 becomes the following.

PROPOSITION 3.31. *Let* C *be a closed and convex subset of* $\mathcal{X}$ *such that* $C \cap \operatorname{dom} f \neq \varnothing$. *Suppose that* $\mathcal{X}$ *is reflexive and that one of the following conditions holds:*

(i) $(\forall y \in \operatorname{int}\operatorname{dom} f)(\forall (x_n)_{n\in\mathbb{N}} \in \mathcal{M}(f_y + \iota_C))\;\; \sup_{n\in\mathbb{N}} \|x_n\| < +\infty$.
(ii) $(\forall y \in \operatorname{int}\operatorname{dom} f)\;\; f_y + \iota_C$ *is coercive.*
(iii) $\operatorname{ran}\nabla f \subset \operatorname{int}\operatorname{dom}(f + \iota_C)^*$.

(iv) $f + \iota_C$ *is cofinite.*
(v) $0 \in \operatorname{int}(\operatorname{dom} f - C)$ *and* $\operatorname{dom} f^* + \operatorname{dom} \iota_C^* = \mathcal{X}^*$.
(vi) f *is essentially strictly convex.*
(vii) $f + \iota_C$ *is supercoercive.*
(viii) f *is supercoercive.*
(ix) C *is bounded.*

Then C *is* D*-proximinal.*

Likewise, Proposition 3.22 with $\varphi = \iota_C$ yields the following.

PROPOSITION 3.32. *Let* C *be a convex subset of* $\mathcal{X}$. *Then the following hold:*
(i) $P_C = \big(\partial(f + \iota_C)\big)^{-1} \circ \nabla f$.
(ii) *If, in addition,* $\operatorname{ran} P_C \subset \operatorname{int} \operatorname{dom} f$, *then*
  (a) $P_C = R_{N_C}$.
  (b) $\operatorname{Fix} P_C = C \cap \operatorname{int} \operatorname{dom} f$.
  (c) P_C *is* D*-firm.*
  (d) C *is* D*-semi-Chebyshev if* $f|_{\operatorname{int} \operatorname{dom} f}$ *is strictly convex.*

The D-viability requirements for the range of P_C are obtained by setting $\varphi = \iota_C$ in Proposition 3.23.

PROPOSITION 3.33. *Let* $C \subset \mathcal{X}$ *be convex such that* $C \cap \operatorname{dom} f \neq \emptyset$. *Assume that one of the following conditions holds:*
(i) $\operatorname{dom} \partial(f + \iota_C) \subset \operatorname{int} \operatorname{dom} f$.
(ii) $C \cap \operatorname{dom} f \subset \operatorname{int} \operatorname{dom} f$.
(iii) $\operatorname{dom} f$ *is open.*
(iv) $C \subset \operatorname{int} \operatorname{dom} f$.
(v) $C \cap \operatorname{int} \operatorname{dom} f \neq \emptyset$ *and one of the following conditions holds:*
  (a) $C \cap \operatorname{dom} \partial f \subset \operatorname{int} \operatorname{dom} f$;
  (b) f *is essentially smooth.*

Then $\operatorname{ran} P_C \subset \operatorname{int} \operatorname{dom} f$.

THEOREM 3.34. *Let* $C \subset \mathcal{X}$ *be a closed convex set such that* $C \cap \operatorname{dom} f \neq \emptyset$. *Suppose that* $\mathcal{X}$ *is reflexive and that one of conditions* (i)–(ix) *in Proposition* 3.31 *holds together with one of conditions* (i)–(v) *in Proposition* 3.33. *Then* $P_C \in \mathfrak{B}$.

Proof. Since Proposition 3.31 parallels Proposition 3.21 and Proposition 3.33 parallels Proposition 3.23, it suffices to set $\varphi = \iota_C$ in Theorem 3.24. □

We conclude this section with the following result.

COROLLARY 3.35. *Suppose that* $\mathcal{X}$ *is reflexive, that* f *is Legendre, and that* C *is a closed convex subset of* $\mathcal{X}$ *such that* $C \cap \operatorname{int} \operatorname{dom} f \neq \emptyset$. *Then*
(i) C *is* D*-Chebyshev and* $P_C \in \mathfrak{B}$;
(ii) *for every* x *and* y *in* $\operatorname{int} \operatorname{dom} f$,

$$x = P_C y \quad \Leftrightarrow \quad \begin{cases} x \in C, \\ C \subset H(y, x). \end{cases} \tag{3.25}$$

Proof. Take $\varphi = \iota_C$ in Corollary 3.25. □

Remark 3.36. Proposition 3.31(vii)–(ix) can be found in [1, Prop. 2.1]. Corollary 3.35(i) covers [8, Cor. 7.9] (see also [7, section 3] in the special case of Euclidean spaces), which was obtained via different arguments. If $\mathcal{X}$ is Hilbertian and $f = \|\cdot\|^2/2$, Corollary 3.35(ii) reduces to the classical characterization of metric projections onto closed convex sets.

3.6. Subgradient *D*-projections. The D-projection onto a closed convex set may be hard to compute. If the set is specified as a lower level set, it can be approximated by the D-projection onto a separating hyperplane, which is much easier

to compute. In the traditional case when $\mathcal{X}$ is Hilbertian and $f = \|\cdot\|^2/2$, this is a standard approach which goes back to [73] (see also [6, 37, 60]). In the context of Bregman distances, we shall define subgradient D-projections as follows (see also [27, 59] for special instances).

DEFINITION 3.37. *Suppose that*

$$\begin{cases} \mathcal{X} \text{ is reflexive and } f \text{ is Legendre,} \\ g\colon \mathcal{X} \to \left]-\infty,+\infty\right] \text{ is lower semicontinuous and convex,} \\ \operatorname{lev}_{\leq 0} g \cap \operatorname{int}\operatorname{dom} f \neq \varnothing \;\text{ and }\; \operatorname{dom} f \subset \operatorname{dom} g. \end{cases} \tag{3.26}$$

For every $x \in \operatorname{int}\operatorname{dom} f$ *and* $x^* \in \partial g(x)$, *set*

$$G(x,x^*) = \big\{y \in \mathcal{X} \mid \langle x-y, x^*\rangle \geq g(x)\big\}. \tag{3.27}$$

The operator

$$Q_g\colon \operatorname{int}\operatorname{dom} f \to \mathcal{X}\colon x \mapsto \big\{P_{G(x,x^*)}x \mid x^* \in \partial g(x)\big\} \tag{3.28}$$

is the subgradient D-projector *onto* $\operatorname{lev}_{\leq 0} g$.

Note that $G(x,x^*)$ is a proper closed half-space if $x^* \neq 0$ and the whole space $\mathcal{X}$ otherwise; the latter may occur only when $x \in \operatorname{Argmin} g$.

PROPOSITION 3.38. *Suppose that* (3.26) *is in force and let* Q_g *be the subgradient* D*-projector onto* $\operatorname{lev}_{\leq 0} g$. *Then*

(i) $\operatorname{Fix} Q_g = \operatorname{lev}_{\leq 0} g \cap \operatorname{int}\operatorname{dom} f$;

(ii) $Q_g \in \mathfrak{B}$.

Proof. Fix $x \in \operatorname{int}\operatorname{dom} f$ and $x^* \in \partial g(x)$. Since $\operatorname{int}\operatorname{dom} f \subset \operatorname{int}\operatorname{dom}\, g \subset \operatorname{dom}\partial g$, $\partial g(x) \neq \varnothing$ and the closed convex set $G(x,x^*)$ is well defined. Moreover, (2.2) yields

$$(\forall y \in \operatorname{lev}_{\leq 0}\, g)\;\; \langle y-x, x^*\rangle \leq g(y) - g(x) \leq -g(x). \tag{3.29}$$

Therefore, $\operatorname{lev}_{\leq 0}\, g \subset G(x,x^*)$ and, in turn, $G(x,x^*) \cap \operatorname{int}\operatorname{dom} f \neq \varnothing$. Hence, Corollary 3.35(i) asserts that $P_{G(x,x^*)}$ is single-valued with $\operatorname{ran} P_{G(x,x^*)} \subset \operatorname{int}\operatorname{dom} f = \operatorname{dom} P_{G(x,x^*)}$, whence $\operatorname{ran} Q_g \subset \operatorname{int}\operatorname{dom} f = \operatorname{dom} Q_g$. (i) Take $y \in \mathcal{X}$. Then it follows from Proposition 3.32(ii)(b) that

$$\begin{aligned} y \in \operatorname{Fix} Q_g &\Leftrightarrow (\exists\, y^* \in \partial g(y))\;\; y = P_{G(y,y^*)}y \\ &\Leftrightarrow (\exists\, y^* \in \partial g(y))\;\; y \in G(y,y^*) \cap \operatorname{int}\operatorname{dom} f \\ &\Leftrightarrow (\exists\, y^* \in \partial g(y))\;\; 0 = \langle y-y, y^*\rangle \geq g(y) \;\text{ and }\; y \in \operatorname{int}\operatorname{dom} f \\ &\Leftrightarrow y \in \operatorname{lev}_{\leq 0}\, g \cap \operatorname{int}\operatorname{dom} f. \end{aligned}$$

Thus, $\operatorname{Fix} Q_g = \operatorname{lev}_{\leq 0}\, g \cap \operatorname{int}\operatorname{dom} f$. (ii) To show that $Q_g \in \mathfrak{B}$ observe that Corollary 3.35(ii) implies that $G(x,x^*) \subset H(x, P_{G(x,x^*)}x)$. Consequently, $\operatorname{Fix} Q_g \subset \operatorname{lev}_{\leq 0} g \subset G(x,x^*) \subset H(x, P_{G(x,x^*)}x)$, where $(x, P_{G(x,x^*)}x)$ is an arbitrary point in $\operatorname{gr} Q_g$. Altogether, $Q_g \in \mathfrak{B}$. ∎

3.7. Relaxed parallel combination of $\mathfrak{B}$-class operators. The following proposition describes a scheme to aggregate $\mathfrak{B}$-class operators in order to create a new $\mathfrak{B}$-class operator.

PROPOSITION 3.39. *Suppose that* $\mathcal{X}$ *is reflexive and that* f *is Legendre. Let* $(T_i)_{i\in I}$ *be a finite family of operators in* $\mathfrak{B}$ *such that* $\bigcap_{i\in I} \operatorname{Fix} T_i \neq \varnothing$, *let* $(\omega_i)_{i\in I}$ *be*

weights in $]0,1]$ *such that* $\sum_{i\in I}\omega_i = 1$, *and let* λ *be a relaxation parameter in* $]0,1]$. *For every* $x \in \operatorname{int}\operatorname{dom} f$, *select* $(u_i)_{i\in I} \in \bigtimes_{i\in I} T_i x$, *put*

$$H(x) = \big\{y \in \mathcal{X} \mid \langle y, x^*\rangle \leq \eta(x)\big\}, \tag{3.30}$$

where

$$\begin{cases} x^* = \nabla f(x) - \sum_{i\in I}\omega_i \nabla f(u_i), \\ \eta(x) = \sum_{i\in I}\omega_i \left\langle x + \lambda(u_i - x), \nabla f(x) - \nabla f(u_i)\right\rangle, \end{cases} \tag{3.31}$$

and define $T\colon \operatorname{int}\operatorname{dom} f \to \mathcal{X}\colon x \mapsto P_{H(x)}x$. *Then the following hold:*

(i) T *is single-valued on* $\operatorname{dom} T = \operatorname{int}\operatorname{dom} f \supset \operatorname{ran} T$.
(ii) *For every* $x \in \operatorname{int}\operatorname{dom} f$, *the following statements are equivalent:*
  (a) $x \in \bigcap_{i\in I}\operatorname{Fix} T_i$.
  (b) $x^* = 0$.
  (c) $H(x) = \mathcal{X}$.
  (d) $x \in H(x)$.
  (e) $x \in \operatorname{Fix} T$.
(iii) $\operatorname{Fix} T = \bigcap_{i\in I}\operatorname{Fix} T_i$.
(iv) $\overline{\operatorname{Fix}}\, T = \bigcap_{i\in I}\overline{\operatorname{Fix}}\, T_i$.
(v) $(\forall x \in \operatorname{int}\operatorname{dom} f)\ H(x) = H(x, Tx)$.
(vi) $T \in \mathfrak{B}$.

Proof. Fix $x \in \operatorname{int}\operatorname{dom} f$. (i) We first observe that the operator T is well defined. Indeed, since $(T_i)_{i\in I}$ lies in $\mathfrak{B}$, x^* and $\eta(x)$ are well defined and we have

$$\begin{aligned} \emptyset \neq \bigcap_{i\in I}\operatorname{Fix} T_i &\subset (\operatorname{int}\operatorname{dom} f) \cap \bigcap_{i\in I} H(x, u_i) \\ &\subset (\operatorname{int}\operatorname{dom} f) \cap \bigcap_{i\in I} \big\{y \in \mathcal{X} \mid \langle y - u_i, \nabla f(x) - \nabla f(u_i)\rangle \\ &\qquad\qquad \leq (1-\lambda)\langle x - u_i, \nabla f(x) - \nabla f(u_i)\rangle\big\} \\ &\subset (\operatorname{int}\operatorname{dom} f) \cap \Big\{y \in \mathcal{X} \mid \sum_{i\in I}\omega_i \langle y - u_i, \nabla f(x) - \nabla f(u_i)\rangle \\ &\qquad\qquad \leq (1-\lambda)\sum_{i\in I}\omega_i\langle x - u_i, \nabla f(x) - \nabla f(u_i)\rangle\Big\} \\ &= (\operatorname{int}\operatorname{dom} f) \cap H(x), \end{aligned} \tag{3.32}$$

where the second inclusion follows from the inequality $\lambda \leq 1$ and the monotonicity of ∇f. Whence, $(\operatorname{int}\operatorname{dom} f) \cap H(x) \neq \emptyset$, and it follows from Corollary 3.35(i) that $P_{H(x)}x$ is a well-defined point in $\operatorname{int}\operatorname{dom} f$. (ii) Since f is essentially strictly convex, it is strictly convex on $\operatorname{int}\operatorname{dom} f$ and it follows from Proposition 3.3(vi) that (a) $\Rightarrow$ $(\forall i \in I)\ u_i = x \Rightarrow$ (b). (b) $\Rightarrow$ (c) Suppose $x^* = 0$ and fix $y \in \bigcap_{i\in I}\operatorname{Fix} T_i$. Then,

since $(T_i)_{i\in I}$ lies in $\mathfrak{B}$,

$$\begin{aligned}0 &\leq \sum_{i\in I}\omega_i\,\langle u_i - y, \nabla f(x) - \nabla f(u_i)\rangle\\ &= \eta(x) - \langle y, x^*\rangle - (1-\lambda)\sum_{i\in I}\omega_i\,\langle x - u_i, \nabla f(x) - \nabla f(u_i)\rangle\\ &\leq \eta(x). \qquad (3.33)\end{aligned}$$

Accordingly, $H(x) = \mathcal{X}$. The implications (c) $\Rightarrow$ (d) $\Rightarrow$ $x = P_{H(x)}x$ $\Rightarrow$ (e) are clear in view of Proposition 3.32(ii)(b). (e) $\Rightarrow$ (a) We have

$$\begin{aligned}x \in \operatorname{Fix} T &\Leftrightarrow x = P_{H(x)}x\\ &\Leftrightarrow x \in H(x)\\ &\Leftrightarrow \langle x, x^*\rangle \leq \eta(x)\\ &\Leftrightarrow \lambda\sum_{i\in I}\omega_i\,\langle x - u_i, \nabla f(x) - \nabla f(u_i)\rangle \leq 0\\ &\Leftrightarrow (\forall i \in I)\ \ x = u_i \in T_i x\\ &\Leftrightarrow x \in \bigcap_{i\in I}\operatorname{Fix} T_i,\end{aligned}$$

where the next to last equivalence follows from the strict monotonicity of ∇f on $\operatorname{int}\operatorname{dom} f$ (f is strictly convex on $\operatorname{int}\operatorname{dom} f$) and the inequalities $\lambda > 0$ and $\min_{i\in I}\omega_i > 0$. (iii) (i) and (ii) yield $\operatorname{Fix} T = (\operatorname{int}\operatorname{dom} f) \cap \operatorname{Fix} T = \bigcap_{i\in I}(\operatorname{Fix} T_i \cap \operatorname{int}\operatorname{dom} f) = \bigcap_{i\in I}\operatorname{Fix} T_i$. (iv) Set $(\forall i \in I)\ F_i = \bigcap_{(x,u)\in \operatorname{gr} T_i} H(x,u)$. Then (iii) and Proposition 3.3(iv) yield $\operatorname{Fix} T = (\operatorname{int}\operatorname{dom} f) \cap \bigcap_{i\in I} F_i$. Therefore, by Lemma 3.2 and Proposition 3.3(vii),

$$\overline{\operatorname{Fix}}\, T = \overline{\operatorname{dom}}\, f \cap \bigcap_{i\in I} F_i = \bigcap_{i\in I}(F_i \cap \overline{\operatorname{dom}}\, f) = \bigcap_{i\in I}\overline{\operatorname{Fix}}\, T_i. \qquad (3.34)$$

(v) By Corollary 3.35(ii), we always have $H(x) \subset H(x, P_{H(x)}x) = H(x, Tx)$. Now suppose $x \in H(x)$. Then (ii) yields $H(x) = \mathcal{X} = H(x,x) = H(x, P_{H(x)}x) = H(x, Tx)$. Next, suppose $x \notin H(x)$. Then (ii) yields $x^* \neq 0$ and $H(x)$ is therefore a proper closed half-space in $\mathcal{X}$. On the other hand, $x \neq P_{H(x)}x = Tx$ and, since ∇f is injective [8, Thm. 5.10], $\nabla f(x) \neq \nabla f(Tx)$. Consequently, $H(x, Tx)$ is also a proper closed half-space in $\mathcal{X}$. Since $Tx \in H(x) \cap \operatorname{bdry} H(x, Tx)$ and $H(x) \subset H(x, Tx)$, we conclude $H(x) = H(x, Tx)$. (vi) It follows successively from (iii), (3.32), and (v) that $\operatorname{Fix} T = \bigcap_{i\in I}\operatorname{Fix} T_i \subset H(x) = H(x, Tx)$. In view of (i), the proof is complete. ∎

4. Bregman monotonicity.

4.1. Properties. *D*-monotonicity was introduced in Definition 1.2. We first collect some elementary properties.

PROPOSITION 4.1. *Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{X}$ which is D-monotone with respect to a set $S \subset \mathcal{X}$. Then the following hold:*

(i) $(\forall x \in S \cap \operatorname{dom} f)\ \big(D(x, x_n)\big)_{n\in\mathbb{N}}$ *converges.*

(ii) $(\forall n \in \mathbb{N})\ \ D_S(x_{n+1}) \leq D_S(x_n)$.

(iii) $\big(D_S(x_n)\big)_{n\in\mathbb{N}}$ *converges.*

(iv) $(\forall (x, x') \in (S \cap \operatorname{dom} f)^2)\ \big(\langle x - x', \nabla f(x_n)\rangle\big)_{n\in\mathbb{N}}$ *converges.*

(v) $(x_n)_{n\in\mathbb{N}}$ *is bounded if, for some* $z \in S \cap \operatorname{dom} f$, *the set* $\operatorname{lev}_{\leq D(z,x_0)} D(z,\cdot)$ *is bounded. This is true in particular if* $S \cap \operatorname{int}\operatorname{dom} f \neq \varnothing$, $\mathcal{X}$ *is reflexive, and one of the following properties is satisfied:*
  (a) f *is supercoercive;*
  (b) $\dim \mathcal{X} < +\infty$ *and* $\operatorname{dom} f^*$ *is open.*

Proof. (i) and (ii) are immediate consequences of Definition 1.2, and (iii) follows from (ii). (iv) Take x and x' in $S \cap \operatorname{dom} f$. By (i), the sequences $\big(f(x_n) + \langle x - x_n, \nabla f(x_n)\rangle\big)_{n\in\mathbb{N}}$ and $\big(f(x_n)+\langle x' - x_n, \nabla f(x_n)\rangle\big)_{n\in\mathbb{N}}$ converge and so does their difference $\big(\langle x - x', \nabla f(x_n)\rangle\big)_{n\in\mathbb{N}}$. (v) By definition, for every $x \in S\cap\operatorname{dom} f$, $(x_n)_{n\in\mathbb{N}}$ lies in $\operatorname{lev}_{\leq D(z,x_0)} D(z,\cdot)$. The second assertion follows from [8, Lemma 7.3(viii) and (ix)], which asserts that $D(z,\cdot)$ is coercive under the stated assumptions if $z \in \operatorname{int}\operatorname{dom} f$. ∎

The following example shows that the conclusion of Proposition 4.1(v) may hold even though the properties (a) and (b) are not satisfied.

EXAMPLE 4.2. *Let* $\mathcal{X} = \ell_2(\mathbb{N})$ *and define*

(4.1)
$$f\colon \mathcal{X} \to \left]-\infty,+\infty\right] : x = (\xi_k)_{k\in\mathbb{N}} \mapsto \begin{cases} \sum_{k\in\mathbb{N}} \xi_k - \ln(1+\xi_k) & \text{if } (\forall k \in \mathbb{N})\ \ \xi_k > -1, \\ +\infty & \text{otherwise.} \end{cases}$$

Then f *is Legendre and* $\operatorname{dom} f$ *is open. Moreover,* $\operatorname{lev}_{\leq \eta} D(0,\cdot)$ *is bounded for* $\eta > 0$ *sufficiently small.*

Proof. We only sketch the arguments, as the example is not utilized elsewhere. Observe that f is separable: $(\forall x \in \mathcal{X})\ f(x) = \sum_{k\in\mathbb{N}} h(\xi_k)$, where

$$(\forall \xi \in \mathbb{R})\ \ h(\xi) = \begin{cases} \xi - \ln(1+\xi) & \text{if } \xi > -1, \\ +\infty & \text{otherwise.} \end{cases} \tag{4.2}$$

Using calculus, one verifies that $\operatorname{dom} f = \big\{x \in \mathcal{X} \mid (\forall k \in \mathbb{N})\ \ \xi_k > -1\big\}$, which is open. Also, f is Gâteaux-differentiable on its domain with $\nabla f(x) = \big(\xi_k/(1+\xi_k)\big)_{k\in\mathbb{N}}$. Hence f is essentially smooth. Now $(\forall x \in \mathcal{X})\ f^*(x) = f(-x)$. Thus f^* is essentially smooth as well. By [8, Thm. 5.4], f is essentially strictly convex. Altogether, f is Legendre. Let $\alpha = \ln(2) - 1/2$. A careful analysis of the Bregman distance D_h associated with h reveals that $D_h(0,\xi) < \alpha \Rightarrow |\xi| < 1 \Rightarrow D_h(0,\xi) \geq \alpha|\xi|^2$. (In passing, we point out that $D_h(0,\cdot)$ is convex precisely on $]-1,+1[$.) Fix $\eta \in [0,\alpha[$ and $x \in \mathcal{X}$ such that $D(0,x) \leq \eta$. Then $(\forall k \in \mathbb{N})\ D_h(0,\xi_k) \geq \alpha|\xi_k|^2$. Summing yields $\eta \geq D(0,x) \geq \alpha\|x\|^2$, whence $x \in B(0;\sqrt{\eta/\alpha})$. ∎

The next two assumptions will be quite helpful in the analysis of the convergence of D-monotone sequences.

CONDITION 4.3. *Given* $S \subset \mathcal{X}$, *for every bounded sequence* $(x_n)_{n\in\mathbb{N}}$ *in* $\operatorname{int}\operatorname{dom} f$, *one has*

$$\begin{cases} x \in \mathfrak{W}(x_n)_{n\in\mathbb{N}} \cap S, \\ x' \in \mathfrak{W}(x_n)_{n\in\mathbb{N}} \cap S, \\ (x_n)_{n\in\mathbb{N}} \text{ is } D\text{-monotone with respect to } S \end{cases} \quad \Rightarrow \quad x = x'. \tag{4.3}$$

CONDITION 4.4. *For all bounded sequences* $(x_n)_{n\in\mathbb{N}}$ *and* $(y_n)_{n\in\mathbb{N}}$ *in* $\operatorname{int}\operatorname{dom} f$, *one has*

$$D(x_n, y_n) \to 0 \quad \Rightarrow \quad x_n - y_n \to 0. \tag{4.4}$$

These two assumptions cover familiar situations, as the following examples show.

EXAMPLE 4.5. *Suppose that S is a subset of $\mathcal{X}$ such that $S \cap \overline{\operatorname{dom}} f$ is a singleton. Then Condition 4.3 is satisfied.*

Proof. Take $(x_n)_{n\in\mathbb{N}}$ in $\operatorname{int}\operatorname{dom} f$. Then $\mathfrak{W}(x_n)_{n\in\mathbb{N}} \subset \overline{\operatorname{dom}} f$ and, therefore, $\mathfrak{W}(x_n)_{n\in\mathbb{N}} \cap S$ is at most a singleton. ∎

EXAMPLE 4.6. *Suppose that $S \subset \operatorname{int}\operatorname{dom} f$ is convex, $f|_S$ is strictly convex, and ∇f is sequentially weak-to-weak$*$ continuous at every point in S. Then Condition 4.3 is satisfied.*

Proof. Let $(x_n)_{n\in\mathbb{N}}$ be a bounded sequence which is D-monotone with respect to S. Then $x_{k_n} \rightharpoonup x \in S$ and $x_{l_n} \rightharpoonup x' \in S$ imply $\nabla f(x_{k_n}) \xrightarrow{*} \nabla f(x)$ and $\nabla f(x_{l_n}) \xrightarrow{*} \nabla f(x')$. Proposition 4.1(iv) therefore forces $\langle x - x', \nabla f(x)\rangle = \langle x - x', \nabla f(x')\rangle$; hence $\langle x - x', \nabla f(x) - \nabla f(x')\rangle = 0$. Since ∇f is strictly monotone on S, we get $x = x'$. ∎

Our next example requires the following lemma.

LEMMA 4.7. *Suppose that $\varepsilon \in \left]0, +\infty\right[$, $x \in \operatorname{dom} f$, and $y \in \operatorname{int}\operatorname{dom} f$. Then there exists $z \in \operatorname{int}\operatorname{dom} f$ such that $\|x - z\| \leq \varepsilon$ and $|D(x, y) - D(z, y)| \leq \varepsilon$.*

Proof. Put $(\forall \alpha \in [0,1[)\ x_\alpha = (1-\alpha)y + \alpha x$. Then $(x_\alpha)_{\alpha\in[0,1[}$ lies in $\operatorname{int}\operatorname{dom} f$, $\lim_{\alpha\uparrow 1^-} x_\alpha = x$ and, by (3.6), $\lim_{\alpha\uparrow 1^-} D(x_\alpha, y) = D(x, y)$. Thus, for α sufficiently close to 1, we can take $z = x_\alpha$. ∎

We now recall the notion of a Bregman/Legendre function in $\mathbb{R}^N$, which covers numerous functions of importance in convex optimization [7]. This notion will allow us to describe a finite-dimensional setting in which Condition 4.3 holds.

DEFINITION 4.8. *Suppose that $\mathcal{X} = \mathbb{R}^N$ and f is Legendre. Then f is* Bregman/Legendre*, if each of the following conditions is satisfied:*

(i) $\operatorname{dom} f^*$ *is open.*

(ii) $(\forall x \in \operatorname{dom} f \smallsetminus \operatorname{int}\operatorname{dom} f)$ $D(x, \cdot)$ *is coercive.*

(iii)
$$\begin{cases} x \in \operatorname{dom} f \smallsetminus \operatorname{int}\operatorname{dom} f, \\ (y_n)_{n\in\mathbb{N}} \text{ in } \operatorname{int}\operatorname{dom} f, \\ y_n \to y \in \operatorname{bdry}\operatorname{dom} f, \\ \big(D(x, y_n)\big)_{n\in\mathbb{N}} \text{ bounded} \end{cases} \quad \Rightarrow \quad D(y, y_n) \to 0.$$

(iv)
$$\begin{cases} (x_n)_{n\in\mathbb{N}} \text{ in } \operatorname{int}\operatorname{dom} f, \\ (y_n)_{n\in\mathbb{N}} \text{ in } \operatorname{int}\operatorname{dom} f, \\ x_n \to x \in \operatorname{dom} f \smallsetminus \operatorname{int}\operatorname{dom} f, \\ y_n \to y \in \operatorname{dom} f \smallsetminus \operatorname{int}\operatorname{dom} f, \\ D(x_n, y_n) \to 0 \end{cases} \quad \Rightarrow \quad x = y.$$

EXAMPLE 4.9. *Suppose that $\mathcal{X} = \mathbb{R}^N$, f is Bregman/Legendre, and S is a subset of $\mathcal{X}$ such that $S \cap \operatorname{dom} f \neq \varnothing$. Then Condition 4.3 is satisfied.*

Proof. Let us start with two useful facts, namely

$$\text{(4.5)} \qquad \begin{cases} x \in \operatorname{dom} f, \\ (y_n)_{n\in\mathbb{N}} \text{ in } \operatorname{int}\operatorname{dom} f, \\ y_n \to y, \\ \big(D(x, y_n)\big)_{n\in\mathbb{N}} \text{ bounded} \end{cases} \quad \Rightarrow \quad \begin{cases} D(y, y_n) \to 0, \\ y \in \operatorname{dom} f \end{cases}$$

and

$$\begin{cases} x \in \operatorname{dom} f, \\ (y_n)_{n\in\mathbb{N}} \text{ in } \operatorname{int}\operatorname{dom} f, \\ y_n \to y \in \operatorname{dom} f, \\ D(x, y_n) \to 0 \end{cases} \quad \Rightarrow \quad x = y. \tag{4.6}$$

If $x \in \operatorname{int}\operatorname{dom} f$, (4.5) follows from [7, Thm. 3.8(ii)]. On the other hand, if $x \in \operatorname{dom} f \smallsetminus \operatorname{int}\operatorname{dom} f$, (4.5) follows from [7, Prop. 3.3] if $y \in \operatorname{int}\operatorname{dom} f$ and from [7, Def. 5.2.BL2] if $y \in \operatorname{bdry}\operatorname{dom} f$. We now turn to (4.6). If x or y belongs to $\operatorname{int}\operatorname{dom} f$, it suffices to apply [7, Thm. 3.9(iii)]. Otherwise, $\{x, y\} \subset \operatorname{dom} f \smallsetminus \operatorname{int}\operatorname{dom} f$ and Lemma 4.7 ensures that, for every $n \geq 1$, we can find a point $x_n \in \operatorname{int}\operatorname{dom} f$ such that $\|x - x_n\| \leq 1/n$ and $|D(x, y_n) - D(x_n, y_n)| \leq 1/n$. Therefore, $x_n \to x$ and, since $D(x, y_n) \to 0$ by assumption, $D(x_n, y_n) \to 0$. It then follows from [7, Def. 5.2.BL3] that $x = y$. Now let $(x_n)_{n\in\mathbb{N}}$ be a bounded sequence which is D-monotone with respect to S and let $z \in S \cap \operatorname{dom} f$. Suppose $x_{k_n} \to x \in S$ and $x_{l_n} \to x' \in S$. Since by D-monotonicity the sequences $\big(D(z, x_{k_n})\big)_{n\in\mathbb{N}}$ and $\big(D(z, x_{l_n})\big)_{n\in\mathbb{N}}$ are bounded, (4.5) yields $D(x, x_{k_n}) \to 0$, $D(x', x_{l_n}) \to 0$, and $\{x, x'\} \subset S \cap \operatorname{dom} f$. However, it follows from Proposition 4.1(i) that $D(x, x_{k_n}) \to 0 \Rightarrow D(x, x_n) \to 0 \Rightarrow D(x, x_{l_n}) \to 0$. In view of (4.6), we conclude $x = x'$, as required. □

Following [25], we say that f is uniformly convex on bounded sets if, for every bounded set $B \subset \mathcal{X}$, one has

$$(\forall t \in]0, +\infty[) \quad \inf \mu\big(B \cap \operatorname{dom} f, t\big) > 0, \tag{4.7}$$

where

$$\mu\colon \operatorname{dom} f \times [0, +\infty[\to [0, +\infty] : (x, t) \mapsto \inf_{\substack{\|x-y\|=t \\ y\in\operatorname{dom} f}} \frac{f(x) + f(y)}{2} - f\left(\frac{x+y}{2}\right). \tag{4.8}$$

Examples of such functions are given in [84].

The next result gives sufficient conditions for Condition 4.4 to hold. (See also [22] and [82] for item (ii).)

EXAMPLE 4.10. *Condition* 4.4 *is satisfied whenever one of the following is true:*

(i) *f is uniformly convex on bounded sets.*

(ii) *$\mathcal{X} = \mathbb{R}^N$, $\operatorname{dom} f$ is closed, and $f|_{\operatorname{dom} f}$ is strictly convex and continuous.*

(iii) *$\mathcal{X} = \mathbb{R}$ and $f|_{\operatorname{dom} f}$ is strictly convex.*

Proof. (i) is a direct consequence of [25, Prop. 4.2]. (ii) and (iii) are special cases of (i) by [85, Prop. 3.6.6(i)]. □

In passing, we note that it follows from [85, Thm. 3.5.13] that item (i) of Example 4.10 forces the underlying space $\mathcal{X}$ to be reflexive.

The above assumptions lead to remarkably simple weak and strong convergence criteria for D-monotone sequences. In the case when $\mathcal{X}$ is Hilbertian and $f = \|\cdot\|^2/2$, Conditions 4.3 and 4.4 are satisfied and these criteria can essentially be found in [53] (see also [6] and [40]). Recall (see section 2) that $\mathfrak{S}$ denotes the set of strong cluster points of a sequence.

THEOREM 4.11. *Let $(x_n)_{n\in\mathbb{N}}$ be a bounded sequence in $\mathcal{X}$ which is D-monotone with respect to a set $S \subset \mathcal{X}$. Suppose that $\mathcal{X}$ is reflexive and Condition* 4.3 *is satisfied. Then*

(i) *$(x_n)_{n\in\mathbb{N}}$ converges weakly to a point in $S \cap \overline{\operatorname{dom}} f$ if and only if $\mathfrak{W}(x_n)_{n\in\mathbb{N}} \subset S$;*

(ii) *supposing that* $x_n \rightharpoonup x \in S \cap \operatorname{int}\operatorname{dom} f$ *and Condition* 4.4 *is satisfied, then* $x_n \to x$ *if and only if* $\mathfrak{S}(x_n)_{n\in\mathbb{N}} \neq \varnothing$.

Proof. (i) Necessity is clear. To prove sufficiency, suppose that $\mathfrak{W}(x_n)_{n\in\mathbb{N}} \subset S$ and take x and x' in $\mathfrak{W}(x_n)_{n\in\mathbb{N}}$, say $x_{k_n} \rightharpoonup x$ and $x_{l_n} \rightharpoonup x'$. Then x and x' lie in S and (4.3) forces $x = x'$. Since $\mathcal{X}$ is reflexive and $(x_n)_{n\in\mathbb{N}}$ is bounded, we conclude $x_n \rightharpoonup x$. Furthermore, since $\overline{\operatorname{dom}} f \ni x_n \rightharpoonup x$ and $\overline{\operatorname{dom}} f$ is weakly closed, $x \in \overline{\operatorname{dom}} f$. (ii) Necessity is clear. To prove sufficiency, suppose that Condition 4.4 is satisfied, $x \in S \cap \operatorname{int}\operatorname{dom} f$, and $\mathfrak{S}(x_n)_{n\in\mathbb{N}} \neq \varnothing$, i.e., some subsequence $(x_{k_n})_{n\in\mathbb{N}}$ converges strongly. Since $x_n \rightharpoonup x$, we must have $x_{k_n} \to x$. In turn, [8, Lemma 7.3(x)] yields $D(x, x_{k_n}) \to 0$ and it follows from Proposition 4.1(i) that $D(x, x_n) \to 0$. In view of (4.4), we conclude $x_n \to x$. ∎

4.2. Construction.

ALGORITHM 4.12. *Starting with* $x_0 \in \operatorname{int}\operatorname{dom} f$, *at every iteration* $n \in \mathbb{N}$, *select first* $T_n \in \mathfrak{B}$ *and then* $x_{n+1} \in T_n x_n$.

PROPOSITION 4.13. *Let* $(x_n)_{n\in\mathbb{N}}$ *be an arbitrary orbit of Algorithm* 4.12. *Suppose that*

$$\bigcap_{n\in\mathbb{N}} \operatorname{Fix} T_n \neq \varnothing, \quad S \subset \bigcap_{n\in\mathbb{N}} \overline{\operatorname{Fix}}\, T_n, \quad \text{and} \quad S \cap \operatorname{dom} f \neq \varnothing. \tag{4.9}$$

Then

(i) *if* $f|_{\operatorname{int}\operatorname{dom} f}$ *is strictly convex,* $(x_n)_{n\in\mathbb{N}}$ *is* D*-monotone with respect to* S;

(ii) $\sum_{n\in\mathbb{N}} D(x_{n+1}, x_n) < +\infty$.

Proof. (i) Proposition 3.3(viii) yields $(\forall n \in \mathbb{N})(\forall y \in \overline{\operatorname{Fix}}\, T_n)\ D(y, x_{n+1}) \leq D(y, x_n)$. (ii) Fix $y \in \bigcap_{n\in\mathbb{N}} \operatorname{Fix} T_n$. Then Proposition 3.3(i) yields the stronger statement

$$(\forall n \in \mathbb{N})\ \ D(y, x_{n+1}) \leq D(y, x_n) - D(x_{n+1}, x_n). \tag{4.10}$$

Therefore $\sum_{n\in\mathbb{N}} D(x_{n+1}, x_n) \leq D(y, x_0)$. ∎

THEOREM 4.14. *Let* $(x_n)_{n\in\mathbb{N}}$ *be an arbitrary bounded orbit of Algorithm* 4.12. *Suppose that* $\mathcal{X}$ *is reflexive, that* $f|_{\operatorname{int}\operatorname{dom} f}$ *is strictly convex, and that* (4.9) *is satisfied. Suppose in addition that Condition* 4.3 *is satisfied and that*

$$\sum_{n\in\mathbb{N}} D(x_{n+1}, x_n) < +\infty \quad \Rightarrow \quad \mathfrak{W}(x_n)_{n\in\mathbb{N}} \subset S. \tag{4.11}$$

Then

(i) $(x_n)_{n\in\mathbb{N}}$ *converges weakly to a point* $x \in S$;

(ii) *the convergence is strong in* (i) *if* $x \in \operatorname{int}\operatorname{dom} f$, *Condition* 4.4 *is satisfied, and*

$$\sum_{n\in\mathbb{N}} D(x_{n+1}, x_n) < +\infty \quad \Rightarrow \quad \mathfrak{S}(x_n)_{n\in\mathbb{N}} \neq \varnothing. \tag{4.12}$$

Proof. Combine Theorem 4.11 and Proposition 4.13. ∎

5. Parallel block-iterative *D*-monotone algorithm.

5.1. Objective. For the remainder of this paper, we assume that

$$\begin{cases} \mathcal{X} \text{ is reflexive and } f \text{ is Legendre,} \\ (S_i)_{i\in I} \text{ is a countable family of closed convex subsets of } \mathcal{X}, \\ (\operatorname{int}\operatorname{dom} f) \cap \bigcap_{i\in I} S_i \neq \varnothing, \\ S = \overline{\operatorname{dom}} f \cap \bigcap_{i\in I} S_i. \end{cases} \tag{5.1}$$

The purpose of this section is to develop a relaxed, parallel, block-iterative algorithm to solve the convex feasibility problem

$$\text{Find } x \in S. \tag{5.2}$$

5.2. Algorithm.

ALGORITHM 5.1. *Starting with $x_0 \in \operatorname{int}\operatorname{dom} f$, take at every iteration n*

① *a nonempty finite index set $I_n \subset I$,*

② *operators $(T_{i,n})_{i\in I_n}$ in $\mathfrak{B}$ such that $(\forall i \in I_n)\ S_i \cap \operatorname{int}\operatorname{dom} f \subset \operatorname{Fix} T_{i,n}$,*

③ *points $(u_{i,n})_{i\in I_n} \in \bigtimes_{i\in I_n} T_{i,n}x_n$,*

④ *weights $(\omega_{i,n})_{i\in I_n}$ in $[0,1]$ such that $\sum_{i\in I_n}\omega_{i,n} = 1$,*

⑤ *a relaxation parameter $\lambda_n \in\,]0,1]$*

and put

⑥ $x_n^* = \nabla f(x_n) - \sum_{i\in I_n}\omega_{i,n}\nabla f(u_{i,n})$,

⑦

$$\eta_n = \left\langle x_n, \nabla f(x_n) - \sum_{i\in I_n}\omega_{i,n}\nabla f(u_{i,n})\right\rangle - \lambda_n \sum_{i\in I_n}\omega_{i,n}\left\langle u_{i,n} - x_n, \nabla f(u_{i,n}) - \nabla f(x_n)\right\rangle,$$

⑧ $H_n = \big\{y \in \mathcal{X} \mid \langle y, x_n^*\rangle \le \eta_n\big\}$.

Then set $x_{n+1} = P_{H_n}x_n$.

We now motivate this algorithm geometrically. At iteration n, x_n is given and a finite block of indices I_n is retained. Set $I_n^+ = \big\{i \in I_n \mid \omega_{i,n} > 0\big\}$. Then, using Lemma 3.2 for the first and last equality, step ② for the third inclusion, and (3.32) for the fourth inclusion,

$$\begin{aligned} S &= \overline{(\operatorname{int}\operatorname{dom} f) \cap \bigcap_{i\in I} S_i} \subset \overline{(\operatorname{int}\operatorname{dom} f) \cap \bigcap_{i\in I_n} S_i} \subset \overline{(\operatorname{int}\operatorname{dom} f) \cap \bigcap_{i\in I_n^+} S_i} \\ &\subset \overline{\bigcap_{i\in I_n^+} \operatorname{Fix} T_{i,n}} \subset \overline{(\operatorname{int}\operatorname{dom} f) \cap H_n} = \overline{\operatorname{dom}}\, f \cap H_n \subset H_n. \end{aligned} \tag{5.3}$$

Thus, H_n acts as an outer approximation to the intersection of the block of constraint sets $(\overline{\operatorname{dom}}\, f \cap S_i)_{i\in I_n}$ and, therefore, to S. More precisely, the block constraint $y \in \overline{\operatorname{dom}}\, f \cap \bigcap_{i\in I_n} S_i$ is replaced by the surrogate affine constraint $\langle y, x_n^*\rangle \le \eta_n$. The update x_{n+1} is then the D-projection of x_n onto H_n, i.e., the D-closest point to x_n which satisfies the surrogate constraint. (x_{n+1} is well defined by virtue of (5.1) and Corollary 3.35(i).) Naturally, such a point is considerably simpler to find than a point in $\overline{\operatorname{dom}}\, f \cap \bigcap_{i\in I_n} S_i$. In spirit, this type of surrogate constraint construction can be found—explicitly or implicitly—in several places in the literature, although not in the context of Bregman distances. (See, for instance, [39, 60] and the references therein.)

The parallel nature of the algorithm stems from the fact that the points $(u_{i,n})_{i\in I_n}$ at step ③ can be computed independently on concurrent processors. In addition, the algorithm has the ability to process variable blocks of constraints, which makes it possible to match closely the computational load of each iteration to the parallel processing architecture at hand. A discussion on the importance of block-processing for task scheduling on parallel architectures can be found in [33].

To shed more light on Algorithm 5.1, we first consider the case when $\mathcal{X}$ is Hilbertian and $f = \|\cdot\|^2/2$. Then, steps ⑥ and ⑦ become

$$\begin{cases} x_n^* = x_n - \sum_{i\in I_n}\omega_{i,n}u_{i,n}, \\ \eta_n = \big\langle x_n, x_n - \sum_{i\in I_n}\omega_{i,n}u_{i,n}\big\rangle - \lambda_n \sum_{i\in I_n}\omega_{i,n}\|u_{i,n} - x_n\|^2. \end{cases} \tag{5.4}$$

Furthermore, the updating step is explicitly given as

$$x_{n+1} = P_{H_n} x_n = x_n + \frac{\eta_n - \langle x_n, x_n^* \rangle}{\|x_n^*\|^2} x_n^* = x_n + \lambda_n L_n \left(\sum_{i \in I_n} \omega_{i,n} (u_{i,n} - x_n) \right), \tag{5.5}$$

where

$$L_n = \begin{cases} \dfrac{\sum_{i \in I_n} \omega_{i,n} \|u_{i,n} - x_n\|^2}{\| \sum_{i \in I_n} \omega_{i,n} (u_{i,n} - x_n)\|^2} & \text{if } x_n \notin \bigcap_{i \in I_n} S_i, \\ 1 & \text{otherwise.} \end{cases} \tag{5.6}$$

This is essentially the algorithm proposed in [41, section 6] (in this setting, the range of λ_n can be extended to $]0, 2[$), which itself contains those of [5, 6, 35, 37, 38, 60, 69] as special cases. In particular, if I is finite, $I_n \equiv I$, $\omega_{i,n} = \omega_i$, and $u_{i,n} = P_i x_n$, where P_i is the metric projector onto S_i, then (5.5)–(5.6) reduces to Pierra's classic extrapolated parallel projection method [72], which in turn can be traced back to Merzlyakov's method [63] for solving systems of linear inequalities in $\mathbb{R}^N$. Since $L_n \geq 1$ in (5.6), large extrapolations are possible in this algorithm by selecting $\lambda_n \approx 1$. It is known that these extrapolations yield significantly accelerated convergence in numerical experiments [36, 37, 50, 72] in comparison with purely averaged iterations, i.e.,

$$x_{n+1} = \sum_{i \in I_n} \omega_{i,n} u_{i,n}, \tag{5.7}$$

which can be derived from (5.5) by setting $\lambda_n = 1/L_n$.

Returning to the standing assumptions, let us now consider the parallel block-iterative update rule

$$\nabla f(x_{n+1}) = \sum_{i \in I_n} \omega_{i,n} \nabla f(u_{i,n}). \tag{5.8}$$

This alternative method for solving (5.2) was recently proposed by Censor and Herman in [29] (see also [31]) for the special case when $\mathcal{X} = \mathbb{R}^N$, I is finite, and $u_{i,n}$ is the D-projection of x_n onto S_i. If we assume that $\mathcal{X}$ is a Hilbert space and $f = \| \cdot \|^2/2$, then (5.8) reduces to (5.7) which, as noted above, is itself a special case of (5.5)–(5.6), hence of Algorithm 5.1. In general, however, we do not know whether (5.8) is always a particularization of Algorithm 5.1.

We now turn to Butnariu and Iusem's algorithmic framework [24] for solving (5.2). (In fact, they study the so-called stochastic convex feasibility problem, which is similar to (5.2) but allows for an uncountable index set I. Their framework requires measure theory for a precise formulation and their assumptions on the underlying function f are different from the ones made here. The reader is referred to [24] for further details.) Let $(R_i)_{i \in I}$ be a family of totally nonexpansive operators in the sense of [24]. (See also the paragraph following Definition 3.1.) Specialized to the case when I is finite, the update step in this algorithm is

$$x_{n+1} = \sum_{i \in I} \omega_i R_i x_n. \tag{5.9}$$

This resembles (5.8), except for notably absent gradients on both sides of the equation and for weights that do not depend on n. (If the R_i's are D-projectors, then

(5.9) can also be interpreted as a sequential algorithm in the product space X^I; see [11].) Note that if $\mathcal{X}$ is a Hilbert space and $f = \|\cdot\|^2/2$, then (5.9) once again corresponds to a parallel Cimmino-type algorithm, which is genuinely more restrictive than Algorithm 5.1 for this set-up.

While a detailed numerical and theoretical comparison of these algorithms lies beyond the scope of this paper, we remark that preliminary experiments suggest that Algorithm 5.1 is more flexible and faster than the one given by (5.8) and that Algorithm 5.1 is genuinely different from the method given by (5.9).

5.3. Convergence. The following notions were introduced in [6, Def. 3.7] and [41, Def. 6.5], respectively, to study the asymptotic behavior of Fejér monotone algorithms in Hilbert spaces. The former can be interpreted as an extension of the notion of demiclosedness at 0 [68] and the latter as an extension of the notion of demicompactness at 0 [70].

DEFINITION 5.2. *Algorithm* 5.1 *is*

- focusing *if for every bounded suborbit* $(x_{k_n})_{n\in\mathbb{N}}$ *it generates and every index* $i \in I$,

$$\begin{cases} i \in \bigcap_{n\in\mathbb{N}} I_{k_n}, \\ x_{k_n} \rightharpoonup x, \\ u_{i,k_n} - x_{k_n} \to 0 \end{cases} \quad \Rightarrow \quad x \in S_i; \tag{5.10}$$

- demicompactly regular *if there exists* $i \in I$, *called an* index of demicompact regularity, *such that for every bounded suborbit* $(x_{k_n})_{n\in\mathbb{N}}$ *it generates,*

$$\begin{cases} i \in \bigcap_{n\in\mathbb{N}} I_{k_n}, \\ u_{i,k_n} - x_{k_n} \to 0 \end{cases} \quad \Rightarrow \quad \mathfrak{S}(x_{k_n})_{n\in\mathbb{N}} \neq \emptyset. \tag{5.11}$$

We now describe the context in which the convergence of Algorithm 5.1 will be investigated.

CONDITION 5.3.

(i) *For some* $z \in \operatorname{dom} f \cap \bigcap_{i\in I} S_i$, $C = \operatorname{lev}_{\leq D(z,x_0)} D(z,\cdot)$ *is bounded.*

(ii) *For all sequences* $(u_n)_{n\in\mathbb{N}}$ *and* $(v_n)_{n\in\mathbb{N}}$ *in* C *such that* $(\forall n \in \mathbb{N})\ u_n \neq v_n$, *one has*

$$\frac{\langle u_n - v_n, \nabla f(u_n) - \nabla f(v_n)\rangle}{\|\nabla f(u_n) - \nabla f(v_n)\|} \to 0 \quad \Rightarrow \quad \nabla f(u_n) - \nabla f(v_n) \to 0. \tag{5.12}$$

CONDITION 5.4.

(i) $(\exists\, \delta_1 \in]0,1[)(\forall n \in \mathbb{N})(\exists\, j \in I_n)$

$$\|\nabla f(u_{j,n}) - \nabla f(x_n)\| = \max_{i\in I_n} \|\nabla f(u_{i,n}) - \nabla f(x_n)\| \ \text{ and } \ \omega_{j,n} \geq \delta_1.$$

(ii) $(\exists\, \delta_2 \in]0,1[)(\forall n \in \mathbb{N})\ \ \lambda_n \geq \delta_2$.

(iii) $(\forall i \in I)(\exists\, M_i \in \mathbb{N} \smallsetminus \{0\})(\forall n \in \mathbb{N})\ \ i \in \bigcup_{k=n}^{n+M_i-1} I_k$.

As will be seen subsequently, the above set of assumptions defines a broad framework which covers numerous practical situations. Note that, by virtue of (5.1), the quotient in (5.12) is well defined since ∇f is injective on $\operatorname{int}\operatorname{dom} f$ [8, Thm. 5.10]. Situations in which Condition 5.3(ii) is satisfied are detailed below. Note also that Condition 5.4(iii) imposes that every index i be activated at least once within any M_i consecutive iterations. This control rule, which has already been used in metric

projection algorithms in Hilbert spaces [35, 37, 38, 60], provides great flexibility in the management of the constraints and the implementation of the algorithm. Condition 5.4(i) provides added flexibility by offering the possibility of setting $\omega_{i,n} = 0$ if the corresponding step size $\|\nabla f(u_{i,n}) - \nabla f(x_n)\|$ is not maximal. It is thereby possible to meet the control condition Condition 5.4(iii) without actually using the ith constraint in the construction of x_{n+1}.

Recall that an operator T from a Banach space $\mathcal{Y}$ to its dual $\mathcal{Y}^*$ is said to be uniformly monotone on $U \subset \operatorname{dom} T$ with modulus c if [86, section 25.3]

$$(\forall x \in U)(\forall y \in U) \quad \langle x - y, Tx - Ty\rangle \geq \|x - y\| \cdot c(\|x - y\|), \tag{5.13}$$

where $c\colon [0, +\infty[\to [0, +\infty[$ is a strictly increasing function such that $c(0) = 0$. In particular, T is said to be strongly monotone on U with constant $\alpha \in]0, +\infty[$ if it is uniformly monotone on U with modulus $c\colon t \mapsto \alpha t$.

PROPOSITION 5.5. *Let z and C be as in Condition* 5.3(i). *Then Condition* 5.3(ii) *is satisfied in each of the following cases:*

(i) *∇f^* is uniformly monotone on $\nabla f(C)$.*
(ii) *∇f is Lipschitz-continuous on $\operatorname{dom} f = \mathcal{X}$.*
(iii) *$\mathcal{X} = \mathbb{R}^N$ and $\overline{C} \subset \operatorname{int}\operatorname{dom} f$.*
(iv) *$\mathcal{X} = \mathbb{R}^N$ and $z \in \operatorname{int}\operatorname{dom} f$.*

Proof. Let $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ be two sequences in C such that $(\forall n \in \mathbb{N})$ $u_n \neq v_n$. (i) Let c be the modulus of uniform monotonicity of ∇f^* on $\nabla f(C)$. Since ∇f is a bijection from $\operatorname{int}\operatorname{dom} f$ to $\operatorname{int}\operatorname{dom} f^*$ with inverse ∇f^* [8, Thm. 5.10] and since $C \subset \operatorname{int}\operatorname{dom} f$, we have $(\forall u \in C)(\forall v \in C)$ $\langle u - v, \nabla f(u) - \nabla f(v)\rangle \geq \|\nabla f(u) - \nabla f(v)\| \cdot c\,(\|\nabla f(u) - \nabla f(v)\|)$. Hence, since c is strictly increasing and $c(0) = 0$,

$$\begin{aligned} \frac{\langle u_n - v_n, \nabla f(u_n) - \nabla f(v_n)\rangle}{\|\nabla f(u_n) - \nabla f(v_n)\|} \to 0 \;&\Rightarrow\; c\big(\|\nabla f(u_n) - \nabla f(v_n)\|\big) \to 0 \\ &\Rightarrow\; \nabla f(u_n) - \nabla f(v_n) \to 0. \end{aligned} \tag{5.14}$$

(ii) $\Rightarrow$ (i) If ∇f is κ-Lipschitz-continuous on $\mathcal{X}$, then it follows from the Baillon–Haddad theorem [4, Cor. 10] that $(\forall x \in \mathcal{X})(\forall y \in \mathcal{X})$ $\langle x - y, \nabla f(x) - \nabla f(y)\rangle \geq \|\nabla f(x) - \nabla f(y)\|^2/\kappa$, i.e., ∇f^* is strongly monotone with constant $1/\kappa$. Consequently, ∇f^* is uniformly monotone on $\nabla f(C)$. (iii) Suppose

$$\frac{\langle u_n - v_n, \nabla f(u_n) - \nabla f(v_n)\rangle}{\|\nabla f(u_n) - \nabla f(v_n)\|} \to 0 \quad \text{and} \quad \nabla f(u_n) - \nabla f(v_n) \not\to 0. \tag{5.15}$$

Then there exists a strictly increasing sequence $(k_n)_{n\in\mathbb{N}}$ in $\mathbb{N}$ and $\varepsilon \in]0, +\infty[$ such that $\inf_{n\in\mathbb{N}} \|\nabla f(u_{k_n}) - \nabla f(v_{k_n})\| \geq \varepsilon$. Since $(u_{k_n})_{n\in\mathbb{N}}$ lies in C, it is bounded and therefore possesses a convergent subsequence, say $u_{k_{l_n}} \to u$. As $(v_{k_{l_n}})_{n\in\mathbb{N}}$ is also bounded, we can assume (passing to a subsequence if necessary) that it converges, say $v_{k_{l_n}} \to v$. Since $\{u, v\} \subset \overline{C} \subset \operatorname{int}\operatorname{dom} f$ and ∇f is continuous at every point in $\operatorname{int}\operatorname{dom} f$ by [77, Thm. 25.5], taking the limit yields $\|\nabla f(u) - \nabla f(v)\| \geq \varepsilon$ and, by injectivity of ∇f on $\operatorname{int}\operatorname{dom} f$ [8, Thm. 5.10], $u \neq v$. On the other hand, (5.15) yields

$$\frac{\big\langle u_{k_{l_n}} - v_{k_{l_n}}, \nabla f(u_{k_{l_n}}) - \nabla f(v_{k_{l_n}})\big\rangle}{\|\nabla f(u_{k_{l_n}}) - \nabla f(v_{k_{l_n}})\|} \to 0, \tag{5.16}$$

and, since $\|\nabla f(u) - \nabla f(v)\| \neq 0$, taking the limit yields $\langle u - v, \nabla f(u) - \nabla f(v)\rangle = 0$. However, $f|_{\operatorname{int}\operatorname{dom} f}$ is strictly convex and therefore ∇f is strictly monotone on $\operatorname{int}\operatorname{dom} f \supset \{u, v\}$. This forces $u = v$ and we reach a contradiction. (iv) In view of (iii), it is enough to show that $\overline{C} \subset \operatorname{int}\operatorname{dom} f$. If the inclusion does not hold, then we can find $y \in \operatorname{bdry}\operatorname{dom} f$ and $(y_n)_{n\in\mathbb{N}}$ in C such that $y_n \to y$. Thus $\sup_{n\in\mathbb{N}} D(z, y_n) \leq D(z, x_0) < +\infty$, and, at the same time, since f is essentially smooth, [7, Thm. 3.8(i)] yields $D(z, y_n) \to +\infty$, which is absurd. □

Remark 5.6. A careful analysis of [85, Corollary 3.4.4"(iii)⇔(iv)"], [85, Proposition 3.5.1], and [85, Proposition 3.6.2] shows that Proposition 5.5(i) holds as soon as ∇f is Lipschitz on bounded sets. In turn, this condition is satisfied in L_p spaces for $f = \|\cdot\|_p^s$, where $\{p, s\} \subset [2, +\infty[$. (The proof relies on the case when $s = 2$; see also Example 5.11 below.)

Examples of Legendre functions f which satisfy Conditions 4.3, 4.4, and 5.3(i)–(ii) will be supplied in section 5.4. Our main convergence result can now be stated and proved.

THEOREM 5.7. *Suppose that Conditions* 4.3, 4.4, 5.3, *and* 5.4 *are satisfied, and let* $(x_n)_{n\in\mathbb{N}}$ *be an arbitrary orbit of Algorithm* 5.1. *Then, for every* $n \in \mathbb{N}$, x_n *and* $(u_{i,n})_{i\in I_n}$ *lie in the bounded set* C. *If, in addition, Algorithm* 5.1 *is focusing, then the following statements hold true:*

(i) $(x_n)_{n\in\mathbb{N}}$ *converges weakly to a point* $x \in S$.

(ii) *If the weak limit* x *from* (i) *belongs to* $\operatorname{int}\operatorname{dom} f$ *and the algorithm is demicompactly regular, then* $(x_n)_{n\in\mathbb{N}}$ *converges strongly.*

Proof. For every $n \in \mathbb{N}$, set $T_n = P_{H_n}$ and $I_n^+ = \big\{i \in I_n \mid \omega_{i,n} > 0\big\}$. Since $x_0 \in \operatorname{int}\operatorname{dom} f$ and, by Proposition 3.39(vi), $T_n \in \mathfrak{B}$, we recognize that

(5.17) Algorithm 5.1 is a special case of Algorithm 4.12.

Our goal is to apply Theorem 4.14 and we must start by verifying (4.9). First, considering (5.1), Algorithm 5.1②, and Proposition 3.39(iii), we obtain

(5.18)
$$(\forall n \in \mathbb{N})\ \ \varnothing \neq (\operatorname{int}\operatorname{dom} f) \cap \bigcap_{i\in I} S_i \subset \bigcap_{i\in I_n^+} (S_i \cap \operatorname{int}\operatorname{dom} f) \subset \bigcap_{i\in I_n^+} \operatorname{Fix} T_{i,n} = \operatorname{Fix} T_n.$$

Hence $\bigcap_{n\in\mathbb{N}} \operatorname{Fix} T_n \neq \varnothing$. In addition, (5.1), Lemma 3.2, and (5.18) yield

$$(\forall n \in \mathbb{N})\ \ S = \overline{\operatorname{dom}}\, f \cap \bigcap_{i\in I} S_i \subset \overline{\operatorname{Fix}}\, T_n. \tag{5.19}$$

Consequently, $S \subset \bigcap_{n\in\mathbb{N}} \overline{\operatorname{Fix}}\, T_n$. Next, we derive from (5.1) that

$$\varnothing \neq (\operatorname{int}\operatorname{dom} f) \cap \bigcap_{i\in I} S_i \subset \operatorname{dom} f \cap \overline{\operatorname{dom}}\, f \cap \bigcap_{i\in I} S_i = \operatorname{dom} f \cap S. \tag{5.20}$$

Thus, (4.9) holds. Now, let z and C be as in Condition 5.3(i). It follows from (5.17) and Proposition 4.13(i) that the sequences $(x_n)_{n\in\mathbb{N}}$ and $(T_n x_n)_{n\in\mathbb{N}}$ are contained in C, which is bounded. In order to verify (4.11), some key facts must be established. Let us

temporarily fix $n \in \mathbb{N}$. The first fact is supplied by the inclusion $x_{n+1} = P_{H_n} x_n \in H_n$, which yields

$$\|x_{n+1} - x_n\| \geq d_{H_n}(x_n). \tag{5.21}$$

Next, it follows from Condition 5.3(i), (5.1), Lemma 3.2, and Algorithm 5.1② that

$$(\forall i \in I_n) \;\; z \in S_i \cap \overline{\operatorname{dom}} f = \overline{S_i \cap \operatorname{int} \operatorname{dom} f} \subset \overline{\operatorname{Fix}} T_{i,n}. \tag{5.22}$$

Hence, for every $i \in I_n$, Algorithm 5.1③ and Proposition 3.3(viii) yield $D(z, u_{i,n}) \leq D(z, x_n) - D(u_{i,n}, x_n) \leq D(z, x_n)$. Therefore,

$$(\forall i \in I_n) \;\; u_{i,n} \in C. \tag{5.23}$$

Now, per Condition 5.4(ii), pick $j_n \in I_n$ such that

$$\|\nabla f(u_{j_n,n}) - \nabla f(x_n)\| = \max_{i \in I_n} \|\nabla f(u_{i,n}) - \nabla f(x_n)\| \quad \text{and} \quad \omega_{j_n,n} \geq \delta_1. \tag{5.24}$$

We claim that

(5.25)

$$\begin{cases} x_n \in \bigcap_{i \in I_n^+} \operatorname{Fix} T_{i,n} & \Leftrightarrow \quad u_{j_n,n} = x_n \quad \Leftrightarrow \quad \|\nabla f(u_{j_n,n}) - \nabla f(x_n)\| = 0, \\ x_n \notin \bigcap_{i \in I_n^+} \operatorname{Fix} T_{i,n} & \Rightarrow \quad d_{H_n}(x_n) \geq \delta_1 \delta_2 \dfrac{\langle u_{j_n,n} - x_n, \nabla f(u_{j_n,n}) - \nabla f(x_n) \rangle}{\|\nabla f(u_{j_n,n}) - \nabla f(x_n)\|}. \end{cases}$$

On the one hand, using Proposition 3.3(vi) and the injectivity of ∇f on $\operatorname{int} \operatorname{dom} f$ [8, Thm. 5.10], since (5.24) forces $j_n \in I_n^+$, we get $x_n \in \bigcap_{i \in I_n^+} \operatorname{Fix} T_{i,n} \Leftrightarrow (\forall i \in I_n^+)$ $u_{i,n} = x_n \Rightarrow u_{j_n,n} = x_n \Rightarrow \|\nabla f(u_{j_n,n}) - \nabla f(x_n)\| = 0 \Rightarrow (\forall i \in I_n)$ $\|\nabla f(u_{i,n}) - \nabla f(x_n)\| = 0 \Leftrightarrow (\forall i \in I_n)$ $u_{i,n} = x_n \Rightarrow (\forall i \in I_n^+)$ $u_{i,n} = x_n$. On the other hand, if $x_n \notin \bigcap_{i \in I_n^+} \operatorname{Fix} T_{i,n}$, then Proposition 3.39(ii) asserts that $x_n \notin H_n$ and $x_n^* \neq 0$, so that

$$\begin{aligned} d_{H_n}(x_n) &= \frac{\langle x_n, x_n^* \rangle - \eta_n}{\|x_n^*\|} \\ &= \lambda_n \frac{\sum_{i \in I_n} \omega_{i,n} \langle u_{i,n} - x_n, \nabla f(u_{i,n}) - \nabla f(x_n) \rangle}{\| \sum_{i \in I_n} \omega_{i,n} \big(\nabla f(u_{i,n}) - \nabla f(x_n) \big) \|} \qquad (5.26) \\ &\geq \delta_2 \frac{\sum_{i \in I_n} \omega_{i,n} \langle u_{i,n} - x_n, \nabla f(u_{i,n}) - \nabla f(x_n) \rangle}{\sum_{i \in I_n} \omega_{i,n} \|\nabla f(u_{i,n}) - \nabla f(x_n)\|} \qquad (5.27) \\ &\geq \delta_1 \delta_2 \frac{\langle u_{j_n,n} - x_n, \nabla f(u_{j_n,n}) - \nabla f(x_n) \rangle}{\|\nabla f(u_{j_n,n}) - \nabla f(x_n)\|}, \qquad (5.28) \end{aligned}$$

where (5.26) follows from [80, Lemma I.1.2] and (5.27) from Condition 5.4(ii). Altogether, (5.25) is verified. The third key fact is derived from (5.23) and Proposition 2.3(i) as follows:

$$\begin{aligned} (\forall i \in I_n) \;\; \operatorname{diam}(C) \|\nabla f(u_{i,n}) - \nabla f(x_n)\| &\geq \langle u_{i,n} - x_n, \nabla f(u_{i,n}) - \nabla f(x_n) \rangle \\ &= D(u_{i,n}, x_n) + D(x_n, u_{i,n}) \\ &\geq D(u_{i,n}, x_n). \end{aligned} \tag{5.29}$$

Let us now verify (4.11). To this end, let us fix $i \in I$ and $x \in \mathfrak{W}(x_n)_{n\in\mathbb{N}}$, say $x_{k_n} \rightharpoonup x$. Because $x \in \overline{\operatorname{dom} f}$, it is sufficient to show

$$D(x_{n+1}, x_n) \to 0 \quad \Rightarrow \quad x \in S_i. \tag{5.30}$$

Let M_i be as in Condition 5.4(iii). After passing to a subsequence of $(x_{k_n})_{n\in\mathbb{N}}$ if necessary, we assume that, for every $n \in \mathbb{N}$, $k_{n+1} \geq k_n + M_i$. This guarantees the existence of a sequence $(p_n)_{n\in\mathbb{N}}$ in $\mathbb{N}$ such that

$$(\forall n \in \mathbb{N}) \;\; k_n \leq p_n \leq k_n + M_i - 1 < k_{n+1} \leq p_{n+1} \quad \text{and} \quad i \in I_{p_n}. \tag{5.31}$$

Now consider the subsequence $(x_{p_n})_{n\in\mathbb{N}}$ of $(x_n)_{n\in\mathbb{N}}$. The triangle inequality yields

$$(\forall n \in \mathbb{N}) \;\; \|x_{p_n} - x_{k_n}\| \leq \sum_{l=k_n}^{k_n+M_i-2} \|x_{l+1} - x_l\| \leq (M_i - 1) \max_{k_n \leq l \leq k_n+M_i-2} \|x_{l+1} - x_l\|. \tag{5.32}$$

Now suppose $D(x_{n+1}, x_n) \to 0$. Then (4.4) yields

$$x_{n+1} - x_n \to 0 \tag{5.33}$$

and it follows from (5.21) that $d_{H_n}(x_n) \to 0$. Consequently, we derive from (5.25), (5.23), and Condition 5.3(ii) that $\max_{j\in I_n} \|\nabla f(u_{j,n}) - \nabla f(x_n)\| \to 0$. In turn, (5.29) implies that $D(u_{i,p_n}, x_{p_n}) \to 0$ and, invoking (4.4) again, we obtain

$$u_{i,p_n} - x_{p_n} \to 0. \tag{5.34}$$

We also derive from (5.32) and (5.33) that $x_{p_n} - x_{k_n} \to 0$, whence $x_{p_n} \rightharpoonup x$. However, since the algorithm is focusing, (5.10) yields $x \in S_i$. Thus (5.30) holds and, consequently, the following conclusions can be drawn:

(i) Theorem 4.14(i) asserts that $(x_n)_{n\in\mathbb{N}}$ converges weakly to $x \in S$.

(ii) Suppose that $x \in \operatorname{int}\operatorname{dom} f$, $i \in I$ is an index of demicompact regularity, and $D(x_{n+1}, x_n) \to 0$. Then it results from (5.34) and (5.11) that (4.12) holds. In view of Condition 4.4, the strong convergence claim therefore follows from Theorem 4.14(ii). □

5.4. When all the assumptions hold. In this subsection, we describe scenarios in which all the assumptions required in Theorem 5.7 on f and on the constraint sets $(S_i)_{i\in I}$ are satisfied.

As a preamble to our first example, recall that if $\mathcal{X}$ is a Hilbert space, the Moreau–Yosida regularization of a proper lower semicontinuous convex function $\varphi\colon \mathcal{X} \to]-\infty, +\infty]$ with parameter $\gamma \in]0, +\infty[$ is the finite continuous convex function ${}_\gamma\varphi = \varphi \,\square\, \big(\|\cdot\|^2/(2\gamma)\big)$. Moreover, Moreau's classic proximal operator associated with φ and γ is given by Definition 3.16 for $f = \|\cdot\|^2/2$ and will be denoted by $\operatorname{Prox}_\gamma^\varphi$. It follows from Proposition 3.21(v) that $\operatorname{Prox}_\gamma^\varphi$ is defined everywhere and, from Proposition 3.22(ii)(d) and (c), that it is single-valued and firmly nonexpansive. Moreover [67, Prop. 7.d],

$$\nabla\, {}_\gamma\varphi = \frac{\operatorname{Id} - \operatorname{Prox}_\gamma^\varphi}{\gamma}. \tag{5.35}$$

EXAMPLE 5.8 (Moreau–Yosida regularization). *Let $\mathcal{X}$ be a Hilbert space, set $w = \|\cdot\|^2/2$, and define $f\colon \mathcal{X} \to \mathbb{R}$ by*

$$f = (1+\gamma)w - {}_1\varphi, \tag{5.36}$$

where $\varphi\colon \mathcal{X} \to]-\infty,+\infty]$ is a proper lower semicontinuous convex function and $\gamma \in]0,+\infty[$. Then

$$\begin{aligned} D\colon (x,y) \mapsto{}& \gamma w(x-y) + w(x-\operatorname{Prox}_1^{\varphi} y) + \varphi(\operatorname{Prox}_1^{\varphi} y) \\ &- \big(w(x-\operatorname{Prox}_1^{\varphi} x) + \varphi(\operatorname{Prox}_1^{\varphi} x)\big) \end{aligned} \tag{5.37}$$

and Conditions 4.4 and 5.3 are satisfied. If $\operatorname{Prox}^{\varphi}_{\frac{\gamma}{1+\gamma}}$ is affine or S is a singleton, then Condition 4.3 is also satisfied.

Proof. The expression (5.37) is derived from (1.5) by simple algebra. Now set $\psi = w - {}_1\varphi$. Then

$$\psi = w - \inf_{x\in\operatorname{dom}\varphi} \varphi(x) + w(\cdot - x) = \sup_{x\in\operatorname{dom}\varphi} \langle x,\cdot\rangle - \varphi(x) - w(x) = (\varphi + w)^*. \tag{5.38}$$

Hence, ψ is a proper lower semicontinuous convex function as the conjugate of one such function. Since ψ is convex, $f = \psi + \gamma w$ is strongly (hence uniformly) convex and, in view of Example 4.10(i), Condition 4.4 is therefore satisfied. On the other hand, (5.35) yields $\operatorname{dom}\nabla f = \mathcal{X}$ and $\nabla f = \operatorname{Prox}_1^{\varphi} + \gamma\operatorname{Id}$. Hence f is essentially smooth by [8, Thm. 5.6]. Furthermore, since $\operatorname{Prox}_1^{\varphi}$ is firmly nonexpansive, it is 1-Lipschitz and therefore ∇f is $(1+\gamma)$-Lipschitz. Accordingly, Proposition 5.5(ii) asserts that Condition 5.3(ii) is satisfied. Next, using standard Hilbertian convex calculus, we obtain

$$\begin{aligned} f^* &= \big(\psi + \gamma w\big)^* = \psi^* \,\square\, (w/\gamma) = (\varphi + w) \,\square\, (w/\gamma) = {}_{\gamma}\big(\varphi + w\big) \\ &= \big({}_{\gamma/(1+\gamma)}\varphi\big)\big(\,\cdot\,/(1+\gamma)\big) + w/(1+\gamma). \end{aligned} \tag{5.39}$$

It therefore follows from (5.35) that

$$\operatorname{dom}\nabla f^* = \mathcal{X} \quad\text{and}\quad \nabla f^* = \frac{\operatorname{Id} - \operatorname{Prox}^{\varphi}_{\gamma/(1+\gamma)}\big(\,\cdot\,/(1+\gamma)\big)}{\gamma}. \tag{5.40}$$

Consequently, f^* is also essentially smooth and it follows from [8, Thm. 5.4] that f is Legendre. Moreover, since $\mathcal{X}$ is a Hilbert space, it is reflexive. We also derive from (5.40) that, since $\operatorname{Id} - \operatorname{Prox}^{\varphi}_{\gamma/(1+\gamma)}$ is (firmly) nonexpansive, ∇f^* is $1/\gamma$-Lipschitz and, thereby, maps bounded sets to bounded sets. It then follows from [8, Thm. 3.3] that f is supercoercive, and Proposition 4.1(v)(a) asserts that Condition 5.3(i) is satisfied. Finally, since ∇f is continuous, it will be weakly continuous when it is affine, i.e., when $\operatorname{Prox}^{\varphi}_{\gamma/(1+\gamma)}$ is. In turn, Example 4.6 implies that Condition 4.3 is satisfied. On the other hand, if S is a singleton, the claim follows from Example 4.5. ∎

If we let φ be the indicator function of a nonempty closed convex set in (5.36), then we obtain the Legendre function studied in [8, Example 7.2]. Specializing even further, we obtain the following examples.

EXAMPLE 5.9 (distance). *In the previous example, set $\varphi = \iota_M$, where M is a closed affine subspace of $\mathcal{X}$, and let P_M be the metric projector onto M. Then Conditions 4.3, 4.4, and 5.3 are satisfied, $f = (1+\gamma)w - d_M^2/2$, and $D\colon (x,y) \mapsto \gamma w(x-y) + w(x - P_M y) - w(x - P_M x)$.*

EXAMPLE 5.10 (energy). *In the previous example, set* $M = \{0\}$ *and* $\gamma = 1$. *Then* $f = \|\cdot\|^2/2$, $\nabla f = \mathrm{Id}$, $D\colon (x,y) \mapsto \|x-y\|^2/2$, *and we recover the usual Fejér monotonicity framework.*

The next example shows that the function $f = \|\cdot\|^2/2$ can also be used outside Hilbert spaces.

EXAMPLE 5.11 (L_p spaces). *Let* $(\Omega, \mathcal{F}, \mu)$ *be a positive measure space and let* $p \in [2, +\infty[$. *Let* $\mathcal{X} = L_p(\Omega, \mathcal{F}, \mu)$, *equipped with its canonical norm, and set* $f = \|\cdot\|^2/2$. *Then Conditions* 4.4 *and* 5.3 *are satisfied. If* S *is a singleton, then Condition* 4.3 *is also satisfied.*

Proof. By [8, Example 6.5], f is Legendre and uniformly convex on closed balls. Hence Condition 4.4 holds by Example 4.10(i). Since f is supercoercive, Condition 5.3(i) follows from [8, Lemma 7.3(viii)]. We now establish Condition 5.3(ii). As $p \in [2, +\infty[$, [45, Corollary V.1.2] implies that $\rho_{\|\cdot\|}$, the modulus of smoothness of $\mathcal{X}$, is of power type 2. We thus obtain $\kappa \in]0, +\infty[$ so that (see [45, section IV.4])

$$(\forall t \in [0, +\infty[) \quad \rho_{\|\cdot\|}(t) \leq \kappa t^2. \tag{5.41}$$

Recall that $\nabla f = J$ and define $j(x) = J(x)/\|x\| = \nabla\|x\|$ for all nonzero $x \in \mathcal{X}$. Now (5.41) and [45, Lemma IV.5.1] yield

$$(\forall u \in S_{\mathcal{X}})(\forall v \in S_{\mathcal{X}}) \quad \|j(u) - j(v)\| \leq \kappa\|u - v\|. \tag{5.42}$$

Fix two nonzero points x and y in $\mathcal{X}$ and assume, without loss of generality, that $\|x\| \geq \|y\|$. Then, using the triangle inequality,

$$\left\| \frac{x}{\|x\|} - \frac{y}{\|y\|} \right\| = \left\| \left(\frac{x}{\|x\|} - \frac{y}{\|x\|} \right) + \frac{\|y\| \cdot y - \|x\| \cdot y}{\|x\| \cdot \|y\|} \right\| \leq \frac{2}{\|x\|}\|x - y\|. \tag{5.43}$$

Thus

$$\|j(x) - j(y)\| = \left\| j\left(\frac{x}{\|x\|}\right) - j\left(\frac{y}{\|y\|}\right) \right\| \leq \kappa \cdot \left\| \frac{x}{\|x\|} - \frac{y}{\|y\|} \right\| \leq \frac{2\kappa}{\|x\|}\|x - y\|, \tag{5.44}$$

where we have used the definition of j for the equality, (5.42) for the first inequality, and (5.43) for the second. Furthermore,

$$\begin{aligned} \|J(x) - J(y)\| &= \big\|\|x\| \cdot j(x) - \|y\| \cdot j(y)\big\| \\ &= \big\|\big(\|x\| \cdot j(x) - \|x\| \cdot j(y)\big) + \big(\|x\| \cdot j(y) - \|y\| \cdot j(y)\big)\big\| \\ &\leq \|x\| \cdot \|j(x) - j(y)\| + \|j(y)\| \cdot \big|\|x\| - \|y\|\big| \\ &\leq (2\kappa + 1) \cdot \|x - y\|, \end{aligned} \tag{5.45}$$

where the last inequality follows from (5.44) and the fact that $\|j(y)\| = 1$. Now (5.45) implies that $J = \nabla f$ is Lipschitz-continuous on $\operatorname{dom} f = \mathcal{X}$, with constant $2\kappa + 1$ (for $x = 0$ or $y = 0$, argue directly). We apply Proposition 5.5(ii) and conclude that Condition 5.3(ii) is satisfied. Finally, if S is a singleton, we employ Example 4.5. ∎

Guaranteeing Condition 4.3 requires some care.

Remark 5.12. As already discussed in Remark 5.6, Proposition 5.5(i) holds as soon as ∇f is Lipschitz on bounded sets. Thus, the assertions of Example 5.11 remain true for $f = \|\cdot\|^s/s$, where $s \in [2, +\infty[$. The case when $s = p$ is particularly interesting because then ∇f becomes J_φ, the duality mapping corresponding to the weight $\varphi\colon t \mapsto t^{p-1}$ (see [34]). If we specialize this further to the space $\ell_p(\mathbb{N})$, then

J_φ is known to be sequentially weakly continuous (see [34, Prop. II.4.14]) and thus Example 4.6 is applicable. To sum up,

> let $\mathcal{X} = \ell_p(\mathbb{N})$ and $f = \|\cdot\|^p/p$, for $p \in [2, +\infty[$;
> then Conditions 4.3, 4.4, and 5.3 are satisfied.

Additional examples can be generated in suitable product spaces such as $\ell_{p_1}(\mathbb{N}) \times \ell_{p_2}(\mathbb{N})$, equipped with the Euclidean product norm and with $\{p_1, p_2\} \subset [2, +\infty[$, or in certain spaces of power type 2. (See [45] for further information about such spaces.)

EXAMPLE 5.13 (closed domain Bregman/Legendre functions). *Let $\mathcal{X} = \mathbb{R}^N$ and let f be a Bregman/Legendre function with closed domain. Then Conditions* 4.3, 4.4, *and* 5.3 *are satisfied.*

Proof. Example 4.9 implies that Condition 4.3 holds. Condition 4.4 follows from [7, Def. 5.2.BL3 and Thm. 3.9(iii)]. It remains to check items (i) and (ii) in Condition 5.3: since $D(z, \cdot)$ is coercive for every $z \in \operatorname{dom} f$ [7, Remark 5.3], (i) holds, whereas (ii) follows from Proposition 5.5(iv). ☐

The class of Bregman/Legendre functions (see Definition 4.8) is large enough to contain many functions important in convex optimization and it is related to the Bregman functions of [30, 33], which require closed domains. We refer the reader to [7] for further information. The following example gives conditions that are easy to verify in practice.

EXAMPLE 5.14 (separable Bregman/Legendre functions). *Let $(\varphi_k)_{1\leq k\leq N} \colon \mathbb{R} \to]-\infty, +\infty]$ be a family of Legendre functions such that $(\operatorname{dom}\varphi_k^*)_{1\leq k\leq N}$ are open. Let $\mathcal{X} = \mathbb{R}^N$, and let $f \colon (\xi_k)_{1\leq k\leq N} \mapsto \sum_{k=1}^N \varphi_k(\xi_k)$. Then Conditions* 4.3, 4.4, *and* 5.3 *are satisfied.*

Proof. By [7, Corollary 5.13], f is Bregman/Legendre. Mimicking the proof of the previous example, we note that it remains to check Condition 4.4. For every $k \in \{1, \ldots, m\}$, since $\varphi_k|_{\operatorname{int}\operatorname{dom}\varphi_k}$ is strictly convex by Legendreness and $\varphi_k|_{\operatorname{dom}\varphi_k}$ is continuous by (3.5), $\varphi_k|_{\operatorname{dom}\varphi_k}$ is strictly convex. Hence, it follows from Example 4.10(iii) that the Bregman distance D_k induced by φ_k on $\mathbb{R}$ satisfies Condition 4.4 and, in turn, so does $D \colon \big((\xi_k)_{1\leq k\leq N}, (\chi_k)_{1\leq k\leq N}\big) \mapsto \sum_{k=1}^N D_k(\xi_k, \chi_k)$. ☐

Unlike the previous examples, the following example does not require that $\mathcal{X}$ be finite-dimensional or that f have full domain.

EXAMPLE 5.15. *Let $\mathcal{X}$ be the Hilbert space $\ell_2(\mathbb{N}) \times \mathbb{R}$ and define*

$$f \colon \mathcal{X} \to]-\infty, +\infty] : (x, \xi) \mapsto \begin{cases} \tfrac{1}{2}\|x\|^2 + \xi\ln(\xi) - \xi & \text{if } \xi > 0, \\ \tfrac{1}{2}\|x\|^2 & \text{if } \xi = 0, \\ +\infty & \text{if } \xi < 0. \end{cases} \tag{5.46}$$

Let $(\forall i \in I)$ $S_i = S = \ell_2(\mathbb{N}) \times [1, +\infty[$. Fix $(z, \zeta) \in S$, $\eta > 0$, and set $C = \operatorname{lev}_{\leq\eta} D\big((z, \zeta), \cdot\big)$. Then Conditions 4.3, 4.4, *and* 5.3 *are satisfied.*

Proof. Let $g = f(\cdot, 0)$ and $h = f(0, \cdot)$. Hence, $(\forall (x, \xi) \in \mathcal{X})$ $f(x, \xi) = g(x) + h(\xi)$. Note that g and h are Legendre, and so is f, with $\operatorname{dom} f = \ell_2(\mathbb{N}) \times [0, +\infty[$. Now, let D_g and D_h be the Bregman distances induced by g on $\ell_2(\mathbb{N})$ and h on $\mathbb{R}$, respectively. Take $(y, \chi) \in \mathcal{X}$ with $D\big((z, \zeta), (y, \chi)\big) = D_g(z, y) + D_h(\zeta, \chi) \leq \eta$. In particular, $D_g(z, y) \leq \eta$ and $D_h(\zeta, \chi) \leq \eta$. Since $D_g(z, \cdot)$ and $D_h(\zeta, \cdot)$ are coercive by Proposition 4.1(v)(a) and (b), C is bounded. Condition 4.3 is a consequence of Example 4.6. Since $D \colon (x, \xi) \mapsto D_g(x) + D_h(\xi)$, Condition 4.4 is immediate by Examples 5.10 and 5.13. Applying [7, Thm. 3.8.(i)] to h and $\zeta \in \operatorname{int}\operatorname{dom} h$, we obtain $\varepsilon \in]0, +\infty[$ such that $(\forall (y, \chi) \in C)$ $\chi \geq \varepsilon$. A straightforward computation shows

that ∇f^* is strongly monotone with constant $\min\{1,\varepsilon\}$. Therefore, using Proposition 5.5(iv), Condition 5.3(ii) holds as well and the proof is complete. ∎

5.5. Applications. A broad class of problems in convex optimization and nonlinear analysis are captured by the mixed convex feasibility problem

$$\text{Find } x \in \overline{\operatorname{dom}} f \text{ such that } \begin{cases} (\forall i \in I^{(1)}) & g_i(x) \le 0, \\ (\forall i \in I^{(2)}) & 0 \in A_i x, \\ (\forall i \in I^{(3)}) & \varphi_i(x) = \inf \varphi_i(\mathcal{X}), \\ (\forall i \in I^{(4)}) & T_i x = x, \end{cases} \tag{5.47}$$

where $(g_i)_{i\in I^{(1)}}$ and $(\varphi_i)_{i\in I^{(3)}}$ are families of proper lower semicontinuous convex functions from $\mathcal{X}$ into $]-\infty,+\infty]$, $(A_i)_{i\in I^{(2)}}$ is a family of maximal monotone operators from $\mathcal{X}$ into $2^{\mathcal{X}^*}$, and $(T_i)_{i\in I^{(4)}}$ is a family of D-firm operators from $\mathcal{X}$ into $\mathcal{X}$. Here, $I^{(1)}$, $I^{(2)}$, $I^{(3)}$, and $I^{(4)}$ are pairwise disjoint, possibly empty, countable index sets such that $I = \bigcup_{k=1}^{4} I^{(k)} \neq \emptyset$. Now let us define

$$(\forall i \in I)\ \ S_i = \begin{cases} \operatorname{lev}_{\le 0} g_i & \text{if } i \in I^{(1)}, \\ A_i^{-1}0 & \text{if } i \in I^{(2)}, \\ \operatorname{Argmin} \varphi_i & \text{if } i \in I^{(3)}, \\ \overline{\operatorname{Fix}} T_i & \text{if } i \in I^{(4)}. \end{cases} \tag{5.48}$$

Throughout this section, the following set of assumptions will be made.

CONDITION 5.16.

(i) *Conditions* 4.3, 4.4, 5.3, *and* 5.4 *are satisfied.*

(ii) *For every* $i \in I^{(1)}$, $\partial g_i(C)$ *is bounded and* $\operatorname{dom} f \subset \operatorname{dom} g_i$.

(iii) *For every* $i \in I^{(2)}$, *one of the following conditions holds:*

(a) $\operatorname{dom} A_i \subset \operatorname{int}\operatorname{dom} f$,

(b) A_i *is* 3^**-monotone.*

(iv) *For every* $i \in I^{(4)}$, $\operatorname{dom} T_i = \operatorname{int}\operatorname{dom} f$ *and* $T_i - \mathrm{Id}$ *is demiclosed at* 0 *in the sense that for every sequence* $(y_n)_{n\in\mathbb{N}}$ *in* $\operatorname{dom} T_i$

$$\begin{cases} y_n \rightharpoonup y, \\ (\forall n \in \mathbb{N})\ \ u_n \in T_i y_n, \\ u_n - y_n \to 0 \end{cases} \quad \Rightarrow \quad y \in \overline{\operatorname{Fix}} T_i. \tag{5.49}$$

Let us observe that the sets $(S_i)_{i\in I}$ are closed and convex. For $i \in I^{(1)} \cup I^{(2)} \cup I^{(3)}$, this follows from well-known facts; for $i \in I^{(4)}$, this follows from Condition 5.16(iv), Propositions 3.5(ii), the essential strict convexity of f, and Proposition 3.3(v). Accordingly, (5.47) is a special case of the convex feasibility problem (5.2) and it can therefore be solved by Algorithm 5.1.

ALGORITHM 5.17 (specific implementation of Algorithm 5.1). *Fix* $(\varepsilon_i)_{i\in I^{(2)}}$ *and* $(\varepsilon_i)_{i\in I^{(3)}}$ *in* $]0,+\infty[$. *Implement Algorithm* 5.1② *by choosing for every* $i \in I_n$

$$T_{i,n} = \begin{cases} Q_{g_i} & \text{if } i \in I^{(1)} \text{ (see Definition 3.37)}, \\ R_{\gamma_{i,n}A_i}, \text{ where } \gamma_{i,n} \in [\varepsilon_i, +\infty[& \text{if } i \in I^{(2)} \text{ (see Definition 3.7)}, \\ \operatorname{prox}^{\varphi_i}_{\gamma_{i,n}}, \text{ where } \gamma_{i,n} \in [\varepsilon_i, +\infty[& \text{if } i \in I^{(3)} \text{ (see Definition 3.16)}, \\ T_i & \text{if } i \in I^{(4)} \text{ (see Definition 3.4)}. \end{cases} \tag{5.50}$$

Thanks to Condition 5.16, (5.50) meets the requirements of Algorithm 5.1② since in each case we have the following:

- $T_{i,n} \in \mathfrak{B}$. This follows from Proposition 3.38(ii) if $i \in I^{(1)}$, from Corollary 3.14(ii) and (iii) if $i \in I^{(2)}$ (since $A_i^{-1}0 \cap \operatorname{int}\operatorname{dom} f \neq \varnothing$, $\operatorname{dom} A_i \cap \operatorname{int}\operatorname{dom} f \neq \varnothing$), from Corollary 3.25(i) if $i \in I^{(3)}$ (since φ_i is proper and $\operatorname{Argmin}\varphi_i \cap \operatorname{int}\operatorname{dom} f \neq \varnothing$, φ_i is bounded below and $\operatorname{dom}\varphi_i \cap \operatorname{int}\operatorname{dom} f \neq \varnothing$), and from Proposition 3.5(ii) if $i \in I^{(4)}$.
- $S_i \cap \operatorname{int}\operatorname{dom} f \subset \operatorname{Fix} T_{i,n}$. (See Proposition 3.38(i), Proposition 3.8(iii), Proposition 3.22(ii)(b), and Proposition 3.3(iv) and (vii), respectively.)

THEOREM 5.18. *Suppose that Condition* 5.16 *is in force and let* $(x_n)_{n\in\mathbb{N}}$ *be an arbitrary orbit of Algorithm* 5.17. *Then* $(x_n)_{n\in\mathbb{N}}$ *converges weakly to a point* $x \in S$. *The convergence is strong if* $x \in \operatorname{int}\operatorname{dom} f$ *and any of the following assumptions is added:*

(i) *For some* $i \in I^{(1)}$ *and some* $\eta \in \left]0, +\infty\right[$, $C \cap \operatorname{lev}_{\leq\eta} g_i$ *is relatively compact.*
(ii) *For some* $i \in I^{(2)}$, $C \cap \operatorname{dom} A_i$ *is relatively compact.*
(iii) *For some* $i \in I^{(3)}$, $C \cap \operatorname{dom} \partial\varphi_i$ *is relatively compact.*
(iv) *For some* $i \in I^{(4)}$, T_i *is demicompact at* 0 *in the sense that for every sequence* $(y_n)_{n\in\mathbb{N}}$ *in* $\operatorname{dom} T_i$

$$\begin{cases} (y_n)_{n\in\mathbb{N}} \text{ bounded}, \\ (\forall n \in \mathbb{N})\ \ u_n \in T_i y_n, \\ u_n - y_n \to 0 \end{cases} \quad \Rightarrow \quad \mathfrak{S}(y_n)_{n\in\mathbb{N}} \neq \varnothing. \tag{5.51}$$

Proof. As seen above, (5.47) is a special case of (5.2), whereas Algorithm 5.17 is a special case of Algorithm 5.1. Invoking Theorem 5.7, we shall prove that Algorithm 5.17 is focusing to establish the weak convergence claim and then that it is demicompactly regular to establish the strong convergence claim. It is recalled that Theorem 5.7 asserts that $(x_n)_{n\in\mathbb{N}}$ and $\big((u_{i,n})_{i\in I_n}\big)_{n\in\mathbb{N}}$ lie in the bounded set C.

To show that Algorithm 5.17 is focusing, let us fix $i \in I$ and a suborbit $(x_{k_n})_{n\in\mathbb{N}}$ such that $i \in \bigcap_{n\in\mathbb{N}} I_{k_n}$, $x_{k_n} \rightharpoonup x$, and $u_{i,k_n} - x_{k_n} \to 0$. According to (5.10), we must show $x \in S_i$. Four cases will be considered:

(1) $i \in I^{(1)}$. We must show $g_i(x) \leq 0$. In view of (5.50), for every $n \in \mathbb{N}$, u_{i,k_n} is the D-projection of x_n onto $G_i(x_{k_n}, x_n^*) = \big\{y \in \mathcal{X} \mid \langle x_{k_n} - y, x_n^*\rangle \geq g_i(x_{k_n})\big\}$ for some $x_n^* \in \partial g_i(x_{k_n})$. Since $u_{i,k_n} \in G_i(x_{k_n}, x_n^*)$, we have

$$\|u_{i,k_n} - x_{k_n}\| \geq d_{G_i(x_{k_n},x_n^*)}(x_{k_n}) = \begin{cases} g_i^+(x_{k_n})/\|x_n^*\| & \text{if } x_n^* \neq 0, \\ 0 & \text{otherwise}, \end{cases} \tag{5.52}$$

where $g_i^+ = \max\{0, g_i\}$ and the last equality follows from [80, Lemma I.1.2]. Since $(x_{k_n})_{n\in\mathbb{N}}$ lies in C, $(x_n^*)_{n\in\mathbb{N}}$ is bounded by Condition 5.16(ii). Therefore, $u_{i,k_n} - x_{k_n} \to 0$ implies $g_i^+(x_{k_n}) \to 0$. However, as g_i^+ is convex and lower semicontinuous, it is weak lower semicontinuous and thus $g_i^+(x) \leq \varliminf g_i^+(x_{k_n}) = 0$. We conclude $g_i(x) \leq 0$.

(2) $i \in I^{(2)}$. We must show $(x, 0) \in \operatorname{gr} A_i$. For every $n \in \mathbb{N}$, (5.50) yields $u_{i,k_n} \in (\nabla f + \gamma_{i,k_n} A_i)^{-1}\big(\nabla f(x_{k_n})\big)$ and we define

$$u_n^* = \frac{\nabla f(x_{k_n}) - \nabla f(u_{i,k_n})}{\gamma_{i,k_n}}. \tag{5.53}$$

Therefore $\big((u_{i,k_n}, u_n^*)\big)_{n\in\mathbb{N}}$ lies in $\operatorname{gr} A_i$ and $u_{i,k_n} - x_{k_n} \to 0 \Rightarrow u_{i,k_n} \rightharpoonup x$. If for all n sufficiently large we have $x_{k_n} = u_{i,k_n}$, then by Proposition 3.8(iii) the

tail of $(x_{k_n})_{n\in\mathbb{N}}$ is in the weakly closed set $A_i^{-1}0$ and therefore $(x,0)\in \operatorname{gr} A_i$. Otherwise, we can extract a subsequence $(x_{k_{l_n}})_{n\in\mathbb{N}}$ such that, for all $n\in\mathbb{N}$, $x_{k_{l_n}}\neq u_{i,k_{l_n}}$. Since, on the one hand, $(x_{k_{l_n}})_{n\in\mathbb{N}}$ and $(u_{i,k_{l_n}})_{n\in\mathbb{N}}$ lie in C and, on the other hand,

$$(5.54)\quad (\forall n\in\mathbb{N})\ \ \|u_{i,k_{l_n}}-x_{k_{l_n}}\|\geq \frac{\langle u_{i,k_{l_n}}-x_{k_{l_n}},\nabla f(u_{i,k_{l_n}})-\nabla f(x_{k_{l_n}})\rangle}{\|\nabla f(u_{i,k_{l_n}})-\nabla f(x_{k_{l_n}})\|},$$

it follows from Condition 5.3(ii), (5.53), and the inequality $\inf_{n\in\mathbb{N}}\gamma_{i,k_{l_n}}\geq\varepsilon_i$ that $u_{i,k_{l_n}}-x_{k_{l_n}}\to 0\Rightarrow \nabla f(u_{i,k_{l_n}})-\nabla f(x_{k_{l_n}})\to 0\Rightarrow u^*_{l_n}\to 0$. Finally, since A_i is maximal monotone, $\operatorname{gr} A_i$ is sequentially closed in the weak $\times$ strong topology of $\mathcal{X}\times\mathcal{X}^*$ and we conclude that $(x,0)\in\operatorname{gr} A_i$, as required.

(3) $i\in I^{(3)}$. We must show $\varphi_i(x)=\inf\varphi_i(\mathcal{X})$, i.e., $(x,0)\in\operatorname{gr}\partial\varphi_i$. Since φ is a proper lower semicontinuous convex function, $A_i=\partial\varphi_i$ is maximal monotone [79, section 29] and 3*-monotone by Lemma 3.10(iv), and, in view of Propositions 3.22(ii)(a) and 3.23(v)(b), the claim follows from case (2).

(4) $i\in I^{(4)}$. We must show $x\in\overline{\operatorname{Fix}}\,T_i$. This follows at once from (5.49).

It remains to show that in each instance (i)–(iv), i is an index of demicompact regularity. Henceforth, $(x_{k_n})_{n\in\mathbb{N}}$ is a suborbit such that $i\in\bigcap_{n\in\mathbb{N}}I_{k_n}$ and $u_{i,k_n}-x_{k_n}\to 0$. By (5.11), we must show $\mathfrak{S}(x_{k_n})_{n\in\mathbb{N}}\neq\varnothing$. (i) Arguing as in case (1), we obtain $\varlimsup g_i(x_{k_n})\leq 0$. Therefore, the tail of $(x_{k_n})_{n\in\mathbb{N}}$ lies in the compact set $\overline{C\cap\operatorname{lev}_{\leq\eta}g_i}$, whence $\mathfrak{S}(x_{k_n})_{n\in\mathbb{N}}\neq\varnothing$. (ii) It follows from (3.16) that for every $n\in\mathbb{N}$

$$\begin{cases} u_{i,k_n}\in C\subset\operatorname{int}\operatorname{dom} f,\\ u_{i,k_n}\in\operatorname{ran}(\nabla f+\gamma_{i,k_n}A_i)^{-1}\circ\nabla f\subset\operatorname{dom}\nabla f\cap\operatorname{dom} A_i=\operatorname{int}\operatorname{dom} f\cap\operatorname{dom} A_i.\end{cases}$$

Therefore, $(u_{i,k_n})_{n\in\mathbb{N}}$ lies in the compact set $\overline{C\cap\operatorname{dom} A_i}$, whence $\mathfrak{S}(u_{i,k_n})_{n\in\mathbb{N}}\neq\varnothing$. Since $u_{i,k_n}-x_{k_n}\to 0$, we conclude $\mathfrak{S}(x_{k_n})_{n\in\mathbb{N}}\neq\varnothing$. (iii) As in case (3), this is a special case of (ii). (iv) This is clear from (5.51). ∎

Theorem 5.18 produces convergence results for various new block-iterative parallel schemes for solving problems, including solving convex inequalities ($I^{(2)}=I^{(3)}=I^{(4)}=\varnothing$), finding common zeros ($I^{(1)}=I^{(3)}=I^{(4)}=\varnothing$), solving systems of variational inequalities ($I^{(1)}=I^{(2)}=I^{(4)}=\varnothing$), finding common fixed points ($I^{(1)}=I^{(2)}=I^{(3)}=\varnothing$), and combinations of these. Note that D-projection methods are also captured by Theorem 5.18 since, in view of Proposition 3.32(ii)(c), one can take, for instance, T_i to be the D-projector onto S_i if $i\in I^{(4)}$ in (5.50).

Naturally, our framework also encompasses relaxed sequential algorithms, which are obtained by taking $(I_n)_{n\in\mathbb{N}}$ to be a sequence of singletons, as in the following example.

EXAMPLE 5.19. *Suppose $\mathcal{X}=\mathbb{R}^N$, $(S_i)_{1\leq i\leq m}$ is a (finite) family of half-spaces with D-projectors $(P_i)_{1\leq i\leq m}$, and, for every $n\in\mathbb{N}$, $I_n=\{n\ (\operatorname{mod} m)+1\}$ and $T_{i,n}=P_i$. Then Algorithm* 5.1 *reduces to the relaxed D-projection method of* [44].

In the case of unrelaxed sequential algorithms, our working assumptions can be loosened. This is discussed next.

5.6. Unrelaxed sequential algorithms. Algorithm 5.1 can be specialized to an unrelaxed sequential algorithm for solving the convex feasibility problem (5.2). Indeed, suppose that at each iteration n only one index, say $\mathrm{i}(n)$, is retained and $\lambda_n=1$. Then Algorithm 5.1⑧ becomes

$$(5.55)\qquad H_n=\big\{y\in\mathcal{X}\mid\langle y-u_n,\nabla f(x_n)-\nabla f(u_n)\rangle\leq 0\big\},$$

where $u_n \in T_n x_n$ for some $T_n \in \mathfrak{B}$ such that $S_{\mathrm{i}(n)} \cap \operatorname{int}\operatorname{dom} f \subset \operatorname{Fix} T_n$. Consequently, since by Corollary 3.35(ii) $P_{H_n} x_n = u_n$, Algorithm 5.1 can be rewritten as follows.

ALGORITHM 5.20. *Starting with $x_0 \in \operatorname{int}\operatorname{dom} f$, take at every iteration n*

① *an index $\mathrm{i}(n) \in I$,*

② *an operator T_n in $\mathfrak{B}$ such that $S_{\mathrm{i}(n)} \cap \operatorname{int}\operatorname{dom} f \subset \operatorname{Fix} T_n$.*

Then select $x_{n+1} \in T_n x_n$.

In this context, Definition 5.2 takes the following form.

DEFINITION 5.21. *Algorithm* 5.20 *is*

- focusing *if for every bounded suborbit $(x_{k_n})_{n\in\mathbb{N}}$ it generates and every index $i \in I$,*

$$\begin{cases} (\forall n \in \mathbb{N}) \;\; i = \mathrm{i}(k_n), \\ x_{k_n} \rightharpoonup x, \\ x_{k_n+1} - x_{k_n} \to 0 \end{cases} \quad \Rightarrow \quad x \in S_i; \tag{5.56}$$

- demicompactly regular *if there exists $i \in I$, called an* index of demicompact regularity*, such that for every bounded suborbit $(x_{k_n})_{n\in\mathbb{N}}$ it generates,*

$$\begin{cases} (\forall n \in \mathbb{N}) \;\; i = \mathrm{i}(k_n), \\ x_{k_n+1} - x_{k_n} \to 0 \end{cases} \quad \Rightarrow \quad \mathfrak{S}(x_{k_n})_{n\in\mathbb{N}} \neq \varnothing. \tag{5.57}$$

Removing item (ii) from Condition 5.3 yields the following set of assumptions for the unrelaxed sequential case.

CONDITION 5.22. *For some $z \in \operatorname{dom} f \cap \bigcap_{i\in I} S_i$, $C = \operatorname{lev}_{\leq D(z,x_0)} D(z,\cdot)$ is bounded.*

CONDITION 5.23. $(\forall i \in I)(\exists\, M_i \in \mathbb{N} \smallsetminus \{0\})(\forall n \in \mathbb{N}) \;\; i \in \big\{\mathrm{i}(n), \ldots, \mathrm{i}(n+M_i-1)\big\}$.

We now show that Algorithm 5.20 converges under this reduced set of assumptions.

THEOREM 5.24. *Suppose that Conditions* 4.3*,* 4.4*,* 5.22*, and* 5.23 *are satisfied and that Algorithm* 5.20 *is focusing. Then the following statements hold true for every orbit $(x_n)_{n\in\mathbb{N}}$ generated by Algorithm* 5.20*:*

(i) *$(x_n)_{n\in\mathbb{N}}$ converges weakly to a point $x \in S$.*

(ii) *If the weak limit x from* (i) *belongs to* $\operatorname{int}\operatorname{dom} f$ *and the algorithm is demicompactly regular, then $(x_n)_{n\in\mathbb{N}}$ converges strongly.*

Proof. In the proof of Theorem 5.7, note that Condition 5.3(ii) is used only to obtain (5.34), i.e., in the present context, $x_{p_n+1} - x_{p_n} \to 0$. However, this property follows directly from(5.33). □

As an example, we revisit Bregman's original cyclic projection method (1.2). (See [1, Thm. 3.1] for a special case.)

COROLLARY 5.25. *Suppose that Conditions* 4.3 *and* 4.4 *are satisfied, that $I = \{1, \ldots, m\}$, and that $C = \operatorname{lev}_{\leq D(z,x_0)} D(z,\cdot)$ is bounded for some $z \in \operatorname{dom} f \cap \bigcap_{1\leq i\leq m} S_i$. Let $(P_i)_{1\leq i\leq m}$ be the D-projectors of $(S_i)_{1\leq i\leq m}$. Then the following statements hold true for every orbit $(x_n)_{n\in\mathbb{N}}$ generated by* (1.2)*:*

(i) *$(x_n)_{n\in\mathbb{N}}$ converges weakly to a point $x \in \overline{\operatorname{dom}} f \cap \bigcap_{1\leq i\leq m} S_i$.*

(ii) *If the weak limit x from* (i) *belongs to* $\operatorname{int}\operatorname{dom} f$ *and $C \cap S_i$ is relatively compact (e.g., S_i is boundedly compact) for some $i \in \{1, \ldots, m\}$, then $(x_n)_{n\in\mathbb{N}}$ converges strongly.*

Proof. In view of Corollary 3.35(i), (1.2) is a special realization of Algorithm 5.20 with $(\forall n \in \mathbb{N})$ $T_n = P_{n \,(\mathrm{mod}\, m)+1}$ (single-valued) and $\lambda_n = 1$. In addition, the index

control rule $i\colon n \mapsto n \,(\mathrm{mod}\, m) + 1$ complies with Condition 5.23. On the other hand, algorithm (1.2) is focusing, as a direct consequence of the weak closedness of the sets $(S_j)_{1\leq j\leq m}$. Finally, i is an index of demicompact regularity since $(x_{nm+i})_{n\in\mathbb{N}}$ lies in $C \cap S_i$. The announced results therefore follow from Theorem 5.24. □

Remark 5.26. Throughout section 5, Legendreness has been imposed on f. This property has been shown to provide a rich and convenient framework in which our results could be derived in a unified manner. Further results can nonetheless be obtained from the analysis of sections 3 and 4 for functions which are not Legendre at the expense of more technical assumptions.

Acknowledgments. We wish to thank Dan Butnariu and Yair Censor for sending us [22, 25, 26, 29], Constantin Zălinescu for sending us [85], and especially Jon Vanderwerff for his help in the derivation of Example 5.11. Two anonymous referees made several helpful comments and suggestions, which led to improvements over the originally submitted version.

REFERENCES

[1] Y. Alber and D. Butnariu, *Convergence of Bregman projection methods for solving consistent convex feasibility problems in reflexive Banach spaces*, J. Optim. Theory Appl., 92 (1997), pp. 33–61.

[2] H. Attouch and H. Brézis, *Duality for the sum of convex functions in general Banach spaces*, in Aspects of Mathematics and Its Applications, North–Holland, Amsterdam, 1986, pp. 125–133.

[3] J.-P. Aubin, *Viability Theory*, Birkhäuser, Boston, 1991.

[4] J.-B. Baillon and G. Haddad, *Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones*, Israel J. Math., 26 (1977), pp. 137–150.

[5] H. H. Bauschke, *Projection Algorithms and Monotone Operators*, Ph.D. thesis, Simon Fraser University, Canada, 1996; available online as preprint 96:080 from http://www.cecm.sfu.ca/preprints.

[6] H. H. Bauschke and J. M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.

[7] H. H. Bauschke and J. M. Borwein, *Legendre functions and the method of random Bregman projections*, J. Convex Anal., 4 (1997), pp. 27–67.

[8] H. H. Bauschke, J. M. Borwein, and P. L. Combettes, *Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces*, Commun. Contemp. Math., 3 (2001), pp. 615–647.

[9] H. H. Bauschke and P. L. Combettes, *A weak-to-strong convergence principle for Fejér-monotone methods in Hilbert spaces*, Math. Oper. Res., 26 (2001), pp. 248–264.

[10] H. H. Bauschke and P. L. Combettes, *Construction of best Bregman approximations in reflexive Banach spaces*, Proc. Amer. Math. Soc., published electronically April 24, 2003.

[11] H. H. Bauschke and P. L. Combettes, *Iterating Bregman retractions*, SIAM J. Optim., 13 (2003), pp. 1159–1173.

[12] H. H. Bauschke, P. L. Combettes, and D. R. Luke, *Phase retrieval, error reduction algorithm, and Fienup variants: A view from convex optimization*, J. Opt. Soc. Amer. A, 19 (2002), pp. 1334–1345.

[13] J. M. Borwein and J. D. Vanderwerff, *Convex functions of Legendre type in general Banach spaces*, J. Convex Anal., 8 (2001), pp. 569–581.

[14] L. M. Bregman, *The method of successive projection for finding a common point of convex sets*, Soviet Math. Dokl., 6 (1965), pp. 688–692.

[15] L. M. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Comput. Math. Math. Phys., 7 (1967), pp. 200–217.

[16] H. Brézis and A. Haraux, *Image d'une somme d'opérateurs monotones et applications*, Israel J. Math., 23 (1976), pp. 165–186.

[17] H. Brézis and P. L. Lions, *Produits infinis de résolvantes*, Israel J. Math., 29 (1978), pp. 329–345.

[18] M. Brohe and P. Tossings, *Perturbed proximal point algorithm with nonquadratic kernel*, Serdica Math. J., 26 (2000), pp. 177–206.

[19] R. E. Bruck and S. Reich, *Nonexpansive projections and resolvents of accretive operators in Banach spaces*, Houston J. Math., 3 (1977), pp. 459–470.

[20] R. S. Burachik and A. N. Iusem, *A generalized proximal point algorithm for the variational inequality problem in a Hilbert space*, SIAM J. Optim., 8 (1998), pp. 197–216.

[21] R. S. Burachik and S. Scheimberg, *A proximal point method for the variational inequality problem in Banach spaces*, SIAM J. Control Optim., 39 (2001), pp. 1633–1649.

[22] D. Butnariu, C. Byrne, and Y. Censor, *Redundant axioms in the definition of Bregman functions,* J. Convex Anal., to appear in (2003).

[23] D. Butnariu and A. N. Iusem, *On a proximal point method for convex optimization in Banach spaces,* Numer. Funct. Anal. Optim., 18 (1997), pp. 723–744.

[24] D. Butnariu and A. N. Iusem, *Totally Convex Functions for Fixed Points Computation and Infinite Dimensional Optimization,* Kluwer Academic Publishers, Boston, MA, 2000.

[25] D. Butnariu, A. N. Iusem, and C. Zălinescu, *On uniform convexity, total convexity and convergence of the proximal point and outer Bregman projection algorithms in Banach spaces*, J. Convex Anal., to appear in (2003).

[26] D. Butnariu, S. Reich, and A. J. Zaslavski, *Asymptotic behavior of relatively nonexpansive operators in Banach spaces*, J. Appl. Anal., 7 (2001), pp. 151–174.

[27] D. Butnariu and E. Resmerita, *The outer Bregman projection method for stochastic feasibility problems in Banach spaces*, in Inherently Parallel Algorithms for Feasibility and Optimization, D. Butnariu, Y. Censor, and S. Reich, eds., Elsevier, New York, 2001, pp. 69–86.

[28] J. A. Cadzow, *Signal enhancement – A composite property mapping algorithm*, IEEE Trans. Acoust. Speech Signal Process., 36 (1988), pp. 49–62.

[29] Y. Censor and G. T. Herman, *Block-iterative algorithms with underrelaxed Bregman projections*, SIAM J. Optim., 13 (2002), pp. 283–297.

[30] Y. Censor and A. Lent, *An iterative row-action method for interval convex programming*, J. Optim. Theory Appl., 34 (1981), pp. 321–353.

[31] Y. Censor and S. Reich, *Iterations of paracontractions and firmly nonexpansive operators with applications to feasibility and optimization*, Optimization, 37 (1996), pp. 323–339.

[32] Y. Censor and S. A. Zenios, *Proximal minimization algorithm with D-functions,* J. Optim. Theory Appl., 73 (1992), pp. 451–464.

[33] Y. Censor and S. A. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications,* Oxford University Press, New York, 1997.

[34] I. Ciorănescu, *Geometry of Banach Spaces, Duality Mappings, and Nonlinear Problems,* Kluwer Academic Publishers, Boston, MA, 1990.

[35] P. L. Combettes, *Construction d'un point fixe commun à une famille de contractions fermes*, C. R. Acad. Sci. Paris Sér. I Math., 320 (1995), pp. 1385–1390.

[36] P. L. Combettes, *The convex feasibility problem in image recovery*, in Advances in Imaging and Electron Physics, Vol. 95, P. Hawkes, ed., Academic Press, New York, 1996, pp. 155–270.

[37] P. L. Combettes, *Convex set theoretic image recovery by extrapolated iterations of parallel subgradient projections*, IEEE Trans. Image Process., 6 (1997), pp. 493–506.

[38] P. L. Combettes, *Hilbertian convex feasibility problem: Convergence of projection methods*, Appl. Math. Optim., 35 (1997), pp. 311–330.

[39] P. L. Combettes, *Strong convergence of block-iterative outer approximation methods for convex optimization*, SIAM J. Control Optim., 38 (2000), pp. 538–565.

[40] P. L. Combettes, *Fejér monotonicity in convex optimization*, in Encyclopedia of Optimization, Vol. 2, C. A. Floudas and P. M. Pardalos, eds., Kluwer Academic Publishers, Boston, MA, 2001, pp. 106–114.

[41] P. L. Combettes, *Quasi-Fejérian analysis of some optimization algorithms*, in Inherently Parallel Algorithms for Feasibility and Optimization, D. Butnariu, Y. Censor, and S. Reich, eds., Elsevier, New York, 2001, pp. 115–152.

[42] P. L. Combettes and T. Pennanen, *Generalized Mann iterates for constructing fixed points in Hilbert spaces*, J. Math. Anal. Appl., 275 (2002), pp. 521–536.

[43] P. L. Combettes and H. J. Trussell, *Method of successive projections for finding a common point of sets in metric spaces*, J. Optim. Theory Appl., 67 (1990), pp. 487–507.

[44] A. R. de Pierro and A. N. Iusem, *A relaxed version of Bregman's method for convex programming*, J. Optim. Theory Appl., 51 (1986), pp. 421–440.

[45] R. Deville, G. Godefroy, and V. Zizler, *Smoothness and Renormings in Banach Spaces*, John Wiley, New York, 1993.

[46] J. Eckstein, *Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming*, Math. Oper. Res., 18 (1993), pp. 202–226.
[47] J. Eckstein, *Approximate iterations in Bregman-function-based proximal algorithms*, Math. Program., 83 (1998), pp. 113–123.
[48] I. Ekeland and R. Témam, *Convex Analysis and Variational Problems,* Classics Appl. Math. 28, SIAM, Philadelphia, PA, 1999.
[49] I. I. Eremin and V. D. Mazurov, *Nonstationary Processes of Mathematical Programming,* Nauka, Moscow, 1979.
[50] U. M. García-Palomares and F. J. González-Castaño, *Incomplete projection algorithms for solving the convex feasibility problem,* Numer. Algorithms, 18 (1998), pp. 177–193.
[51] K. Goebel and S. Reich, *Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings,* Marcel Dekker, New York, 1984.
[52] K. M. Grigoriadis and R. E. Skelton, *Low-order control design for LMI problems using alternating projection methods*, Automatica J. IFAC, 32 (1996), pp. 1117–1125.
[53] L. G. Gubin, B. T. Polyak, and E. V. Raik, *The method of projections for finding the common point of convex sets*, USSR Comput. Math. Math. Phys., 7 (1967), pp. 1–24.
[54] C. D. Ha, *A generalization of the proximal point algorithm*, SIAM J. Control Optim., 28 (1990), pp. 503–512.
[55] A. Iusem and R. G. Otero, *Inexact versions of proximal point and augmented Lagrangian algorithms in Banach spaces*, Numer. Funct. Anal. Optim., 22 (2001), pp. 609–640.
[56] A. Iusem and R. G. Otero, *Erratum: "Inexact versions of proximal point and augmented Lagrangian algorithms in Banach spaces"*, Numer. Funct. Anal. Optim., 23 (2002), pp. 227–228.
[57] G. Kassay, *The proximal points algorithm for reflexive Banach spaces,* Studia Univ. Babeş-Bolyai Math., 30 (1985), pp. 9–17.
[58] K. C. Kiwiel, *Proximal minimization methods with generalized Bregman functions*, SIAM J. Control Optim., 35 (1997), pp. 1142–1168.
[59] K. C. Kiwiel, *Generalized Bregman projections in convex feasibility problems*, J. Optim. Theory Appl., 96 (1998), pp. 139–157.
[60] K. C. Kiwiel and B. Łopuch, *Surrogate projection methods for finding fixed points of firmly nonexpansive mappings*, SIAM J. Optim., 7 (1997), pp. 1084–1102.
[61] B. Lemaire, *On the convergence of some iterative methods for convex minimization*, in Recent Developments in Optimization, Lect. Notes Econ. Math. Syst. 429, Springer-Verlag, Berlin, 1995, pp. 252–268.
[62] A. Levi and H. Stark, *Image restoration by the method of generalized projections with application to restoration from magnitude*, J. Opt. Soc. Amer. A, 1 (1984), pp. 932–943.
[63] Y. I. Merzlyakov, *On a relaxation method of solving systems of linear inequalities*, USSR Comput. Math. Math. Phys., 2 (1963), pp. 504–510.
[64] J.-J. Moreau, *Fonctions convexes duales et points proximaux dans un espace hilbertien*, C. R. Acad. Sci. Paris Sér. A Math., 255 (1962), pp. 2897–2899.
[65] J.-J. Moreau, *Propriétés des applications prox*, C. R. Acad. Sci. Paris Sér. A Math., 256 (1963), pp. 1069–1071.
[66] J.-J. Moreau, *Sur la fonction polaire d'une fonction semi-continue supérieurement*, C. R. Acad. Sci. Paris Sér. A Math., 258 (1964), pp. 1128–1130.
[67] J.-J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
[68] Z. Opial, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.
[69] N. Ottavy, *Strong convergence of projection-like methods in Hilbert spaces*, J. Optim. Theory Appl., 56 (1988), pp. 433–461.
[70] W. V. Petryshyn, *Construction of fixed points of demicompact mappings in Hilbert space*, J. Math. Anal. Appl., 14 (1966), pp. 276–284.
[71] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, 2nd ed., Lecture Notes in Math. 1364, Springer-Verlag, New York, 1993.
[72] G. Pierra, *Decomposition through formalization in a product space*, Math. Program., 28 (1984), pp. 96–115.
[73] E. Raik, *Fejér type methods in Hilbert space*, Esti NSV Tead. Akad. Toimetised Füüs.-Mat., 16 (1967), pp. 286–293.
[74] S. Reich, *The range of sums of accretive and monotone operators*, J. Math. Anal. Appl., 68 (1979), pp. 310–317.
[75] S. Reich, *A weak convergence theorem for the alternating method with Bregman distances*, in Theory and Applications of Nonlinear Operators of Accretive and Monotone Type, A. G.

Kartsatos, ed., Marcel Dekker, New York, 1996, pp. 313–318.
[76] R. T. Rockafellar, *Level sets and continuity of conjugate convex functions*, Trans. Amer. Math. Soc., 123 (1966), pp. 46–63.
[77] R. T. Rockafellar, *Convex Analysis,* Princeton University Press, Princeton, NJ, 1970.
[78] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[79] S. Simons, *Minimax and Monotonicity,* Lecture Notes in Math. 1693, Springer-Verlag, New York, 1998.
[80] I. Singer, *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces,* Springer-Verlag, New York, 1970.
[81] I. Singer, *The Theory of Best Approximation and Functional Analysis,* CBMS-NSF Regional Conf. Ser. in Appl. Math. 13, SIAM, Philadelphia, PA, 1974.
[82] M. V. Solodov and B. F. Svaiter, *An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions,* Math. Oper. Res., 25 (2000), pp. 214–230.
[83] M. Teboulle, *Entropic proximal mappings with applications to nonlinear programming*, Math. Oper. Res., 17 (1992), pp. 670–690.
[84] C. Zălinescu, *On uniformly convex functions*, J. Math. Anal. Appl., 95 (1983), pp. 344–374.
[85] C. Zălinescu, *Convex Analysis in General Vector Spaces,* World Scientific Publishing, River Edge, NJ, 2002.
[86] E. Zeidler, *Nonlinear Functional Analysis and Its Applications.* II*/B. Nonlinear Monotone Operators,* Springer-Verlag, New York, 1990.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 637–668

OPTIMAL COORDINATED MOTIONS OF MULTIPLE AGENTS MOVING ON A PLANE*

JIANGHAI HU†, MARIA PRANDINI‡, AND SHANKAR SASTRY†

Abstract. We address the problem of optimal coordinated motions of multiple agents moving in the same planar region. The agents' motions must satisfy a separation constraint throughout the encounter to be conflict-free. The objective is to determine the conflict-free maneuvers (motions) with the least combined energy, while taking into account the fact that agents may have different priorities. A formal classification of conflict-free maneuvers into homotopy types is introduced by using their braid representation. Various local and global optimality conditions are derived through variational analysis in the presence of the separation constraint. In the case of two agents, these optimality conditions allow us to construct the optimal maneuvers geometrically. For the general multi-agent case, a convex optimization algorithm is proposed to compute within each homotopy type a solution to the optimization problem restricted to the class of multilegged maneuvers. Since the number of types grows explosively with the number of agents, a stochastic algorithm is suggested as the "type chooser," thus leading to a randomized optimization algorithm.

Key words. cooperative motion planning, braids, calculus of variation with constraints, convex optimization

AMS subject classifications. 65K10, 93C85, 90C25

PII. S0363012901387562

1. Introduction.

In this paper, the problem of designing coordinated maneuvers for multiple agents moving on a plane is studied. The joint maneuver has to be chosen so as to guide each agent from its starting position to its target position, while avoiding *conflicts*, that is, situations where the Euclidean distance between any two agents is smaller than some fixed threshold $R > 0$. Among all the conflict-free joint maneuvers, we aim to determine the one with the least overall cost. Here the cost of a single agent's maneuver is its energy, and the overall cost is a weighted sum of the maneuver energies of all individual agents, with the weights representing the priorities of the agents. A precise formulation of the problem is given in section 3.

This problem is of great interest since it is actually encountered in many different practical areas. For example, in the air traffic control (ATC) context, aircraft flying at the same altitude must maintain a minimal horizontal separation R of at least 3 nautical miles (nmi) inside the terminal radar approach control facilities and 5 nmi in the en-route airspace [35]. In this case, the energy is closely related to practical aspects such as travel distance, fuel consumption, passenger comfort, etc. Numerous approaches have been proposed in the literature on aircraft conflict resolution, including optimal control theory [5], semidefinite programming [10], sequential quadratic programming [29], game theory [39, 40], parallel coordinates representation

*Received by the editors April 9, 2001; accepted for publication (in revised form) January 7, 2003; published electronically May 29, 2003. Part of the results in this paper have appeared in [12, 14, 17]. These references are available in electronic format upon request. This research has been supported by DARPA under grant F33615-98-C-3614, by the National Science Foundation under grant EIA-0122599, and by MIUR under the project "New techniques for the identification and adaptive control of industrial systems."

http://www.siam.org/journals/sicon/42-2/38756.html

†Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 (jianghai@robotics.eecs.berkeley.edu, sastry@robotics.eecs.berkeley.edu).

‡Dipartimento di Elettronica e Informazione, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy (prandini@elet.polimi.it).

[19], and genetic algorithms [28], to name a few. Readers are referred to [15, 23] for a survey on aircraft conflict resolution. Similar problems have been studied in other transportation systems as well, such as [30].

The problem of optimal multi-agent coordinated motions also finds applications in robotics. For example, for multiple cooperating mobile robots moving in a common workspace, the requirement that there be no collision among them can be reformulated as that their joint maneuver be conflict-free, with R being twice the robot radius. The literature on the general problem of robot motion planning with static or dynamic obstacles is vast (see, e.g., [4, 9, 11, 24, 38] and the survey [18]), and it is impossible to survey them in this paper. Here we limit our review to those contributions more relevant to our work. A large portion of the treatments focus on the feasibility and the algorithmic complexity aspects of the problem. Some of them indeed deal with the multiple robots case using certain optimality criteria. To name a few, [5] studies the problem of time-optimal cooperative motions of multiple Dubin vehicles moving at constant speed with bounded curvature, while in [25] each robot minimizes its own independent cost function by using techniques from multi-objective optimization and game theory. [7] addresses the problem of optimal motion planning for multiple nonholonomic manipulators transporting a grasped object.

The distinguishing feature of our approach to coordinated motion planning consists in the interpretation of maneuvers as braids. Besides giving a complete homotopic classification of conflict-free maneuvers, this also provides us insights on the derivation of optimality conditions. Although the space-time representation of motions is not new in the literature (see, e.g., [9, 37]), to our knowledge, however, it has never been used to such an extent in the optimality analysis of coordinated motions.

Due to the many different interpretations of conflict-free maneuvers (not only as braids, but also as, e.g., solutions to mechanical systems or geodesics in a manifold with boundary), many of the results in this paper can be derived in more than one way. For example, some of the local optimality conditions in section 3 can be derived by using the symmetry reduction method in [3, 26]. In most cases, we choose our approaches with an emphasis on their geometric appealing and their relevancy to the braid point of view. As a result, they may not always be the most elegant and efficient ones. In addition, although we focus exclusively on the case when the state space is $\mathbb{R}^2$, extensions to general state spaces are possible [16]. These possible extensions, as well as the remaining open issues, will be pointed out in the paper wherever possible.

This paper is organized as follows. In section 2, we introduce a formal classification of conflict-free maneuvers into homotopy types by using the notion of pure braids group. Inspired by the braid representation of conflict-free maneuvers, we define various transformations of joint maneuvers that preserve the minimum separation condition. Such transformations are used in the variational analysis in section 3 to derive local and global necessary conditions on optimal conflict-free maneuvers. In particular, the optimal conflict-free maneuvers for the 2-agent case are derived in section 3.3. Two mechanical interpretations of the problem are given in section 3.8.

As the number of agents increases, it is difficult in practice to derive analytically the optimal conflict-free maneuvers. By focusing on those maneuvers specified by a set of waypoints, we are able to use convex optimization techniques to obtain multilegged approximated solutions to the constrained optimization problem within each homotopy type (section 4). A stochastic algorithm is proposed in section 4.4 to address the problem of selecting the homotopy type, thus leading to a randomized convex optimization algorithm.

The paper is concluded in section 5 with some general remarks and the outline of some possible extensions of this research.

2. Classification of conflict-free maneuvers. In this section, we introduce a qualitative classification of conflict-free maneuvers involving multiple agents. Roughly speaking, two conflict-free maneuvers are classified as of the same "type" if there exists a continuous conflict-free deformation of one to the other. Hence switching between different types cannot be done smoothly without causing a conflict.

Consider n agents (numbered from 1 to n) moving in $\mathbb{R}^2$, where each agent, say agent i, starts at position $a_i \in \mathbb{R}^2$ at time t_0 and ends in position $b_i \in \mathbb{R}^2$ at time t_f. Let $T \triangleq [t_0, t_f]$ be the time interval of the encounter. Denote by $\mathbf{P}_i \triangleq \{\alpha_i \in C(T, \mathbb{R}^2) : \alpha_i(t_0) = a_i,\ \alpha_i(t_f) = b_i\}$ the set of maneuvers for agent i consisting of all the continuous maps from T to $\mathbb{R}^2$ that take the values a_i and b_i at times t_0 and t_f, respectively. Set $\mathbf{P}(\mathbf{a}, \mathbf{b}) \triangleq \prod_{i=1}^{n} \mathbf{P}_i$, where $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_n)$. Each element $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbf{P}(\mathbf{a}, \mathbf{b})$ is called a *joint maneuver* (n-maneuver or simply maneuver when there is no ambiguity) for the n-agent system. The *minimum separation over encounter* (MSE) for a joint maneuver α is defined to be the minimum Euclidean distance between any pair of agents during the whole time interval T, i.e.,

$$\Delta(\alpha) \triangleq \min_{1 \le i < j \le n} \inf_{t \in T} \|\alpha_i(t) - \alpha_j(t)\|.$$

The set of *conflict-free maneuvers* is then defined as

$$\mathbf{P}(R, \mathbf{a}, \mathbf{b}) \triangleq \{\alpha \in \mathbf{P}(\mathbf{a}, \mathbf{b}) : \Delta(\alpha) > R\},$$

where R is a positive number representing, for example, the radius of the protection zone surrounding an aircraft or twice the radius of a circular robot. We assume that the minimum distance between any pair of starting positions in the n-tuple $\langle a_i \rangle_{i=1}^{n}$ and any pair of ending positions in the n-tuple $\langle b_i \rangle_{i=1}^{n}$ is strictly greater than R, so that $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$ is nonempty.

We distinguish different maneuvers in $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$ according to the following equivalence relation.

DEFINITION 2.1 (R-homotopy). *Two conflict-free maneuvers in* $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$ *are* R-homotopic *if there exists a continuous deformation of one to the other in* $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$*, or, equivalently, if there exists a continuous deformation of one to the other in* $\mathbf{P}(\mathbf{a}, \mathbf{b})$ *such that the joint maneuvers obtained throughout the deformation are conflict-free.*

The objective of this section is to characterize the structure of the equivalence classes of $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$ induced by the R-homotopy relation. With this purpose in mind, we now recall the concept of braids [6, 32].

DEFINITION 2.2 (braids). *A* braid *joining* $\mathbf{a} = (a_1, \ldots, a_n)$ *to* $\mathbf{b} = (b_1, \ldots, b_n)$ *is an* n*-tuple* $\langle \gamma_i \rangle_{i=1}^{n}$ *of continuous curves in* $\mathbb{R}^2 \times T \subset \mathbb{R}^3$ *satisfying the following conditions:*

- *Each point* (a_i, t_0), $i = 1, \ldots, n$, *is joined by exactly one curve in* $\langle \gamma_i \rangle_{i=1}^{n}$ *to one of the points* (b_j, t_f), $1 \le j \le n$.
- *The plane* $t = \tau$ *intersects each curve at exactly one point for all* $\tau \in T$.
- $\gamma_i \cap \gamma_j = \emptyset$ *whenever* $i \neq j$.

In the following, we shall occasionally use the term n*-braid* to indicate the number of curves in the braid. The set of all braids joining $\mathbf{a}$ to $\mathbf{b}$ is denoted by $\mathbf{B}(\mathbf{a}, \mathbf{b})$. If i and j are required to be identical in the first condition of Definition 2.2, the corresponding braid is called a *pure braid*. The set of all pure braids joining $\mathbf{a}$ to $\mathbf{b}$ is denoted by $\mathbf{PB}(\mathbf{a}, \mathbf{b})$. An example of a pure 3-braid is shown in the right-hand side of Figure 2.1.

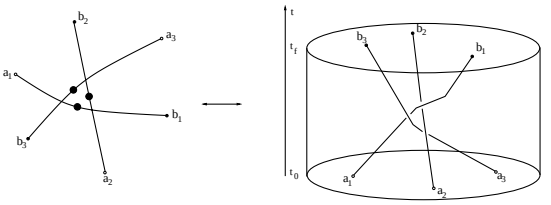


FIG. 2.1. *A 3-maneuver in* $\mathbf{P}(0, \mathbf{a}, \mathbf{b})$ *and its braid representation.*

There is a simple equivalence relation defined on $\mathbf{B}(\mathbf{a}, \mathbf{b})$ and hence on $\mathbf{PB}(\mathbf{a}, \mathbf{b})$ as well [32].

DEFINITION 2.3 (string isotopy). *Two braids in* $\mathbf{B}(\mathbf{a}, \mathbf{b})$ *are said to be* string isotopic *if the n curves of one of them can be continuously deformed to those of the other such that the n curves in* $\mathbb{R}^2 \times T$ *obtained throughout the deformation satisfy all the conditions in Definition* 2.2.

The reason for introducing the notion of braids is that there exists a very natural one-to-one correspondence between joint maneuvers in $\mathbf{P}(0, \mathbf{a}, \mathbf{b})$ and pure braids in $\mathbf{PB}(\mathbf{a}, \mathbf{b})$. To see this, for each joint maneuver $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbf{P}(0, \mathbf{a}, \mathbf{b})$, let $\hat{\alpha}_i$ be the curve in $\mathbb{R}^2 \times T$ joining (a_i, t_0) to (b_i, t_f) defined as the image of the map $t \mapsto (\alpha_i(t), t)$, $t \in T$. Then, it is clear from the definition of $\mathbf{P}(0, \mathbf{a}, \mathbf{b})$ that the n-tuple $\langle \hat{\alpha}_i \rangle_{i=1}^n$ of curves is indeed a pure braid in $\mathbf{PB}(\mathbf{a}, \mathbf{b})$, which we shall denote by $\hat{\alpha}$. (See Figure 2.1 for a 3-maneuver in $\mathbf{P}(0, \mathbf{a}, \mathbf{b})$ and its braid representation.) The map $\alpha \mapsto \hat{\alpha}$ can be verified to be a bijection between $\mathbf{P}(0, \mathbf{a}, \mathbf{b})$ and $\mathbf{PB}(\mathbf{a}, \mathbf{b})$. Furthermore, the following result is an immediate consequence of the above definitions.

PROPOSITION 2.4 (equivalence of 0-homotopy and string isotopy). *α and $\beta \in \mathbf{P}(0, \mathbf{a}, \mathbf{b})$ are 0-homotopic if and only if $\hat{\alpha}$ and $\hat{\beta}$ are string isotopic in $\mathbf{PB}(\mathbf{a}, \mathbf{b})$.*

As a result of Proposition 2.4, there is a one-to-one correspondence between the 0-homotopy classes of $\mathbf{P}(0, \mathbf{a}, \mathbf{b})$ and the (string) isotopy classes of $\mathbf{PB}(\mathbf{a}, \mathbf{b})$.

We next show that the isotopy classes of braids with identical starting and ending points, say $\mathbf{B}(\mathbf{a}, \mathbf{a})$, form a group under a suitable product operation. For each $\hat{\alpha} \in \mathbf{B}(\mathbf{a}, \mathbf{b})$ and $\hat{\beta} \in \mathbf{B}(\mathbf{b}, \mathbf{c})$, define the product $\hat{\gamma} \triangleq \hat{\alpha} \cdot \hat{\beta}$ as the braid $\hat{\gamma} \in \mathbf{B}(\mathbf{a}, \mathbf{c})$ obtained by first concatenating the n curves of $\hat{\alpha}$ with those of $\hat{\beta}$ and then renormalizing the t axis linearly such that the resultant n curves connect $\langle (a_i, t_0) \rangle_{i=1}^n$ to $\langle (c_i, t_f) \rangle_{i=1}^n$ via $\langle (b_i, \frac{t_0 + t_f}{2}) \rangle_{i=1}^n$. Note that the ending points of $\hat{\alpha}$ and the starting points of $\hat{\beta}$ have to coincide for the product to be well defined. It can be easily checked that this product operation preserves string isotopy, i.e., if $\hat{\alpha}'$ is string isotopic to $\hat{\alpha}$ in $\mathbf{B}(\mathbf{a}, \mathbf{b})$ and $\hat{\beta}'$ is string isotopic to $\hat{\beta}$ in $\mathbf{B}(\mathbf{b}, \mathbf{c})$, then $\hat{\alpha}' \cdot \hat{\beta}'$ is string isotopic to $\hat{\alpha} \cdot \hat{\beta}$ in $\mathbf{B}(\mathbf{a}, \mathbf{c})$. Therefore, it induces a product operation on the isotopy classes of braids. This induced product operation makes the isotopy classes of $\mathbf{B}(\mathbf{a}, \mathbf{a})$ into a group, with the inverse operation being the reflection of the n curves across the plane $t = \frac{t_0 + t_f}{2}$. We denote this group by $\mathbf{B}_n$. Similarly the isotopy classes of pure braids $\mathbf{PB}(\mathbf{a}, \mathbf{a})$ form under the same induced product operation a group, which we denote by $\mathbf{PB}_n$. $\mathbf{PB}_n$ is a normal subgroup of $\mathbf{B}_n$. Readers are referred to [12] or [32] for a detailed derivation of the above claims.

Now if we fix a braid $\hat{\beta}$ in $\mathbf{PB}(\mathbf{b}, \mathbf{a})$, then $\hat{\alpha} \mapsto \hat{\alpha} \cdot \hat{\beta}$ defines a map from $\mathbf{PB}(\mathbf{a}, \mathbf{b})$ to $\mathbf{PB}(\mathbf{a}, \mathbf{a})$. Since this map preserves string isotopy, it induces a map from the isotopy classes of $\mathbf{PB}(\mathbf{a}, \mathbf{b})$ to the isotopy classes of $\mathbf{PB}(\mathbf{a}, \mathbf{a})$, i.e., $\mathbf{PB}_n$. The induced map is easily verified to be a bijection. This fact combined with the result in Proposition 2.4

FIG. 2.2. 2-*agent encounter. Left: Maneuver* 1. *Right: Maneuver* 2.

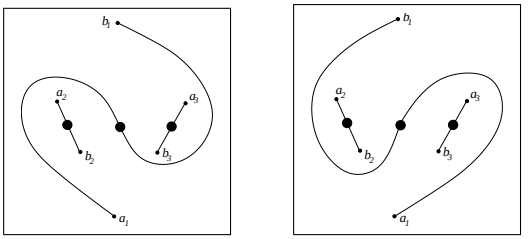


FIG. 2.3. *Two* 3-*maneuvers with the same turning angle but belonging to different types.*

implies that there exists a bijection between the 0-homotopy classes of $\mathbf{P}(0, \mathbf{a}, \mathbf{b})$ and the elements of $\mathbf{PB}_n$.

The above conclusions remain valid for the case of an arbitrary $R > 0$. Hence, the following theorem.

THEOREM 2.5 (classification of conflict-free n-maneuvers). *The R-homotopy classes of conflict-free maneuvers in $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$ have a one-to-one correspondence with the elements of the group of pure n-braids $\mathbf{PB}_n$.*

In Remark 1 of section 3, we give an alternative interpretation of the above result. For a discussion on the use of braid groups to classify motions on a graph, see [1].

The group $\mathbf{PB}_n$ is described by a set of generators together with a set of relations defined on them [32, 33]. Therefore, Theorem 2.5 completely characterizes the structure of the homotopy types of conflict-free maneuvers for n-agent encounters. On the other hand, the characterization is unsatisfactory in practical terms since the description of $\mathbf{PB}_n$ is very complicated. However, when n is small, the result in Theorem 2.5 may have simple interpretations. Consider, for example, the 2-agent encounter shown in Figure 2.2. Theorem 2.5 says that each maneuver in $\mathbf{P}(0, \mathbf{a}, \mathbf{b})$ is 0-homotopic to maneuver 1, or maneuver 2, or one of the following two maneuvers:

- Maneuver 1 followed by the motions where agent 2 stays at b_2 and agent 1 starts from b_1, circles around agent 2 counterclockwise k times for some integer $k \geq 1$, and returns to b_1.
- Maneuver 2 followed by the motions where agent 2 stays at b_2 and agent 1 starts from b_1, circles around agent 2 clockwise k times for some integer $k \geq 1$, and returns to b_1.

The angle that one agent turns with respect to the other during T plays a decisive role in determining the homotopy type of the conflict-free 2-maneuvers. Maneuver 1 and maneuver 2 are representatives of the only two types for which the absolute values of this angle do not exceed 360°. We shall call such types *fundamental*. Then there are exactly two fundamental types for any 2-agent encounter.

It is tempting to extend this definition to the n-agent case and conclude that there are exactly $2^{\frac{n(n-1)}{2}}$ fundamental types of conflict-free maneuvers, since there are two fundamental types for each of the $\frac{n(n-1)}{2}$ agent pairs. Unfortunately this is not the case. Shown in Figure 2.3 are the plots of two conflict-free maneuvers for a 3-agent

encounter that have the same turning angle within the range $(-360°, 360°)$ between any pair of agents, but in fact belong to different types.

3. Optimal conflict-free maneuvers. In this section, the problem of finding "optimal" conflict-free maneuvers for multi-agent encounters is formulated and studied. To ensure that the problem is well defined and admits a solution, we modify some of the notation introduced in the previous section. In particular, the set of maneuvers for agent i, $\mathbf{P}_i$, is redefined to be the set of all continuous *and piecewise* C^2 maps[1] from T to $\mathbb{R}^2$ that take the values a_i and b_i at times t_0 and t_f, respectively. The set of joint maneuvers $\mathbf{P}(\mathbf{a}, \mathbf{b})$ and the MSE $\Delta(\alpha)$, $\alpha \in \mathbf{P}(\mathbf{a}, \mathbf{b})$, are defined in section 2, whereas $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$ is redefined to be the set of all joint maneuvers with an MSE greater than *or equal to* R. Note that the results in section 2 on the qualitative classification of conflict-free maneuvers still hold for the newly defined $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$ by the compactness of T.

3.1. Statement of the problem. Consider a maneuver of a single agent, say $\alpha_i \in \mathbf{P}_i$, $i \in \{1, \ldots, n\}$. The *energy* of α_i is defined as

$$J(\alpha_i) = \frac{1}{2} \int_{t_0}^{t_f} ||\dot{\alpha}_i(t)||^2 dt. \tag{3.1}$$

Let $L(\alpha_i)$ be the arc length of the curve α_i, i.e., $L(\alpha_i) = \int_{t_0}^{t_f} ||\dot{\alpha}_i(t)|| dt$. Then the application of the Cauchy–Schwarz inequality to (3.1) yields [31]

$$J(\alpha_i) \geq \frac{1}{2} \frac{L(\alpha_i)^2}{(t_f - t_0)}, \tag{3.2}$$

where the equality holds if and only if $\|\dot{\alpha}_i(t)\|$ is constant. This implies that if agent i is forced to move along some fixed curve and if we ignore the presence of other agents temporarily, then of all the different parameterizations, the one with a constant speed has the minimal energy, and the minimal energy is proportional to the square of the curve length. Therefore, in the presence of *static* obstacles, the maneuver of agent i with the least energy between two points is the shortest curve between them parameterized proportionally to the arc length. In particular, if there are no obstacles, the energy-minimizing maneuver of agent i is the constant speed motion along the line segment from a_i to b_i. It follows from this discussion that the energy-minimizing maneuvers tend to be straighter and smoother, which has practical implications, for example, in terms of passenger comfort, brake erosion, fuel consumption, etc.

The *μ-energy* of a joint maneuver $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbf{P}(\mathbf{a}, \mathbf{b})$ is defined as

$$J_\mu(\alpha) \triangleq \sum_{i=1}^{n} \mu_i J(\alpha_i), \tag{3.3}$$

where $\mu_1, \ldots, \mu_n$ are n positive numbers adding up to 1 (i.e., $\sum_{i=1}^{n} \mu_i = 1$) representing the priorities of the agents.

Our goal is to find the conflict-free maneuver with the least μ-energy, i.e.,

$$\text{minimize } J_\mu(\alpha) \text{ subject to } \alpha \in \mathbf{P}(R, \mathbf{a}, \mathbf{b}). \tag{3.4}$$

[1] Piecewise C^2 means that there is a finite subdivision of T such that the map is continuously differentiable till the second order on each (open) subinterval. In what follows, when we use $\dot{\alpha}_i(t), \ddot{\alpha}_i(t)$, we shall mean at those t where they are well defined, i.e., except at a finite set of time instants t.

If α is required to belong to a certain type in $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$, then we get a restricted version of problem (3.4). All the necessary conditions obtained in this section remain valid for the restricted problem, with the only exception of Proposition 3.8.

Remark 1 (geodesics in a manifold with boundary). Problem (3.4) can be formulated in an alternative way. By viewing $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbf{P}(\mathbf{a}, \mathbf{b})$ as a curve in $\mathbb{R}^{2n}$, and $\mathbf{a}, \mathbf{b}$ as two points in $\mathbb{R}^{2n}$, a conflict-free maneuver in $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$ corresponds to a curve in $\mathbb{R}^{2n}$ joining $\mathbf{a}$ to $\mathbf{b}$ and avoiding the obstacle W defined by

$$W = \{(p_1, \ldots, p_n) \in \mathbb{R}^{2n} : p_i \in \mathbb{R}^2, 1 \le i \le n, \text{ and } \|p_j - p_k\| < R \text{ for some } j \ne k\}.$$

If the coefficients μ_i, $i = 1, \ldots, n$, are identical, then the μ-energy of a joint maneuver is proportional to the energy of the corresponding curve in $\mathbb{R}^{2n}$. Therefore, problem (3.4) is equivalent to finding the curve in $\mathbb{R}^{2n} \setminus W$ joining $\mathbf{a}$ to $\mathbf{b}$ with the least energy, which is a minimizing geodesic of $\mathbb{R}^{2n} \setminus W$ connecting $\mathbf{a}$ to $\mathbf{b}$. Note that $\mathbb{R}^{2n} \setminus W$ is a manifold with nonsmooth boundary whose fundamental group is isomorphic to $\mathbf{PB}_n$ by Theorem 2.5. The general case of arbitrary $\langle \mu_i \rangle_{i=1}^n$ can be reduced to this special case by scaling the p_i axes of $\mathbb{R}^{2n}$ by a factor of $\sqrt{\mu_i}$, $i = 1, \ldots, n$. The interested readers are referred to [17] for further details.

The rest of this section is devoted to the solution of problem (3.4), a variational problem with complicated and nonsmooth constraints. Inspired by the braid representation introduced in section 2, we propose various transformations of joint maneuvers that preserve the MSE and use these transformations in the variational analysis to obtain necessary conditions for a maneuver $\alpha^* \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$ to be optimal.

3.2. μ-alignment of optimal conflict-free maneuvers. As explained in section 2, each conflict-free maneuver $\alpha \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$ has a natural braid representation $\hat{\alpha} \in \mathbf{PB}(\mathbf{a}, \mathbf{b})$, whose n strings are determined by the images of the maps $t \mapsto (\alpha_i(t), t)$, $t \in T$, $i = 1, \ldots, n$. Furthermore, $\hat{\alpha}$ satisfies the *R-separation property* in that the intersection of $\hat{\alpha}$ with the plane $t = \tau$ for any $\tau \in T$ consists of n points whose pairwise minimum distance is at least R. All the operations on conflict-free maneuvers we shall introduce in the following preserve this separation property in the braid representation; hence they are indeed transformations of conflict-free maneuvers.

For each $w \in \mathbb{R}^2$, denote by $\mathbf{b} + w$ the n-tuple $(b_1 + w, \ldots, b_n + w)$.

DEFINITION 3.1 (tilt operator $\mathcal{T}_w$). *The tilt operator* $\mathcal{T}_w : \mathbf{P}(R, \mathbf{a}, \mathbf{b}) \to \mathbf{P}(R, \mathbf{a}, \mathbf{b} + w)$ *is a map such that for any* $\alpha \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$, $\beta = \mathcal{T}_w(\alpha)$ *is defined by*

$$\beta_i(t) = \alpha_i(t) + \frac{t - t_0}{t_f - t_0} w, \quad t \in T, \quad i = 1, \ldots, n.$$

It is easily seen that $\mathcal{T}_w$ is *MSE-preserving* in the sense that α and $\mathcal{T}_w(\alpha)$ have the same MSE. Hence $\mathcal{T}_w$ maps $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$ into $\mathbf{P}(R, \mathbf{a}, \mathbf{b} + w)$. In fact, $\mathcal{T}_w$ is a bijection from $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$ to $\mathbf{P}(R, \mathbf{a}, \mathbf{b} + w)$ since $\mathcal{T}_w \circ \mathcal{T}_{-w} = \mathcal{T}_{-w} \circ \mathcal{T}_w = \mathrm{id}$. In the braid representation, $\hat{\beta}$ is obtained by tilting $\hat{\alpha}$ linearly; hence the name for the operator $\mathcal{T}_w$. More precisely, in order to get $\hat{\beta}$ from $\hat{\alpha}$, the plane $t = t_0$ is kept invariant (shifted by 0), the plane $t = t_f$ is shifted by w, and each intermediate plane $t = \tau$, $\tau \in (t_0, t_f)$, is shifted by an amount determined by the linear interpolation of 0 and w according to the position of τ in T. Figure 3.1 illustrates the effect of the $\mathcal{T}_w$ operator on the braid representation of a 2-maneuver.

The importance of introducing $\mathcal{T}_w$ lies in the following result.

PROPOSITION 3.2. *Suppose that* α^* *is a conflict-free maneuver in* $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$ *with the least* μ*-energy. Fix* $w \in \mathbb{R}^2$. *Then* $\beta^* = \mathcal{T}_w(\alpha^*)$ *is a conflict-free maneuver in* $\mathbf{P}(R, \mathbf{a}, \mathbf{b} + w)$ *with the least* μ*-energy.*

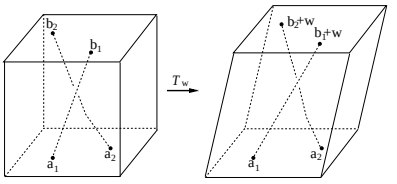


FIG. 3.1. *Tilt operation $\mathcal{T}_w$ on a 2-maneuver.*

Proof. For any $\beta \in \mathbf{P}(R, \mathbf{a}, \mathbf{b}+w)$, let $\alpha = \mathcal{T}_{-w}(\beta)$. Then $\alpha \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$ and $J_\mu(\beta)$ can be expressed as

$$\begin{aligned} J_\mu(\beta) &= \frac{1}{2}\int_{t_0}^{t_f} \sum_{i=1}^n \mu_i \|\dot\beta_i(t)\|^2 dt = \frac{1}{2}\int_{t_0}^{t_f} \sum_{i=1}^n \mu_i \left\|\dot\alpha_i(t) + \frac{w}{t_f - t_0}\right\|^2 dt \\ &= \frac{1}{2}\int_{t_0}^{t_f} \sum_{i=1}^n \mu_i \|\dot\alpha_i(t)\|^2 dt + \int_{t_0}^{t_f} \frac{w^T}{t_f - t_0} \sum_{i=1}^n \mu_i \dot\alpha_i(t) dt + \frac{\|w\|^2}{2(t_f - t_0)} \\ &= J_\mu(\alpha) + \frac{w^T[\sum_{i=1}^n \mu_i(b_i - a_i) + w/2]}{t_f - t_0}. \end{aligned} \tag{3.5}$$

Note that the second term in (3.5) is a constant independent of β. Denote it by C. It follows by (3.5) and the optimality of α^* that $J_\mu(\beta) \geq J_\mu(\alpha^*) + C$ for all $\beta \in \mathbf{P}(R, \mathbf{a}, \mathbf{b}+w)$, with the equality if $\alpha = \alpha^*$, i.e., $\beta = \beta^*$. □

Consider arbitrary starting and destination positions $\mathbf{a}$ and $\mathbf{b}$, and set $\mathbf{b}' \triangleq \mathbf{b} + w$ where $w = \sum_{i=1}^n \mu_i(a_i - b_i)$. Then $\mathbf{a}$ and $\mathbf{b}'$ are *μ-aligned* in the sense that they have the same μ-centroid, i.e.,

$$\sum_{i=1}^n \mu_i a_i = \sum_{i=1}^n \mu_i b_i'. \tag{3.6}$$

By Proposition 3.2, solutions to problem (3.4) for general $\mathbf{a}$ and $\mathbf{b}$ can be obtained from solutions to problem (3.4) for μ-aligned $\mathbf{a}$ and $\mathbf{b}'$ by applying the tilt operator $\mathcal{T}_{-w}$ with $w = \sum_{i=1}^n \mu_i(a_i - b_i)$. This is the reason why we shall focus on the special case of μ-aligned $\mathbf{a}$ and $\mathbf{b}$.

The next transformation we shall introduce is the *drift operation*. Let $\gamma : T \to \mathbb{R}^2$ be a continuous and piecewise C^2 map such that $\gamma(t_0) = \gamma(t_f) = 0$.

DEFINITION 3.3 (drift operator $\mathcal{D}_\gamma$). *The drift operator $\mathcal{D}_\gamma : \mathbf{P}(R, \mathbf{a}, \mathbf{b}) \to \mathbf{P}(R, \mathbf{a}, \mathbf{b})$ is a map such that for any $\alpha \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$, $\beta = \mathcal{D}_\gamma(\alpha)$ is defined by*

$$\beta_i(t) = \alpha_i(t) + \gamma(t), \quad t \in T, \quad i = 1, \ldots, n.$$

In the braid representation, $\hat\beta$ is obtained from $\hat\alpha$ by drifting each plane $t = \tau$, $\tau \in T$, by an offset $\gamma(\tau) \in \mathbb{R}^2$. It can be verified that $\mathcal{D}_\gamma$ is MSE-preserving and a bijection of $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$ onto itself since $\mathcal{D}_\gamma \circ \mathcal{D}_{-\gamma} = \mathcal{D}_{-\gamma} \circ \mathcal{D}_\gamma = \mathrm{id}$. By using the drift operator, we can prove the following result.

PROPOSITION 3.4. *Suppose that $\mathbf{a}$ and $\mathbf{b}$ are μ-aligned and $\alpha^* \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$ is an optimal solution to problem* (3.4). *Then*

$$\sum_{i=1}^n \mu_i \alpha_i^*(t) = \sum_{i=1}^n \mu_i a_i = \sum_{i=1}^n \mu_i b_i \qquad \forall t \in T.$$

Proof. For each $\lambda \in \mathbb{R}$ define $\beta_\lambda \triangleq \mathcal{D}_{\lambda\gamma}(\alpha^*)$. Note that $\beta_\lambda \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$ and $\beta_0 = \alpha^*$. Moreover,

$$\begin{aligned} J_\mu(\beta_\lambda) &= \frac{1}{2}\int_{t_0}^{t_f} \sum_{i=1}^{n} \mu_i \|\dot{\alpha}_i^*(t) + \lambda\dot{\gamma}(t)\|^2 dt \\ &= J_\mu(\alpha^*) + \frac{\lambda^2}{2}\int_{t_0}^{t_f} \|\dot{\gamma}(t)\|^2 dt + \lambda \int_{t_0}^{t_f} \dot{\gamma}(t)^T \sum_{i=1}^{n} \mu_i \dot{\alpha}_i^*(t) dt. \end{aligned}$$

The difference $J_\mu(\beta_\lambda) - J_\mu(\alpha^*)$ is a quadratic function of λ, which, by the optimality of α^*, must be nonnegative for all λ. Hence we have $\int_{t_0}^{t_f} \dot{\gamma}(t)^T \sum_{i=1}^{n} \mu_i \dot{\alpha}_i^*(t) dt = 0$, which must hold for any choice of γ such that $\gamma(t_0) = \gamma(t_f) = 0$. Since $\mathbf{a}$ and $\mathbf{b}$ are μ-aligned, we can choose $\gamma(t) = \sum_{i=1}^{n} \mu_i \alpha_i^*(t) - \sum_{i=1}^{n} \mu_i a_i$. Given that α^* is piecewise C^2, this leads to $\sum_{i=1}^{n} \mu_i \dot{\alpha}_i^*(t) = 0$ for almost all $t \in T$, and hence, by integration, to the desired conclusion. □

We can now use Proposition 3.2 to get the formulation of Proposition 3.4 for arbitrary $\mathbf{a}$ and $\mathbf{b}$.

COROLLARY 3.5. *Suppose that $\alpha^* \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$ is an optimal solution to problem* (3.4). *Then*

$$\sum_{i=1}^{n} \mu_i \alpha_i^*(t) = \sum_{i=1}^{n} \mu_i a_i + \frac{t - t_0}{t_f - t_0}\left(\sum_{i=1}^{n} \mu_i b_i - \sum_{i=1}^{n} \mu_i a_i\right) \qquad \forall t \in T.$$

In other words, the μ-centroid of $\langle \alpha_i^*(t) \rangle_{i=1}^n$ moves from the μ-centroid of $\mathbf{a}$ at time t_0 to the μ-centroid of $\mathbf{b}$ at time t_f with constant velocity.

Remark 2. The results in Proposition 3.2, Proposition 3.4, and Corollary 3.5 for conflict-free maneuvers in $\mathbb{R}^2$ are still valid when the underlying space is $\mathbb{R}^k$ with $k > 2$. These can be proved by following exactly the same procedure as in the $\mathbb{R}^2$ case.

Remark 3. A geometric interpretation of Corollary 3.5 can be given in the case when the μ_i's are identical. Let W be the obstacle in $\mathbb{R}^{2n}$ defined as in Remark 1. An important observation is that W is cylindrical in the direction of the two-dimensional subspace N spanned by vectors $(1, 0, 1, 0, \ldots, 1, 0)^T$ and $(0, 1, 0, 1, \ldots, 0, 1)^T$ in $\mathbb{R}^{2n}$, in the sense that for any $x \in \mathbb{R}^{2n}$, $x \in W$ if and only if $x + N \subset W$. Let V be the orthogonal complement of N in $\mathbb{R}^{2n}$. Then $\mathbf{a}$ and $\mathbf{b}$ are μ-aligned if and only if $\mathbf{a}$ and $\mathbf{b}$ are on the same V-slice in $\mathbb{R}^{2n}$, i.e., if and only if $\mathbf{a} - \mathbf{b} \in V$. The conclusions of Proposition 3.2 and Corollary 3.5 say that for $\mathbf{a}$ and $\mathbf{b}$ that are not necessarily μ-aligned, the shortest geodesic in $\mathbb{R}^{2n} \setminus W$ from $\mathbf{a}$ to $\mathbf{b}$ can be decomposed into two parts: its projection onto N, which is a constant speed motion along the straight line from $\pi_N(\mathbf{a})$ to $\pi_N(\mathbf{b})$, where $\pi_N : \mathbb{R}^{2n} \to N$ denotes the orthogonal projection map onto N; and its projection onto V, which is the shortest geodesic in $V \cap W^c$ connecting $\pi_V(\mathbf{a})$ and $\pi_V(\mathbf{b})$, where $\pi_V : \mathbb{R}^{2n} \to V$ denotes the orthogonal projection map onto V. Since V is of dimension $2n - 2$, this effectively reduces the dimension of the problem by 2.

3.3. Optimal conflict-free maneuvers for two agents. We now show that the solution to problem (3.4) in the case when there are only two agents follows directly from Corollary 3.5.

Assume that $\mathbf{a} = (a_1, a_2)$ and $\mathbf{b} = (b_1, b_2)$ are μ-aligned, and denote by c their common μ-centroid. If $\alpha^* = (\alpha_1^*, \alpha_2^*) \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$ is an optimal solution to problem

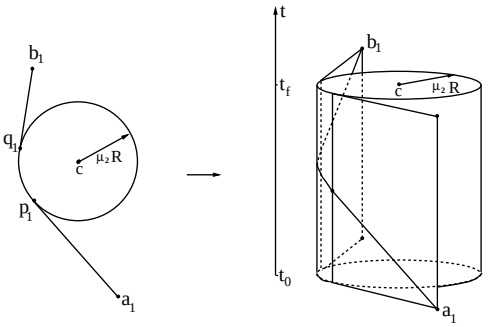


FIG. 3.2. *Optimal 2-maneuver and its braid representation.*

(3.4), then, by Proposition 3.4, the μ-centroid of $\alpha_1^*(t)$ and $\alpha_2^*(t)$ is equal to c for any $t \in T$, or equivalently,

$$\alpha_1^*(t) - c = -\frac{\mu_2}{\mu_1}(\alpha_2^*(t) - c) \qquad \forall t \in T. \tag{3.7}$$

From (3.7), it then follows that the energies of α_1^* and α_2^* are related by $\mu_1^2 J(\alpha_1^*) = \mu_2^2 J(\alpha_2^*)$ and that the separation constraint $\|\alpha_1^*(t) - \alpha_2^*(t)\| \geq R$ is equivalent to $\|\alpha_1^*(t) - c\| \geq \mu_2 R$. Therefore, problem (3.4) can be reduced to

$$\text{minimize } J(\alpha_1) \text{ subject to } \alpha_1 \in \mathbf{P}_1 \text{ and } \alpha_1 : T \to B^c(c, \mu_2 R), \tag{3.8}$$

where $B^c(c, \mu_2 R)$ denotes the complement in $\mathbb{R}^2$ of the open disk of radius $\mu_2 R$ centered at c. Thus the problem becomes finding the minimum energy maneuver for a single agent in the presence of the static obstacle $B(c, \mu_2 R)$.

By assumption, both a_1 and b_1 belong to $B^c(c, \mu_2 R)$ since otherwise the problem is infeasible. From the discussion at the beginning of section 3.1, we know that the optimal solution to problem (3.8) is a constant speed motion along the shortest curve joining a_1 to b_1 while avoiding the obstacle $B(c, \mu_2 R)$. Let ∂B be the boundary of the disk $B(c, \mu_2 R)$. The geometric construction of the shortest curve within a given fundamental type is shown in Figure 3.2. The curve is composed of three pieces: first from a_1 to $p_1 \in \partial B$ along a straight line tangent to ∂B, then from p_1 to q_1 along ∂B, and finally from q_1 to b_1 along another straight line tangent to ∂B. Here choosing a fundamental type is equivalent to choosing a side of the cylinder in the braid representation. The globally optimal solution α_1^* is the one of the two locally optimal solutions with shorter arc length (or any one of them if they have the same length). α_2^* is then obtained from α_1^* by (3.7). This is for the μ-aligned case. Denote by $\gamma_i^*(\mathbf{a}, \mathbf{b})$, $i = 1, 2$, the obtained optimal maneuvers. For the general case when $\mathbf{a}$ and $\mathbf{b}$ are not necessarily μ-aligned, we have by Proposition 3.2 the following theorem.

THEOREM 3.6 (optimal conflict-free 2-maneuver). *If $n = 2$, then the optimal solution $\alpha^* \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$ to problem* (3.4) *is given by*

$$\begin{cases} \alpha_1^*(t) = \gamma_1^*(\mathbf{a}, \mathbf{b} + w)(t) - \frac{t - t_0}{t_f - t_0} w, \\ \alpha_2^*(t) = \gamma_2^*(\mathbf{a}, \mathbf{b} + w)(t) - \frac{t - t_0}{t_f - t_0} w \end{cases} \qquad \forall t \in T, \tag{3.9}$$

where $w = \mu_1 a_1 - \mu_1 b_1 + \mu_2 a_2 - \mu_2 b_2$.

Consider the case when the priority of agent 1 is much higher than that of agent 2, which can be modeled by $\mu_2 \simeq 0$. In the μ-aligned case, this implies $a_1 \simeq b_1 \simeq c$

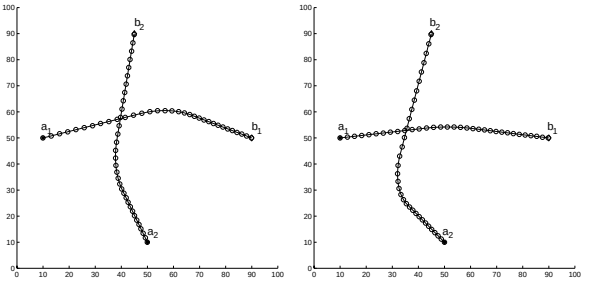


FIG. 3.3. *Optimal 2-maneuvers ($R = 30$). Left: $\mu_1 = \mu_2 = 0.5$. Right: $\mu_1 = 0.8$, $\mu_2 = 0.2$.*

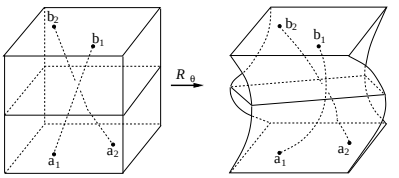


FIG. 3.4. *Twist operation $\mathcal{R}_\theta$ on a 2-maneuver.*

and that the radius of the disk $B(c, \mu_2 R)$ is about 0. Therefore, γ_1^* is nearly a zero motion. For general $\mathbf{a}$ and $\mathbf{b}$, it follows from Theorem 3.6 that the optimal maneuver for agent 1 is almost a constant speed motion along the line segment from a_1 to b_1. Hence, as expected, agent 2 is the one assuming most of the responsibility of avoiding conflicts.

Shown in Figure 3.3 are the plots of optimal conflict-free maneuvers for a typical 2-agent encounter with two different sets of priorities. The circles represent the positions of the two agents at evenly distributed time instants. The plots show that, in the case when $\mathbf{a}$ and $\mathbf{b}$ are not μ-aligned, the speeds of the agents in the optimal maneuvers are not constant. As the priority of agent 1 increases, however, its optimal maneuver gets closer to the constant speed motion along the straight line connecting a_1 to b_1.

3.4. Twist optimality. Another MSE-preserving operator can be introduced as follows. Suppose that $\theta : T \to \mathbb{R}$ is a continuous and piecewise C^2 map satisfying $\theta(t_0) = 0$, $\theta(t_f) = 2k\pi$ for some $k \in \mathbb{Z}$.

DEFINITION 3.7 (twist operator $\mathcal{R}_\theta$). *The twist operator $\mathcal{R}_\theta : \mathbf{P}(R, \mathbf{a}, \mathbf{b}) \to \mathbf{P}(R, \mathbf{a}, \mathbf{b})$ is a map such that for any $\alpha \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$, $\beta = \mathcal{R}_\theta(\alpha)$ is defined by*

$$\beta_i(t) = T_{\theta(t)}\alpha_i(t), \quad t \in T, \quad i = 1, \ldots, n,$$

where $T_{\theta(t)}$ is the matrix corresponding to a rotation of $\theta(t)$ counterclockwise:

$$T_{\theta(t)} = \begin{pmatrix} \cos[\theta(t)] & -\sin[\theta(t)] \\ \sin[\theta(t)] & \cos[\theta(t)] \end{pmatrix}.$$

The constraints on $\theta(t_0)$ and $\theta(t_f)$ ensure that $\mathcal{R}_\theta(\alpha)$ and α have the same starting and ending positions. It is easy to see that $\mathcal{R}_\theta$ is MSE-preserving and hence has its image in $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$. Figure 3.4 shows the effect of $\mathcal{R}_\theta$ ($k = 0$) on the braid representation of a 2-maneuver, which motivates the name "twist operator" for it.

By considering the perturbed maneuvers generated by $\mathcal{R}_\theta$, we have the following proposition.

PROPOSITION 3.8. *Suppose that* $\alpha^* \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$ *is an optimal solution to problem* (3.4). *Fix* $s \in \mathbb{R}^2$. *Then*

$$\frac{1}{2}\sum_{i=1}^{n} \mu_i (\alpha_i^*(t) - s)^T T_{-\frac{\pi}{2}} \dot{\alpha}_i^*(t) = C \qquad \forall t \in T, \tag{3.10}$$

where C *is a constant belonging to* $[-\frac{\pi}{z}, \frac{\pi}{z}]$, *with* $z \triangleq 2\int_{t_0}^{t_f} \left[\sum_{i=1}^{n} \mu_i \|\alpha_i^*(t) - s\|^2\right]^{-1} dt$.

Proof. See Appendix A. □

If $k \neq 0$, then the operator $\mathcal{R}_\theta$ changes the homotopy type of conflict-free maneuvers in $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$, thus enabling us to compare the performance of conflict-free maneuvers of different types. In this sense, the result in Proposition 3.8 is global. We illustrate this statement by the following example.

Example 1. Assume that $n = 2$ and $\mu_1 = \mu_2 = \frac{1}{2}$. Let $t_0 = 0$ and $t_f = \tau$ for some $\tau \in (0, 2\pi)$. Set $a_1 = \frac{R}{2}(1, 0)^T$, $b_1 = \frac{R}{2}(\cos\tau, \sin\tau)^T$, $a_2 = -a_1$, and $b_2 = -b_1$. Consider the conflict-free maneuvers α and β in $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$ defined by $\alpha_1(t) = \frac{R}{2}(\cos t, \sin t)^T$, $\alpha_2(t) = -\alpha_1(t)$, and $\beta_1(t) = \frac{R}{2}(\cos(\frac{\tau - 2\pi}{\tau}t), \sin(\frac{\tau-2\pi}{\tau}t))^T$, $\beta_2(t) = -\beta_1(t)$ for all $t \in [0, \tau]$. The two agents under maneuver α (β) rotate around the origin at constant angular velocity counterclockwise (clockwise) during $[0, \tau]$. Note that β can be obtained from α by applying the twist operator $\mathcal{R}_\theta$ with $\theta(t) = -2\pi t/\tau$ satisfying $\theta(\tau) = -2\pi$ and that α and β belong to different types. Since $\mathbf{a}$ and $\mathbf{b}$ are μ-aligned, the results in section 3.3 imply that α and β are the optimal solutions to problem (3.4) restricted to the two fundamental types. The global optimal solution is the one of them with smaller arc length, which can be easily seen to be α if $\tau \in (0, \pi)$ and β if $\tau \in (\pi, 2\pi)$. This conclusion can also be reached directly by an application of Proposition 3.8. In fact, if we choose $s = 0$ and compute C and z defined in Proposition 3.8 with α in the place of α^*, we get $C = R^2/8$ and $z = 8\tau/R^2$, and the inequality $|C| \leq \pi/z$ becomes $\tau \leq \pi$, which implies that α is not globally optimal for $\tau \in (\pi, 2\pi)$. If we compute C and z with β in the place of α^*, we get $C = R^2(\tau - 2\pi)/8\tau$ and $z = 8\tau/R^2$, and the inequality $|C| \leq \pi/z$ becomes $\tau \geq \pi$. Hence β is not globally optimal for $\tau \in (0, \pi)$.

Note that by choosing different $s \in \mathbb{R}^2$, Proposition 3.8 provides a family of inequalities of the form $-\frac{\pi}{z} \leq C \leq \frac{\pi}{z}$ that an optimal solution α^* to problem (3.4) must satisfy, where C and z are functions of s and α^*. In the case when $\mathbf{a}$ and $\mathbf{b}$ are μ-aligned, by Proposition 3.4, we have $\sum_{i=1}^{n} \mu_i s^T T_{-\frac{\pi}{2}} \dot{\alpha}_i^*(t) \equiv 0$. Hence the inequality becomes

$$\left| \frac{1}{2}\sum_{i=1}^{n} \mu_i \alpha_i^*(t)^T T_{-\frac{\pi}{2}} \dot{\alpha}_i^*(t) \right| \leq \frac{\pi}{2} \left\{ \int_{t_0}^{t_f} \left[\sum_{i=1}^{n} \mu_i \|\alpha_i^*(t) - s\|^2 \right]^{-1} dt \right\}^{-1}.$$

The most restrictive bound is obtained by setting s equal to the common μ-centroid of $\mathbf{a}$ and $\mathbf{b}$, which minimizes the right-hand side of the above equation. Moreover, one can derive further optimality conditions by applying Proposition 3.8 to $\mathcal{T}_w(\alpha^*)$ for any $w \in \mathbb{R}^2$, since by Proposition 3.2, $\mathcal{T}_w(\alpha^*)$ is optimal in $\mathbf{P}(R, \mathbf{a}, \mathbf{b} + w)$.

3.5. Analysis by partial operators. Further optimality conditions can be derived by considering those transformations that change the maneuvers of only a subset of the n agents (partial operators).

Let α be an arbitrary conflict-free maneuver in $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$. At each time $t \in T$, we can construct an undirected graph $G_\alpha(t)$ as follows: $G_\alpha(t)$ has n vertices, numbered from 1 to n, corresponding to the n agents, and an edge connects vertices i and j if

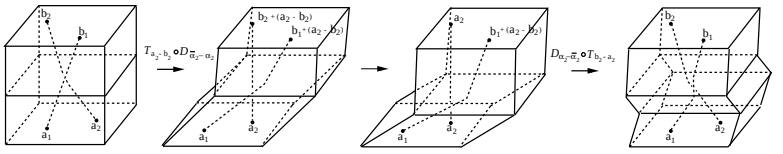


FIG. 3.5. *Slide operation* $\mathcal{L}_h^{12}$ *on braids.*

and only if $\|\alpha_i(t) - \alpha_j(t)\| = R$. If there does exist an edge between vertex i and vertex j in $G_\alpha(t)$, we say that agent i and agent j *contact* at time t. $G_\alpha(t)$ is then called the *contact graph* of α at time t.

We start from a very special case. Assume that α is a conflict-free maneuver in $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$ such that during the whole encounter the distance of agent 1 from any of the other agents is strictly greater than R except possibly from agent 2, i.e., $\|\alpha_1(t) - \alpha_i(t)\| > R$ for all $t \in T$, $i = 3, \ldots, n$. We shall introduce operators that leave α_i unchanged for $i = 2, 3, \ldots, n$ and perturb α_1 slightly, so that the perturbed α_1 has the same minimum distance from α_2 in the time interval T. If such a perturbation is small enough, then the perturbed α_1 does not cause a conflict between agent 1 and any of the agents with index $i \geq 3$, given that their original minimum distance in the time interval T was strictly greater than R.

Let $h : T \to T$ be a reparameterization of T, i.e., a bijection such that both h and h^{-1} are continuous and piecewise C^2, and $h(t_0) = t_0$ and $h(t_f) = t_f$.

DEFINITION 3.9 (partial slide operator $\mathcal{L}_h^{12}$). *The partial slide operator* $\mathcal{L}_h^{12} : \mathbf{P}(R, \mathbf{a}, \mathbf{b}) \to \mathbf{P}(\mathbf{a}, \mathbf{b})$ *is a map such that for any* $\alpha \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$, $\beta = \mathcal{L}_h^{12}(\alpha)$ *is defined by*

$$\begin{cases} \beta_1(t) = \alpha_1[h(t)] - \alpha_2[h(t)] + \alpha_2(t), & t \in T, \\ \beta_i(t) = \alpha_i(t), & t \in T, \quad i = 2, \ldots, n. \end{cases} \tag{3.11}$$

Note that $\inf_{t \in T} \|\beta_1(t) - \beta_2(t)\| = \inf_{t \in T} \|\alpha_1(t) - \alpha_2(t)\|$, and that for h sufficiently close to the identity map, the minimum distance in the time interval T between β_1 and β_i is greater than R for $i \geq 3$ by our assumption on α. These two conditions together imply that $\beta \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$.

Figure 3.5 shows how β is constructed geometrically. First, the operator $\mathcal{D}_{\bar{\alpha}_2 - \alpha_2}$ is performed on (α_1, α_2) to "straighten" the string corresponding to α_2, where $\bar{\alpha}_2$ denotes the constant velocity motion along the straight line between a_2 and b_2. Next, the operator $\mathcal{T}_{a_2 - b_2}$ is applied to the resulting 2-maneuver to get a 2-maneuver $\gamma = (\gamma_1, \gamma_2)$ with $\gamma_1 = \alpha_1 - \alpha_2 + a_2$ and $\gamma_2 \equiv a_2$. Then, γ is reparameterized by h to obtain $\eta = (\eta_1, \eta_2)$ with $\eta_1 = (\alpha_1 \circ h) - (\alpha_2 \circ h) + a_2$ and $\eta_2 \equiv a_2$. Finally, the reverse procedures of the second and first steps are applied subsequently to obtain (β_1, β_2) from η. Roughly speaking, $\hat{\beta}$ is obtained by "sliding" $\hat{\alpha}_1$ along $\hat{\alpha}_2$; hence the name "slide operator" for $\mathcal{L}_h^{12}$. Note that the superscript and the subscript in $\mathcal{L}_h^{12}$ indicate, respectively, the two strings the operator works on and the reparameterization used.

By using the partial slide operator to generate the perturbation in the variational analysis, we get the following proposition. (See [12] for the detailed proof.)

PROPOSITION 3.10. *Suppose that* $\alpha^* \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$ *is an optimal solution to problem* (3.4) *and that there exists a subinterval* $(t_0', t_f') \subset T$ *such that* $\|\alpha_1^*(t) - \alpha_i^*(t)\| > R$, $i = 3, \ldots, n$, *for all* $t \in (t_0', t_f')$. *Then* α^* *satisfies*

$$\ddot{\alpha}_1^*(t)^T (\dot{\alpha}_1^*(t) - \dot{\alpha}_2^*(t)) \equiv 0 \qquad \forall t \in (t_0', t_f'). \tag{3.12}$$

Instead of sliding α_1 along α_2, we can rotate it. Let $\theta : T \to \mathbb{R}$ be a continuous and piecewise C^2 map with $\theta(t_0) = \theta(t_f) = 0$.

DEFINITION 3.11 (partial rotation operator $\mathcal{R}^{12}_\theta$). *The partial rotation operator* $\mathcal{R}^{12}_\theta : \mathbf{P}(R, \mathbf{a}, \mathbf{b}) \to \mathbf{P}(\mathbf{a}, \mathbf{b})$ *is a map such that for any* $\alpha \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$, $\beta = \mathcal{R}^{12}_\theta(\alpha)$ *is defined by*

$$\begin{cases} \beta_1(t) = T_\theta(t)[\alpha_1(t) - \alpha_2(t)] + \alpha_2(t), & t \in T, \\ \beta_i(t) = \alpha_i(t), & t \in T, \quad i = 2, \dots, n. \end{cases}$$

In the braid representation, $\hat{\beta}$ is obtained by rotating the string $\hat{\alpha}_1$ around the string $\hat{\alpha}_2$. If θ is close enough to the zero map, $\beta = \mathcal{R}^{12}_\theta(\alpha) \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$. Similarly to the proof of Proposition 3.10, by using the partial rotation operator, we get the following proposition [12].

PROPOSITION 3.12. *Under the hypotheses of Proposition* 3.10, α^* *satisfies*

$$\ddot{\alpha}^*_1(t)^T T_{\frac{\pi}{2}}(\alpha^*_1(t) - \alpha^*_2(t)) \equiv 0 \qquad \forall t \in (t'_0, t'_f). \tag{3.13}$$

It can be verified that the optimal solution for the 2-agent case obtained in Theorem 3.6 indeed satisfies both conditions (3.12) and (3.13). Moreover, if one of the two agents has a predetermined maneuver throughout T, equations (3.12) and (3.13) will govern the motion of the other agent. Note also that if, in addition, $\|\alpha^*_1 - \alpha^*_2\| = R$ on (t'_0, t'_f), then these two equations are equivalent, since in this case $\|\alpha^*_1 - \alpha^*_2\|^2 \equiv R^2$ implies that $(\dot{\alpha}^*_1 - \dot{\alpha}^*_2)^T(\alpha^*_1 - \alpha^*_2) \equiv 0$, i.e., $\dot{\alpha}^*_1 - \dot{\alpha}^*_2$ and $T_{\frac{\pi}{2}}(\alpha^*_1 - \alpha^*_2)$ have the same direction. The intuitive understanding is that, in the braid representation, the slide and rotation operations of a string on the surface of a cylinder lead to the same orthogonal perturbation.

The above idea can be carried out even further. Suppose that the contact graph of an optimal maneuver $\alpha^* \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$ remains constant on some subinterval $(t'_0, t'_f) \subset T$. We can perturb α^* by sliding (rotating) slightly the maneuvers of a subset of the n agents with respect to that of agent i in the time subinterval (t'_0, t'_f). To ensure that the perturbed joint maneuver belongs to $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$, any agent in this subset should have a minimum distance strictly greater than R from any of the agents not belonging to the subset, except possibly from agent i, in the time interval (t'_0, t'_f). Since α^* is optimal, its μ-energy cannot be decreased by such a perturbation. By using the same arguments leading to Proposition 3.10 and Proposition 3.12, we then have the following proposition [12].

PROPOSITION 3.13. *Suppose that* $\alpha^* \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$ *is an optimal solution to problem* (3.4) *and that its contact graph remains constant on some subinterval* $(t'_0, t'_f) \subset T$. *Pick any agent, say, agent* i, *and let* $\mathcal{I} \subset \{1, 2, \dots, n\} \setminus \{i\}$ *be a subset of the remaining agents that corresponds to a maximal connected component of the graph obtained by removing node* i *and all the edges connected with it from the contact graph during* (t'_0, t'_f). *Then for all* $t \in (t'_0, t'_f)$,

$$\begin{aligned} &\sum_{j \in \mathcal{I}} \mu_j \ddot{\alpha}^*_j(t)^T(\dot{\alpha}^*_j(t) - \dot{\alpha}^*_i(t)) \equiv 0, \\ &\sum_{j \in \mathcal{I}} \mu_j \ddot{\alpha}^*_j(t)^T T_{\frac{\pi}{2}}(\alpha^*_j(t) - \alpha^*_i(t)) \equiv 0. \end{aligned} \tag{3.14}$$

Note that (3.12) and (3.13) are special cases of (3.14) when $i = 2$ and $\mathcal{I} = \{1\}$.

Proposition 3.13 is the most comprehensive optimality condition we have obtained so far. Next, in section 3.6 we will show by a simple example how it can be used (together with the global optimality conditions) to determine the optimal maneuver with a particular contact graph. This example will also serve as a counterexample to the conjecture that for each multi-agent encounter, there is a unique optimal conflict-free maneuver within each homotopy type, which is true for $n = 2$ by Theorem 3.6.

Remark 4. All the optimality conditions we have obtained so far admit mechanical interpretations, as will be shown in section 3.8. However, it should be pointed out that, in general, they cannot completely characterize the optimal maneuver with an arbitrary contact graph. A complete set of local optimality conditions can be derived by considering all possible local perturbations of maneuvers that preserve the contact graph, or in the light of Remark 1, by writing down the geodesics equation in a suitable Riemannian manifold.

3.6. An interesting example. Consider three agents with equal priorities $\mu_1 = \mu_2 = \mu_3 = \frac{1}{3}$ and $R = 1$. Suppose that α^* is an optimal conflict-free maneuver for some starting position $\mathbf{a} = (a_1, a_2, a_3)$ and destination position $\mathbf{b} = (b_1, b_2, b_3)$ that are μ-aligned with a common μ-centroid at the origin, and suppose that on some subinterval of T (which we may assume without loss of generality to be T itself), its contact graph $G_{\alpha^*}(t)$ is constant with edges between vertices 1 and 3 and between vertices 2 and 3 but no edges between vertices 1 and 2. Then, by Corollary 3.5 and Proposition 3.13, $\alpha^* = (\alpha_1^*, \alpha_2^*, \alpha_3^*)$ must satisfy for $t \in T$

$$\begin{cases} \sum_{i=1}^{3} \alpha_i^*(t) = 0, \\ \ddot{\alpha}_1^*(t)^T T_{\frac{\pi}{2}}(\alpha_1^*(t) - \alpha_3^*(t)) = 0, \\ \ddot{\alpha}_2^*(t)^T T_{\frac{\pi}{2}}(\alpha_2^*(t) - \alpha_3^*(t)) = 0, \\ \|\alpha_1^*(t) - \alpha_3^*(t)\| = \|\alpha_2^*(t) - \alpha_3^*(t)\| = 1. \end{cases} \tag{3.15}$$

We now show that (3.15) is equivalent to the geodesics equation of a suitable Riemannian manifold (a differential manifold together with a smoothly varying positive definite quadratic form on its tangent bundle [8]). Hence, for any set of initial conditions $\alpha_i^*(t_0), \dot{\alpha}_i^*(t_0)$, $i = 1, 2, 3$, it has a unique solution for t belonging to a neighborhood of t_0. First, notice that α^* as a curve in $\mathbb{R}^6$ lies in the submanifold Q of $\mathbb{R}^6$ determined by the first and the last equations of (3.15), namely the set of all those points $(x_1, y_1, x_2, y_2, x_3, y_3)$ in $\mathbb{R}^6$ such that $\sum_{i=1}^{3} x_i = \sum_{i=1}^{3} y_i = 0$, $(x_1 - x_3)^2 + (y_1 - y_3)^2 = 1$, and $(x_2 - x_3)^2 + (y_2 - y_3)^2 = 1$. Q is a compact two-dimensional submanifold of $\mathbb{R}^6$ and admits a global coordinate (θ_1, θ_2) defined by

$$\theta_1 = \arctan \frac{y_1 - y_3}{x_1 - x_3}, \qquad \theta_2 = \arctan \frac{y_2 - y_3}{x_2 - x_3}.$$

(θ_1, θ_2) takes values in the rectangle $[0, 2\pi] \times [0, 2\pi]$ with opposite edges identified, i.e., the 2-torus $\mathbb{T}^2$. In order to satisfy our assumption that the distance between agent 1 and agent 2 is greater than R during T, α^* must lie in an open subset Q_0 of Q consisting of all those points $(x_1, y_1, x_2, y_2, x_3, y_3)$ in Q such that $(x_1 - x_2)^2 + (y_1 - y_2)^2 > 1$. In the (θ_1, θ_2) coordinate, Q_0 corresponds to an open subset $\mathbb{T}_0^2$ of $\mathbb{T}^2$ obtained by removing from $\mathbb{T}^2$ the shaded region shown in Figure 3.6. Hence topologically Q_0 is homeomorphic to $S^1 \times (0, 1)$, an untwisted ribbon whose boundary consists of two disjoint circles.

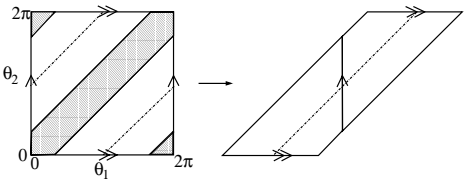


FIG. 3.6. $\mathbb{T}^2_0$ *as a subset of* $\mathbb{T}^2$ *in the* (θ_1, θ_2) *coordinate.*

Each $(\theta_1, \theta_2) \in \mathbb{T}^2$ determines a unique point $f(\theta_1, \theta_2)$ in Q by

$$f(\theta_1,\theta_2) = \frac{1}{3}(2\cos\theta_1 - \cos\theta_2, 2\sin\theta_1 - \sin\theta_2, -\cos\theta_1 + 2\cos\theta_2, \\ -\sin\theta_1 + 2\sin\theta_2, -\cos\theta_1 - \cos\theta_2, -\sin\theta_1 - \sin\theta_2)^T, \tag{3.16}$$

which is an embedding of $\mathbb{T}^2$ (respectively, $\mathbb{T}^2_0$) into $\mathbb{R}^6$ whose image is Q (respectively, Q_0).

By using f as the coordinate map, it can be verified that in the (θ_1, θ_2) coordinate, (3.15) is reduced to the following second order ODE:

$$\begin{cases} 2\ddot{\theta}_1 - \cos(\theta_1 - \theta_2)\ddot{\theta}_2 = \sin(\theta_1 - \theta_2)(\dot{\theta}_2)^2, \\ 2\ddot{\theta}_2 - \cos(\theta_1 - \theta_2)\ddot{\theta}_1 = -\sin(\theta_1 - \theta_2)(\dot{\theta}_1)^2. \end{cases} \tag{3.17}$$

Equation (3.17) is the geodesics equation of $\mathbb{T}^2$ with a suitably chosen metric g. In fact, let $\mathbb{R}^6$ be equipped with the standard Riemannian metric. Q as a submanifold inherits from $\mathbb{R}^6$ a metric by restriction. Let g be the corresponding metric on $\mathbb{T}^2$ obtained by pulling back the metric on Q via f, so that f becomes an isometry. Then, it can be proved (see Appendix B) that (3.17) is indeed the equation for geodesics of $\mathbb{T}^2$ under the metric g. As a result, each solution α^* of (3.15) is a geodesic of Q, which is not surprising by Remark 1. Since $\mathbb{T}^2$ (hence Q) is compact, a solution to (3.15) is defined for all duration of time, provided that it stays inside Q_0. Equation (3.17) can be solved by two integrals; see [13] for details.

Deeper optimality conditions of conflict-free maneuvers can be obtained in this interpretation. For example, it is computed in Appendix B that at each point $(\theta_1, \theta_2) \in \mathbb{T}^2$ the sectional curvature of the tangent plane spanned by the basis $\frac{\partial}{\partial\theta_1}$ and $\frac{\partial}{\partial\theta_2}$ at that point is

$$K(\theta_1,\theta_2) = \frac{-9\cos(\theta_1 - \theta_2)}{[4 - \cos^2(\theta_1 - \theta_2)]^2}. \tag{3.18}$$

Now consider the curve θ in $\mathbb{T}^2$ defined by $\theta(t) = (\theta_1(t), \theta_2(t)) = (t, \pi + t)$ for $t \in [0, \tau]$, where τ is positive. θ is a trivial solution to (3.17), hence a geodesic of $\mathbb{T}^2$ that is contained completely in $\mathbb{T}^2_0$. θ determines a 3-maneuver $\alpha^* = f \circ \theta$, i.e.,

$$\alpha_1^*(t) = (\cos t, \sin t)^T, \quad \alpha_2^*(t) = (-\cos t, -\sin t)^T, \quad \alpha_3^*(t) = (0,0)^T,\ t \in [0,\tau]. \tag{3.19}$$

In the motions specified by α^*, agent 3 stays at the origin, while agent 1 and agent 2 are at unit distance from agent 3 but on the opposite side of it so that three of them are always collinear, and both agent 1 and agent 2 rotate at the same constant angular velocity around agent 3. α^* thus defined is a solution to (3.15). An application of

Proposition 3.8 implies that α^* is no longer optimal if $\tau > \pi$, for otherwise a better maneuver can be obtained by rotating agent 1 and agent 2 the opposite way around agent 3. The following proposition improves this result.

PROPOSITION 3.14. *Maneuver α^* defined by* (3.19) *is not optimal if $\tau > \frac{\sqrt{2}}{2}\pi$.*

Proof. Since f is an isometry, we need only to prove that the geodesic θ is no longer distance-minimizing between its end points $\theta(0) = (0, \pi)$ and $\theta(\tau) = (\tau, \pi + \tau)$ once $\tau > \tau_0 = \frac{\sqrt{2}}{2}\pi$. To this end, it suffices to prove that $\theta(\tau_0)$ is a conjugate point of $\theta(0)$ along θ; in other words, there exists a nontrivial Jacobi field X along θ that vanishes at both $\theta(0)$ and $\theta(\tau_0)$ [20].

Define two vector fields along θ by $W_1 = \frac{\partial}{\partial\theta_1} + \frac{\partial}{\partial\theta_2}$ and $W_2 = \frac{\partial}{\partial\theta_1} - \frac{\partial}{\partial\theta_2}$. Then, it is easy to verify that W_1 and W_2 are orthogonal and that W_1 coincides with the velocity field $\dot\theta$ of the geodesic θ. Moreover, using the Christoffel symbols calculated in Appendix B, we conclude that $\nabla_{\dot\theta} W_2 \equiv 0$; hence, W_2 is parallel along θ.

A Jacobi field X along θ and orthogonal to $\dot\theta$ is necessarily of the form $X(t) = h(t)W_2(t)$ for some function h defined on $[0, \tau]$ and satisfies the Jacobi equation $\nabla_{\dot\theta}\nabla_{\dot\theta} X + R(\dot\theta, X)\dot\theta = 0$, where R is the curvature tensor of $\mathbb{T}^2$. Since $\nabla_{\dot\theta}\nabla_{\dot\theta} X = \ddot h W_2$ and $R(\dot\theta, X)\dot\theta$ are both orthogonal to $\dot\theta$, the Jacobi equation is equivalent to $\langle \ddot h W_2, W_2\rangle + \langle R(\dot\theta, hW_2)\dot\theta, W_2\rangle = 0$. By (3.18), the sectional curvature K of $\mathbb{T}^2$ along θ is constant 1. Using the relation $\langle R(\dot\theta, hW_2)\dot\theta, W_2\rangle = hK[\langle\dot\theta,\dot\theta\rangle\langle W_2, W_2\rangle - \langle\dot\theta, W_2\rangle^2]$, we have $\ddot h + 2h = 0$. A solution of h vanishing at 0 is $h(t) = \sin(\sqrt{2}t)$, so $X(t) = \sin(\sqrt{2}t)W_2(t)$ is a Jacobi field along θ vanishing at $t = 0$ and $t = \frac{\sqrt{2}}{2}\pi = \tau_0$. Therefore, $\theta(\tau_0)$ is a conjugate point of $\theta(0)$ along θ. □

A more intuitive way of obtaining the conclusion of Proposition 3.14 is through variational analysis of α^* using perturbations of the following form. Recall that $\theta(t) = (\theta_1(t), \theta_2(t)) = (t, \pi + t)$, $t \in T = [0, \tau]$, is the curve in $\mathbb{T}^2_0$ that α^* corresponds to. Let $\xi_1 : T \times (-\epsilon, \epsilon) \to \mathbb{R}$ be a *proper variation* of the map $\theta_1 : T \to \mathbb{R}$; i.e., ξ_1 is a smooth map such that $\xi_1(t, 0) = \theta_1(t)$, $\xi_1(0, s) = \theta_1(0)$, $\xi_1(\tau, s) = \theta_1(\tau)$ for $t \in T$ and $s \in (-\epsilon, \epsilon)$, where ϵ is a small positive number. Let $\xi_2 : T \times (-\epsilon, \epsilon) \to \mathbb{R}$ be a proper variation of the map $\theta_2 : T \to \mathbb{R}$. Consider joint maneuvers β_s defined in the (θ_1, θ_2) coordinate by $(\xi_1(\cdot, s), \xi_2(\cdot, s))$ for $s \in (-\epsilon, \epsilon)$, which all start from $\alpha^*(0)$ and end in $\alpha^*(\tau)$. In the braid representation, $\hat\beta_s$ is obtained from $\hat\alpha^*$ by rotating the strings $\hat\alpha_1^*$ and $\hat\alpha_2^*$ by certain angles with respect to the string $\hat\alpha_3^*$ and then realigning the three strings to the origin. β_s is conflict-free if the variations ξ_1 and ξ_2 are small enough. Then, a necessary condition for α^* to be optimal is that the μ-energy of β_s be minimized at $s = 0$ for all possible ξ_1 and ξ_2. After a lengthy calculation, this will lead to the conclusion of Proposition 3.14.

If we consider only conflict-free maneuvers with this particular contact graph, then it is proved in [17] that, after τ passes the critical value $\frac{\sqrt{2}}{2}\pi$, the optimal conflict-free maneuver from $\alpha^*(0)$ to $\alpha^*(\tau)$ bifurcates from α^* into two conflict-free maneuvers with identical energy. Shown in the first row of Figure 3.7 are the plots of α^* for some $\tau > \frac{\sqrt{2}}{2}\pi$. The middle column is its plot in the (θ_1, θ_2) coordinate, and the right column is its braid representation. In the second and third rows, we plot by numerical simulations the two bifurcated optimal conflict-free maneuvers with this contact graph, which in the (θ_1, θ_2) coordinate are the mirror image of each other with respect to the line $\theta_1 - \theta_2 = \frac{\pi}{2}$. For more details on the above claims and the general problem of conjugate points in manifolds with boundary, see [17].

One can also consider $n \geq 3$ agents with equal priorities, which are originally in a straight line with distance between successive agents being R and which rotate at a

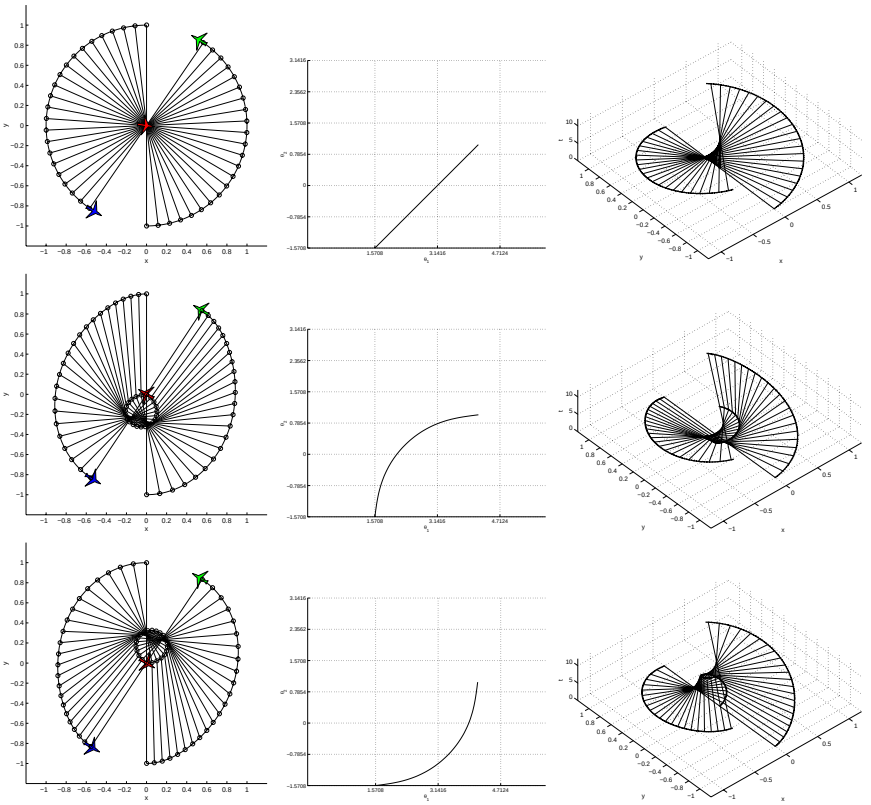


FIG. 3.7. *Bifurcation of minimizing geodesics in* $\mathbb{T}^2$. *Left column:* 3*-maneuvers. Middle column:* (θ_1, θ_2) *phase plots. Right column: braids.*

constant angular velocity around their centroid. This defines a geodesic in a certain submanifold of $\mathbb{R}^{2n}$ as we have discussed before. The maximal angle they can rotate before the first conjugate point of this geodesic is encountered in the submanifold is denoted by τ_n^*. It can be expected that τ_n^* decreases with n. We conjecture that $\tau_n^* = \pi/\sqrt{\frac{n(n-1)}{2} - 1}$. The case $n = 3$ is proved in Proposition 3.14. The cases $n = 4, 5, 6, 7$ are verified symbolically using MAPLE.

It is worthwhile at this point to summarize the optimality conditions we have derived so far. All of them, with the exception of Proposition 3.8 and Proposition 3.14, are *local* in the sense that they can be obtained by using spike-like perturbations in the variational analysis, which only change maneuvers in a neighborhood of a fixed time epoch. Proposition 3.14 is *semiglobal* in that its conclusion can be reached only by perturbations that change maneuvers throughout a subinterval of the encounter with positive length. Proposition 3.8 is the only global one, in the sense that it enables us to compare the performance of maneuvers belonging to different homotopy types.

3.7. Regularity of optimal conflict-free maneuvers. The regularity of optimal conflict-free maneuvers is a tricky issue. For example, it is unknown whether for each optimal α^* there exists a *finite* subdivision of T, $t_0 < t_1 < \cdots < t_m = t_f$, such that the contact graph $G_{\alpha^*}(t)$ remains constant during each subinterval (t_k, t_{k+1}) and contiguous subintervals correspond to different contact graphs. It is proved in [2]

that, in a Euclidean space under the presence of open obstacles with locally analytic boundary, a geodesic can have, in any segment of finite arc length, only a finite number of *switch points* where it switches from an interior segment to a segment on the boundary of an obstacle or vice versa. Unfortunately, this result does not apply in our case, since the obstacle W as defined in Remark 1 has nonsmooth boundary.

On the other hand, it can be proved that an optimal α^* is always C^1; i.e., there are no sharp turns in the optimal conflict-free maneuvers. In fact, this follows from a general result proved in [13], which states that if a manifold M with (nonsmooth) boundary is a subset of $\mathbb{R}^k$ obtained by removing from $\mathbb{R}^k$ a finite union of open convex subsets, each of which has a smooth boundary, then any geodesic of M is of class C^1. Note that the convex subsets are not required to be disjoint for this conclusion. In our case, by Remark 1, the obstacle is the union of $\frac{n(n-1)}{2}$ convex cylinders in $\mathbb{R}^{2n}$.

3.8. Two mechanical analogies. We now give two mechanical analogies of the above results. It should be pointed out that they serve only as analogies to gain more insights into the results obtained, and are not rigorous proofs themselves.

First, consider the following experiment. Instead of n agents, we have n particles of mass $\mu_1, \ldots, \mu_n$ on a horizontal plane with no external forces acting on them. At time t_0, they are at the initial positions $a_1, \ldots, a_n$ with certain initial velocities. Each particle i moves with constant velocity until the distance between it and some other particle j becomes R. Then a rigid rod of zero mass is introduced between particle i and particle j to prevent their distance from further decreasing, and the two particles move together with the rod at velocities determined by the law of conservation of momentum and angular momentum. We refer to the above process where a rigid rod is introduced between two particles as a (two-particle) *join*. There are two types of joins: *tangential* and *nontangential*. A join is tangential if the time derivative of the distance between the two particles at the time of join is zero, otherwise the join is nontangential. It is evident that some kinetic energy is lost for a nontangential join since there is a collision between the two particles along the direction of the rod. As time goes on, more particles can join to form larger groups. In addition to joins, a group of particles connected by rods can *split* at any time, in the sense that some or all of the rods disappear instantly at that time. So when a split occurs, neither the positions nor the velocities of the particles change, but the group separates into several independent subgroups.

It is claimed that by appropriately choosing the initial velocities, time, and order of the joins and splits, one can get from such an experiment the optimal maneuver α^*. In fact, during any time interval I in which there are neither joins nor splits, the system of particles naturally corresponds to a contact graph with edges between vertices representing rods between particles. Moreover, if I is sufficiently small, the motions of the particles correspond to the optimal conflict-free maneuver associated with such a contact graph. To see this, recall that by the principle of least action [3], the motion of the interconnected particles system is an extremal of the action integral $\int_I (E - U)dt$. Here $E = \frac{1}{2}\sum_{i=1}^n \mu_i v_i^2$ represents the kinetic energy, and U is the potential, which is zero by our assumption on the absence of external forces. So, for a sufficiently small time interval I, the motions of the interconnected particles minimize $\frac{1}{2}\sum_{i=1}^n \mu_i \int_I v_i^2 dt$; hence they specify precisely the optimal maneuver over I by definition. Equation (3.15) determines, for example, the motions of three particles connected by two rigid rods with zero masses. For discussions on the general problem of kinematically coupled structures composed of rigid and flexible bodies, see [22] and other references in the same book.

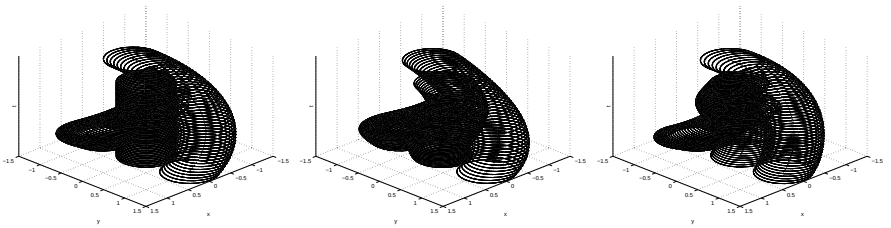


FIG. 3.8. *Examples of elastic (enlarged) braids in equilibrium positions. Left: unstable. Center and right: stable.*

In this mechanical interpretation, the conclusion of Proposition 3.2 is simply the invariance of the motions of a mechanical system with respect to changes of inertial coordinates. Since the total momentum and the total angular momentum of the system are conserved in each time interval with constant configuration (contact graph) and do not change during joins or splits, they are constant during the whole time interval T, which are the conclusions of Corollary 3.5 and Proposition 3.8, respectively. Proposition 3.8 further imposes an upper bound on the total angular momentum, implying that the whole system cannot spin "too fast." In addition, the assertion in section 3.7 that α^* is C^1 implies that all the joins should be tangential; hence there is no kinetic energy lost during joints and splits and the total kinetic energy $\frac{1}{2}\sum_{i=1}^{n}\mu_i v_i^2$ is also conserved, as it is shown in [12] by using a reparameterization operator.

In mechanics, there is a systematic way of using symmetry on the configuration space to reduce the degree of freedom [3, 26]. In our case, the symmetry is $\mathbf{SE}_2$, the group of rigid motions in $\mathbb{R}^2$, acting on $\mathbb{R}^2$. Hence the analysis leading to Corollary 3.5 and Proposition 3.8 (except the bound in Proposition 3.8) is simply the application of the symmetry reduction method uniformly to all the configuration spaces of a system with time-varying configurations. Compared with more advanced techniques such as those based on the Hamiltonian, symplectic, and Poisson viewpoints, our approach, which is Lagrangian in nature, deals with the nonsmoothness of the boundary constraints directly, thus avoiding the trouble of solving for each smooth component of the boundary constraints individually before piecing them together properly to get the final solution. In [16], the corresponding method is generalized to an arbitrary Riemannian manifold with a group of isometries. For application of Lagrangian reduction to holonomic and nonholonomic mechanical systems, see [21].

A major drawback of the above mechanical model is that it is local; hence little insight can be obtained about the global optimality conditions. In this sense, the second model we are going to present is more "faithful" and once again demonstrates the advantage of adopting the braid point of view. As we have shown in section 2, each conflict-free maneuver α of the n agents corresponds to an n-braid $\hat{\alpha}$, whose intersection with any *horizontal* plane $t = \tau$ ($\tau \in T$) consists of n points satisfying the R-separation property. Therefore, if we enlarge the radius of strings in $\hat{\alpha}$ to $R/2$, or more precisely if we think of each of the n strings in $\hat{\alpha}$ as consisting of an infinite number of horizontal disks of radius $R/2$ and height 0 mounting vertically, with each disk confined to move in a fixed horizontal plane $t = t_1$ for some $t_1 \in T$, then the condition that α is conflict-free is equivalent to that the n enlarged strings in $\hat{\alpha}$ do not overlap. Examples of such enlarged braids are shown in Figure 3.8 for the three conflict-free maneuvers in Figure 3.7.

Assume that, for each $i = 1, \ldots, n$, the enlarged string $\hat{\alpha}_i$ in $\hat{\alpha}$ is elastic with

elasticity coefficient μ_i and has a smooth surface so that any two strings can slide along each other without frictions. Under these assumptions, the elastic energy of this n-string system is proportional to the μ-energy of the corresponding conflict-free maneuver. If we fix the strings in $\hat{\alpha}$ at both the bottom ($t = t_0$) and the top ($t = t_f$) horizontal planes and leave free the remaining parts, then for certain choices of α this elastic n-string system will be in an equilibrium (stationary) position. The optimal conflict-free maneuvers have minimal energy, hence necessarily correspond to equilibrium positions.

Suppose that $\hat{\alpha}$ is in an equilibrium position. Pick any disk in $\hat{\alpha}$ that belongs to the string $\hat{\alpha}_i$ and lies on the horizontal plane $t = t_1$ for some $i = 1, \ldots, n$ and $t_0 < t_1 < t_f$. Denote this disk by $D_i(t_1)$. Then $D_i(t_1)$ is subject to two types of forces: forces enacted by disks in the same string that are immediately above and below $D_i(t_1)$, i.e., $D_i(t_1^+)$ and $D_i(t_1^-)$; and forces enacted by disks in the same horizontal plane $t = t_1$ but belonging to different strings, i.e., $D_j(t_1)$ with $j \neq i$. Since $D_i(t_1)$ is confined to move on the plane $t = t_1$, we are concerned with only the projection of the forces onto this plane. The contribution of the forces of the first type is easily seen to be proportional to $\mu_i \ddot{\alpha}_i(t_1)$. As for the forces of the second type, say, the force enacted by disk $D_j(t_1)$ ($j \neq i$) that contacts $D_i(t_1)$, by our assumption of no frictions this force is directed from the center of $D_j(t_1)$ to the center of $D_i(t_1)$, i.e., from $(\alpha_j(t_1), t_1)$ to $(\alpha_i(t_1), t_1)$. Now the conclusion of Proposition 3.13 can be explained as follows. Let $\mathcal{I}$ be a subset of $\{1, \ldots, n\} \setminus \{i\}$ that corresponds to a maximal connected component of the graph obtained by removing node i and all the edges connected with it from the contact graph of α at time t_1. Since $\hat{\alpha}$ is in an equilibrium position, the subsystem $D_{\mathcal{I}}(t_1)$ consisting of disks $D_j(t_1)$ for $j \in \mathcal{I}$ is stationary. So the total moment (torque) of external forces acting on $D_{\mathcal{I}}(t_1)$ is zero, which is exactly the conclusion of Proposition 3.13. Note that here we choose $(\alpha_i(t_1), t_1)$ as the origin and use the fact that torque of forces enacted by $D_i(t_1)$ on disks in $D_{\mathcal{I}}(t_1)$ is zero by our above analysis.

Other optimality conditions can also be explained in this model. For example, the conclusion of Corollary 3.5 is, after differentiation with respect to t twice, simply that on any horizontal plane $t = t_1$, $t_0 < t_1 < t_f$, the combined external forces acting on the subsystem consisting of disks $D_i(t_1)$, $i = 1, \ldots, n$, is zero. For the example in section 3.6, the semiglobal conclusion of Proposition 3.14 can be intuitively understood as that, after a rotation of more than $\frac{\pi}{\sqrt{2}}$, the cumulative force of the two neighboring strings on the central one exceeds the critical value so that the equilibrium position of $\hat{\alpha}^*$ becomes unstable. Any slight perturbation will then render the system to settle in one of the two bifurcated positions with minimal elastic energy (see Figure 3.8), provided that there exists very small but nonzero air frictions to avoid persistent oscillation.

4. Optimal multilegged conflict-free maneuvers. Due to the difficulty in computing analytically the optimal conflict-free maneuver when the number n of agents is greater than two, we now restrict our attention to those maneuvers specified by a set of waypoints, which might well be the only feasible form of joint maneuvers that a central controller can specify to the participating agents in practice.

To be precise, consider n agents with starting position $\mathbf{a} = (a_1, \ldots, a_n)$ and destination position $\mathbf{b} = (b_1, \ldots, b_n)$. Assume that a set of epochs $\{t_j\}_{j=0}^m$, $t_0 < t_1 < \cdots < t_{m-1} < t_m = t_f$, where m is a positive integer, has been fixed. For each agent i, choose a set of waypoints $\{c_{i,j}\}_{j=0}^m$ in $\mathbb{R}^2$ such that $c_{i,0} = a_i$ and $c_{i,m} = b_i$. Then, an *m-legged maneuver* of agent i is a maneuver consisting of m stages, where at each

stage $j \in \{0, 1, \ldots, m-1\}$ agent i starts from $c_{i,j}$ at time t_j and reaches $c_{i,j+1}$ at time t_{j+1} with constant velocity. Denote by $\mathbf{P}_i^m$ the set of all m-legged maneuvers of agent i and by $\mathbf{P}^m(\mathbf{a}, \mathbf{b}) = \prod_{i=1}^n \mathbf{P}_i^m$ the set of all m-legged joint maneuvers. In the braid representation, an m-legged joint maneuver corresponds to n strings, each one consisting of m line segments pieced together. The set of *m-legged conflict-free maneuvers* consists of all elements of $\mathbf{P}^m(\mathbf{a}, \mathbf{b})$ with MSE at least R and is denoted by $\mathbf{P}^m(R, \mathbf{a}, \mathbf{b})$.

In this section, we shall try to solve the following version of problem (3.4):

$$\text{minimize } J_\mu(\alpha) \text{ subject to } \alpha \in \mathbf{P}^m(R, \mathbf{a}, \mathbf{b}). \tag{4.1}$$

By using similar arguments, one can show that some of the optimality conditions in section 3, such as Corollary 3.5, still apply for solutions to problem (4.1). In general, a solution to problem (4.1) is only suboptimal for problem (3.4).

4.1. Optimal 2-legged conflict-free maneuver for two agents. We start from the simplest case when $n = 2$ and $m = 2$. Consider two agents with starting position $\mathbf{a} = (a_1, a_2)$ and destination position $\mathbf{b} = (b_1, b_2)$. Let $\alpha = (\alpha_1, \alpha_2)$ be a 2-legged conflict-free maneuver in $\mathbf{P}^2(R, \mathbf{a}, \mathbf{b})$ with three waypoints $c_{i,j}$, $j = 0, 1, 2$, for each agent $i = 1, 2$. Since $c_{i,0} = a_i$ and $c_{i,2} = b_i$ are fixed for each agent i, the middle waypoints $c_{i,1}$ will be denoted by c_i to simplify the notation. Let $t_c \in (t_0, t_f)$ be the epoch corresponding to the middle waypoints. Then, the motions of the two agents are described by

$$\alpha_i(t) = \begin{cases} a_i + (c_i - a_i)\frac{t-t_0}{t_c-t_0}, & t_0 \le t \le t_c, \\ b_i + (c_i - b_i)\frac{t-t_f}{t_c-t_f}, & t_c \le t \le t_f, \end{cases} \quad i = 1, 2.$$

After some calculations, the μ-energy of a maneuver $\alpha \in \mathbf{P}^2(\mathbf{a}, \mathbf{b})$ as the function of c_1 and c_2 can be expressed as follows:

$$J_\mu(\alpha) = \frac{t_f - t_0}{(t_f - t_c)(t_c - t_0)}[\mu_1 \|c_1 - c_1^u\|^2 + \mu_2 \|c_2 - c_2^u\|^2] + C, \tag{4.2}$$

where C is a constant and c_i^u, $i = 1, 2$, are defined by

$$c_i^u = \frac{(t_f - t_c)a_i + (t_c - t_0)b_i}{t_f - t_0}, \quad i = 1, 2. \tag{4.3}$$

Note that c_1^u and c_2^u are the optimal waypoints when minimizing $J_\mu(\alpha)$ without the MSE constraint. In the braid representation, c_1^u and c_2^u correspond to the intersections of the plane $t = t_c$ with the lines joining (a_i, t_0) to (b_i, t_f) for $i = 1$ and 2, respectively.

The MSE constraint can be simplified as well. The minimal distance d_l between the two agents during the time interval $[t_0, t_c]$ is given by

$$d_l = \begin{cases} \|c_1 - c_2\| & \text{if } \lambda < -\|c_1 - c_2 - a_1 + a_2\|^2, \\ \sqrt{\|a_1 - a_2\|^2 - \lambda^2/\|c_1 - c_2 - a_1 + a_2\|^2} & \text{if } -\|c_1 - c_2 - a_1 + a_2\|^2 \le \lambda \le 0, \\ \|a_1 - a_2\| & \text{if } \lambda > 0, \end{cases}$$

where $\lambda \triangleq (a_1 - a_2)^T(c_1 - c_2 - a_1 + a_2)$. Note that d_l is a function of the relative positions $a_1 - a_2$ and $c_1 - c_2$ only and is independent of the epoch t_c. We then use $d_l(a_1 - a_2, c_1 - c_2)$ to denote it explicitly. Similarly, the minimum distance between the two agents during the time interval $[t_c, t_f]$ is $d_l(c_1 - c_2, b_1 - b_2)$.

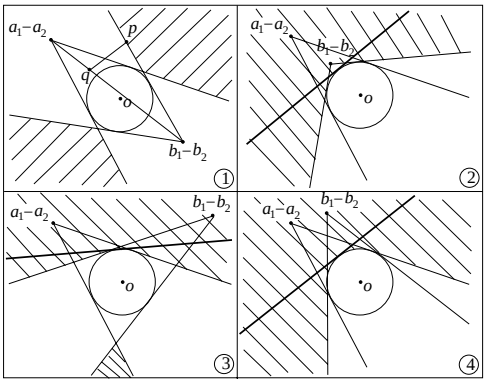


FIG. 4.1. *The four configurations of the feasible set A for $c_1 - c_2$.*

For α to be a conflict-free maneuver, both $d_l(a_1 - a_2, c_1 - c_2)$ and $d_l(c_1 - c_2, b_1 - b_2)$ have to be at least R, yielding two constraints on $c_1 - c_2$. Depending on the relative position of $a_1 - a_2$ and $b_1 - b_2$, the feasible set A for $c_1 - c_2$ has four possible configurations, which are numbered from 1 to 4 and represented by shaded regions in Figure 4.1. Notice that A consists of two connected components in configurations 1 and 3, which correspond to the two fundamental types of conflict-free maneuvers. In configurations 2 and 4, however, only one fundamental type can be achieved by 2-legged maneuvers.

Remark 5. The feasible set A for $c_1 - c_2$ can be characterized as the subset of $\mathbb{R}^2$ consisting of all those points that are "visible" to both $a_1 - a_2$ and $b_1 - b_2$ in the presence of the open disk $B(0, R)$ as an obstacle. In fact, by applying an appropriate tilt operator $\mathcal{T}_w$ that preserves the MSE and $c_1 - c_2$, one can assume that $c_2 = a_2$, i.e., agent 2 stays at a_2 during $[t_0, t_c]$. Thus the MSE constraint during $[t_0, t_c]$ is equivalent to the constraint that the line segment from a_1 to c_1 does not intersect $B(a_2, R)$, or alternatively, the line segment from $a_1 - a_2$ to $c_1 - c_2$ does not intersect $B(0, R)$. Similar arguments apply to the second stage of α.

As a result of the above simplifications, problem (4.1) is reduced to

$$\text{minimize } \mu_1\|c_1 - c_1^u\|^2 + \mu_2\|c_2 - c_2^u\|^2 \text{ subject to } c_1 - c_2 \in A. \tag{4.4}$$

THEOREM 4.1. *Define $q \triangleq c_1^u - c_2^u = \frac{t_f - t_c}{t_f - t_0}(a_1 - a_2) + \frac{t_c - t_0}{t_f - t_0}(b_1 - b_2)$. Let p be a point in A at minimum distance from q. An optimal solution to problem* (4.4) *is then given by*

$$c_1^* = \mu_1 c_1^u + \mu_2 c_2^u + \mu_2 p, \quad c_2^* = \mu_1 c_1^u + \mu_2 c_2^u - \mu_1 p.$$

Moreover, if problem (4.4) *is restricted to one of the two fundamental types of conflict-free maneuvers that is achievable by 2-legged maneuvers, then c_1^* and c_2^* are unique.*

Proof. Set $\Delta c = c_1 - c_2$. Then we have

$$\begin{aligned}
&\min\{\mu_1\|c_1 - c_1^u\|^2 + \mu_2\|c_2 - c_2^u\|^2 : c_1, c_2 \text{ such that } \Delta c \in A\} \\
&\quad = \min_{\Delta c \in A} \min_{c_2} \{\mu_1\|c_2 + \Delta c - c_1^u\|^2 + \mu_2\|c_2 - c_2^u\|^2\} \\
&\quad = \min_{\Delta c \in A} \min_{c_2} \{\|c_2 - \mu_1(c_1^u - \Delta c) - \mu_2 c_2^u\|^2 + \mu_1\mu_2\|c_1^u - c_2^u - \Delta c\|^2\} \\
&\quad = \min_{\Delta c \in A} \mu_1\mu_2\|q - \Delta c\|^2 \\
&\quad = \mu_1\mu_2\|q - p\|^2,
\end{aligned}$$

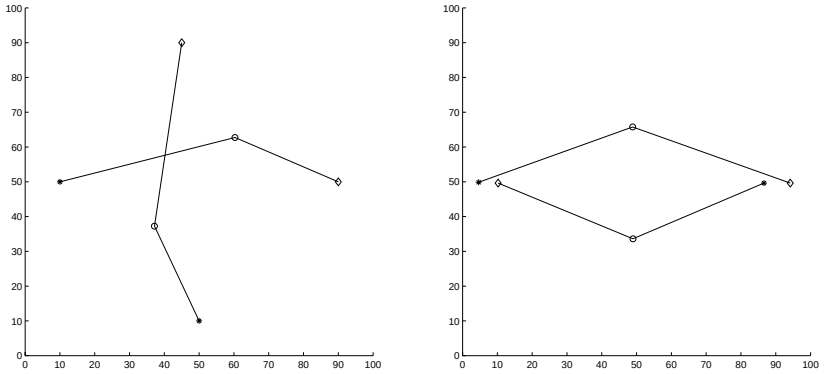


FIG. 4.2. 2-*legged optimal conflict-free maneuvers for* 2-*agent encounters* ($\mu_1 = \mu_2 = 0.5$, $R = 30$).

where the last two equalities follow by choosing $c_2 = \mu_1(c_1^u - \Delta c) + \mu_2 c_2^u$ and $\Delta c = p$. Together they imply the desired expressions of c_1^* and c_2^*. The uniqueness of c_1^* and c_2^* given a particular fundamental type is a consequence of the fact that p is unique, since either the connected component of A corresponding to that type is convex, or q is contained in it since it lies on the line segment connecting $a_1 - a_2$ to $b_1 - b_2$. □

Note that in configurations 2, 3, and 4, $p = q$ since q lies on the line segment connecting $a_1 - a_2$ and $b_1 - b_2$ that is contained entirely in A. Hence c_1^* and c_2^* are equal to c_1^u and c_2^u, respectively. In configuration 1, the set A is the union of two disjoint convex sets, so there might be up to two points in A nearest to q, with two being the case when there is an exact collision for the unconstrained optimal joint maneuver. In this case, we can choose either of the two points as p.

Figure 4.2 shows the optimal 2-legged conflict-free maneuvers for some typical 2-agent encounters when the agents have equal priorities. In each plot, the starting points are marked with stars and the ending points with diamonds. The circles are the waypoints specified by Theorem 4.1.

4.2. Optimal 2-legged conflict-free maneuver for multiple agents. Consider the case $m = 2$ and $n \geq 3$. Roughly speaking, the nature of problem (4.1) is mainly combinatorial in that the major task is to choose the type of conflict-free maneuvers in which one can find the optimal solution. In this section, we deal only with the problem of finding the optimal conflict-free maneuver within a *given* type. We postpone to section 4.4 the discussion on how to choose the maneuver type.

Fix $t_c \in (t_0, t_f)$ and denote by A_{ij} the feasible set for $c_i - c_j$ when only the agent pair (i, j) is present. A_{ij} is computed as set A in the last subsection with a_i, b_i, a_j, b_j in the place of a_1, b_1, a_2, b_2. Suppose that we have chosen a type of conflict-free maneuver. Then, the problem is to find the waypoints $c_1, \ldots, c_n$ that

$$\text{minimize } \sum_{i=1}^{n} \mu_i \|c_i - c_i^u\|^2 \text{ subject to } c_i - c_j \in A_{ij}^{\pm}, \quad 1 \leq i < j \leq n, \tag{4.5}$$

where c_i^u is defined as in (4.3) for $i = 1, \ldots, n$, and $A_{ij}^{\pm}$ denotes the connected component of the set A_{ij} matching the desired type. Note that only a finite subset of types of conflict-free maneuvers can be represented in this way, and we assume that the given type belongs to this subset.

Notice that in all but the first configuration shown in Figure 4.1 representing

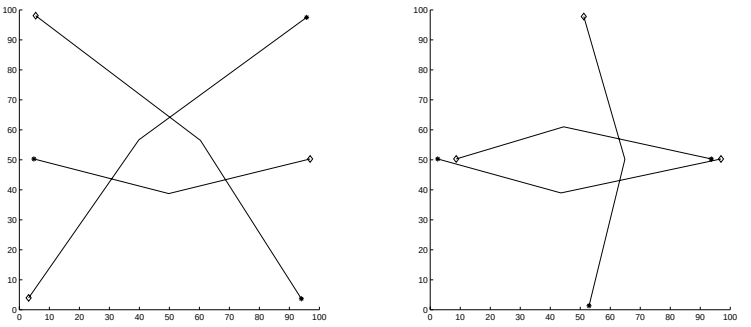


FIG. 4.3. *Globally optimal* 2*-legged conflict-free* 3*-maneuvers* ($\mu_1 = \mu_2 = 0.5, R = 20$).

A_{ij} for $i = 1$ and $j = 2$, one of the connected components of A_{ij} is nonconvex, posing a great challenge for the efficient solution of problem (4.5). Therefore, in configurations 2, 3, and 4, we linearize the nonconvex component of A_{ij} by using a half-space inner approximation, as it is shown in Figure 4.1 by the black lines tangential to the boundary of $B(0, R)$. The choice of the black line may not be unique, and one should ensure that the inner approximated feasible region of $c_i - c_j$ contains the unconstrained optimal value $c_i^u - c_j^u$.

Remark 6. Problem (4.5) is a linearly constrained convex optimization problem in the special case when any pair of agents is in the first configuration, i.e., when the unconstrained optimal joint maneuver will cause a conflict between any pair of agents. Therefore, our linear approximation scheme is tight for the most critical encounters.

After the linearization, if necessary, we have a linearly constrained quadratic optimization problem that can be solved efficiently. In the case when the number of agents is relatively small, we can afford the luxury of running the optimization algorithm for each type achievable by 2-legged maneuvers so as to find the globally optimal 2-legged conflict-free maneuver. Simulation results using MATLAB are shown in Figure 4.3 for two 3-agent encounters. In both cases, each pair of agents is in the first configuration, so linearizations are not necessary and the obtained maneuvers are actually the globally optimal 2-legged conflict-free maneuvers.

4.3. Optimal m-legged conflict-free maneuver for multiple agents. The algorithm described in section 4.2 can be used in an iterative way in the general case when the number m of legs is greater than two. Fix a set of epochs $t_0 < t_1 < \cdots < t_{m-1} < t_m = t_f$. A necessary condition for a set of waypoints $c_{i,j}$, $i = 1, \ldots, n$, $j = 0, \ldots, m$, with $c_{i,0} = a_i$, $c_{i,m} = b_i$, to be an optimal solution to problem (4.1) is that

$$c_{i,j} = c_i^*((c_{1,j-1}, \ldots, c_{n,j-1}), (c_{1,j+1}, \ldots, c_{n,j+1}), t_{j-1}, t_j, t_{j+1}) \tag{4.6}$$

for $1 \le j \le m - 1$. Here $c_i^*((c_{1,j-1}, \ldots, c_{n,j-1}), (c_{1,j+1}, \ldots, c_{n,j+1}), t_{j-1}, t_j, t_{j+1})$ denotes the waypoint of agent i for the optimal 2-legged maneuver when the starting and destination positions of the agents are $(c_{1,j-1}, \ldots, c_{n,j-1})$ and $(c_{1,j+1}, \ldots, c_{n,j+1})$, and the starting, middle, and ending epochs are t_{j-1}, t_j, t_{j+1}, respectively. This condition inspires the following algorithm.

ALGORITHM 1.

1. *Pick any feasible set of waypoints* $c_{i,j}^{(0)}$, $1 \le i \le n$, $0 \le j \le m$, *such that* $c_{i,0}^{(0)} = a_i, c_{i,m}^{(0)} = b_i$ *for* $1 \le i \le n$ *and such that the MSE constraint is*

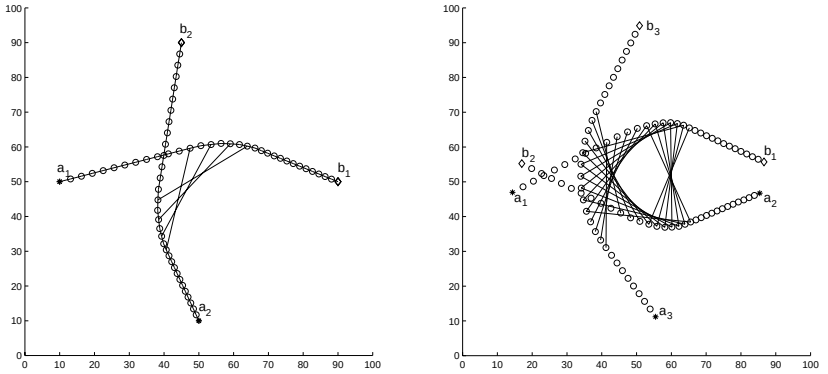


FIG. 4.4. *Simulation results of Algorithm* 1 *for two and three agents encounters* ($R = 30$).

satisfied over T.

2. *For* $j = 1, \ldots, m-1$ *compute for* $i = 1, \ldots, n$

$$c_{i,j}^{(k+1)} = c_i^*((c_{1,j-1}^{(k)}, \ldots, c_{n,j-1}^{(k)}), (c_{1,j+1}^{(k)}, \ldots, c_{n,j+1}^{(k)}), t_{j-1}, t_j, t_{j+1}).$$

3. *Repeat procedure* 2 *with* $k := k + 1$ *until the decrease in* μ*-energy is below some threshold* ε.

It is easily seen that the μ-energy of the conflict-free maneuvers obtained by Algorithm 1 is nonincreasing as a function of the iteration number k and is strictly decreasing whenever condition (4.6) is not satisfied. Therefore, the iteration procedure converges asymptotically to a conflict-free maneuver satisfying condition (4.6). A convergence analysis of Algorithm 1 is yet to be achieved. Besides the issue of local minima suggested by the example in section 3.6, the situation is further complicated by the fact that the convex optimization procedure introduced in section 4.2 yields only an approximation of c_i^*. Another open issue is the suboptimality of optimal m-legged maneuvers in $\mathbf{P}^m(R, \mathbf{a}, \mathbf{b})$ with respect to optimal solutions in $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$. Although in theory the performance gap decreases to zero as $m \to \infty$, in practice, it is not easy to quantify the performance degradation for a finite m.

In Figure 4.4, some simulation results for Algorithm 1 when the agents have identical priorities and $R = 30$ are shown. The epochs are chosen to evenly divide $[t_0, t_f]$, and the corresponding waypoints are marked with small circles. In the plots, whenever two agents are at distance R, their positions are joined by a line segment. Note that the result shown in the left figure is a good approximation to the optimal maneuvers plotted in Figure 3.3.

4.4. Randomized optimization. In [12, 34], a decentralized algorithm for multi-agent conflict resolution is proposed in the context of ATC. By modeling the agent motion as a Brownian motion with drift, the probability of conflict between two agents is estimated and then used to generate repulsive forces between the agents, inspired by the potential and vortex field methodology for path planning [27, 36]. Compared with traditional potential field methods that use only the positions of the agents, this algorithm considers also their headings and speeds, and hence generates maneuvers with less abrupt turns.

Although the stochastic algorithm can be run in real time regardless of the number n of agents involved, one of its drawbacks is that absolute safety cannot be guaranteed with probability one. On the other hand, the convex optimization algorithm we

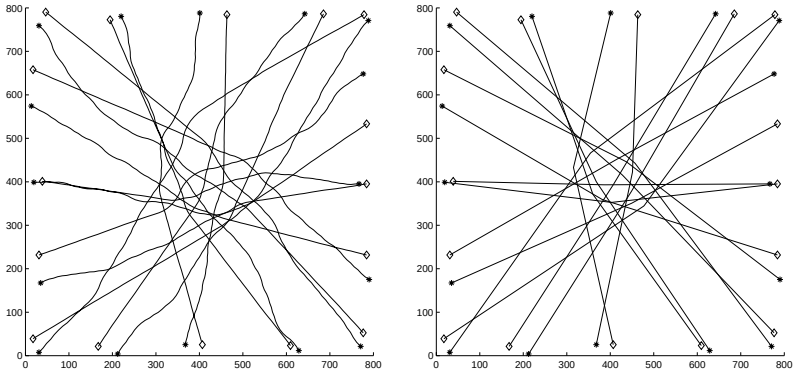


FIG. 4.5. 16-*maneuvers generated by stochastic (left) and convex optimization algorithms (right).*

propose in this paper can ensure absolute safety, but it cannot handle the explosively increasing number of types when n is large. We then suggest a solution that combines the positive features of these two algorithms; hence it both guarantees safety and is computationally feasible. The proposed algorithm uses the stochastic algorithm as the random "type chooser." More specifically, for a given multi-agent encounter, first the stochastic algorithm is run to generate a joint maneuver corresponding to a particular type, and then the convex optimization algorithm is utilized to obtain an approximation of the optimal multilegged maneuver within the type selected by the stochastic algorithm.

Simulation results for a 16-agent symmetric encounter are shown in Figure 4.5, in which 16 agents with identical priorities pass approximatively through a common point at angles evenly distributed in $[0, 2\pi]$ and $R = 30$. The one on the left is the joint maneuver generated by the stochastic algorithm, whereas the one on the right is the optimal 2-legged conflict-free maneuver within this type generated by the convex optimization algorithm.

Remark 7. When the number of agents is small, say, $n = 2, 3$, experiments show that the stochastic algorithm tends to choose with higher probability those types with lower μ-energy. However, when n is large, such as in the previous example, it is hard to evaluate the performance of the randomized algorithm, since currently no theoretical result exists that can exhaust the explosively increasing number of resolution types and find the optimal one (or ones). Much more work is needed in this respect.

5. Conclusions and future directions. In this paper, the problem of optimal coordinated motion planning for multiple agents moving on a plane is studied. After a classification of the homotopy types of conflict-free maneuvers, a weighted energy is proposed as the cost function to select the optimal one. Various local and global optimality conditions are derived. For 2-agent encounters, analytical solutions are obtained both for the optimal continuous and piecewise-C^2 maneuvers and the optimal 2-legged maneuvers. For the general multi-agent case, a randomized convex optimization algorithm is proposed to find the optimal multilegged maneuvers numerically.

To completely characterize the optimal conflict-free maneuvers, many issues remain to be solved. The results in this paper could serve as a good starting point. Possible directions of future research include the analysis of the proposed numerical algorithm in terms of its performance and its robustness with respect to uncertainty

on the agents' positions and velocities and the study of more realistic (and more complicated) models for the agent dynamics than the kinematic one adopted in this paper. Some contribution in this direction can be found in [15], which focuses exclusively on air traffic management systems.

Appendix A. Proof of Proposition 3.8. Consider first the case when $s = 0$. For each $\alpha \in \mathbf{P}(R, \mathbf{a}, \mathbf{b})$, let $\beta = \mathcal{R}^0_\theta(\alpha)$. Then

$$\dot{\beta}_i(t) = T_{\theta(t)}\dot{\alpha}_i(t) + \frac{d}{dt}T_{\theta(t)}\alpha_i(t) = T_{\theta(t)}\dot{\alpha}_i(t) + \dot{\theta}(t)T_{\frac{\pi}{2}+\theta(t)}\alpha_i(t), \quad i = 1, \dots, n.$$

Since $T_{\theta(t)}$ and $T_{\frac{\pi}{2}+\theta(t)}$ are orthonormal matrices and $T^T_{\frac{\pi}{2}+\theta(t)} = T_{-\frac{\pi}{2}-\theta(t)}$, we have

$$\|\dot{\beta}_i(t)\|^2 = \|\dot{\alpha}_i(t)\|^2 + \|\alpha_i(t)\|^2|\dot{\theta}(t)|^2 + 2\dot{\theta}(t)\alpha_i^T(t)T_{-\frac{\pi}{2}}\dot{\alpha}_i(t), \quad i = 1, \dots, n.$$

Integrating and summing over i, we can write the cost difference $\Delta J_\mu(\theta)$ as

$$\Delta J_\mu(\theta) = J_\mu(\beta) - J_\mu(\alpha) = \int_{t_0}^{t_f} [f(t)|\dot{\theta}(t)|^2 + 2g(t)\dot{\theta}(t)]dt, \tag{A.1}$$

where f and g are functions defined by

$$f(t) \triangleq \frac{1}{2}\sum_{i=1}^{n} \mu_i \|\alpha_i(t)\|^2, \quad g(t) \triangleq \frac{1}{2}\sum_{i=1}^{n} \mu_i \alpha_i^T(t)T_{-\frac{\pi}{2}}\dot{\alpha}_i(t) \qquad \forall t \in T. \tag{A.2}$$

Note that we use the notation $\Delta J_\mu(\theta)$ to indicate that it is a function of θ. We next compute the optimal twist θ^* such that $\Delta J_\mu(\theta)$ is minimized. θ is subject to the constraint that $\theta(t_0) = 0$, $\theta(t_f) = 2k\pi$ for some *fixed* $k \in \mathbb{Z}$. For $\dot{\theta}$, this translates into $\int_{t_0}^{t_f} \dot{\theta}(t)dt = 2k\pi$. We can then write the Lagrangian function for this problem as

$$\begin{aligned}\mathcal{L}(\theta, \lambda) &\triangleq \Delta J_\mu(\theta) + \lambda\left[\int_{t_0}^{t_f} \dot{\theta}(t)dt - 2k\pi\right] \\ &= \int_{t_0}^{t_f} \left\{ f(t)\left[\dot{\theta}(t) + \frac{g(t) + \frac{\lambda}{2}}{f(t)}\right]^2 - \frac{[g(t) + \frac{\lambda}{2}]^2}{f(t)} \right\} dt - 2\lambda k\pi.\end{aligned}$$

Thus $\dot{\theta}^*(t) = -[g(t) + \lambda^*/2]/f(t)$ where, since $\int_{t_0}^{t_f} \dot{\theta}(t)dt = 2k\pi$, λ^* is given by

$$\lambda^* = -2\left[\int_{t_0}^{t_f} \frac{g(t)}{f(t)}dt + 2k\pi\right] \Big/ \int_{t_0}^{t_f} \frac{1}{f(t)}dt.$$

Then, we have the following expression for $\dot{\theta}^*(t)$:

$$\dot{\theta}^*(t) = -\frac{g(t)}{f(t)} + \left[\int_{t_0}^{t_f} \frac{g(t)}{f(t)}dt + 2k\pi\right] \Big/ \left[f(t)\int_{t_0}^{t_f} \frac{1}{f(t)}dt\right].$$

Substituting this into (A.1), we get the minimal $\Delta J_\mu(\theta)$:

$$\Delta J_\mu(\theta^*) = \left[\int_{t_0}^{t_f} \frac{g(t)}{f(t)}dt + 2k\pi\right]^2 \Big/ \int_{t_0}^{t_f} \frac{1}{f(t)}dt - \int_{t_0}^{t_f} \frac{g^2(t)}{f(t)}dt.$$

If $\alpha = \alpha^*$ is an optimal maneuver, then $\Delta J_\mu(\theta^*) \geq 0$. Hence,

$$\left[\int_{t_0}^{t_f} \frac{g(t)}{f(t)}dt + 2k\pi\right]^2 \geq \int_{t_0}^{t_f} \frac{1}{f(t)}dt \cdot \int_{t_0}^{t_f} \frac{g^2(t)}{f(t)}dt. \tag{A.3}$$

In the case when $k = 0$, the equality holds in (A.3) since the lower bound $\Delta J_\mu(\theta^*) \geq 0$ can be strictly achieved by choosing $\theta^*(t) \equiv 0$. Therefore,

$$\left[\int_{t_0}^{t_f} \frac{g(t)}{f(t)}dt\right]^2 = \int_{t_0}^{t_f} \frac{1}{f(t)}dt \cdot \int_{t_0}^{t_f} \frac{g^2(t)}{f(t)}dt.$$

Applying the Cauchy–Schwarz inequality to functions $1/\sqrt{f(t)}$ and $g(t)/\sqrt{f(t)}$, we have that the above equality holds if and only if $g(t)/\sqrt{f(t)} = C/\sqrt{f(t)}$ for some constant C, i.e., if and only if $g(t) \equiv C$. In this case, (A.3) degenerates into

$$(Cz + 2k\pi)^2 \geq C^2 z^2 \qquad \forall k \in \mathbb{Z},$$

where $z = \int_{t_0}^{t_f} 1/f(t)dt$, or equivalently $k\pi Cz + k^2\pi^2 \geq 0$ for all $k \in \mathbb{Z}$. This is possible if and only if $-\pi \leq Cz \leq \pi$, thus completing the proof for the case $s = 0$.

The general case when $s \neq 0$ can be reduced to the above case by first noticing that the optimality of α^* in $\mathbf{P}(R, \mathbf{a}, \mathbf{b})$ implies the optimality of $\alpha^* - s = (\alpha_1^* - s, \ldots, \alpha_n^* - s)$ in $\mathbf{P}(R, \mathbf{a} - s, \mathbf{b} - s)$ and then applying the results proved for the case $s = 0$ to the optimal maneuver $\alpha^* - s$.

Appendix B. Geometry of $\mathbb{T}^2$ under metric g. In section 3.6, we define a Riemannian metric g on the 2-torus $\mathbb{T}^2$. Here we will derive some useful quantities characterizing its geometry.

At each point $(\theta_1, \theta_2) \in \mathbb{T}^2$, a basis $\frac{\partial}{\partial\theta_1}$ and $\frac{\partial}{\partial\theta_2}$ of the tangent space of $\mathbb{T}^2$ is mapped by the differential of the coordinate map f defined in (3.16) to

$$\begin{cases} df(\frac{\partial}{\partial\theta_1}) = \frac{1}{3}(-2\sin\theta_1, 2\cos\theta_1, \sin\theta_1, -\cos\theta_1, \sin\theta_1, -\cos\theta_1)^T, \\ df(\frac{\partial}{\partial\theta_2}) = \frac{1}{3}(\sin\theta_2, -\cos\theta_2, -2\sin\theta_2, 2\cos\theta_2, \sin\theta_2, -\cos\theta_2)^T, \end{cases} \tag{B.1}$$

which is a basis of the tangent space of Q at $f(\theta_1, \theta_2)$. Here we have identified the tangent space of $\mathbb{R}^6$ at $f(\theta_1, \theta_2)$ with $\mathbb{R}^6$ itself, and the tangent space of Q at $f(\theta_1, \theta_2)$ becomes a subspace of $\mathbb{R}^6$. The standard metric of $\mathbb{R}^6$ induces by f isometrically the metric g on $\mathbb{T}^2$ of the form

$$g = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 2 & -\cos(\theta_1 - \theta_2) \\ -\cos(\theta_1 - \theta_2) & 2 \end{bmatrix}, \tag{B.2}$$

where $g_{ij} \triangleq \langle \frac{\partial}{\partial\theta_i}, \frac{\partial}{\partial\theta_j} \rangle$ for $i = 1, 2$, $j = 1, 2$. The inverse of g can be written as

$$g^{-1} = \begin{bmatrix} g^{11} & g^{12} \\ g^{21} & g^{22} \end{bmatrix} = \frac{3}{4 - \cos^2(\theta_1 - \theta_2)}\begin{bmatrix} 2 & \cos(\theta_1 - \theta_2) \\ \cos(\theta_1 - \theta_2) & 2 \end{bmatrix}.$$

The covariant derivative ∇ of $\mathbb{T}^2$ with respect to the Levi–Civita connection is defined by [8]

$$\nabla_{\frac{\partial}{\partial\theta_i}} \frac{\partial}{\partial\theta_j} = \sum_{m=1}^{2} \Gamma_{ij}^m \frac{\partial}{\partial\theta_m} \qquad \forall 1 \leq i, j \leq 2,$$

where Γ_{ij}^m, $1 \le i, j, m \le 2$, are the *Christoffel symbols* that can be computed by

$$\Gamma_{ij}^m = \frac{1}{2}\sum_{k=1}^{2}\left\{\frac{\partial g_{jk}}{\partial \xi_i} + \frac{\partial g_{ki}}{\partial \xi_j} - \frac{\partial g_{ij}}{\partial \xi_k}\right\} g^{km}, \qquad 1 \le i, j, m \le 2.$$

It is easy to verify that

$$\Gamma_{11}^1 = -\Gamma_{22}^2 = \frac{\sin(\theta_1-\theta_2)\cos(\theta_1-\theta_2)}{4-\cos^2(\theta_1-\theta_2)}, \qquad \Gamma_{11}^2 = -\Gamma_{22}^1 = \frac{2\sin(\theta_1-\theta_2)}{4-\cos^2(\theta_1-\theta_2)},$$

and $\Gamma_{12}^m = \Gamma_{21}^m = 0$ for $m = 1, 2$. The equations for geodesics in $\mathbb{T}^2$ are $\ddot{\xi}_k + \sum_{i,j}\Gamma_{ij}^k \dot{\xi}_i \dot{\xi}_j = 0$, $k = 1, 2$, which yield

$$\begin{aligned}
[4-\cos^2(\theta_1-\theta_2)]\ddot{\theta}_1 &= -\sin(\theta_1-\theta_2)\cos(\theta_1-\theta_2)(\dot{\theta}_1)^2 + 2\sin(\theta_1-\theta_2)(\dot{\theta}_2)^2,\\
[4-\cos^2(\theta_1-\theta_2)]\ddot{\theta}_2 &= -2\sin(\theta_1-\theta_2)(\dot{\theta}_1)^2 + \sin(\theta_1-\theta_2)\cos(\theta_1-\theta_2)(\dot{\theta}_2)^2.
\end{aligned}$$

The above equations are readily seen to be equivalent to (3.17).

Next, we will compute the curvature of $\mathbb{T}^2$. Let R be the curvature tensor of $\mathbb{T}^2$. Let R_{ijkl} be its value in basis $\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}$ defined by [8]

$$\begin{aligned}
R_{ijkl} &\triangleq \left\langle R\left(\frac{\partial}{\partial\theta_i}, \frac{\partial}{\partial\theta_j}\right)\frac{\partial}{\partial\theta_k}, \frac{\partial}{\partial\theta_l}\right\rangle\\
&= \left\langle \left(\nabla_{\frac{\partial}{\partial\theta_j}}\nabla_{\frac{\partial}{\partial\theta_i}} - \nabla_{\frac{\partial}{\partial\theta_i}}\nabla_{\frac{\partial}{\partial\theta_j}} + \nabla_{[\frac{\partial}{\partial\theta_i}, \frac{\partial}{\partial\theta_j}]}\right)\frac{\partial}{\partial\theta_k}, \frac{\partial}{\partial\theta_l}\right\rangle
\end{aligned}$$

for all $1 \le i, j, k, l \le 2$. Then $R_{ijkl} = \sum_{s=1}^{2} R_{ijk}^s g_{sl}$, where R_{ijk}^s can be computed by

$$R_{ijk}^s = \sum_{m=1}^{2}\Gamma_{ik}^m\Gamma_{jm}^s - \sum_{m=1}^{2}\Gamma_{jk}^m\Gamma_{im}^s + \frac{\partial}{\partial\theta_j}\Gamma_{ik}^s - \frac{\partial}{\partial\theta_i}\Gamma_{jk}^s.$$

In our case, calculation shows that

$$R_{121}^1 = R_{122}^2 = \frac{-3\cos^2(\theta_1-\theta_2)}{[4-\cos^2(\theta_1-\theta_2)]^2}, \qquad R_{121}^2 = R_{122}^1 = \frac{-6\cos(\theta_1-\theta_2)}{[4-\cos^2(\theta_1-\theta_2)]^2},$$

and $R_{21k}^s = -R_{12k}^s$, $R_{11k}^s = R_{22k}^s = 0$ for all $1 \le k, s \le 2$. Hence,

$$R_{1212} = \frac{-\cos(\theta_1-\theta_2)}{4-\cos^2(\theta_1-\theta_2)}.$$

Therefore, the sectional curvature of $\mathbb{T}^2$ is

$$K = \frac{R_{1212}}{g_{11}g_{22} - g_{12}^2} = \frac{-9\cos(\theta_1-\theta_2)}{[4-\cos^2(\theta_1-\theta_2)]^2}. \tag{B.3}$$

K depends only on $\theta_1 - \theta_2$ since the map $(\theta_1, \theta_2) \mapsto (\theta_1 + \xi, \theta_2 + \xi) \bmod 2\pi$ is an isometry of $\mathbb{T}^2$ for each ξ. In the special case when $\theta_1 - \theta_2 = \pi$, we have $K = 1$. For further analysis, see [17].

Acknowledgment. The first author would like to thank Alan Weinstein for his insightful comments and discussions on the results of this paper.

REFERENCES

[1] A. Abrams, *Configuration Spaces and Braid Groups of Graphs*, Ph.D. thesis, University of California, Berkeley, CA, 2000.
[2] F. Albrecht and I. Berg, *Geodesics in Euclidean space with analytic obstacle*, Proc. Amer. Math. Soc., 113 (1991), pp. 201–207.
[3] V. I. Arnold, K. Vogtmann, and A. Weinstein, *Mathematical Methods of Classical Mechanics*, 2nd ed., Springer-Verlag, New York, 1989.
[4] B. Aronov, M. de Berg, A. van der Stappen, P. Svestka, and J. Vleugels, *Motion planning for multiple robots*, Discrete Comput. Geom., 22 (1999), pp. 502–525.
[5] A. Bicchi and L. Pallottino, *On optimal cooperative conflict resolution for air traffic management systems*, IEEE Trans. Intelligent Transport. Systems, 1 (2000), pp. 221–231.
[6] J. Birman, *Braids, Links, and Mapping Class Groups*, Princeton University Press, Princeton, NJ, 1974.
[7] J. P. Desai and V. Kumar, *Nonholonomic motion planning for multiple mobile manipulators*, in Proceedings of the IEEE International Conference on Robotics and Automation, Vol. 4, IEEE Press, Piscataway, NJ, 1997, pp. 3409–3414.
[8] M. P. do Carmo, *Riemannian Geometry*, Birkhäuser, Boston, MA, 1992.
[9] M. Erdmann and T. Lozano-Perez, *On multiple moving objects (motion planning)*, Algorithmica, 2 (1987), pp. 477–521.
[10] E. Frazzoli, Z.-H. Mao, J.-H. Oh, and E. Feron, *Resolution of conflicts involving many aircraft via semidefinite programming*, J. Guidance Control Dynam., 24 (2001), pp. 79–86.
[11] K. Fujimura, *Motion Planning in Dynamic Environments*, Springer-Verlag, Tokyo, 1991.
[12] J. Hu, *A Study of Conflict Detection and Resolution in Free Flight*, Master's thesis, University of California, Berkeley, CA, 1999.
[13] J. Hu, M. Prandini, K. H. Johansson, and S. Sastry, *Hybrid geodesics as optimal solutions to the collision-free motion planning problem*, in Hybrid Systems: Computation and Control. 4th International Workshop (HSCC01), Rome, Italy, 2001, Lecture Notes in Comput. Sci. 2034, M. Domenica, D. Benedetto, and A. Sangiovanni-Vincentelli, eds., Springer-Verlag, Berlin, pp. 305–318.
[14] J. Hu, M. Prandini, and S. Sastry, *Optimal maneuver for multiple aircraft conflict resolution: A braid point of view*, in Proceedings of the 39th IEEE International Conference on Decision and Control, Vol. 4, IEEE Press, Piscataway, NJ, 2000, pp. 4164–4169.
[15] J. Hu, M. Prandini, and S. Sastry, *Optimal coordinated maneuvers for three dimensional aircraft conflict resolution*, J. Guidance Control Dynam., 25 (2002), pp. 888–900.
[16] J. Hu and S. Sastry, *Optimal collision avoidance and formation switching on Riemannian manifolds*, in Proceedings of the 40th IEEE International Conference on Decision and Control, Vol. 2, IEEE Press, Piscataway, NJ, 2001, pp. 1071–1076.
[17] J. Hu and S. Sastry, *Hybrid Geodesic Flows on Manifolds with Boundary*, University of California, Berkeley, CA, manuscript.
[18] Y. K. Hwang and N. Ahuja, *Gross motion planning*, ACM Comput. Surveys, 24 (1992), pp. 219–291.
[19] A. Inselberg and B. Dimsdale, *Multidimensional lines* II: *Proximity and applications*, SIAM J. Appl. Math., 54 (1994), pp. 578–596.
[20] J. Jost, *Riemannian Geometry and Geometric Analysis*, 2nd ed., Springer-Verlag, Berlin, 1998.
[21] W.-S. Koon and J. E. Marsden, *Optimal control for holonomic and nonholonomic mechanical systems with symmetry and Lagrangian reduction*, SIAM J. Control Optim., 35 (1997), pp. 901–929.
[22] P. S. Krishnaprasad, *Eulerian many-body problems*, in Dynamics and Control of Multibody Systems (Brunswick, ME, 1988), AMS, Providence, RI, 1989, pp. 187–208.
[23] J. Kuchar and L. C. Yang, *Survey of conflict detection and resolution modeling methods*, IEEE Trans. Intelligent Transport. Systems, 1 (2000), pp. 179–189.
[24] J. Latombe, *Robot Motion Planning*, Kluwer Academic Publishers, Boston, MA, 1991.
[25] S. LaValle and S. Hutchinson, *Optimal motion planning for multiple robots having independent goals*, IEEE Trans. Robotics Automation, 14 (1998), pp. 912–925.
[26] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, 2nd ed., Springer-Verlag, New York, 1999.

[27] C. D. Medio and G. Oriolo, *Robot obstacle avoidance using vortex fields*, in Advances in Robot Kinematics, S. Stifter and J. Lenarcic, eds., Springer-Verlag, Vienna, 1991, pp. 227–235.

[28] F. Medioni, N. Durand, and J. M. Alliot, *Air traffic conflict resolution by genetic algorithms*, in Artificial Evolution, European Conference (AE '95), Springer-Verlag, Berlin, 1995, pp. 370–383.

[29] P. Menon, G. Sweriduk, and B. Sridhar, *Optimal strategies for free-flight air traffic conflict resolution*, J. Guidance Control Dynam., 22 (1999), pp. 202–211.

[30] A. Miele, T. Wang, C. Chao, and J. Dabney, *Optimal control of a ship for collision avoidance maneuvers*, J. Optim. Theory Appl., 103 (1999), pp. 495–519.

[31] J. W. Milnor, *Morse Theory*, Princeton University Press, Princeton, NJ, 1963.

[32] S. Morgan, *The Mathematical Theory of Knots and Braids: An Introduction*, North–Holland, Amsterdam, 1991.

[33] K. Murasugi and B. I. Kurpita, *A Study of Braids*, Kluwer Academic Publishers, Boston, MA, 1999.

[34] M. Prandini, J. Hu, J. Lygeros, and S. Sastry, *A probabilistic approach to aircraft conflict detection*, IEEE Trans. Intelligent Transport. Systems, 1 (2000), pp. 199–220.

[35] Radio Technical Commission for Aeronautics, *Minimum Aviation System Performance Standards for Automatic Dependent Surveillance-Broadcast (ADS-B)*, Technical report, RTCA-186, Draft 4.0, RTCA, Inc., Washington, DC, 1997.

[36] E. Rimon and D. Koditschek, *Exact robot navigation using artificial potential functions*, IEEE Trans. Robotics Automation, 8 (1993), pp. 501–519.

[37] M. Rude, *Collision avoidance by using space-time representations of motion processes*, Autonomous Robots, 4 (1997), pp. 101–119.

[38] M. Sharir and S. Sifrony, *Coordinated motion planning for two independent robots*, Ann. Math. Artificial Intelligence, 3 (1991), pp. 107–130.

[39] C. Tomlin, *Hybrid Control of Air Traffic Management Systems*, Ph.D. thesis, University of California, Berkeley, CA, 1998.

[40] C. Tomlin, G. Pappas, and S. Sastry, *Conflict resolution for air traffic management: A study in multi-agent hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 509–521.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 669–685

EXISTENCE OF OPTIMAL CONTROLS FOR A GENERAL CLASS OF IMPULSIVE SYSTEMS ON BANACH SPACES*

N. U. AHMED†

Abstract. This paper presents some results on the question of existence of optimal controls for a large class of semilinear impulsive systems in infinite dimensional spaces with admissible controls from the space of vector measures. This also includes, as a special case, the class of purely impulsive controls. Two physical examples are presented for illustration.

Key words. impulsive systems, Banach spaces, vector measures, optimal control, existence

AMS subject classifications. 34G20, 34K30, 34K35, 49K27, 93C25

PII. S0363012901391299

1. Introduction.

In this paper we present some results on existence of optimal controls for systems governed by nonlinear impulsive evolution equations on Banach spaces. The general description of such systems was proposed in [1, 2, 3] as given below:

$$dx = Axdt + f(t,x)dt + g(t,x)\nu(dt) + C(t,x)u(dt), \qquad t \in I, \tag{1}$$

$$x(0) = x_0. \tag{2}$$

In general the operator A is the infinitesimal generator of a C_0-semigroup in a Banach space E, f and g are nonlinear operators mapping $I \times E$ to E, ν is a countably additive bounded signed measure on I, C is an operator valued function mapping $I \times E$ into $\mathcal{L}(F,E)$, and u is an F-valued vector measure representing the control.

This model includes all the standard models used by many authors in the field [9, 10, 11, 12, 13, 14] (including the references therein). It also includes the models considered in [1, 2, 4, 5, 6]. Inclusion of delay in our model by considering f as a functional mapping $I \times C([-\tau,0],E)$ to E and specifying initial data $\zeta \in C([-\tau,0],E)$ does not complicate the problem in any substantial way.

Returning to the control problem, we assume that the objective functional is given by

$$J(u) = \int_I \ell(t,x(t))dt + \Psi(x(T)) + \varphi(u), \tag{3}$$

where ℓ, Ψ, φ are suitable functions to be defined later. In a recent paper of the author, some results on the question of existence of optimal controls were proved in [1] under the assumption that C is independent of x. In a more recent paper, necessary conditions of optimality also were developed for this class of systems. In [1], the control appears linearly, that is, the control operator C is independent of the state. Once this operator is dependent on the state x, the problem becomes more difficult. Our concern here is to study the question of existence of optimal controls

*Received by the editors June 22, 2001; accepted for publication (in revised form) December 2, 2002; published electronically May 29, 2003. This work was partially supported by the National Science and Engineering Research Council of Canada under grant A7109.
http://www.siam.org/journals/sicon/42-2/39129.html
†School of Information Technology and Engineering and Department of Mathematics, University of Ottawa, Ottawa, ON, Canada K15 6N5 (ahmed@site.uottawa.ca).

for this nonlinear case. This was stated as an open problem in our paper [6], where necessary conditions of optimality were developed.

The rest of the paper is organized as follows. In section 2, some basic notations and terminologies are presented. In section 3, we present a result on existence and uniqueness of solutions for the system (1)–(2). In section 4, we develop our main results on the question of existence of optimal controls. The first result given in Theorem 4.2 assumes a decomposability property which is somewhat restrictive. Conditions guaranteeing this property are stated in the remark following the introduction of admissible controls. This assumption is disposed of in Theorem 4.3 and Corollary 4.4 by imposing some conditions on the control operator C. In section 5, we present some comments and questions on open problems. The article concludes with several physical examples.

2. Some notations and terminologies. Let X be a Banach space with dual X^*, and let $\mathcal{B}$ denote the sigma algebra of Borel subsets of the interval $I \equiv [0,T]$. Let $\mathcal{M}_c(I,X)$ denote the space of bounded countably additive X-valued vector measures on the sigma algebra $\mathcal{B}$ having bounded total variation as defined below. That is, for each $\mu \in \mathcal{M}_c(I,X)$, we write

$$|\mu|_v \equiv |\mu|(I) \equiv \sup_{\pi}\Big\{\sum_{J\in\pi} \parallel \mu(J) \parallel_X\Big\},$$

where the supremum is taken over all partitions π of the interval I into a finite number of disjoint members of $\mathcal{B}$. With respect to this topology, $\mathcal{M}_c(I,X)$ is a Banach space. For any $J \in \mathcal{B}$, define the variation of μ on J by

$$V(\mu)(J) \equiv V(\mu,J) \equiv |\mu|(J).$$

Since μ is countably additive and bounded, this defines a countably additive bounded positive measure on $\mathcal{B}$. In case $X = R$, the real line, we have the space of real valued signed measures. We denote this by simply $\mathcal{M}_c(I)$ in place of $\mathcal{M}_c(I,R)$. Clearly for $\nu \in \mathcal{M}_c(I)$, $V(\nu)$ is also a countably additive bounded positive measure. For uniformity of notation we use λ to denote the Lebesgue measure. Strong convergence of a sequence $\{\xi_n\} \in X$ to an element $\xi \in X$ is denoted by $\xi_n \overset{s}{\longrightarrow} \xi$, and its weak convergence (weak star convergence) by $\xi_n \overset{w(w^*)}{\longrightarrow} \xi$. For any pair of Banach spaces X, Y, $\mathcal{L}(X,Y)$ will denote the space of bounded linear operators from X to Y.

For any Banach space X, we use $B(I,X)$ to denote the linear space of all bounded X-valued functions on the interval I, and $C(I,X)$ the space of continuous functions with values in X. The space $B(I,X)$, furnished with the sup norm topology,

$$\parallel z \parallel_0 \equiv \sup\{\parallel z(t) \parallel_X, t \in I\}, \qquad z \in B(I,X),$$

is a Banach space, and $C(I,X)$ is a closed subspace of $B(I,X)$; hence it is also a Banach space. It is clear that, with respect to the above topology, $B(I,X)$ is a normed vector space. To see that it is complete, let $\{x_n\}$ be a Cauchy sequence. Then, for every $\varepsilon > 0$, there exists an integer $m(\varepsilon) \in N$ such that $\parallel x_n - x_m \parallel_0 < \varepsilon$ for all $m, n \geq m(\varepsilon)$. For each $t \in I$, let $x(t) = s - \lim x_n(t)$. Since E is a Banach space, the limit indicated does exist. Then, for each $t \in I$, there is an integer $r \geq m(\varepsilon)$ with $\parallel x(t) - x_r(t) \parallel < \varepsilon$, and thus, for $n \geq m(\varepsilon)$,

$$\parallel x(t) - x_n(t) \parallel \leq \parallel x(t) - x_r(t) \parallel + \parallel x_r(t) - x_n(t) \parallel < 2\varepsilon \text{ for all } t \in I.$$

Since $\varepsilon > 0$ is arbitrary, this proves that every Cauchy sequence in $B(I, X)$ has a limit with respect to the norm topology and hence is a Banach space.

This is a very general Banach space. For application we need measurability. Let $B_\infty(I, X)$ denote the space of bounded strongly measurable X-valued functions on I. Furnished with the same sup norm topology, it is a closed subspace of $B(I, X)$ and hence a Banach space. This is the Banach space used throughout this paper.

The space of bounded piecewise continuous functions (with the sup norm topology), denoted by $PWC(I, X)$, is certainly a linear subspace of $B_\infty(I, X)$, but it is not clear if it is a closed subspace. In any case we can choose the space $B_\infty(I, X)$ for the space of solutions of all our impulsive systems. However, in many cases the solutions may posses stronger regularity properties such as piecewise continuity. This certainly happens in case all the measures including the control measures are all purely atomic. Further regularity is expected if ν is absolutely continuous with respect to Lebesgue measure λ, and given that both F and F^* satisfy the Radon–Nikodým property, the control measures are λ continuous. In this case the solutions are in $C(I, X)$.

3. Existence and regularity of solutions. For study of the question of existence of optimal controls, it is certainly necessary to guarantee the existence (and possibly uniqueness) of solutions of the controlled evolution equation for each and every control from the admissible class. Here in this section, we present a simple existence and uniqueness theorem for the purpose. First let us recall that by a mild solution of the system (1)–(2) we mean a function $x \in B_\infty(I, E)$ that satisfies the integral equation

$$x(t) = S(t)x_0 + \int_0^t S(t-s)f(s, x(s))ds + \int_0^t S(t-s)g(s, x(s))\nu(ds) + \int_0^t S(t-s)C(s, x(s))u(ds), \qquad t \in I. \tag{4}$$

We can prove the following result along the same lines as in [1, 3].

THEOREM 3.1. *Consider the system* (1)–(2) *and suppose that* E, F *are Banach spaces,* A *is the infinitesimal generator of a* C_0*-semigroup* $S(t), t \geq 0$*, in* E*,* $\nu \in \mathcal{M}_c(I)$*, and* $u \in \mathcal{M}_c(I, F)$*. Suppose that both* f *and* g *map* $I \times E$ *into* E *and are measurable in* t *on* I *and continuous in* x *on* E *and that there exist two nonnegative functions* $K \in L_1^+(I, \lambda)$ *and* $L \in L_1^+(I, V(\nu))$ *such that, for all* $x, y \in E$*,*

$$\| f(t,x) - f(t,y) \|_E \leq K(t) \| x - y \|_E, \qquad \| f(t,x) \| \leq K(t)(1 + \| x \|), \tag{5}$$

$$\| g(t,x) - g(t,y) \|_E \leq L(t) \| x - y \|_E, \qquad \| g(t,x) \| \leq L(t)(1 + \| x \|), \tag{6}$$

where the first set of inequalities hold λ *a.e. and those of the second set hold* $V(\nu)$ *a.e. The operator valued function* C *mapping* $I \times E$ *to* $\mathcal{L}(F, E)$ *is measurable in* t *on* I *and continuous in* x *on* E*, all with respect to the uniform operator topology of* $\mathcal{L}(F, E)$*; and there exists an* $R \in L_1^+(I, V(u))$ *(possibly dependent on* u*) so that*

$$\| C(t,x) - C(t,y) \|_{\mathcal{L}(F,E)} \leq R(t) \| x - y \|_E, \quad \| C(t,x) \|_{\mathcal{L}(F,E)} \leq R(t)(1 + \| x \|). \tag{7}$$

Then, for each $x_0 \in E$ *and* $u \in \mathcal{M}_c(I, F)$*, the system* (1) *has a unique mild solution in* $B_\infty(I, E)$*.*

Proof. The proof is similar to that of [1, Theorem 1]. We present a very brief outline. Since S is a C_0 semigroup on E, there exists a finite number $M \geq 1$ such that

$$\sup\{\| S(t) \|_{\mathcal{L}(E)}, t \in I\} = M.$$

Define the operator G as follows:

$$(Gx)(t) \equiv S(t)x_0 + \int_0^t S(t-s)f(s,x(s))ds + \int_0^t S(t-s)g(s,x(s))\nu(ds) + \int_0^t S(t-s)C(s,x(s))u(ds). \tag{8}$$

Using the growth and continuity assumptions of the operators f, g, C with respect to the state $x \in E$, it is easy to verify that G maps $B_\infty(I,E)$ into itself. Define the scalar measure α on $\mathcal{B}$ by

$$\alpha(\sigma) \equiv \int_\sigma K(s)ds + \int_\sigma L(s)V(\nu,ds) + \int_\sigma R(s)V(u,ds), \qquad \sigma \in \mathcal{B}. \tag{9}$$

Note that this is a countably additive positive measure. For any $x, y \in B_\infty(I,E)$ define

$$\rho_t(x,y) \equiv \sup_{0\le s\le t} \| x(s) - y(s) \| \tag{10}$$

and set $\rho(x,y) \equiv \rho_T(x,y)$. Clearly $\rho(x,y)$ defines a metric on $B_\infty(I,E)$, and with respect to this metric $B_\infty(I,E)$ is a complete metric space. For any pair of elements $x, y \in B_\infty(I,E)$, it follows from the expressions (8), (9), and (10) that

$$\rho_t(Gx,Gy) \le M\int_0^t \rho_s(x,y)\alpha(ds), \qquad t \in I. \tag{11}$$

Define

$$W(t) \equiv \int_0^t \alpha(ds), \qquad t \in I. \tag{12}$$

By repeated substitution, it follows from (11) and (12) that

$$\rho_t(G^n x, G^n y) \le \big(M^n W^n(t)/\Gamma(n+1)\big)\rho_t(x,y), \qquad t \in I.$$

This implies that

$$\rho(G^n x, G^n y) \le \big(M^n W^n(T)/\Gamma(n+1)\big)\rho(x,y).$$

Hence, for sufficiently large n, G^n is a contraction, and by the Banach fixed point theorem it has a unique fixed point $x^* \in B_\infty(I,E)$ which is also the unique fixed point of the operator G itself. Hence x^* is the unique mild solution of the evolution equation (1)–(2). This completes the outline of our proof. □

Remark. The conclusions of Theorem 3.1 also remain valid under local Lipschitz conditions, given that the growth assumption remains in force.

Remark. This result has been substantially generalized in [3], where continuous dependence of solutions on initial data and control measures also is studied. However, the result presented above is sufficient for our purpose.

In [6], a set of necessary conditions of optimality was developed for similar systems. In that the operator C was assumed to be independent of the state. By imposing appropriate regularity property for C with respect to the state $x \in E$, it is possible to develop similar necessary conditions of optimality once we have the existence result. Thus our primary objective here is to prove existence of optimal controls. This we do in the next section. This problem was left open in [6].

4. Existence of optimal controls. For convenience of the reader we state a result from Diestel and Uhl [7] that gives the necessary and sufficient conditions for weak compactness of subsets of the space $\mathcal{M}_c(I,F)$. This is a celebrated result due to Bartle, Dunford, and Schwartz. An elegant proof of this result is given in Diestel and Uhl [7]. We present here a special version of their result.

THEOREM 4.1 (Bartle–Dunford–Schwartz [7]). *Let $\mathcal{B}$ be a sigma-field of subsets of the set I. Suppose that F is a Banach space such that both F and its dual F^* satisfy the Radon–Nikodým property. A subset $\mathcal{U}$ of $\mathcal{M}_c(I,F)$ is weakly compact if and only if*

(i) *$\mathcal{U}$ is bounded,*

(ii) *there exists a nonnegative countably additive finite scalar valued measure μ on $\mathcal{B}$ such that $\lim_{\mu(\sigma)\to 0}|u|(\sigma)=0$ uniformly with respect to $u\in\mathcal{U}$,*

(iii) *for each $\sigma\in\mathcal{B}$, the set $\{u(\sigma), u\in\mathcal{U}\}$ is a relatively weakly compact subset of F.*

For details on Banach spaces satisfying the Radon–Nikodým property see Diestel and Uhl [7]. We simply mention here that all Hilbert spaces, reflexive Banach spaces, Gelfand spaces, and many more satisfy the Radon–Nikodým property. The spaces L_1, L_∞, C do not satisfy the Radon–Nikodým property. For a comprehensive list of Banach spaces that satisfy the Radon–Nikodým property see [7, p. 218].

Let $L_1(\mu,F)$ denote the space of μ-measurable F-valued functions on I which are Lebesgue–Bochner integrable with respect to the measure μ, and suppose this is furnished with the standard norm topology making it into a Banach space. Define the map Γ from $L_1(\mu,F)$ to $\mathcal{M}_c(I,F)$ by

$$\Gamma(g)(\sigma)\equiv\int_\sigma g(t)\mu(dt),\qquad \sigma\in\mathcal{B}.$$

Clearly Γ is a linear operator continuous with respect to the norm topologies and hence also continuous with respect to the weak topologies. Thus Γ maps weakly compact subsets of $L_1(\mu,F)$ into weakly compact subsets of $\mathcal{M}_c(I,F)$.

Admissible controls. Now we are prepared to prove our main results. First, let us recall that a set $K\subset L_1(\mu,F)$ is said to be decomposable if, for every measurable set $\sigma\subset I$ and every pair $f_1,f_2\in K$, the function

$$f\equiv\chi_\sigma f_1+\chi_{I\setminus\sigma}f_2\in K,$$

where χ_σ denotes the indicator function of the set σ. In general, this definition is also used for the larger space $L_0(\mu,F)\supset L_1(\mu,F)$. For admissible controls we choose a set $\mathcal{U}_{ad}\subset\mathcal{M}_c(I,F)$ satisfying the following properties:

(P1) $\mathcal{U}_{ad}$ is weakly compact satisfying Theorem 4.1.

(P2) $\Gamma^{-1}(\mathcal{U}_{ad})$ is a decomposable subset of $L_1(\mu,F)$.

These assumptions are relaxed later in Theorem 4.3 and Corollary 4.4 by imposing stronger assumptions on the operator C.

Remark. Even though the class of admissible controls chosen is somewhat restrictive, it covers a fairly large class. For example, take any nonnegative countably additive finite measure μ and choose any decomposable weakly compact set K_{ad} of $L_1(\mu,F)$ and define $\mathcal{U}_{ad}=\Gamma(K_{ad})$. Clearly this set satisfies the properties (P1) and (P2).

For our purpose we introduce the following basic assumptions.

Assumption A1. The operator L with values

$$L_t(u) \equiv \int_0^t S(t-s)C(s,x(s))u(ds), \qquad t \in I, \tag{13}$$

maps every weakly convergent sequence $\{u_n\} \subset \mathcal{M}_c(I,F)$ into a strongly convergent sequence in E for each $t \in I$ and for any fixed but arbitrary $x \in B(I,E)$.

Assumption A2. The function ℓ is measurable in t on I and lower semicontinuous in x on E, and Ψ is lower semicontinuous on E. There exist $h \in L_1(I), c_2 \in R$ so that

$$\ell(t,x) \geq h(t) \text{ a.e. for all } x \in E, \tag{14}$$

$$\Psi(x) \geq c_2 \text{ for all } \ x \in E, \tag{15}$$

and the functional φ is weakly lower semicontinuous on $\mathcal{M}_c(I,F)$.

THEOREM 4.2. *Consider the system given by* (1)–(2) *with the cost functional* (3) *and admissible controls* $\mathcal{U}_{ad}$ *satisfying properties* (P1) *and* (P2) *where both* F *and its dual* F^* *satisfy the Radon–Nikodým property. Suppose Assumptions* A1 *and* A2 *and further that the assumptions of Theorem* 3.1 *hold with* $R(t) = R$ *a constant positive number possibly dependent on the set* $\mathcal{U}_{ad}$. *Then there exists an optimal control, that is, a control* $u_o \in \mathcal{U}_{ad}$ *that minimizes the functional* $J(u)$ *given by* (3).

Proof. If $J(u) \equiv +\infty$ for all $u \in \mathcal{U}_{ad}$, there is nothing to prove. Again, by virtue of Assumption A2, we have

$$\inf\{J(u), u \in \mathcal{U}_{ad}\} = m > -\infty. \tag{16}$$

Let $\{u_n\} \subset \mathcal{U}_{ad}$ be a minimizing sequence so that

$$\lim_{n\to\infty} J(u_n) = m.$$

Since $\mathcal{U}_{ad}$ is weakly compact, by the Eberlein–Smulian theorem [8, Theorem V.6.1], it is also weakly sequentially compact. Hence there exists a subsequence of the sequence $\{u_n\}$, relabeled as $\{u_n\}$, and an element $u_o \in \mathcal{U}_{ad}$ so that $u_n \overset{w}{\longrightarrow} u_o$. We show that u_o is an optimal control. Let $\{x_n, x_o\} \in B_\infty(I,E)$ denote the mild solutions of the evolution equation (1)–(2) corresponding to the controls $\{u_n, u_o\}$, respectively. In other words, these functions satisfy the following integral equations:

$$\begin{aligned} x_n(t) = S(t)x_0 &+ \int_0^t S(t-s)f(s,x_n(s))ds + \int_0^t S(t-s)g(s,x_n(s))\nu(ds) \\ &+ \int_0^t S(t-s)C(s,x_n(s))u_n(ds), \qquad n \in N, \end{aligned} \tag{17}$$

$$\begin{aligned} x_o(t) = S(t)x_0 &+ \int_0^t S(t-s)f(s,x_o(s))ds + \int_0^t S(t-s)g(s,x_o(s))\nu(ds) \\ &+ \int_0^t S(t-s)C(s,x_o(s))u_o(ds). \end{aligned} \tag{18}$$

First we show that there exists a ball $B_r \subset E$ of finite radius $r > 0$, centered at the origin, so that $x_n(t), x_o(t) \in B_r$ for all $t \in I$. Considering the first equation, and recalling the bound M of the semigroup S on I, we have

$$\begin{aligned} \| x_n(t) \| \leq M \| x_0 \| &+ M \int_0^t K(s)(1+ \| x_n(s) \|)ds \\ &+ M \int_0^t L(s)(1+ \| x_n(s) \|)V(\nu, ds) + M \int_0^t R(1+ \| x_n(s) \|)V(u_n, ds). \end{aligned} \tag{19}$$

Introducing the sequence of measures $\{\gamma_n\}$ as

$$\begin{aligned}\gamma_n(\sigma) &\equiv \int_\sigma K(s)ds + \int_\sigma L(s)V(\nu, ds) + \int_\sigma RV(u_n, ds),\\ &\equiv \int_\sigma K(s)ds + \int_\sigma L(s)V(\nu, ds) + RV(u_n, \sigma),\end{aligned} \tag{20}$$

it follows from the preceding inequality that

$$(1+ \parallel x_n(t) \parallel) \leq (1 + M \parallel x_0 \parallel) + M \int_0^t (1+ \parallel x_n(s) \parallel)\gamma_n(ds). \tag{21}$$

Since ν and $\{u_n\}$ are countably additive measures of bounded total variation, it follows from our assumptions on K and L that $\{\gamma_n\}$ is a sequence of countably additive positive measures of bounded total variation. By virtue of an inequality recently proved by the author [2, Lemma 5, p. 268], it follows from (21) that

$$(1+ \parallel x_n(t) \parallel) \leq (1 + M \parallel x_0 \parallel)\exp\left\{M \int_0^t \gamma_n(ds)\right\}, \qquad t \in I. \tag{22}$$

Similarly, using (18), we obtain

$$(1+ \parallel x_o(t) \parallel) \leq (1 + M \parallel x_0 \parallel)\exp\left\{M \int_0^t \gamma_o(ds)\right\}, \qquad t \in I, \tag{23}$$

where the measure γ_o is given by

$$\gamma_o(\sigma) \equiv \int_\sigma K(s)ds + \int_\sigma L(s)V(\nu, ds) + RV(u_o, \sigma), \qquad \sigma \in \mathcal{B}. \tag{24}$$

Since $\mathcal{U}_{ad}$ is bounded (Theorem 4.1 (i)), it follows from (22), (23), and (24) that there exists a finite positive number r such that $x_o(t), x_n(t) \in B_r(E) \equiv \{e \in E :\parallel e \parallel_E \leq r\}$ for all $t \in I$ and for all $n \in N$. Now define

$$e_n(t) \equiv \parallel x_n(t) - x_o(t) \parallel, \qquad t \in I, \quad n \in N, \tag{25}$$

$$y_n(t) \equiv \int_0^t S(t-s)C(s, x_0(s))(u_0 - u_n)(ds), \qquad t \in I, \quad n \in N, \tag{26}$$

$$z_n(t) \equiv \parallel y_n(t) \parallel, \qquad t \in I, \quad n \in N. \tag{27}$$

Using (17) and (18), subtracting one from the other, and utilizing the above notations, one can easily verify that

$$\begin{aligned}e_n(t) \leq z_n(t) &+ M \int_0^t K(s)e_n(s)ds + \int_0^t L(s)e_n(s)V(\nu, ds)\\ &+MR \int_0^t e_n(s)V(u_n, ds), \qquad t \in I, \quad n \in N.\end{aligned} \tag{28}$$

Introduce the nonnegative set function

$$\beta(\sigma) \equiv \sup\{V(u, \sigma) \equiv |u|(\sigma), u \in \mathcal{U}_{ad}\}, \qquad \sigma \in \mathcal{B}. \tag{29}$$

Since $\mathcal{U}_{ad}$ is weakly compact, it follows from the Bartle–Dunford–Schwartz theorem—in particular Theorem 4.1(iii)—that the set $\{u(\sigma), u \in \mathcal{U}_{ad}\}$ is relatively weakly compact and thus a bounded subset of F for each $\sigma \in \mathcal{B}$. Thus the set function β is

well defined for each $\sigma \in \mathcal{B}$. In fact it is a finitely additive measure. This follows from (i) and (ii) of Theorem 4.1. Indeed, since the elements of $\mathcal{U}_{ad}$ are μ continuous (F-valued) vector measures and both F and F^* satisfy the Radon–Nikodým property, for every $u \in \mathcal{U}_{ad}$ there exists a $g \in L_1(\mu, F)$ such that

$$u(\sigma) = \int_\sigma g(t)\mu(dt),$$

and thus the operator Γ, as defined above, is an isometric isomorphism of $L_1(\mu, F)$ onto a subspace of $\mathcal{M}_c(I, F)$ with Range$(\Gamma) \supset \mathcal{U}_{ad}$. Since $\mathcal{U}_{ad}$ is weakly compact, it is clear that $\Gamma^{-1}(\mathcal{U}_{ad})$ is a weakly compact subset of $L_1(\mu, F)$.

Hence, for any disjoint members $\sigma_1, \sigma_2 \in \mathcal{B}$, it follows from the decomposability assumption (P2) that

$$\begin{aligned}
\beta(\sigma_1 \cup \sigma_2) &= \sup\{V(u, \sigma_1 \cup \sigma_2), u \in \mathcal{U}_{ad}\} \\
&= \sup\left\{\int_{\sigma_1\cup\sigma_2} \| g(t) \| \mu(dt), g \in \Gamma^{-1}(\mathcal{U}_{ad})\right\} \\
&= \sup\left\{\int_{\sigma_1} \| g(t) \| \mu(dt), g \in \Gamma^{-1}(\mathcal{U}_{ad})\right\} \\
&\quad + \sup\left\{\int_{\sigma_2} \| g(t) \| \mu(dt), g \in \Gamma^{-1}(\mathcal{U}_{ad})\right\} \\
&= \beta(\sigma_1) + \beta(\sigma_2). \qquad (30)
\end{aligned}$$

Clearly this identity holds for any finite family of disjoint members from the sigma-field $\mathcal{B}$. Thus β is a finitely additive positive measure. Since I is a compact interval, it follows from the extension theorem for finitely additive set functions [8, Theorem III.5.13, III.5.14] that β has a countably additive extension which we denote by β again. Using this fact, it follows from (28) that

$$\begin{aligned}
e_n(t) \le z_n(t) &+ M\int_0^t K(s)e_n(s)ds + \int_0^t L(s)e_n(s)V(\nu, ds) \\
&+MR\int_0^t e_n(s)\beta(ds), \qquad t \in I, \quad n \in N. \qquad (31)
\end{aligned}$$

Defining

$$\beta_o(\sigma) \equiv \int_\sigma K(s)ds + \int_\sigma L(s)V(\nu, ds) + R\beta(\sigma) \qquad (32)$$

and recalling that $M \ge 1$, (31) reduces to

$$e_n(t) \le z_n(t) + M\int_0^t e_n(s)\beta_o(ds), \qquad t \in I. \qquad (33)$$

Again using Lemma 5 of [2], it follows from this that

$$e_n(t) \le z_n(t) + \int_0^t M\exp\left\{\int_s^t M\beta_o(dr)\right\} z_n(s)\beta_o(ds). \qquad (34)$$

Since $x_o \in B(I, E)$ is bounded with $x_o(t) \in B_r$, and $u_n \xrightarrow{w} u_o \in \mathcal{U}_{ad}$, it follows from Assumption A1 and (26) that $y_n(t) \xrightarrow{s} 0$ in E pointwise in t for each $t \in I$. Hence $z_n(t) \longrightarrow 0$ pointwise in t on I. Further, it follows from the boundedness of $\mathcal{U}_{ad}$ that

$$\sup_{n\in N}\sup\{\| y_n(t) \|, t \in I\} = \sup_{n\in N}\sup\{z_n(t) : t \in I\} < \infty.$$

Thus, by virtue of the dominated convergence theorem, it follows from (34), upon letting $n \to \infty$, that

$$\lim_{n\to\infty} e_n(t) = 0$$

for each $t \in I$. In other words, $x_n(t) \xrightarrow{s} x_o(t)$ in E for each $t \in I$. By using this result, Fatou's lemma, and Assumption A2, we conclude that

$$\liminf_{n\to\infty} J(u_n) \geq J(u_o).$$

Thus J is weakly lower semicontinuous on $\mathcal{U}_{ad}$. Since $u_o \in \mathcal{U}_{ad}$ and $\{u_n\}$ is a minimizing sequence, we have

$$m \leq J(u_o) \leq \liminf_{n\to\infty} J(u_n) \leq \lim_{n\to\infty} J(u_n) = m.$$

This shows that J attains its minimum at $u_o \in \mathcal{U}_{ad}$ and hence proves the existence. This completes the proof. □

By imposing a stronger assumption on the operator C we can eliminate the requirement of decomposability of the set $\Gamma^{-1}(\mathcal{U}_{ad})$. These results are stated in the following theorem and its corollary.

THEOREM 4.3. *Consider the system given by* (1)–(2) *with the cost functional* (3) *and admissible controls* $\mathcal{U}_{ad}$ *a weakly compact subset of* $\mathcal{M}_c(I, F)$ *satisfying the Bartle–Dunford–Schwartz theorem, Theorem* 4.1. *Suppose that all the other assumptions of Theorem* 4.2 *hold except that* $R(\cdot)$ *is uniformly bounded and Assumption* A1 *is replaced by the following.*

Assumption B1. The operator L_t, given by

$$L_t(u) \equiv \int_0^t S(t-s)C(s,\xi)u(ds), \qquad t \in I,$$

maps every weakly convergent sequence $\{u_n\} \subset \mathcal{M}_c(I, F)$ into a strongly convergent sequence in E for each $t \in I$, uniformly with respect to $\xi \in B_r(E)$. Then there exists an optimal control, that is, a control $u_o \in \mathcal{U}_{ad}$ that minimizes the functional $J(u)$ given by (3).

Proof. The proof is essentially identical. We give a brief outline. Subtracting (18) from (17) and writing

$$\begin{aligned}\int_0^t S(t-s)C(s,x_n(s))u_n(ds) - \int_0^t S(t-s)C(s,x_o(s))u_0(ds)\\ = \int_0^t S(t-s)C(s,x_n(s))(u_n(ds) - u_o(ds))\\ + \int_0^t S(t-s)\{C(s,x_n(s)) - C(s,x_o(s))\}u_o(ds),\end{aligned}$$

the inequality (34) takes the form

$$e_n(t) \leq \tilde{z}_n(t) + M\int_0^t \exp\Big\{M\int_s^t \tilde{\beta}_o(dr)\Big\}\tilde{z}_n(s)\tilde{\beta}_o(ds), \qquad t \in I, \tag{35}$$

where

$$\tilde{z}_n(t) \equiv \| \tilde{y}_n(t) \|_E, \quad \tilde{y}_n(t) \equiv \int_0^t S(t-s)C(s,x_n(s))(u_n(ds) - u_o(ds)),$$

and the measure $\tilde{\beta}_o$ is given by

$$\tilde{\beta}_o(\Delta) \equiv \int_\Delta K(s)ds + \int_\Delta L(s)V(\nu, ds) + \int_\Delta R(s)V(u_o, ds), \qquad \Delta \in \mathcal{B}.$$

Since R is uniformly bounded and $V(\nu)$ and $V(u_o)$ are countably additive bounded positive measures, it is clear that $\tilde{\beta}_o$ is also a countably additive bounded positive measure. Recalling that $x_n(t) \in B_r$ for all integers n and all $t \in I$, it follows from the current assumption on the operator $L_t, t \in I$, that $\tilde{z}_n(t) \to 0$ for each $t \in I$. Again using dominated convergence theorem, it follows from inequality (35) that $e_n(t) \longrightarrow 0$ for each $t \in I$. The rest of the proof is identical to that of Theorem 4.2. □

In case the operator C is independent of the state, system (1) reduces to

$$dx = Axdt + f(t,x)dt + g(t,x)\nu(dt) + C(t)u(dt), \qquad t \in I, \tag{36}$$

$$x(0) = x_0. \tag{37}$$

In this case also we do not require the decomposability condition, and we have the following result.

COROLLARY 4.4. *Consider the system given by* (36)–(37) *with the cost functional* (3) *and admissible controls* $\mathcal{U}_{ad}$ *a weakly compact subset of* $\mathcal{M}_c(I,F)$ *satisfying the Bartle–Dunford–Schwartz theorem, Theorem* 4.1. *Suppose all the other assumptions of Theorem* 4.3 *hold except those related to the operator* C. *Suppose the operator* $L_t, t \in I$, *given by*

$$L_t(u) \equiv \int_0^t S(t-s)C(s)u(ds), \qquad t \in I,$$

maps every weakly convergent sequence $\{u_n\} \subset \mathcal{M}_c(I,F)$ *into a strongly convergent sequence in* E *for each* $t \in I$. *Then there exists an optimal control, that is, a control* $u_o \in \mathcal{U}_{ad}$ *that minimizes the functional* $J(u)$ *given by* (3).

By imposing a stronger continuity condition on both ℓ and Ψ, as functions of x on E, it is possible to relax the requirements (14) and (15). This is presented below as a corollary after we have introduced Assumption A3, replacing A2.

Assumption A3. For each $x \in E$, the function ℓ is measurable in t on I and, for almost all $t \in I$, is continuous and bounded in x on bounded sets of E, and the function Ψ is also continuous and bounded on bounded sets of E satisfying the following conditions: For each finite $r > 0$, there exists an $h_r \in L_1(I)$ and $C_r \in R$ such that

$$\ell(t,x) \geq h_r(t) \text{ for all } x \in B_r \subset E,$$
$$\Psi(x) \geq C_r \text{ for all } \ x \in B_r \subset E,$$

and the functional φ is weakly lower semicontinuous on $\mathcal{M}_c(I,F)$.

COROLLARY 4.5. *Suppose that all the assumptions of Theorem* 4.2, *except* A2, *hold and that* A2 *is replaced by Assumption* A3 *as stated above. Then there exists an optimal control.*

Remarks on Assumptions A1 *and* A2.

(R1) The lower semicontinuity of ℓ and Ψ on E, as stated in Theorem 4.2, is always satisfied if they are convex and once Gâteaux differentiable.

(R2) One choice of the cost functional φ is

$$\varphi(u) \equiv \Phi(|u|_v)$$

with $\Phi : [0,\infty] \longrightarrow [0,\infty]$, a nondecreasing extended real valued function. With this hypothesis, it is easy to verify [1] that the functional φ is weakly lower semicontinuous on $\mathcal{M}_c(I,F)$. This is a good measure of impulsive energy spent in the control process. Another possibility is

$$\varphi(u) \equiv \Phi(u(\phi_1), u(\phi_2), \ldots, u(\phi_m)),$$

where $\Phi \in C(R^m)$ is a real valued lower semicontinuous function on R^m bounded away from $-\infty$; and $\phi_i \in C_b(I,F^*)$, the space of bounded continuous functions on I with values in F^*, or $\phi_i \in B_m(I,F^*)$, the space of bounded Borel measurable maps from I to F^*. The action of u on ϕ_i is given by

$$u(\phi_i) \equiv \int_I \langle \phi_i(t), u(dt)\rangle.$$

This is a useful functional if one wishes to minimize or maximize control efforts over selected periods of time.

(R3) Assumption A1 is critical. It is satisfied if $C_0(s) \equiv C(s, x_o(s))$ is a compact operator from F to E for μ-almost all $s \in I$. This can be verified as follows. Assume for simplicity that the compactness property holds for all $s \in I$ and not just for μ-a.e. Let $u_n \xrightarrow{w} u^o$. Define $v_n \equiv u_n - u^o$. Clearly then $v_n \xrightarrow{w} 0$. It suffices to show that

$$\langle L_T v_n, e^*\rangle \longrightarrow 0 \tag{38}$$

uniformly with respect to $e^* \in B_1(E^*)$, the unit ball of E^*. This can be proved by establishing a contradiction. Suppose this is false. Then there exists a sequence $\{e_n^*\} \in B_1(E^*)$ and a $\delta > 0$ such that

$$\begin{aligned} |\langle L_T v_n, e_n^*\rangle| &= \left| \int_I \langle C_0^*(t) S^*(T-t) e_n^*, v_n(dt)\rangle \right| \\ &\equiv \left| \int_I \langle \Gamma(t) e_n^*, v_n(dt)\rangle \right| \ \geq \delta. \end{aligned} \tag{39}$$

Since C_0 and hence C_0^* are compact operator valued functions and the semigroup $S(t), t \in I$, is a family of bounded linear operators, Γ is a compact operator valued function on I. Thus Γe_n^* has a Cauchy subsequence (relabeled as such) in the space $L_\infty(I,F^*)$ which is the dual of $L_1(I,F)$ since F^* has the Radon–Nikodým property. Hence, for any $\varepsilon > 0$, there exists an integer $n(\varepsilon)$ such that, for all $n, m > n(\varepsilon)$,

$$\| \Gamma e_n^* - \Gamma e_m^* \|_{L_\infty(I,F^*)} < \varepsilon. \tag{40}$$

Since v_n converges weakly, it is bounded in the norm, and let this bound be $\sup_n |v_n|_v \leq \tilde{M}$. Using this it follows from (39) and (40) that, for $n, m > n(\varepsilon)$, we have

$$0 < \delta \leq \left| \int_I \langle \Gamma(t) e_n^*, v_n(dt)\rangle \right| \leq \tilde{M}\varepsilon + \left| \int_I \langle \Gamma(t) e_m^*, v_n(dt)\rangle \right|. \tag{41}$$

For fixed m, letting $n \to \infty$, we arrive at the inequality $0 < \delta \leq \tilde{M}\varepsilon$. Since $\varepsilon > 0$ is completely arbitrary, this leads to the contradiction that $\delta > 0$, thereby proving the statement. A similar proof applies in the general case, provided that we subtract a μ-null set from the set I, on which Γ lies outside the space of compact operators, and use the fact that by BDS the sequence $\{u_n\}$ is μ continuous.

Apparently, another possibility is that the semigroup $S(t), t > 0$, is compact. But this condition does not seem to lead to compactness of the integral operator L_t as defined in A1. The problem arises from the fact that the measure μ appearing in the Bartle–Dunford–Schwartz theorem (Theorem 4.1) need not be absolutely continuous with respect to the Lebesgue measure. If it were, the compactness of the operator L_t would follow from that of the semigroup $S(t), t > 0$. But this absolute continuity assumption would rule out inclusion of Dirac measures in the class of admissible controls.

A third possibility obtains if the semigroup $S(t), t > 0$, is nuclear (and thus compact). Suppose the dual pair of Banach spaces $\{E, E^*\}$ admits a biorthogonal basis $\{e_n, e_n^*\}$. If the semigroup $S(t), t > 0$, is nuclear, it has the representation

$$S(t) = \sum_{k=1}^{\infty} \lambda_k(t) e_k \bigotimes e_k^*, \tag{42}$$

where $e_k \bigotimes e_k^* \in \mathcal{L}(E)$ and for each $\xi \in E$, $(e_k \bigotimes e_k^*)(\xi) = (e_k^*, \xi)e_k$. The sequence $\{\lambda_k\}$ is a family of continuous scalar valued functions satisfying the functional equation

$$\lambda_k(t+s) = \lambda_k(t)\lambda_k(s), \lambda_k(0) = 1 \text{ for all } t, s \geq 0, k \geq 1. \tag{43}$$

The reader can easily verify that these conditions guarantee that $S(t), t \geq 0$, is a C_0-semigroup on E. Now returning to the operator L_t (see Assumption A1), we have

$$L_t(u_n - u_o) = \int_0^t S(t-s)C(s, x(s))(u_n - u_o)(ds) \tag{44}$$

$$= \sum_{k\geq 1} \left\{ \int_0^t \lambda_k(t-s) \langle C^*(s, x(s))e_k^*, u_n(ds) - u_o(ds) \rangle_{F^*,F} \right\} e_k \tag{45}$$

$$= \sum_{k\geq 1} \alpha_{k,n}(t)e_k = \sum_{k=1}^{m} \alpha_{k,n}(t)e_k + \sum_{k\geq m+1} \alpha_{k,n}(t)e_k \tag{46}$$

for any sequence $\{u_n, u_o\} \in \mathcal{U}_{ad}$. Since the semigroup is nuclear, it is clear that $\sum_{k\geq 1} |\lambda_k(t)| < \infty$ for $t > 0$. In view of the functional equation (43), we must have

$$\lambda_k(t) = e^{\varpi_k t}, \qquad \varpi_k \in (R/C).$$

Nuclearity of the semigroup demands that the exponents $\{\varpi_k\}$ have only finitely many terms with positive real parts and that $\varpi_k \to -\infty$. From this it follows that

$$\sum_{k\geq 1} \sup_{t\in(0,T]} |\lambda_k(t)| = \sum_{k\geq 1} \sup_{t\in(0,T]} \lambda_k(t) < \infty. \tag{47}$$

Further, for each $x \in B(I, E)$, the operator valued function $C(t, x(t))$ is also bounded in the sense that

$$\sup\{\| C(t, x(t)) \|_{\mathcal{L}(F,E)}, t \in I\} < \infty.$$

Since $\mathcal{U}_{ad}$ is bounded, the sequence $\{u_n, u_o\}$ is also bounded. Suppose $u_n \xrightarrow{w} u_o$ in $\mathcal{M}_c(I, F)$. Then in view of (47), for each $\varepsilon > 0$ one can choose $m \in N$ sufficiently large but finite so that the second term in the expression (46) has E-norm less than ε. Then letting $n \to \infty$, the first term of (46) converges to zero strongly in E. Since $\varepsilon > 0$, is arbitrary, this shows that nuclearity of the semigroup is a sufficient condition for weak to strong continuity of the operator L_t from $\mathcal{M}_c(I, F)$ to E.

5. Some open questions.

(Q1) Our main results, Theorems 4.2 and 4.3 and Corollary 4.4, are based on the assumption that the operator $L_t, t \in I$, maps a weakly convergent sequence of $\mathcal{M}_c(I,F)$ into a strongly convergent sequence in E. We have presented some remarks with regards to sufficient conditions that guarantee this hypothesis. To the knowledge of the author, this seems to be an open problem in the theory of vector measures. The general problem can be stated as follows. Under what conditions a weakly compact operator $L \in \mathcal{L}(\mathcal{M}_c(I,F),E)$ is actually compact in the sense that it maps every weakly convergent sequence from the Banach space $\mathcal{M}_c(I,F)$ into a strongly convergent sequence in the Banach space E. For closely related problems see [7, Chapters IV, VI]. For similar problems with respect to operators $T \in \mathcal{L}(L_1(S,\Sigma,\mu),E)$ see Dunford and Schwartz [8, Theorem VI.8.12].

(Q2) Our first result on existence of optimal control given by Theorem 4.2 depends on the decomposability property (P2) of the set $\Gamma^{-1}(\mathcal{U}_{ad})$. Thus the full generality of the Bartle–Dunford–Schwartz theorem is not utilized, although we have stated a sufficient condition under which this property holds. The decomposability assumption is disposed of in Theorem 4.3 and Corollary 4.4 by imposing some stronger conditions on the operator C. It would be desirable to relax these assumptions further.

(Q3) Frequent switching of controls is costly. It would be interesting to include the cost of switching. This requires a functional that maps the control space into the set of nonnegative integers signifying the number of jumps in the control used. In general this seems to be a difficult problem. As mentioned by one of the reviewers, controls cannot be changed too frequently since otherwise it should be possible to stabilize any system. This is certainly true. In optimization problems, however, given the maximum admissible number of switchings, one may find an optimal control with fewer switchings, depending on the relative weights given to cost of switchings and the running cost.

6. Examples. For illustration we present two examples.

Example 1. The dynamics of transverse vibration of a mast attached to a rigid body satellite on one end and carrying an antenna at the free end is described by the following system of equations [15, Li, p. 138]:

$$
\begin{aligned}
&W_{tt}(t,\xi) + D_\xi^4 W + f_1(t,\xi,W,W_t) = g_1(t,\xi,W), \qquad \xi \in \Omega \equiv (0,1),\\
&W(t,0) = D_\xi W(t,1) = 0,\\
&W_{tt}(t,1) = D_\xi^3 W(t,1) + C_1 u_1,\\
&(D_\xi W)_{tt} = -D_\xi^2 W(t,1) + C_2 u_2,
\end{aligned}
\tag{48}
$$

where D_ξ^k denotes the spatial derivative and $()_t$ and $()_{tt}$ denote the first and second partials with respect to the time variable. The function W denotes the transverse displacement. This is a normalized system where system parameters such as the mass density, the flexural rigidity, and the length of the mast have all been set equal to one. This is purely for convenience of presentation. Here f_1 takes into account all nonconservative forces, including damping. The functions $\{g_1, C_1u_1, C_2u_2\}$ are external forces: the first one distributed along the body of the mast, and the second and the third representing the shear and torsional forces applied at the end of the mast. We shall model them as impulsive forces as explained later. The operators C_1 and C_2 may depend nonlinearly on the state at the boundary and determine the effectiveness of control forces on the motion of the system. Define $y(t) \equiv (y_1(t), y_2(t), y_3(t))' \equiv$

$(W(t,\cdot), W(t,1), D_\xi W(t,1))'$ and A as the differential operator

$$Ay = (-D_\xi^4 W, D_\xi^3 W(t,1), -D_\xi^2 W(t,1))'$$

subject to the clamped boundary conditions at the fixed end attached to the spacecraft body, as given by the second set of expressions above. Using this operator, we can write this system as a second order evolution equation,

$$(d^2y/dt^2) = Ay + \tilde{f}_1(t,y,\dot{y}) + \tilde{g}_1(t,y) + \tilde{C}(t,y)u, \tag{49}$$

in the Hilbert space $H \equiv L_2(\Omega) \times R \times R$, where $\tilde{f}_1 \equiv (f_1, 0, 0)'$, $\tilde{g}_1 \equiv (g_1, 0, 0)'$, and $\tilde{C}u \equiv (0, C_1u_1, C_2u_2)$. Note that the domain of the operator A is given by

$$D(A) = \{\phi \in H^4(\Omega) \times R \times R : \phi(0) = D_\xi\phi(0) = 0, \phi(1) = \phi_2, D_\xi\phi(1) = \phi_3\}$$

and that $-A$ is a positive unbounded self-adjoint operator in the Hilbert space H. Using the standard approach, we may convert this into a first order system. Defining $x_1 = y$ and $x_2 = \dot{y}$ and $x \equiv (x_1, x_2)'$, we can rewrite this as a first order evolution equation:

$$(dx/dt) = \mathcal{A}x + f(t,x) + g(t,x) + C(t,x)u, \tag{50}$$

with the operators $\{\mathcal{A}, f, g, C\}$ given by

$$\mathcal{A} = \begin{pmatrix} 0 & I \\ A & 0 \end{pmatrix}, \quad f = \begin{pmatrix} 0 \\ \tilde{f} \end{pmatrix}, \quad g = \begin{pmatrix} 0 \\ \tilde{g} \end{pmatrix}, \quad Cu = \begin{pmatrix} 0 \\ \tilde{C}u \end{pmatrix}. \tag{51}$$

In view of the physical nature of the problem we may choose the energy space as the state space

$$E \equiv D(\sqrt{-A}) \times H$$

with the natural scalar product given by

$$(x,z)_E = (x_1, z_1)_{D(\sqrt{-A})} + (x_2, z_2)_H.$$

The domain of the operator $\mathcal{A}$ is given by

$$D(\mathcal{A}) = D(A) \times D(\sqrt{-A}).$$

It is easy to show that $\mathcal{A}$ is skew adjoint and hence that $i\mathcal{A}$ is self-adjoint, and it follows from Stone's theorem that $\mathcal{A}$ generates a unitary group of operators $S(t), t \in R$, in E. Returning to the problem, it is well known that satellites in space are subject to disturbances caused by forces such as solar pressure, geomagnetic forces, and impulsive forces imparted by bombardment of micrometeorites. To control vibration, thrusters or jets are used for producing shear force and torsional moments (a twister). These thrusters are fired on and off for very short intervals of time, like impulses imparting a burst of energy to the mast. Thus a more appropriate formulation of the system, replacing (50), is given by

$$dx = \mathcal{A}xdt + f(t,x)dt + g(t,x)\nu(dt) + C(t,x)u(dt), \tag{52}$$

where ν may represent the impulsive forces caused by micrometeorites and u is the impulsive control generated by thruster firings. Here $F = R^2$, and $C(t,x) \in \mathcal{L}(F,E)$. The objective functional may be taken as

$$J(u) = (1/2)\int_I \| x(t) \|_E^2 \, dt + \Phi(|u|_v),$$

where Φ is any monotone nondecreasing function on the positive half of the real line with $\Phi(0) = 0$. Minimizing this cost functional is equivalent to minimizing vibration. Clearly this is a special case and our theory applies.

Example 2. Phased array pulsed radars are used to search and track targets simultaneously. This is possible because the phased array radars use an array of transmit-receive antennas which are distributed around the operating site. Thus it can switch back and forth from one antenna to the other and scan the whole horizon in a matter of seconds, and all these actions are done by electronic switching, not mechanically (as in the old system). Hence one radar station can search, detect, and track multiple targets at the same time. The antennas transmit electromagnetic energy in the form of sharp pulses which are generated at the site, and also receive the signals reflected back from the targets. These electrical signals are governed by Maxwell's equations. Using the vector and scalar potentials denoted by (a,ϕ) and the Lorentz gauge, the Maxwell's equation is given by a system of wave equations:

$$\partial^2 a/\partial t^2 - (1/\mu\epsilon)\triangle a = (1/\epsilon)i, \qquad t \geq 0, \quad \xi \in \Omega \equiv R^3, \tag{53}$$

$$\partial^2 \phi/\partial t^2 - (1/\mu\epsilon)\triangle \phi = (1/\mu\epsilon^2)\rho, \qquad t \geq 0, \quad \xi \in \Omega, \tag{54}$$

where i and ρ are the sources, the first denoting the current density (vector) and the second the charge density. These are the sources that can be controlled by radar operators in terms of short bursts of pulses, thereby producing sharp and short bursts of electromagnetic energy directed to the target. Define $H \equiv L_2(\Omega, R^3) \times L_2(\Omega, R)$, denote $y \equiv (a,\phi)'$, and define the formal differential operator B by $By \equiv -(1/\mu\epsilon)(\triangle a, \triangle\phi)'$. Then introduce the operator A as follows:

$$\begin{aligned} &D(A) \equiv \{y \in H : By \in H\} = H^2(\Omega, R^3) \times H^2(\Omega, R) \\ &\text{and set } Az = Bz \text{ for } z \in D(A). \end{aligned} \tag{55}$$

Again, one can verify that $-A$ is an unbounded positive self-adjoint operator in H.

Define the state space as $E \equiv D(\sqrt{-A}) \times H$ and the state as $x = (y, \dot{y})'$. Then following exactly the procedure in the first example, one defines the operator $\mathcal{A}$ as in (51). The control operator C is a constant (8×4) matrix with the first four rows being all zero, while the last (4×4) matrix is a diagonal matrix with the first three entries being equal to $(1/\mu\epsilon)$ and the last one being $(1/\mu\epsilon^2)$. This leads to the abstract differential equation

$$dx = \mathcal{A}x\,dt + C(t)u(dt), \qquad t \geq 0, \tag{56}$$

on the Hilbert space E. Again $\mathcal{A}$ is the infinitesimal generator of a unitary group of operators $S(t), t \in R$, on E.

For the control space F, one may choose $F \equiv L_2(\Omega_0, R^3) \times L_2(\Omega_0, R)$, where Ω_0 is any open bounded (connected) subset of Ω representing the domain of the electromagnetic source and the wave guides. Let $\mathcal{T}(t)$ denote the surface area of the target as seen by the radar at any moment of time $t \in I$ and σ denote the surface

measure. We may assume that $t \longrightarrow \mathcal{T}(t)$ is continuous in the Hausdorff metric. Then the electromagnetic energy hitting the target at time t is given by the functional

$$\begin{aligned}\ell(x(t)) &\equiv \int_{\mathcal{T}(t)} \{|E(t,\xi)|^2 + |B(t,\xi)|^2\} d\sigma(\xi) \\ &\equiv \int_{\mathcal{T}(t)} \{|\dot{a} + \nabla\phi|^2 + |\nabla \times a|^2\} d\sigma(\xi),\end{aligned}$$

where in the last expression we have used the fact that $E = -\dot{a} - \nabla\phi$ and $B = \nabla \times a$ [17]. Using the energy norm, one can easily verify that

$$\ell(x(t)) \leq \| x(t) \|_E^2.$$

The objective is to maximize the energy delivered to the target so that the reflected energy received by the receiving antenna is maximized. Hence the cost functional for this problem can be taken as

$$J(u) \equiv \int_I -\ell(x(t))dt + \Phi(|u|_v).$$

Our problem is to find a control that minimizes this functional. Since ℓ is continuous, the existence follows from Corollary 4.4.

Similar models involving Maxwell's equation arise in the field of optical communication [16, 17], where the optical beams are modulated by message signals to be transmitted by optical fibers. Since optical spectrum is extraordinarily wide, extra-high-speed data transmission is now possible. These data bits entering the fiber network could be considered as a train of impulses. The response of the optical devices and the network to such ultra-high-speed data traffic is an interesting area where the theory of impulsive systems may find interesting applications.

Similar examples can be found in population dynamics involving reaction, convection, and diffusion, where unexpected events may occur and last for a very short time. Considering biochemical time scales, such events may be considered impulsive and may require impulsive controls.

Acknowledgments. The author would like to thank the anonymous reviewers for their valuable comments and suggestions, which led to significant improvement of the original presentation of this paper.

REFERENCES

[1] N.U. Ahmed, *Vector measures for optimal control of impulsive systems in Banach spaces*, Nonlinear Funct. Anal. Appl., 5(2) (2000), pp. 95–106.

[2] N.U. Ahmed, *Some remarks on the dynamics of impulsive systems in Banach spaces*, Dyn. Contin. Discrete Impuls. Syst. Ser. A Math. Anal., 8 (2001), pp. 261–274.

[3] N.U. Ahmed, *State dependent vector measures as feedback controls for impulsive systems in Banach spaces*, Dyn. Contin. Discrete Impuls. Syst. Ser. B Appl. Algorithms, 8 (2001), pp. 251–261.

[4] N.U. Ahmed, *Measure solutions for impulsive systems in Banach spaces and their control*, Dynam. Contin. Discrete Impuls. Systems, 6 (1999), pp. 519–535.

[5] N.U. Ahmed, *Optimal impulse control for impulsive systems in Banach spaces*, Int. J. Differ. Equ. Appl., 1 (2000), pp. 37–52.

[6] N.U. Ahmed, *Necessary conditions of optimality for impulsive systems on Banach spaces*, Nonlinear Anal., 51 (2002), pp. 409–424.

[7] J. Diestel and J.J. Uhl, Jr., *Vector Measures*, Math. Surveys Monogr. 15, AMS, Providence, RI, 1977.

[8] N. Dunford and J.T. Schwartz, *Linear Operators*, Part I, John Wiley and Sons, New York, London, 1988.
[9] X. Fu and B. Yan, *The global solutions of impulsive functional differential equations in Banach spaces*, Nonlinear Stud., 1 (2000), pp. 1–17.
[10] V. Lakshmikantham, D.D. Bainov, and P.S. Simenov, *Theory of Impulsive Differential Equations*, World Scientific, Singapore, London, 1989.
[11] J.H. Liu, *Nonlinear impulsive evolution equations*, Dynam. Contin. Discrete Impuls. Systems, 6 (1999), pp. 77–85.
[12] K. Nakagawa, *Existence of a global solution for an impulsive semi linear parabolic equation and its asymptotic behavior*, Commun. Appl. Anal., 4 (2000), pp. 403–409.
[13] A.M. Samoilenko and N.A. Perestyuk, *Impulsive Differential Equations*, World Scientific, Singapore, 1995.
[14] T. Yang, *Impulsive Control Theory*, Springer-Verlag, Berlin, 2001.
[15] L. Peng, *Controllability, Stability, and Stabilizability of Distributed Parameter Systems*, Ph.D thesis, SITE, University of Ottawa, Ottawa, ON, Canada, 1991.
[16] A.Yariv, *Optical Electronics in Modern Communications*, The Oxford Series in Electrical and Computer Engineering, Oxford University Press, New York, Oxford, 1997.
[17] S.L. Chuang, *Physics of Optoelectronic Devices*, Wiley Ser. Pure Appl. Optics, John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore, 1995.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 686–708

VARIATIONAL INEQUALITIES FOR COMBINED CONTROL AND STOPPING*

HIROAKI MORIMOTO†

Abstract. We study the variational inequality associated with the combined stochastic control problem and establish the existence of a unique viscosity solution without uniform ellipticity. An application to the quasi-variational inequality is given.

Key words. variational inequality, viscosity solution, combined control, quasi-variational inequality

AMS subject classifications. 49C20, 49J40, 49L25, 60G40, 93E20

PII. S0363012901369041

1. Introduction.

We are concerned with the variational inequality in the combined control problem and stopping for diffusions. We are given two bounded continuous functions f, g, and $a > 0$. The variational inequality we deal with can be written as

$$\begin{aligned} &-Lv + f - \frac{a}{4}|Dv|^2 \geq 0, \\ &v \leq g \qquad\qquad \text{in } \mathbf{R^N}, \\ &\left(-Lv + f - \frac{a}{4}|Dv|^2\right)(v-g)^- = 0, \end{aligned} \tag{1}$$

where $L = L_0 + \alpha$, L_0 denotes the second order differential operator

$$L_0 = -\frac{1}{2}\mathrm{tr}(\sigma\sigma^* D^2) - b \cdot D$$

for two Lipschitz continuous functions $b(x)$ and $\sigma(x)$ on $\mathbf{R^N}$, taking values in $\mathbf{R^N}$ and $\mathbf{R^N} \otimes \mathbf{R^N}$, α is a positive constant, $|\cdot|$ is the Euclidean norm, σ^* is the transpose of σ, $x^- = \max(-x, 0)$, and $D = (\partial/\partial x_1, \ldots, \partial/\partial x_N)$.

If $-L_0$ is uniformly elliptic and $a = 0$, the variational inequality (1) has been studied by Bensoussan [2], Bensoussan and Lions [3], Friedman [6], and Kinderlehrer and Stampacchia [9]. We also refer to Menaldi [10, 11] for the variational inequality associated with the stopping time problem for degenerate diffusions and Karatzas and Wang [8] for the combined control problem applied to mathematical finance.

The purpose of this paper is to show the existence of a unique viscosity solution of the variational inequality (1) without uniform ellipticity and then characterize the solution v. From the point of view of Bensoussan and Lions [3] and a simple relation

$$\min_c(a^{-1}|c|^2 + c \cdot Dv) = -\frac{a}{4}|Dv|^2,$$

this variational inequality is relevant to the combined stochastic control problem to minimize the cost:

$$J(\theta, c) = E\left[\int_0^\theta e^{-\alpha t}\left\{f(x_t) + a^{-1}|c_t|^2\right\}dt + e^{-\alpha\theta}g(x_\theta)\right] \tag{2}$$

*Received by the editors November 27, 2001; accepted for publication (in revised form) October 26, 2002; published electronically May 29, 2003.
http://www.siam.org/journals/sicon/42-2/36904.html

†Department of Mathematical Sciences, Faculty of Science, Ehime University, Matsuyama 790-0826, Japan (morimoto@sci.sci.ehime-u.ac.jp).

over $\mathcal{S} \times \mathcal{A}$ subject to the stochastic differential equation

$$dx_t = [b(x_t) + c_t]dt + \sigma(x_t)dW_t, \qquad x_0 = x \in \mathbf{R^N}, \tag{3}$$

where W_t is an N-dimensional standard Brownian motion defined on a complete probability space $(\Omega, \mathcal{F}, P)$ endowed with the natural filtration $\mathcal{F}_t$ generated by $\sigma(W(s), s \le t)$, $\mathcal{S}$ is the class of all stopping times θ, and $\mathcal{A}$ denotes the class of all N-dimensional $\mathcal{F}_t$-progressively measurable processes $c = (c_t)$ with $E[\int_0^\infty e^{-\alpha s}|c_s|^2 ds] < \infty$. For simplicity, we take $a = 1$ throughout the paper.

The content of this paper is as follows. In section 2 we study the penalized problem. We show that the penalty equation admits a unique viscosity solution u_ε. In section 3 we give the definition of the viscosity solutions to the variational inequality (1). It is shown that u_ε converges to a unique viscosity solution of (1). In section 4 we study the quasi-variational inequality associated with impulsive control.

2. Penalized problem.

2.1. Existence.

We consider the equation of the form

$$u(x) = \inf_c E\left[\int_0^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}\left\{\left(f + \frac{1}{\varepsilon}u \wedge g\right)(x_t) + |c_t|^2\right\}dt\right], \qquad \varepsilon > 0. \tag{4}$$

Let $\mathcal{C} = \mathcal{C}(\mathbf{R^N})$ denote the Banach space of all bounded uniformly continuous functions h on $\mathbf{R^N}$ with norm $\|h\| = \sup_x |h(x)|$, and $\mathcal{C}_+ = \{h \in \mathcal{C} : \ h \ge 0\}$.

We assume that

$$b,\ \sigma : \text{ Lipschitz continuous, bounded,} \tag{5}$$

$$\alpha > \nu := \sup\left\{\operatorname{tr}\left[\frac{(\sigma(x) - \sigma(y))(\sigma(x) - \sigma(y))^*}{|x-y|^2}\right] + \frac{2(x-y)\cdot(b(x) - b(y))}{|x-y|^2} : \ x,\ y \in \mathbf{R^N},\ x \ne y\right\}, \tag{6}$$

and

$$f,\ g \in \mathcal{C}_+. \tag{7}$$

THEOREM 2.1. *Under* (5), (6), *and* (7), *equation* (4) *admits a unique solution* $u \in \mathcal{C}_+$.

Proof. We first note that $\mathcal{C}_+$ is a closed subset of $\mathcal{C}$. Define

$$Tw(x) = \inf_c E\left[\int_0^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}\left\{\left(f + \frac{1}{\varepsilon}w \wedge g\right)(x_t) + |c_t|^2\right\}dt\right] \quad \text{for } w \in \mathcal{C}. \tag{8}$$

We shall show

$$T : \ \mathcal{C}_+ \ \to \ \mathcal{C}_+. \tag{9}$$

It is easy to see that

$$\begin{aligned} 0 \le Tw(x) &\le E\left[\int_0^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}\left\{\left(f + \frac{1}{\varepsilon}g\right)(\bar{x}_t)\right\}dt\right] \\ &\le \frac{\varepsilon}{\alpha\varepsilon + 1}\left(\|f\| + \frac{1}{\varepsilon}\|g\|\right), \qquad w \in \mathcal{C}_+, \end{aligned}$$

for the response $\bar{x}_t$ to $c_t = 0$. Moreover, it follows from (8) that

$$\begin{aligned}
&|Tw(x) - Tw(y)| \\
&\quad \le \sup_c E\left[\int_0^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}\left\{|f(x_t) - f(y_t)| + \frac{1}{\varepsilon}(|w(x_t) - w(y_t)| + |g(x_t) - g(y_t)|)\right\} dt\right] \\
&\quad \le I_f + \frac{1}{\varepsilon}(I_w + I_g),
\end{aligned}$$

where y_t is the solution of (3) with $y_0 = y$, and

$$I_h = \sup_c E\left[\int_0^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}|h(x_t) - h(y_t)|dt\right] \quad \text{for} \quad h \in \mathcal{C}.$$

By (6) we can choose $\eta > 0$ such that $-\alpha + \nu + \eta < 0$, and Ito's formula applied to the function $|x|^2 e^{(-\alpha+\eta)t}$ gives

$$E[|x_t - y_t|^2 e^{(-\alpha+\eta)t}] \le |x - y|^2. \tag{10}$$

Also, we note that there exists a constant $C_{\zeta,h} > 0$, for any $\zeta > 0$, such that

$$|h(x) - h(y)| \le \zeta + C_{\zeta,h}|x - y|, \qquad x, y \in \mathbf{R}^{\mathbf{N}}. \tag{11}$$

Then

$$\begin{aligned}
I_h &\le \sup_c E\left[\int_0^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}(\zeta + C_{\zeta,h}|x_t - y_t|)dt\right] \\
&\le \zeta/\alpha + 2C_{\zeta,h}|x - y|/(\alpha + \eta). \qquad (12)
\end{aligned}$$

Thus, letting $\delta \to 0$ and $\zeta \to 0$, we get

$$\lim_{\delta\to 0} \sup_{|x-y|<\delta} I_h = 0, \tag{13}$$

and hence

$$\lim_{\delta\to 0} \sup_{|x-y|<\delta} |Tw(x) - Tw(y)| = 0,$$

which implies (9).

Now, we have by (8)

$$\begin{aligned}
|Tw_1(x) - Tw_2(x)| &\le \sup_c E\left[\int_0^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}\left(\frac{1}{\varepsilon}|w_1 \wedge g(x_t) - w_2 \wedge g(x_t)|\right) dt\right] \\
&\le \sup_c E\left[\int_0^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}\left(\frac{1}{\varepsilon}|w_1(x_t) - w_2(x_t)|\right) dt\right] \\
&\le \frac{1}{\alpha\varepsilon + 1}\|w_1 - w_2\|.
\end{aligned}$$

Therefore T is a contraction mapping, which completes the proof.

2.2. Viscosity solutions. We are here concerned with the penalty equation:

$$-Lu + f - \frac{1}{\varepsilon}(u-g)^+ - \frac{1}{4}|Du|^2 = 0 \quad \text{in } \mathbf{R^N}. \tag{14}$$

DEFINITION 2.2. *$w \in \mathcal{C}$ is called a viscosity solution of* (14) *if the following assertions are satisfied:*

For any $\varphi \in C^2$ and any local maximum point z of $w - \varphi$,

$$-\alpha w(z) - L_0\varphi(z) + f(z) - \frac{1}{\varepsilon}(w-g)^+(z) - \frac{1}{4}|D\varphi(z)|^2 \geq 0. \tag{15}$$

For any $\varphi \in C^2$ and any local minimum point z of $w - \varphi$,

$$-\alpha w(z) - L_0\varphi(z) + f(z) - \frac{1}{\varepsilon}(w-g)^+(z) - \frac{1}{4}|D\varphi(z)|^2 \leq 0. \tag{16}$$

As is well known in Fleming and Soner [5] and Crandall, Ishii, and Lions [4], Definition 2.2 turns out to be equivalent to the following.

DEFINITION 2.3. *$w \in \mathcal{C}$ is called a viscosity solution of* (14) *if the following assertions are satisfied:*

$$-\alpha w + \frac{1}{2}\mathrm{tr}(\sigma\sigma^* X) + b\cdot p + f - \frac{1}{\varepsilon}(w-g)^+ - \frac{1}{4}|p|^2 \geq 0 \quad \forall (p,X) \in J^{2,+}w(x), \quad \forall x \in \mathbf{R^N}, \tag{17}$$

$$-\alpha w + \frac{1}{2}\mathrm{tr}(\sigma\sigma^* X) + b\cdot p + f - \frac{1}{\varepsilon}(w-g)^+ - \frac{1}{4}|p|^2 \leq 0 \quad \forall (p,X) \in J^{2,-}w(x), \quad \forall x \in \mathbf{R^N}, \tag{18}$$

where $J^{2,+}$ and $J^{2,-}$ are the second order superjets and subjets defined by

$$J^{2,+}w(x) = \left\{ (p,X) \in \mathbf{R^N} \times \mathbf{S^N} : \limsup_{y\to x} \frac{w(y) - w(x) - p\cdot(y-x) - \frac{1}{2}X(y-x)\cdot(y-x)}{|y-x|^2} \leq 0 \right\},$$

$$J^{2,-}w(x) = \left\{ (p,X) \in \mathbf{R^N} \times \mathbf{S^N} : \liminf_{y\to x} \frac{w(y) - w(x) - p\cdot(y-x) - \frac{1}{2}X(y-x)\cdot(y-x)}{|y-x|^2} \geq 0 \right\},$$

and $\mathbf{S^N}$ is the space of symmetric $N \times N$ matrices.

In order to show that u is a viscosity solution of (14), we define $u_k \in \mathcal{C}$ by

$$u_k(x) = \inf\left\{ E\left[\int_0^\infty e^{-(\alpha+\frac{1}{\varepsilon})t} \left\{ F(x_t) + |c_t|^2 \right\} dt \right] :\ c \in \mathcal{A}_k \right\} \tag{19}$$

for every $k > 0$, where $\mathcal{A}_k = \{c \in \mathcal{A} :\ |c_t| \leq k\ \forall t\}$ and

$$F = f + \frac{1}{\varepsilon} u \wedge g \in \mathcal{C}_+. \tag{20}$$

LEMMA 2.4. *Under* (5), (6), *and* (7), *u_k is a viscosity solution of*

$$-\left(\alpha + \frac{1}{\varepsilon}\right) u_k - L_0 u_k + F + \min_{|c|\leq k}(|c|^2 + c\cdot Du_k) = 0. \tag{21}$$

Proof. We recall a standard result of the theory of viscosity solutions in Fleming and Soner [5, Thm. 3.1, p. 220] and Soner [13, pp. 149–151]. To prove (15) and (16), it is sufficient to show that the dynamic programming principle holds, i.e.,

$$u_k(x) = \inf\left\{ E\left[\int_0^\theta e^{-(\alpha+\frac{1}{\varepsilon})t}\{F(x_t)+|c_t|^2\}dt + e^{-(\alpha+\frac{1}{\varepsilon})\theta}u_k(x_\theta)\right]:\ c\in\mathcal{A}_k \right\} \tag{22}$$

for any $\theta \in \mathcal{S}$, which may depend on $c \in \mathcal{A}_k$. Indeed, we denote by $u^r(x)$ the right-hand side of (22) and we may suppose that θ is bounded. By (19) we have

$$\begin{aligned} u_k(x) &= \inf\Bigg(E\Bigg[\int_0^\theta e^{-(\alpha+\frac{1}{\varepsilon})t}\Big\{F(x_t)+|c_t|^2\Big\}dt \\ &\qquad + \int_\theta^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}\Big\{F(x_t)+|c_t|^2\Big\}dt\Bigg]\Bigg) \\ &= \inf\Bigg(E\Bigg[\int_0^\theta e^{-(\alpha+\frac{1}{\varepsilon})t}\Big\{F(x_t)+|c_t|^2\Big\}dt \\ &\qquad + e^{-(\alpha+\frac{1}{\varepsilon})\theta}\int_0^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}\Big\{F(\tilde{x}_t)+|\tilde{c}_t|^2\Big\}dt\Bigg]\Bigg) \\ &\geq u^r(x), \end{aligned}$$

where

$$\begin{aligned} d\tilde{x}_t &= [b(\tilde{x}_t)+\tilde{c}_t]dt + \sigma(\tilde{x}_t)d\tilde{W}_t, \qquad \tilde{x}_0 = x_\theta, \\ \tilde{c}_t &= c_{t+\theta}. \end{aligned}$$

To prove the reverse inequality, let $\zeta > 0$ be arbitrary and we set

$$U_c(x) = E\left[\int_0^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}\Big\{F(x_t)+|c_t|^2\Big\}dt\right], \qquad c\in\mathcal{A}_k. \tag{23}$$

By analogy with (12) and (13), we observe that there exists $\delta > 0$ (independent of c) such that

$$|x-y|<\delta \quad\Rightarrow\quad |u_k(x)-u_k(y)| \leq \sup_c |U_c(x)-U_c(y)| \leq \zeta. \tag{24}$$

We consider a sequence $\{S_i\}$ of disjoint subsets of $\mathbf{R^N}$ such that

$$\operatorname{diam}(S_i) < \delta \quad \text{and} \quad \bigcup_i S_i = \mathbf{R^N}.$$

For any i, we take $x_i \in S_i$ and $c^{(i)} \in \mathcal{A}_k$ such that

$$U_{c^{(i)}}(x_i) \leq \inf_c U_c(x_i) + \zeta. \tag{25}$$

Define $c^\theta \in \mathcal{A}_k$ by

$$c^\theta_t = c_t 1_{\{t<\theta\}} + c^{(i)}_{t-\theta} 1_{\{t\geq\theta\}}$$

for $x_\theta \in S_i$. Hence, by (24) and (25)

$$\begin{aligned} U_{c^{(i)}}(x_\theta) &= U_{c^{(i)}}(x_\theta) - U_{c^{(i)}}(x_i) + U_{c^{(i)}}(x_i) \\ &\le \zeta + U_{c^{(i)}}(x_i) \\ &\le 2\zeta + \inf_c U_c(x_i) \\ &= 2\zeta + u_k(x_i) \\ &\le 3\zeta + u_k(x_\theta). \end{aligned}$$

Now, we can find $c \in \mathcal{A}_k$ such that

$$u^r(x) + \zeta \ge E\left[\int_0^\theta e^{-(\alpha+\frac{1}{\varepsilon})t}\{F(x_t) + |c_t|^2\}dt + e^{-(\alpha+\frac{1}{\varepsilon})\theta}u_k(x_\theta)\right].$$

Thus

$$\begin{aligned} u^r(x) + \zeta &\ge \sum_i E\left[\int_0^\theta e^{-(\alpha+\frac{1}{\varepsilon})t}\{F(x_t) + |c_t|^2\}dt + e^{-(\alpha+\frac{1}{\varepsilon})\theta}(U_{c^{(i)}}(x_\theta) - 3\zeta) : x_\theta \in S_i\right] \\ &\ge E\left[\int_0^\theta e^{-(\alpha+\frac{1}{\varepsilon})t}\{F(x_t^\theta) + |c_t^\theta|^2\}dt + \int_\theta^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}\{F(x_t^\theta) + |c_t^\theta|^2\}dt\right] - 3\zeta \\ &= U_{c^\theta}(x) - 3\zeta \\ &\ge u_k(x) - 3\zeta, \end{aligned}$$

where

$$dx_t^\theta = [b(x_t^\theta) + c_t^\theta]dt + \sigma(x_t^\theta)dW_t, \qquad x_0^\theta = x.$$

Consequently we deduce $u^r(x) \ge u_k(x)$, completing the proof of (22).

LEMMA 2.5. *Under* (5), (6), *and* (7), *we have*

$$u_k \quad \to \quad u \quad \textit{locally uniformly in } \mathbf{R^N}.$$

Proof. By (4) and (19), it is clear that $u_k \ge u$. By Dini's theorem, it suffices to show that

$$u_k(x) \quad \downarrow \quad u(x) \quad \text{as } k \to \infty \ \text{ for each } x. \tag{26}$$

Let $c_t^k = c_t 1_{\{|c_t| \le k\}}$ for any $c \in \mathcal{A}$ and x_t^k be the solution of

$$dx_t^k = [b(x_t^k) + c_t^k]dt + \sigma(x_t^k)dW_t, \qquad x_0^k = x.$$

Applying Ito's formula and localizing the stochastic integral, we observe by (5) and (6) that for any $\theta \in \mathcal{S}$,

$$\begin{aligned} E[e^{-\alpha\theta}|x_\theta|^2] &\le E\left[\int_0^\theta e^{-\alpha t}\left\{-\alpha|x|^2 + 2x\cdot(b(x)+c_t) + \mathrm{tr}(\sigma\sigma^*(x))\right\}\Big|_{x=x_t} dt\right] \\ &\le E\left[\int_0^\theta e^{-\alpha t}\left\{-\frac{\alpha}{2}|x_t|^2 + \frac{4}{\alpha}(\sup_x |b(x)|^2 + |c_t|^2) + \sup_x |\sigma(x)|^2\right\}dt\right] < \infty \end{aligned}$$

and

$$E[e^{-\alpha\theta}|x_\theta^k - x_\theta|^2] \le E\left[\int_0^\theta e^{-\alpha t}\{(-\alpha+\nu)|x_t^k - x_t|^2 + 2(x_t^k - x_t)\cdot(c_t^k - c_t)\}dt\right]$$
$$\le E\left[\int_0^\theta e^{-\alpha t}\{-\eta|x_t^k - x_t|^2 + 2(x_t^k - x_t)\cdot(c_t^k - c_t)\}dt\right]$$
$$\le E\left[\int_0^\theta e^{-\alpha t}\left\{-\frac{\eta}{2}|x_t^k - x_t|^2 + \frac{2}{\eta}|c_t^k - c_t|^2\right\}dt\right].$$

By (4) there exists $c \in \mathcal{A}$, for any $\zeta > 0$, such that $u(x) + \zeta > U_c(x)$. Then, by (11), (23), and (20)

$$|U_c(x) - U_{c^k}(x)| \le E\left[\int_0^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}\{\zeta + C_{\zeta,F}|x_t - x_t^k| + |c_t|^2 - |c_t^k|^2\}dt\right]$$
$$\le \zeta/\alpha + C_{\zeta,F}\left(E\left[\int_0^\infty e^{-\alpha t}|x_t - x_t^k|^2 dt\right]\right)^{1/2}\left(\int_0^\infty e^{-\alpha t}dt\right)^{1/2}$$
$$+E\left[\int_0^\infty e^{-\alpha t}(|c_t|^2 - |c_t^k|^2)dt\right]$$
$$\le (1/\alpha + 2)\zeta$$

for sufficiently large k. Thus

$$u(x) + \zeta \ge U_{c^k}(x) - [U_{c^k}(x) - U_c(x)|$$
$$\ge u_k(x) - (1/\alpha + 2)\zeta.$$

Letting $k \to \infty$ and $\zeta \to 0$, we obtain (26).

Remark. Taking into account the proof of Lemma 2.5, we notice that for any $\theta \in \mathcal{S}$,

$$E[e^{-\alpha\theta}|u_k(x_\theta^k) - u(x_\theta)|] \le E[e^{-\alpha\theta}|u_k(x_\theta^k) - u(x_\theta^k)|] + E[e^{-\alpha\theta}u(x_\theta^k) - u(x_\theta)|]$$
$$\le \sup_{|x|\le R}|u_k(x) - u(x)| + 2\|u_1\|E[e^{-\alpha\theta}|x_\theta^k|^2]/R^2$$
$$+ E[\zeta + C_{\zeta,u}e^{-\alpha\theta}|x_\theta^k - x_\theta|] \quad \to \quad 0,$$

as $k \to \infty$ and $R, \zeta^{-1} \to \infty$. Then, passing to the limit in (22), we can show that the dynamic programming principle holds for u.

THEOREM 2.6. *We make the assumptions of Theorem* 2.1. *Then the solution* u *of* (4) *is a viscosity solution of* (14).

Proof. Combining Lemma 2.4 with 2.5, we get the assertion by the stability result [5, Lem. 6.2, p. 73] as follows.

Let $\varphi \in C^2$ and z be the maximizer of $u - \varphi$ such that

$$u(z) - \varphi(z) > u(x) - \varphi(x)$$

in the closed ball $\bar{B}(z,\delta)$ with radius δ of $z \ne x$. By Lemma 2.5, $u_k - \varphi$ attains a local maximum at some $z_k \in \bar{B}(z,\delta)$. Take a subsequence $z_{k'}$ of z_k such that

$$z_{k'} \quad \to \quad z' \in \bar{B}(z,\delta).$$

By Lemma 2.5

$$(u_{k'} - \varphi)(z_{k'}) \quad \to \quad (u - \varphi)(z').$$

Since

$$(u_{k'} - \varphi)(z_{k'}) > (u_{k'} - \varphi)(x), \qquad x \in \bar{B}(z, \delta),$$

we have

$$(u - \varphi)(z') \geq (u - \varphi)(x), \qquad x \in \bar{B}(z, \delta).$$

Thus we deduce

$$z = z'$$

and the convergence of the whole sequence.

Now, it follows from Lemma 2.4 that

$$-\left(\alpha + \frac{1}{\varepsilon}\right) u_k(z_k) - L_0\varphi(z_k) + F(z_k) + \min_{|c| \leq k}(|c|^2 + c \cdot D\varphi(z_k)) \geq 0.$$

Note that

$$\min_{|c| \leq k}(|c|^2 + c \cdot \xi) \to \min_{c}(|c|^2 + c \cdot \xi) \quad \text{ locally uniformly in } \mathbf{R^N} \text{ as } k \to \infty.$$

Letting $k \to \infty$, we get

$$-\left(\alpha + \frac{1}{\varepsilon}\right) u(z) - L_0\varphi(z) + F(z) + \min_{c}(|c|^2 + c \cdot D\varphi(z)) \geq 0.$$

By a simple relation

$$u \wedge g = u - (u - g)^+,$$

we see that u satisfies (15). By a similar argument, we conclude that u fulfills (16).

2.3. Another representation of u. In this subsection, we shall show that the unique solution u of (4) admits another representation

$$u(x) = \inf_{c} E\left[\int_0^\theta e^{-\alpha t}\left\{f(x_t) - \frac{1}{\varepsilon}(u - g)^+(x_t) + |c_t|^2\right\}dt + e^{-\alpha\theta}u(x_\theta)\right] \tag{27}$$

for any $\theta \in \mathcal{S}$. We set $H(x) = f(x) - \frac{1}{\varepsilon}(u - g)^+(x)$ and consider

$$-L\xi + H(x) - \frac{1}{4}|D\xi|^2 = 0, \qquad x \in \mathbf{R^N}. \tag{28}$$

Define

$$\xi(x) = \inf_{c} E\left[\int_0^\infty e^{-\alpha t}\left\{H(x_t) + |c_t|^2\right\}\, dt\right], \tag{29}$$

which belongs to $\mathcal{C}$. By the same arguments as in subsection 2.2, we can see that ξ satisfies the dynamic programming principle

$$\xi(x) = \inf_{c} E\left[\int_0^\theta e^{-\alpha t}\{H(x_t) + |c_t|^2\}dt + e^{-\alpha\theta}\xi(x_\theta)\right], \qquad \theta \in \mathcal{S},$$

and then ξ is a viscosity solution of (28).

THEOREM 2.7. *Under the assumptions of Theorem* 2.1, *we have* (27).

Proof. By Theorem 2.6, the relation (27) follows from the uniqueness of viscosity solutions of (28). Dividing the proof into several steps, we claim

$$\xi_1 \le \xi_2 \tag{30}$$

for two viscosity solutions $\xi_i \in \mathcal{C}$, $i = 1, 2$, of (28).

Step 1. Suppose there exists $\bar{x} \in \mathbf{R^N}$ such that

$$\xi_1(\bar{x}) - \xi_2(\bar{x}) > 0,$$

which implies

$$\xi_1(\bar{x}) - \xi_2(\bar{x}) \ge \delta \tag{31}$$

for some $\delta > 0$.

Define

$$\Psi_k(x, y) = \xi_1(x) - \xi_2(y) - \frac{k}{2}|x - y|^2 - \frac{1}{k}(\psi(x) + \psi(y)), \tag{32}$$

where $\psi(x) = \frac{1}{2}\log(1 + |x|^2)$ and $k > 0$. Since

$$\Psi_k(x, y) \quad \to \quad -\infty \quad \text{as} \quad |x|, |y| \to \infty,$$

we find $(x_k, y_k) \in \mathbf{R^N} \times \mathbf{R^N}$ such that

$$\begin{aligned}
\Psi_k(x_k, y_k) &= \sup \Psi_k(x, y) \\
&\ge \Psi_k(\bar{x}, \bar{x}) \\
&= \xi_1(\bar{x}) - \xi_2(\bar{x}) - \frac{2}{k}\psi(\bar{x}) \\
&\ge \delta - \frac{2}{k}\psi(\bar{x}) \\
&\ge \frac{\delta}{2} \quad \text{for} \quad k \ge \exists k_0 > 0.
\end{aligned}$$

Thus

$$\begin{aligned}
\frac{\delta}{2} &\le \xi_1(x_k) - \xi_2(y_k) - \frac{k}{2}|x_k - y_k|^2 - \frac{1}{k}(\psi(x_k) + \psi(y_k)) \\
&\le \xi_1(x_k) - \xi_2(y_k).
\end{aligned} \tag{33}$$

Step 2. By the definition of (x_k, y_k), we have

$$2\Psi_k(x_k, y_k) \ge \Psi_k(x_k, x_k) + \Psi_k(y_k, y_k),$$

or equivalently

$$\begin{aligned}
2\Big[\xi_1(x_k) - \xi_2(y_k) - \frac{k}{2}|x_k - y_k|^2 - \frac{1}{k}(\psi(x_k) + \psi(y_k))\Big] \\
\ge \xi_1(x_k) - \xi_2(x_k) - \frac{2}{k}\psi(x_k) \\
+ \xi_1(y_k) - \xi_2(y_k) - \frac{2}{k}\psi(y_k).
\end{aligned}$$

Hence

$$k|x_k - y_k|^2 \le \xi_1(x_k) - \xi_2(y_k) + \xi_2(x_k) - \xi_1(y_k) \le C \qquad (C > 0).$$

Thus

$$|x_k - y_k| \le (C/k)^{1/2}. \tag{34}$$

By the uniform continuity of ξ_i, we deduce

$$\begin{aligned} k|x_k - y_k|^2 &\le \sup_{|x-y|\le (C/k)^{1/2}} |\xi_1(x) - \xi_1(y)| + |\xi_2(x) - \xi_2(y)| \\ &\to \quad 0 \quad \text{as} \quad k \to \infty. \end{aligned} \tag{35}$$

Step 3. We here invoke the following lemma (see Crandall, Ishii, and Lions [4, Thm. 3.2], Fleming and Soner [5, Lem. 6.1, p. 238], and Ishii [7, Lem. 1, p. 149] for the proof).

LEMMA 2.8 (Ishii [7]). *Let $U, -V$ be upper semicontinuous in an open domain, and set*

$$W(x,y) = U(x) - V(y) - \frac{k}{2}|x-y|^2.$$

Let $(\hat{x}, \hat{y})$ be the local maximizer of W. Then there exist $X, Y \in \mathbf{S^N}$ such that

$$(k(\hat{x} - \hat{y}), X) \in \bar{J}^{2,+}U(\hat{x}), \tag{36}$$

$$(k(\hat{x} - \hat{y}), Y) \in \bar{J}^{2,-}V(\hat{y}), \tag{37}$$

$$-3k \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \le \begin{pmatrix} X & 0 \\ 0 & -Y \end{pmatrix} \le 3k \begin{pmatrix} I & -I \\ -I & I \end{pmatrix}, \qquad I = \textit{identity}, \tag{38}$$

where

$$\begin{aligned} \bar{J}^{2,\pm}U(x) = \{(p,X) : \exists x_n \to x, \exists (p_n, X_n) \in J^{2,\pm}U(x_n), \\ (U(x_n), p_n, X_n) \to (U(x), p, X)\}. \end{aligned}$$

Now we take

$$\begin{aligned} U(x) &= \xi_1(x) - \frac{1}{k}\psi(x), \\ V(y) &= \xi_2(y) + \frac{1}{k}\psi(y) \end{aligned}$$

and note that

$$J^{2,+}\xi_1(x) = \left\{ (\underline{p}, \underline{X}) + \left(\frac{1}{k} D\psi(x), \frac{1}{k} D^2\psi(x) \right) : \ (\underline{p}, \underline{X}) \in J^{2,+}U(x) \right\},$$

$$J^{2,-}\xi_2(y) = \left\{ (\underline{p}, \underline{X}) - \left(\frac{1}{k} D\psi(y), \frac{1}{k} D^2\psi(y) \right) : \ (\underline{p}, \underline{X}) \in J^{2,-}V(y) \right\}.$$

Then it follows from (36), (37), and the definition of $\bar{J}^{2,+}\xi_1(x_k), \bar{J}^{2,-}\xi_2(y_k)$ that

$$(p_1, \bar{X}) := (k(x_k - y_k), X) + \left(\frac{1}{k} D\psi(x_k), \frac{1}{k} D^2\psi(x_k) \right) \in \bar{J}^{2,+}\xi_1(x_k), \tag{39}$$

$$(p_2, \bar{Y}) := (k(x_k - y_k), Y) - \left(\frac{1}{k} D\psi(y_k), \frac{1}{k} D^2\psi(y_k) \right) \in \bar{J}^{2,-}\xi_2(y_k). \tag{40}$$

Step 4. We have by (28) and (40) that

$$-\alpha\xi_2(y_k) + \frac{1}{2}\mathrm{tr}(\sigma\sigma^*(y_k)\bar{Y}) + b(y_k)\cdot p_2 + H(y_k) - \frac{1}{4}|p_2|^2 \le 0. \tag{41}$$

Also, by (28) and (39)

$$\alpha\xi_1(x_k) \le \frac{1}{2}\mathrm{tr}(\sigma\sigma^*(x)\bar{X}) + b(x_k)\cdot p_1 + H(x_k) - \frac{1}{4}|p_1|^2. \tag{42}$$

Thus, adding (42) to (41), we obtain

$$\begin{aligned}\alpha(\xi_1(x_k) - \xi_2(y_k)) &\le \mathrm{tr}(\sigma\sigma^*(x_k)X - \sigma\sigma^*(y_k)Y)/2 \\ &\quad + \mathrm{tr}(\sigma\sigma^*(x_k)D^2\psi(x_k) + \sigma\sigma^*(y_k)D^2\psi(y_k))/2k \\ &\quad + [b(x_k)\cdot p_1 - b(y_k)\cdot p_2] \\ &\quad + [H(x_k) - H(y_k)] \\ &\quad - [|p_1|^2 - |p_2|^2]/4. \\ &\equiv I_1/2 + I_2/2 + I_3 + I_4 - I_5/4.\end{aligned} \tag{43}$$

Step 5. We claim that

$$I_j \quad \to \quad 0 \quad \text{as} \quad k \to \infty \qquad (j = 1, 2, \ldots, 5),$$

which leads to a contradiction with (33).

According to Fleming and Soner [5, Lem. 6.2, p. 240], we know that (38) implies

$$\mathrm{tr}(\sigma\sigma^*(x)X - \sigma\sigma^*(y)Y) \le 3k|\sigma(x) - \sigma(y)|^2.$$

Hence, we have by the Lipschitz continuity of $\sigma(x)$ and (35)

$$I_1 \le 3kC|x_k - y_k|^2 \quad \to \quad 0 \quad \text{as} \quad k \to \infty.$$

By a simple manipulation, we get

$$|D\psi(x)| = \frac{|x|}{1+|x|^2}, \quad |D^2\psi(x)| \le \frac{C}{1+|x|^2}, \quad C > 0.$$

Then

$$\begin{aligned}I_2 &\le \frac{1}{k}\left[|\sigma\sigma^*(x_k)D^2\psi(x_k)| + |\sigma\sigma^*(y_k)D^2\psi(y_k)|\right] \\ &\le \frac{C}{k}\left[\frac{(|\sigma(0)| + |x_k|)^2}{1+|x_k|^2} + \frac{(|\sigma(0)| + |y_k|)^2}{1+|y_k|^2}\right] \quad \to \quad 0 \quad \text{as} \quad k \to \infty.\end{aligned}$$

By the Lipschitz continuity of $b(x)$, we see that

$$\begin{aligned}|I_3| &\le kC|x_k - y_k|^2 + (|b(x_k)||D\psi(x_k)| + |b(y_k)||D\psi(y_k)|)/k \\ &\le kC|x_k - y_k|^2 + C\sup_x\left(\frac{(|b(0)| + |x|)|x|}{1+|x|^2}\right)/k \\ &\quad \to \quad 0 \quad \text{as} \quad k \to \infty.\end{aligned}$$

By (34), we have

$$|I_4| \le \sup_{|x-y|\le(C/k)^{1/2}} |H(x) - H(y)| \to 0.$$

Finally, by (35)

$$\begin{aligned}|I_5| &\le \frac{2}{k}|k(x_k - y_k)\cdot(D\psi(x_k) + D\psi(y_k))| \\ &\quad + \left| |\tfrac{1}{k}D\psi(x_k)|^2 - |\tfrac{1}{k}D\psi(y_k)|^2 \right| \\ &\le \frac{2}{k}|k(x_k - y_k)| + \frac{C}{k^2} \quad \to \quad 0 \ \text{ as } \ k \to \infty.\end{aligned}$$

Therefore the theorem is established.

3. Viscosity solutions of variational inequalities.

3.1. Convergence as $\varepsilon \to 0$. We study the convergence of $u_\varepsilon = u$ as $\varepsilon = \varepsilon_n = 2^{-n} \to 0$. Define

$$G_\beta h(x) = E\left[\int_0^\infty e^{-\beta t} h(\bar{x}_t)dt\right], \qquad \beta > 0, \tag{44}$$

and

$$\mathcal{D} = \{G_\beta(\beta h) : \ h \in \mathcal{C}, \ \beta > \alpha\}, \tag{45}$$

where $\bar{x}_t$ is the unique solution of

$$d\bar{x}_t = b(\bar{x}_t)dt + \sigma(\bar{x}_t)dW_t, \qquad \bar{x}_0 = x. \tag{46}$$

LEMMA 3.1. *Under* (5) *and* (6), $\mathcal{D}$ *is dense in* $\mathcal{C}$.

Proof. We claim that

$$\mathcal{D} \subset \mathcal{C}. \tag{47}$$

Let $h \in \mathcal{C}$ be arbitrary. It is clear that $\|G_\beta(\beta h)\| \le \|h\|$. Further, (10) gives

$$E[|\bar{x}_t - \bar{y}_t|^2 e^{(-\beta+\eta)t}] \le |x - y|^2$$

for the solution $\bar{y}_t$ of (46) with $\bar{y}_0 = y$. By the same calculations as (12) and (13), we can get

$$|G_\beta(\beta h)(x) - G_\beta(\beta h)(y)| \le \zeta + 2C_{\zeta,h}|x - y|\beta/(\beta + \eta) \qquad \forall \zeta > 0,$$

and hence

$$\lim_{\delta \to 0} \sup_{|x-y|<\delta} |G_\beta(\beta h)(x) - G_\beta(\beta h)(y)| = 0,$$

which implies (47). Moreover, by (11), (46), and (5)

$$\begin{aligned}E[h(\bar{x}_t) - h(x)|] &\le \zeta + C_{\zeta,h}E[|\bar{x}_t - x|] \\ &\le \zeta + \bar{C}(t + \sqrt{t}),\end{aligned}$$

where $\bar{C} = C_{\zeta,h}[\sup_x |b(x)| + \sup_x |\sigma(x)|]$. Therefore, letting $\beta \to \infty$ and then $\zeta \to 0$, we deduce

$$\begin{aligned}\|G_\beta(\beta h) - h\| &\le \zeta + \bar{C}\int_0^\infty \beta e^{-\beta t}(t + \sqrt{t})dt \\ &= \zeta + \bar{C}\int_0^\infty e^{-s}(s/\beta + \sqrt{s/\beta})ds \quad \to \quad 0,\end{aligned}$$

which completes the proof.

LEMMA 3.2. *Let $\tilde{u}_\varepsilon$ be the solution of* (4) *corresponding to $\tilde{g} \in \mathcal{C}_+$. Then we have, under* (5), (6), *and* (7),

$$\|u_\varepsilon - \tilde{u}_\varepsilon\| \le \|g - \tilde{g}\|. \tag{48}$$

Proof. We shall show the assertion by the same line as Bensoussan [2, Lem. 5.5, p. 320]. Let $w, \tilde{w} \in \mathcal{C}$ verify

$$\|w - \tilde{w}\| \le \|g - \tilde{g}\|.$$

Then, it is easy to check that

$$\|w \wedge g - \tilde{w} \wedge \tilde{g}\| \le \|g - \tilde{g}\|,$$

and then, by (8),

$$\begin{aligned} |Tw(x) - \tilde{T}\tilde{w}(x)| &\le \sup_c E\left[\int_0^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}\frac{1}{\varepsilon}|w \wedge g - \tilde{w} \wedge \tilde{g}|(x_t)dt\right] \\ &\le \|g - \tilde{g}\|, \end{aligned}$$

where $\tilde{T}$ denotes T with $\tilde{g}$ replacing g. But

$$\begin{aligned} |T0 - \tilde{T}0| &\le \sup_c E\left[\int_0^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}\frac{1}{\varepsilon}|g^- - \tilde{g}^-|(x_t)dt\right] \\ &\le \|g - \tilde{g}\|. \end{aligned}$$

Taking $w = T0$, $\tilde{w} = \tilde{T}0$, we have

$$\|T^2 0 - \tilde{T}^2 0\| \le \|g - \tilde{g}\|,$$

and then

$$\|T^n 0 - \tilde{T}^n 0\| \le \|g - \tilde{g}\|, \qquad n = 1, 2, \ldots.$$

Letting $n \to \infty$, we deduce (48) by Theorem 2.1.

LEMMA 3.3. *Under* (5), (6), *and* (7), *we have*

$$u_\varepsilon(x) = \inf_c \inf_\theta E\left[\int_0^\theta e^{-\alpha t}\{f(x_t) + |c_t|^2\}dt + e^{-\alpha\theta}\{g + (u_\varepsilon - g)^+\}(x_\theta)\right]. \tag{49}$$

Proof. By Theorem 2.7

$$\begin{aligned} u_\varepsilon(x) &= \inf_c E\left[\int_0^\theta e^{-\alpha t}\left\{f(x_t) - \frac{1}{\varepsilon}(u - g)^+(x_t) + |c_t|^2\right\}dt + e^{-\alpha\theta}u_\varepsilon(x_\theta)\right] \\ &\le \inf_c E\left[\int_0^\theta e^{-\alpha t}\left\{f(x_t) + |c_t|^2\right\}dt + e^{-\alpha\theta}u_\varepsilon \vee g(x_\theta)\right] \qquad \forall\theta \in \mathcal{S}. \end{aligned}$$

On the other hand, we take $\theta = \tau = \inf\{t : \ u_\varepsilon(x_t) \ge g(x_t)\}$. Since

$$e^{-\alpha\tau}g(x_\tau) = e^{-\alpha\tau}u_\varepsilon \vee g(x_\tau) = e^{-\alpha\tau}[g + (u_\varepsilon - g)^+](x_\tau),$$

we have

$$\begin{aligned} u_\varepsilon(x) &= \inf_c E\left[\int_0^\tau e^{-\alpha t}\left\{f(x_t) - \frac{1}{\varepsilon}(u - g)^+(x_t) + |c_t|^2\right\}dt + e^{-\alpha\tau}u_\varepsilon(x_\tau)\right] \\ &= \inf_c E\left[\int_0^\tau e^{-\alpha t}\left\{f(x_t) + |c_t|^2\right\}dt + e^{-\alpha\tau}g(x_\tau)\right], \end{aligned}$$

which completes the proof.

THEOREM 3.4. *Under* (5), (6), *and* (7), *we have*

$$u_{\varepsilon_n} \quad \to \quad v \in \mathcal{C}, \tag{50}$$

where $\varepsilon_n = 2^{-n}$.

Proof. Let $g = G_\beta(\beta\hat{h}) \in \mathcal{D}$ for some $\hat{h} \in \mathcal{C}$. In view of Theorems 2.6 and 2.7, we observe that g is a unique viscosity solution of

$$-\beta g - L_0 g + \beta\hat{h} = 0 \quad \text{in} \quad \mathbf{R^N},$$

or equivalently

$$-\left(\alpha + \frac{1}{\varepsilon}\right) g - L_0 g + \beta h + \frac{1}{\varepsilon} g = 0 \quad \text{in} \quad \mathbf{R^N},$$

where $\beta h = \beta\hat{h} + (\alpha - \beta)g$. Hence we have $g = G_{\alpha+\frac{1}{\varepsilon}}(\beta h + \frac{1}{\varepsilon} g)$. Therefore

$$\begin{aligned} u_\varepsilon - g &\le u_\varepsilon - \inf_c E\left[\int_0^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}\left\{\beta h(x_t) + \frac{1}{\varepsilon} g(x_t) + |c_t|^2\right\} dt\right] \\ &\le \sup_c E\left[\int_0^\infty e^{-(\alpha+\frac{1}{\varepsilon})t}\left\{f - \beta h + \frac{1}{\varepsilon}(u_\varepsilon \wedge g - g)\right\}(x_t) dt\right] \\ &\le \varepsilon \|f - \beta h\|. \end{aligned} \tag{51}$$

Applying (49) to $u_{\varepsilon_{n+1}}(x)$ and $u_{\varepsilon_n}(x)$, we have by (51)

$$\begin{aligned} |u_{\varepsilon_{n+1}}(x) - u_{\varepsilon_n}(x)| &\le \sup_c \sup_\theta E[e^{-\alpha\theta} |(u_{\varepsilon_{n+1}} - g)^+ - (u_{\varepsilon_n} - g)^+|(x_\theta)] \\ &\le (\varepsilon_{n+1} + \varepsilon_n)\|f - \beta h\|. \end{aligned}$$

Thus

$$\sum_{n=1}^\infty \|u_{\varepsilon_{n+1}} - u_{\varepsilon_n}\| \le \sum_{n=1}^\infty (\varepsilon_{n+1} + \varepsilon_n)\|f - \beta h\| < \infty.$$

This implies that $\{u_{\varepsilon_n}\}$ is a Cauchy sequence in $\mathcal{C}$, and we get (50).

In the case $g \in \mathcal{C}_+$, there exists a sequence $\{g_m\} \subset \mathcal{D}$ such that $g_m \to g$ by Lemma 3.1. Let u_ε^m be the solution of (4) corresponding to g_m. By the above argument, we see that

$$u_{\varepsilon_n}^m \quad \to \quad v^m \in \mathcal{C} \quad \text{as} \quad n \to \infty. \tag{52}$$

By (48),

$$\|u_{\varepsilon_n}^m - u_{\varepsilon_n}^{m'}\| \le \|g_m - g_{m'}\|.$$

Letting $n \to \infty$, we have

$$\|v^m - v^{m'}\| \le \|g_m - g_{m'}\|.$$

Hence $\{v^m\}$ is a Cauchy sequence, and

$$v^m \quad \to \quad v \in \mathcal{C}. \tag{53}$$

Thus

$$\begin{aligned} \|u_{\varepsilon_n} - v\| &\le \|u_{\varepsilon_n} - u_{\varepsilon_n}^m\| + \|u_{\varepsilon_n}^m - v^m\| + \|v^m - v\| \\ &\le \|g - g_m\| + \|u_{\varepsilon_n}^m - v^m\| + \|v^m - v\|. \end{aligned}$$

Letting $n \to \infty$ and then $m \to \infty$, we obtain (50).

3.2. Existence. In this subsection we consider the variational inequality:

$$
\begin{aligned}
&-Lv + f - \frac{1}{4}|Dv|^2 \geq 0, \\
&v \leq g \qquad\qquad \text{in } \mathbf{R^N}, \\
&\left(-Lv + f - \frac{1}{4}|Dv|^2\right)(v-g)^- = 0.
\end{aligned} \tag{54}
$$

We present the definition of the viscosity solutions of the variational inequality in the following.

DEFINITION 3.5. *$w \in \mathcal{C}$ is called a viscosity solution of* (54) *if the following assertions are satisfied:*

$$
\begin{aligned}
&\text{For any } \varphi \in C^2 \text{ and any local maximum point } z \text{ of } w - \varphi, \\
&-\alpha w(z) - L_0\varphi(z) + f(z) - \frac{1}{4}|D\varphi(z)|^2 \geq 0, \\
&w(x) \leq g(x), \qquad x \in \mathbf{R^N}. \\
&\text{For any } \varphi \in C^2 \text{ and any local minimum point } z \text{ of } w - \varphi, \\
&\left(-\alpha w(z) - L_0\varphi(z) + f(z) - \frac{1}{4}|D\varphi(z)|^2\right)(w-g)^-(z) \leq 0.
\end{aligned} \tag{55}
$$

THEOREM 3.6. *We make the assumptions of Theorem* 3.4. *Then the limit v of* (50) *is a viscosity solution of* (54).

Proof. Let $\varphi \in C^2$ and z be the maximizer of $v - \varphi$ such that

$$v(z) - \varphi(z) > v(x) - \varphi(x), \quad x \in \bar{B}(z,\delta), \quad z \neq x.$$

By the uniform convergence in Theorem 3.4, $u_{\varepsilon_n} - \varphi$ attains a local maximum at $x_n \in \bar{B}(z,\delta)$. By the same argument as Theorem 2.6, we deduce

$$x_n \quad \to \quad z.$$

Now, we have by Theorem 2.6 and (15)

$$-\alpha u_{\varepsilon_n}(x_n) - L_0\varphi(x_n) + f(x_n) - \frac{1}{\varepsilon_n}(u_{\varepsilon_n} - g)^+(x_n) - \frac{1}{4}|D\varphi(x_n)|^2 \geq 0,$$

from which

$$-\alpha u_{\varepsilon_n}(x_n) - L_0\varphi(x_n) + f(x_n) - \frac{1}{4}|D\varphi(x_n)|^2 \geq 0.$$

Letting $n \to \infty$, we get

$$-\alpha v(z) - L_0\varphi(z) + f(z) - \frac{1}{4}|D\varphi(z)|^2 \geq 0. \tag{56}$$

Next, by (51)

$$(u^m_{\varepsilon_n} - g_m)^+ \leq \varepsilon_n \|f - h_m\|,$$

where $g_m = G_\beta h_m$ for some $h_m \in \mathcal{C}$. Letting $n \to \infty$, we have by (52)

$$v^m \leq g_m,$$

and then by (53)

$$v \le g. \tag{57}$$

Finally, let $\bar{z}$ be the minimizer of $v-\varphi$, and $\bar{x}_n$ be the local minimizer of $u_{\varepsilon_n}-\varphi$ with $\bar{x}_n \to \bar{z}$. Then, by Theorem 2.6 and (16)

$$-\alpha u_{\varepsilon_n}(\bar{x}_n) - L_0\varphi(\bar{x}_n) + f(\bar{x}_n) - \frac{1}{\varepsilon_n}(u_{\varepsilon_n}-g)^+(\bar{x}_n) - \frac{1}{4}|D\varphi(\bar{x}_n)|^2 \le 0.$$

Multiply both sides by $(u_{\varepsilon_n}-g)^-$ to obtain

$$\left(-\alpha u_{\varepsilon_n}(\bar{x}_n) - L_0\varphi(\bar{x}_n) + f(\bar{x}_n) - \frac{1}{4}|D\varphi(\bar{x}_n)|^2\right)(u_{\varepsilon_n}-g)^-(\bar{x}_n) \le 0.$$

Letting $n \to \infty$, we deduce

$$\left(-\alpha v(\bar{z}) - L_0\varphi(\bar{z}) + f(\bar{z}) - \frac{1}{4}|D\varphi(\bar{z})|^2\right)(v-g)^-(\bar{z}) \le 0. \tag{58}$$

The assertions (56), (57), and (58) verify that v is a viscosity solution of (54) in the sense of (55).

3.3. Uniqueness.

THEOREM 3.7. *The assumptions are those of Theorem* 3.4. *Let* $v_i \in \mathcal{C}$, $i=1,2$, *be two viscosity solutions of* (54). *Then we have*

$$v_1 = v_2.$$

Proof. We shall show that

$$\left(-\alpha v_2(y) + \frac{1}{2}\mathrm{tr}(\sigma\sigma^*(y)\tilde{X}) + b(y)\cdot p + f(y) - \frac{1}{4}|p|^2\right)(v_2-v_1)^-(y) \le 0 \tag{59}$$
$$\forall (p,\tilde{X}) \in \bar{J}^{2,-}v_2(y), \quad \forall y \in \mathbf{R^N},$$

or equivalently

$$\left(-\alpha v_2(\tilde{z}) - L_0\varphi(\tilde{z}) + f(\tilde{z}) - \frac{1}{4}|D\varphi(\tilde{z})|^2\right)(v_2-v_1)^-(\tilde{z}) \le 0, \tag{60}$$

where $\varphi \in C^2$ and $\tilde{z}$ is the minimizer of $v_2-\varphi$. If $v_2 \ge v_1$, then $(v_2-v_1)^- = 0$. If $v_2 < v_1$, then $v_2 < v_1 \le g$, and thus $(v_2-g)^- > 0$. By (58) we recall

$$\left(-\alpha v_2(\tilde{z}) - L_0\varphi(\tilde{z}) + f(\tilde{z}) - \frac{1}{4}|D\varphi(\tilde{z})|^2\right)(v_2-g)^-(\tilde{z}) \le 0.$$

Then we deduce

$$-\alpha v_2(\tilde{z}) - L_0\varphi(\tilde{z}) + f(\tilde{z}) - \frac{1}{4}|D\varphi(\tilde{z})|^2 \le 0,$$

which implies (60).

Now, we prove the theorem by the same line as Theorem 2.7. Suppose there exists $\bar{x} \in \mathbf{R^N}$ such that

$$v_1(\bar{x}) - v_2(\bar{x}) > 0,$$

which implies

$$v_1(\bar{x}) - v_2(\bar{x}) \geq \delta \tag{61}$$

for some $\delta > 0$. We define

$$\Phi_k(x,y) = v_1(x) - v_2(y) - \frac{k}{2}|x-y|^2 - \frac{1}{k}(\psi(x) + \psi(y))$$

as in (32). Then the maximizer (x_k, y_k) of $\Phi_k(x,y)$ satisfies

$$\frac{\delta}{2} \leq v_1(x_k) - v_2(y_k) \tag{62}$$

for sufficiently large k. As in Steps 2 and 3 of Theorem 2.7 we have

$$k|x_k - y_k|^2 \quad \to \quad 0 \quad \text{as} \quad k \to \infty$$

and

$$(\hat{p}_1, \hat{X}) := (k(x_k - y_k), X) + \left(\frac{1}{k}D\psi(x_k), \frac{1}{k}D^2\psi(x_k)\right) \in \bar{J}^{2,+}v_1(x_k),$$

$$(\hat{p}_2, \hat{Y}) := (k(x_k - y_k), Y) - \left(\frac{1}{k}D\psi(y_k), \frac{1}{k}D^2\psi(y_k)\right) \in \bar{J}^{2,-}v_2(y_k).$$

By (62) we get

$$\begin{aligned} v_1(y_k) - v_2(y_k) &\geq v_1(x_k) - v_2(y_k) - |v_1(x_k) - v_1(y_k)| \\ &\geq \frac{\delta}{2} - |v_1(y_k) - v_1(x_k)| \\ &\geq \frac{\delta}{4} \quad \text{for sufficiently large } k, \end{aligned}$$

which implies

$$(v_2(y_k) - v_1(y_k))^- > 0.$$

From (59) it follows that

$$-\alpha v_2(y_k) + \frac{1}{2}\mathrm{tr}(\sigma\sigma^*(y_k)\hat{Y}) + b(y_k)\cdot\hat{p}_2 + f(y_k) - \frac{1}{4}|\hat{p}_2|^2 \leq 0. \tag{63}$$

Also, by (56)

$$\alpha v_1(x_k) \leq \frac{1}{2}\mathrm{tr}(\sigma\sigma^*(x_k)\hat{X}) + b(x_k)\cdot\hat{p}_1 + f(x_k) - \frac{1}{4}|\hat{p}_1|^2. \tag{64}$$

Hence we see that (63) and (64) are the similar relations to (41) and (42). Thus, we deduce by the same calculations as in Steps 4 and 5 of Theorem 2.7

$$\alpha(v_1(x_k) - v_2(y_k)) \quad \to \quad 0 \ \text{ as } \ k \ \to \infty,$$

which is contrary with (62). The proof is complete.

3.4. A stochastic interpretation of v. We here give a stochastic interpretation of the viscosity solution v of (54).

THEOREM 3.8. *The assumptions are those of Theorem* 3.4. *Then we have*

$$v(x) = \inf_c \inf_\theta E\left[\int_0^\theta e^{-\alpha t}\{f(x_t) + |c_t|^2\}dt + e^{-\alpha\theta}g(x_\theta)\right]. \tag{65}$$

Proof. We extend the technique of Morimoto [12] for the proof. Let $\hat{v}$ denote the right-hand side of (65). By (27) we have

$$\begin{aligned} u_{\varepsilon_n}(x) &= \inf_c E\left[\int_0^\theta e^{-\alpha t}\left\{f(x_t) - \frac{1}{\varepsilon_n}(u_{\varepsilon_n} - g)^+(x_t) + |c_t|^2\right\}dt + e^{-\alpha\theta}u_{\varepsilon_n}(x_\theta)\right] \\ &\le \inf_c E\left[\int_0^\theta e^{-\alpha t}\{f(x_t) + |c_t|^2\}dt + e^{-\alpha\theta}u_{\varepsilon_n}(x_\theta)\right] \qquad \forall\theta \in \mathcal{S}. \end{aligned}$$

Letting $n \to \infty$, we have by (50)

$$\begin{aligned} v(x) &\le \inf_c E\left[\int_0^\theta e^{-\alpha t}\{f(x_t) + |c_t|^2\}dt + e^{-\alpha\theta}v(x_\theta)\right] \\ &\le \inf_c E\left[\int_0^\theta e^{-\alpha t}\{f(x_t) + |c_t|^2\}dt + e^{-\alpha\theta}g(x_\theta)\right], \end{aligned}$$

which yields $v \le \hat{v}$. For the reverse inequality, we take any $c \in \mathcal{A}$ and set

$$R_m = \inf\left\{t : \ v(x_t) + \frac{1}{m} \ge g(x_t)\right\}.$$

Since

$$v(x_t) + \frac{1}{m} < g(x_t) \ \text{ on } \ \{t < R_m\},$$

we get

$$\begin{aligned} E\left[\int_0^{R_m} e^{-\alpha t}(u_{\varepsilon_n} - g)^+(x_t)dt\right] &\le E\left[\int_0^{R_m} e^{-\alpha t}\left(u_{\varepsilon_n} - \left(v + \frac{1}{m}\right)\right)^+(x_t)dt\right] \\ &\le E\left[\int_0^{R_m} e^{-\alpha t}\left(\|u_{\varepsilon_n} - v\| - \frac{1}{m}\right)^+(x_t)dt\right] \\ &= 0 \end{aligned}$$

for sufficiently large n. Hence, by (27)

$$\begin{aligned} u_{\varepsilon_n}(x) &= \inf_c E\left[\int_0^{R_m} e^{-\alpha t}\{f(x_t) - \frac{1}{\varepsilon_n}(u_{\varepsilon_n} - g)^+(x_t) + |c_t|^2\}dt + e^{-\alpha R_m}u_{\varepsilon_n}(x_{R_m})\right] \\ &= \inf_c E\left[\int_0^{R_m} e^{-\alpha t}\{f(x_t) + |c_t|^2\}dt + e^{-\alpha R_m}u_{\varepsilon_n}(x_{R_m})\right]. \end{aligned}$$

Letting $n \to \infty$, we have by (50) and the definition of R_m

$$\begin{aligned} v(x) &= \inf_c E\left[\int_0^{R_m} e^{-\alpha t}\{f(x_t) + |c_t|^2\}dt + e^{-\alpha R_m} v(x_{R_m})\right] \\ &= \inf_c E\left[\int_0^{R_m} e^{-\alpha t}\{f(x_t) + |c_t|^2\}dt + e^{-\alpha R_m}\left\{g(x_{R_m}) - \frac{1}{m}\right\}\right] \\ &\geq \inf_c E\left[\int_0^{R_m} e^{-\alpha t}\{f(x_t) + |c_t|^2\}dt + e^{-\alpha R_m} g(x_{R_m})\right] - \frac{1}{m} \\ &\geq \hat{v}(x) - \frac{1}{m}. \end{aligned}$$

Letting $m \to \infty$, we deduce $v(x) \geq \hat{v}(x)$, which completes the proof.

4. Impulsive control.

4.1. Setting of the problem. We consider an application of the variational inequality to the impulsive control problem. Impulsive control is described by a set (Θ, γ):

$$\begin{aligned} &\Theta = \{\theta_n\}, \qquad \theta_n \in \mathcal{S} \ \uparrow \infty, \\ &\gamma = \{\gamma_n\}, \qquad \gamma_n \in \mathbf{R}_+^{\mathbf{N}} : \ \mathcal{F}_{\theta_n} - \text{measurable}. \end{aligned}$$

The controlled equation becomes

$$d\chi_t = [b(\chi_t) + c_t]dt + \sigma(\chi_t)dW_t + \sum_{n=1}^{\infty} \delta(t - \theta_n)\gamma_n, \qquad \chi_0 = x,$$

where $\delta(\cdot)$ is the Dirac measure. More precisely we define a sequence of processes:

$$\begin{aligned} &d\chi_t^n = [b(\chi_t^n) + c_t]dt + \sigma(\chi_t^n)dW_t, \quad \chi^n(\theta_n) = \chi^{n-1}(\theta_n) + \gamma_n, \quad n \geq 1, \\ &d\chi_t^0 = [b(\chi_t^0) + c_t]dt + \sigma(\chi_t^0)dW_t, \quad \chi^0(0) = x. \end{aligned}$$

Then we write

$$\chi_t = \chi_t^n, \qquad \theta_n \leq t < \theta_{n+1}.$$

The aim is to minimize the cost

$$J(c, \theta, \gamma) = E\left[\int_0^{\infty} e^{-\alpha t}\{f(\chi_t) + |c_t|^2\}dt + \sum_{n=1}^{\infty} e^{-\alpha\theta_n}\rho(\gamma_n)\right],$$

where f and ρ are assumed to satisfy

$$\begin{aligned} &f \in \mathcal{C}_+, \\ &\rho(x) = k + \rho_0(x) \quad \text{for} \quad x \in \mathbf{R}_+^{\mathbf{N}}, \\ &k > 0, \quad \rho_0 \in \mathcal{C}_+(\mathbf{R}_+^{\mathbf{N}}), \quad \rho_0 > (0) = 0. \end{aligned} \tag{66}$$

Now, the quasi-variational inequality associated with the impulsive control problem is given by

$$\begin{aligned} &-Lv + f - \frac{1}{4}|Dv|^2 \geq 0, \\ &v \leq Mv \qquad\qquad \text{in} \quad \mathbf{R}^{\mathbf{N}}, \\ &\left(-Lv + f - \frac{1}{4}|Dv|^2\right)(v - Mv)^- = 0, \end{aligned} \tag{67}$$

where $Mv(x) := \inf_{\gamma \in \mathbf{R}_+^{\mathbf{N}}}[v(x + \gamma) + \rho(\gamma)]$.

4.2. Quasi-variational inequalities. In this subsection we show the existence of a unique viscosity solution of the quasi-variational inequality (67).

Define

$$Qw(x) = \inf_c \inf_\theta E\left[\int_0^\theta e^{-\alpha t}\{f(x_t) + |c_t|^2\}dt + e^{-\alpha\theta} Mw(x_\theta)\right], \qquad w \in \mathcal{C}_+, \tag{68}$$

where x_t is as in (3).

LEMMA 4.1. *For all $w, \tilde{w} \in \mathcal{C}_+$ we have, under* (5), (6), *and* (66),

$$0 \le Qw \le \|f\|/\alpha, \tag{69}$$

$$Qw \in \mathcal{C}_+, \tag{70}$$

$$w \le \tilde{w} \quad \Rightarrow \quad Qw \le Q\tilde{w}, \tag{71}$$

$$Q(\mu w + (1-\mu)\tilde{w}) \ge \mu Qw + (1-\mu)Q\tilde{w}, \qquad \mu \in [0,1]. \tag{72}$$

Proof. As is shown in Bardi and Capuzzo-Dolcetta [1], we can easily see that

$$\begin{aligned}
&0 \le Mw \le \|w\| + k,\\
&Mw \in \mathcal{C}_+,\\
&\|Mw - M\tilde{w}\| \le \|w - \tilde{w}\|,\\
&w \le \tilde{w} \quad \Rightarrow \quad Mw \le M\tilde{w},\\
&M(\mu w + (1-\mu)\tilde{w}) \ge \mu Mw + (1-\mu)M\tilde{w}, \qquad \mu \in [0,1].
\end{aligned}$$

Hence (69) and (71) are obvious. (72) is an easy consequence of the concavity of M. (70) follows from Theorems 3.6–3.8.

LEMMA 4.2. *If $w, \tilde{w} \in \mathcal{C}_+$ satisfy $w - \tilde{w} \le \lambda w$ for some $\lambda \in [0,1]$, then we have, under* (5), (6), *and* (66),

$$Qw - Q\tilde{w} \le \lambda(1-\mu)Qw \qquad \forall \mu \in \left(0, \frac{k}{\|v^0\|} \wedge 1\right), \tag{73}$$

where

$$v^0(x) := \inf_c E\left[\int_0^\infty e^{-\alpha t}\left\{f(x_t) + |c_t|^2\right\}dt\right].$$

Proof. By (72) we see

$$Q((1-\lambda)w + \lambda 0) \ge (1-\lambda)Qw + \lambda Q0.$$

Since

$$(1-\lambda)w \le \tilde{w},$$

we have

$$Q\tilde{w} \ge (1-\lambda)Qw + \lambda Q0,$$

or equivalently

$$Qw - Q\tilde{w} \le \lambda(Qw - Q0).$$

By virtue of (11) and (12) we note that

$$v^0 \in \mathcal{C}_+,$$

and by (68)

$$Qw \le v^0.$$

To complete the proof, it suffices to show that

$$Q0 \ge \mu v^0 \qquad \forall \mu \in \left(0, \frac{k}{\|v^0\|} \wedge 1\right). \tag{74}$$

By (66) we have $M0 = k$, and then

$$Q0 = \inf_c \inf_\theta E\left[\int_0^\theta e^{-\alpha t}\{f(x_t) + |c_t|^2\}dt + e^{-\alpha\theta}k\right].$$

It is obvious that

$$\mu v^0(x_t) \le \mu\|v^0\| \le \frac{k}{\|v^0\|}\|v^0\| = k.$$

Thus, we get by the dynamic programming principle for $v^0(x)$

$$\begin{aligned} Q0 &\ge \inf_c \inf_\theta E\left[\int_0^\theta e^{-\alpha t}\{f(x_t) + |c_t|^2\}dt + e^{-\alpha\theta}\mu v^0(x_\theta)\right] \\ &\ge \mu \inf_c \inf_\theta E\left[\int_0^\theta e^{-\alpha t}\{f(x_t) + |c_t|^2\}dt + e^{-\alpha\theta} v^0(x_\theta)\right] \\ &= \mu v^0(x), \end{aligned}$$

which implies (74).

THEOREM 4.3. *We assume* (5), (6), *and* (66). *Then there exists one and only one viscosity solution* $\bar{v} \in \mathcal{C}_+$ *of* (67).

Proof. Set $v^n = Q^n v^0 \in \mathcal{C}_+$. It is clear that

$$0 \le v^1 = Qv^0 \le v^0,$$

and then

$$0 \le v^n \le v^{n-1} \le v^0.$$

Moreover,

$$v^1 - v^0 \le v^1.$$

By (73) we have

$$Qv^1 - Qv^0 \le (1-\mu)Qv^1 \qquad \forall \mu \in \left(0, \frac{k}{\|v^0\|} \wedge 1\right),$$

from which

$$v^2 - v^1 \le (1-\mu)v^2.$$

By iteration, we get

$$v^{n+1} - v^n \le (1-\mu)^n v^{n+1} \le (1-\mu)^n v^0.$$

Thus we deduce

$$v^n \quad \to \quad \bar{v} \quad \text{in } \mathcal{C}_+.$$

We recall by Theorems 3.6–3.8 that v^n is a unique viscosity solution of

$$\begin{aligned} &-Lv^n + f - \frac{1}{4}|Dv^n|^2 \ge 0, \\ &v^n \le Mv^{n-1} \qquad\qquad \text{in } \mathbf{R^N}, \\ &\left(-Lv^n + f - \frac{1}{4}|Dv^n|^2\right)(v^n - Mv^{n-1})^- = 0, \end{aligned}$$

and we apply the stability result on viscosity solutions as in Theorem 2.6. Letting $n \to \infty$, we see that $\bar{v}$ is a viscosity solution of (67) in the sense of Definition 3.5 with $M\bar{v}$ replacing g.

To prove uniqueness, let $v_i \in \mathcal{C}_+$, $i = 1, 2$, be two viscosity solutions of (67). By Theorems 3.7 and 3.8, we see

$$v_i = Qv_i, \qquad i = 1, 2.$$

Clearly

$$v_1 - v_2 \le v_1.$$

Applying (73) with $\lambda = 1$, we have

$$Qv_1 - Qv_2 \le (1-\mu)Qv_1 \qquad \forall \mu \in \left(0, \frac{k}{\|v^0\|} \wedge 1\right).$$

Hence

$$v_1 - v_2 \le (1-\mu)v_1.$$

By iteration,

$$v_1 - v_2 \le (1-\mu)^n v_1, \qquad n = 1, 2, \ldots.$$

Letting $n \to \infty$, we get $v_1 \le v_2$, which completes the proof.

Concluding remarks. In this paper we have investigated the nonlinear variational inequality associated with the combined control problem with discretionary stopping. We have shown the existence of a unique viscosity solution to the variational inequality by the improved methods of the theory of linear variational inequalities. Further we have applied the technique of viscosity solutions to solve quasi-variational inequalities.

This paper presents the definition of viscosity solutions to the variational inequality, which is different from the ordinary one in [1, p. 196; 4]. As seen in Definition 3.5, the viscosity solution w coincides with a viscosity solution V of the boundary value problem

$$\begin{aligned} -LV + f(x) - \frac{1}{4}|DV|^2 = 0 \quad &\text{in} \quad \mathcal{O}, \\ V = w \qquad\qquad\qquad\quad &\text{on} \quad \partial\mathcal{O}, \end{aligned}$$

on any open domain $\mathcal{O} \subset \{x : w(x) < g(x)\}$. It seems possible to have the smoothness of w from the classical results on the smoothness of V together with the properties of g and the boundary $\partial\mathcal{O}$. We need to study the regularity of viscosity solutions of the variational inequality under milder conditions on b, σ, f, g for the optimization problem (2), (3).

Acknowledgments. The author would like to thank two anonymous referees for their valuable and helpful comments that led to an improved version of the paper.

REFERENCES

[1] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Boston, MA, 1997.
[2] A. Bensoussan, *Stochastic Control by Functional Analysis Methods*, North–Holland, Amsterdam, 1982.
[3] A. Bensoussan and J. L. Lions, *Applications des Inéquations Variationnelles en Contrôle Stochastique*, Dunod, Paris, 1978.
[4] M. G. Crandall, H. Ishii, and P. L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.
[5] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[6] A. Friedman, *Stochastic Differential Equations and Applications, Vol.* 2, Academic Press, New York, 1976.
[7] H. Ishii, *Viscosity solutions to nonlinear partial differential equations*, Sugaku, 46 (1994), pp. 144–157 (in Japanese).
[8] I. Karatzas and H. Wang, *Utility maximization with discretionary stopping*, SIAM J. Control Optim., 39 (2000), pp. 306–329.
[9] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.
[10] J. L. Menaldi, *On the optimal stopping time problem for degenerate diffusions*, SIAM J. Control Optim., 18 (1980), pp. 697–721.
[11] J. L. Menaldi, *On the optimal impulse control problem for degenerate diffusions*, SIAM J. Control Optim., 18 (1980), pp. 722–739.
[12] H. Morimoto, *Dynkin games and martingale methods*, Stochastics, 13 (1984), pp. 213–228.
[13] H. M. Soner, *Controlled Markov Processes, Viscosity Solutions and Applications to Mathematical Finance*, Lecture Notes in Math. 1660, Springer-Verlag, Berlin, 1997, pp. 134–185.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 709–728

REGIONAL NULL CONTROLLABILITY OF A LINEARIZED CROCCO-TYPE EQUATION*

P. MARTINEZ†, J.-P. RAYMOND†, AND J. VANCOSTENOBLE†

Abstract. We are interested in controllability problems of equations coming from a boundary layer model. We simplify the problem by considering only equations with constant coefficients. The problem is described by a degenerate parabolic equation (a linearized Crocco-type equation) where phenomena of diffusion and transport are coupled.

First we give a geometric characterization of the influence domain of a locally distributed control. Then we prove *regional* null controllability results on this domain. The proof is based on an adequate observability inequality for the homogeneous adjoint problem. This inequality is obtained by decomposition of the space-time domain and Carleman-type estimates along characteristics.

In the second part of this paper, we treat the case of a boundary control.

Key words. regional null controllability, parabolic degenerate equation, transport equation, heat equation, Carleman estimates

AMS subject classifications. 93B05, 93C20, 93B07, 35K65

DOI. 10.1137/S0363012902403547

1. Introduction.

The velocity field of a laminar flow on a flat plate can be described by the Prandtl equations [23]. For a two dimensional flow, these equations are stated in an unbounded domain $(0, L) \times (0, \infty)$, where $(0, L)$ represents the part of the plate where the flow is laminar, and $(0, \infty)$ represents the "thickness" of the boundary layer. The matching conditions with the external flow are stated at $+\infty$.

By using the so-called Crocco transformation, these equations are transformed into a nonlinear degenerate parabolic equation (the Crocco equation; see [23]) which is stated in a bounded domain $\Omega = (0, L) \times (0, 1)$. The linearization of the Crocco equation around a stationary solution is an equation of the form

$$\begin{cases} u_t + au_x - bu_{yy} + cu = g, & (x, y, t) \in \Omega \times (0, T), \\ u(x, 0, t) = 0, & (x, t) \in (0, L) \times (0, T), \\ u_y(x, 1, t) = \chi_{(x_0, x_1)}(x) f(x, t), & (x, t) \in (0, L) \times (0, T), \\ u(0, y, t) = u_1(y, t), & (y, t) \in (0, 1) \times (0, T), \\ u(x, y, 0) = u_0(x, y), & (x, y) \in \Omega, \end{cases} \tag{1.1}$$

where g and u_1 depend on the incident velocity of the flow, and where the function f is the control used to stabilize the velocity in the boundary layer. The coefficients a, b, and c are regular, but degenerate, and have the following behavior [6, 5]:

$$0 < a_1 \leq \frac{a(y)}{y} \leq a_2, \quad 0 < b_1 \leq \frac{b(x, y)}{-(y-1)^2 \ln(\mu(1-y))} \leq b_2, \quad c(x, y) \geq 0,$$

where $0 < \mu < 1$. Since the coefficient b is degenerate, the Dirichlet boundary condition at $y = 1$ has to be correctly interpreted (see [5]).

*Received by the editors March 4, 2002; accepted for publication (in revised form) January 1, 2003; published electronically June 12, 2003.

http://www.siam.org/journals/sicon/42-2/40354.html

†Laboratoire M.I.P., U.M.R. 5640, Université Paul Sabatier Toulouse III, 118 route de Narbonne, 31 062 Toulouse Cedex 4, France (martinez@mip.ups-tlse.fr, raymond@mip.ups-tlse.fr, vancoste@mip.ups-tlse.fr).

This linearized model has been used to study stabilization problems of boundary layers in [4]. The perturbations of the velocity field in the boundary layer are controlled by a suction velocity f through the plate, localized on a slot (x_0, x_1).

In this paper we are interested in the null controllability problem for an equation of the type (1.1), but with constant coefficients. For simplicity, we first study a problem with homogeneous Dirichlet boundary conditions and a locally distributed control (see section 2). The cases of the other kinds of boundary conditions (Neumann or mixed Dirichlet–Neumann conditions) together with the case of a boundary control are treated in a second part of the paper (see sections 3 and 4).

Let $\omega = (x_0, x_1) \times \omega_y$, where $0 < x_0 < x_1 < L$ and ω_y is an open subset of $(0, 1)$, and let χ_ω be the characteristic function of ω.

For $u_0 \in L^2(\Omega)$, $u_1 \in L^2((0,1) \times (0,T))$, and $f \in L^2(\omega \times (0,T))$, we consider the following control problem:

$$
\begin{cases}
u_t + u_x - u_{yy} = \chi_\omega(x,y) f(x,y,t), & (x,y,t) \in \Omega \times (0,T), \\
u(x,0,t) = u(x,1,t) = 0, & (x,t) \in (0,L) \times (0,T), \\
u(0,y,t) = u_1(y,t), & (y,t) \in (0,1) \times (0,T), \\
u(x,y,0) = u_0(x,y), & (x,y) \in \Omega.
\end{cases} \tag{1.2}
$$

First one can prove that the problem is well-posed: for all $u_0 \in L^2(\Omega)$, $u_1 \in L^2((0,1)\times(0,T))$, and $f \in L^2(\omega\times(0,T))$, problem (1.2) has a unique solution that satisfies $u \in \mathcal{C}^0([0,T]; L^2(\Omega)) \cap \mathcal{C}^0([0,L]; L^2((0,T)\times(0,1))) \cap L^2((0,T)\times(0,L); H_0^1(0,1))$.

Then we study the following *regional* null controllability problem: *for all* $u_0 \in L^2(\Omega)$, $u_1 \in L^2((0,1) \times (0,T))$, *does there exist* $f \in L^2(\omega \times (0,T))$ *such that the solution* u *of* (1.2) *satisfies* $u(x,y,T) = 0$ *for* $(x,y) \in \Omega_C(T)$, *where* $\Omega_C(T)$ *denotes a part of* Ω*?*

Note that for nondegenerate parabolic equations stated in a bounded domain Ω, (*global*) null controllability is by now well known: for all nonempty $\omega \subset \Omega$ and for all $T > 0$, there exists $f \in L^2((0,T) \times \omega)$ such that the corresponding solution of the heat equation satisfies $u(T) \equiv 0$ in the *whole* domain Ω. These results are, in general, obtained via Carleman estimates (see, for example, [16, 14, 2]). But to our knowledge it seems that no result was known concerning a degenerate case.

Note also that, in the case studied in this paper, due to transport phenomenon, the influence domain of the control $\chi_\omega f$ is not the whole domain Ω at time $T > 0$. Thus (global) null controllability does not occur. For this reason, we introduce the notion of *regional* null controllability.

As a first step, we give a geometric characterization of the influence domain of the control $\chi_\omega f$ in order to determine the region $\Omega_C(T)$ of Ω on which it will be possible to control $u(\cdot, T)$ (see section 2.1).

Then we prove a result of *regional* null controllability on the domain $\Omega_C(T)$ (see section 2.3). This result is obtained via the introduction of an adequate penalized problem and the obtaining of a suitable observability inequality (given in section 2.2) for the homogeneous adjoint problem. This inequality is obtained by decomposition of the space-time domain and by using Carleman-type estimates along characteristics.

In the second part of this paper, we treat the case of a Dirichlet *boundary control*. We also prove a *regional* null controllability result via the obtaining of a similar boundary observability inequality. However, due to a lack of regularity of the solutions of the boundary value problem, the penalized problem that we introduce has to be modified.

2. The case of a localized distributed control.

2.1. Domain of influence of the control. Let $T > 0$ be fixed and define

$$\Omega_C(T) := \begin{cases} (x_0, x_1 + T) \times (0,1) & \text{if } T < L - x_1, \\ (x_0, L) \times (0,1) & \text{if } T > L - x_1. \end{cases}$$

Using spectral decomposition of the solution of (1.2), one can prove that the domain of influence of $\chi_\omega f$ at time T is the domain $\Omega_C(T)$ represented in Figure 2.1 (in the case $T < L - x_1$). Indeed, due to the phenomenon of diffusion in the direction y, the region of influence in y at time T of a control supported in y in ω_y is the whole interval $(0,1)$. On the other hand, due to the transport phenomenon (at speed equal to 1) in the x-direction, the region of influence in x at time T of a control supported in x in (x_0, x_1) is only $(x_0, x_1 + T)$ in the case $T < L - x_1$ and is only (x_0, L) in the case $T > L - x_1$. This means that a control localized in $\omega \times (0,T)$ has no influence at time T on the solution in $\Omega \setminus \Omega_c(T)$.

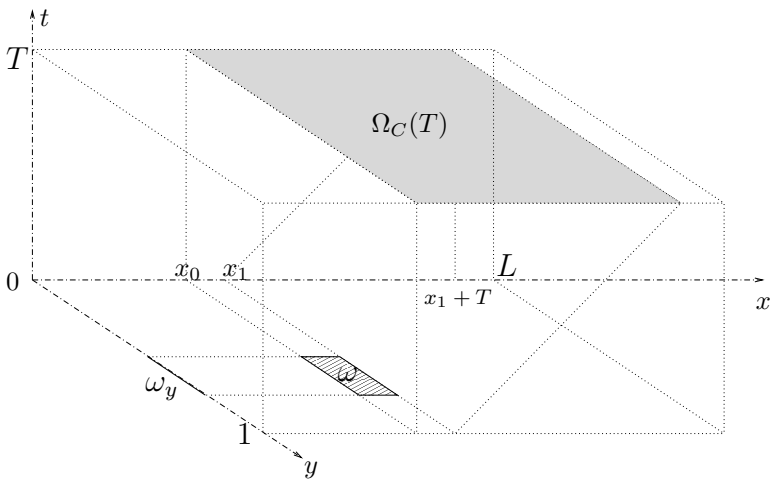


FIG. 2.1.

More precisely, we prove the following proposition.

PROPOSITION 2.1. *Assume that $\tilde{\Omega}$ is such that $\Omega_C(T) \subset \tilde{\Omega} \subset \Omega$ and $\Omega_C(T) \neq \tilde{\Omega}$. Then there exists $u_0 \in L^2(\Omega)$, $u_1 \in L^2((0,1) \times (0,T))$ such that for all $f \in L^2(\omega \times (0,T))$, the solution u of (1.2) is not identically equal to zero in $\tilde{\Omega}$.*

Thus we will finally prove a *regional* null controllability result in a set $\Omega_C(T, \delta)$ that will be arbitrarily close to $\Omega_C(T)$. For this, we first need an adequate observability estimate for the homogeneous adjoint problem.

2.2. Observability inequality. Assume that $T > 0$ and for all δ such that $0 < \delta < (x_1 - x_0)/2$ define

$$\Omega_C(T, \delta) := \begin{cases} (x_0 + \delta, x_1 + T - \delta) \times (0,1) & \text{if } 0 < T < L - x_1 + \delta, \\ (x_0 + \delta, L) \times (0,1) & \text{if } T > L - x_1 + \delta. \end{cases} \tag{2.1}$$

Then we prove the following observability estimate.

THEOREM 2.1. *Under the previous assumptions, there exists* $C(T,\delta,\omega_y)>0$ *such that the solutions* v *of the adjoint equation*

$$v_t + v_x + v_{yy} = 0, \quad (x,y,t)\in\Omega\times(0,T), \tag{2.2}$$

belonging to $\mathcal{C}^0([0,T];L^2((0,L)\times(0,1)))\cap\mathcal{C}^0([0,L];L^2((0,T)\times(0,1)))\cap L^2((0,L)\times(0,T);H_0^1(0,1))$ *satisfy*

$$\begin{aligned}\iint_{(0,L)\times(0,1)} v(x,y,0)^2\,dydx \;+\; \iint_{(0,1)\times(0,T)} v(0,y,t)^2\,dtdy \\ \le\; C\,(T,\delta,\omega_y)\left(\iiint_{\omega\times(0,T)} v(x,y,t)^2\,dtdydx\right. \\ \left. +\iint_{\Omega\setminus\Omega_C(T,\delta)} v(x,y,T)^2\,dydx \;+\; \iint_{(0,1)\times(0,T)} v(L,y,t)^2\,dtdy\right).\end{aligned} \tag{2.3}$$

Remark. The proof is first based on a decomposition of the domain $\Omega\times(0,T)$. Moreover, we notice that, along characteristics, v is solution of a nondegenerate parabolic equation for which Carleman's estimates are known. Then, on each subdomains and along characteristics, we apply Carleman's estimates.

2.3. Null controllability result. Finally we deduce the following result of null controllability.

THEOREM 2.2. *Under the previous assumptions, for all* $u_0\in L^2(\Omega)$, $u_1\in L^2((0,1)\times(0,T))$, *there exists* $f\in L^2(\omega\times(0,T))$ *such that the solution* u^f *of* (1.2) *satisfies*

$$u^f(x,y,T)=0 \quad \textit{for } (x,y)\in\Omega_C(T,\delta). \tag{2.4}$$

Remarks. 1. In the proof, we use the following penalized problem

$$\inf_{f\in L^2(\omega\times(0,T))}\Big(\frac{1}{2}\iiint_{\omega\times(0,T)} f^2\,dtdydx \;+\; \frac{1}{2\varepsilon}\iint_{\Omega_C(T,\delta)} u^f(x,y,T)^2\,dydx\Big),$$

where u^f is the solution of (1.2) associated with f.

2. In the case $T>L-x_1+\delta$, it is easy to see that there exists $f\in L^2(\omega\times(0,T))$ such that the solution u^f of (1.2) satisfies the stronger property

$$u^f(x,y,t)=0 \quad \text{for } (x,y,t)\in\Big(\Omega_C(T,\delta)\times\{T\}\Big)\cup\Big(\{L\}\times(0,1)\times(L-x_1+\delta,T)\Big).$$

This means that if we extend the solution to $[0,L']$ with $L'>T-x_1+\delta$, the corresponding domain on which the null controllability result is true at time T is the domain $(x_0+\delta,x_1+T-\delta)\times(0,1)$.

3. Lot of results are known concerning nondegenerate parabolic equations. But to our knowledge it seems that very few results were known concerning a degenerate case.

First, a quite similar problem was recently studied in [1]. The authors consider a model of age-dependent population dynamics with diffusion. The structure of the equation is quite similar to the structure of (1.2): there is a diffusion phenomenon in one direction while there is a transport phenomenon in the other one. For this model, the authors prove a null controllability result on the whole domain. This comes from the fact that the control is globally distributed in the direction of transport (while in our case, the control is locally distributed in (x_0,x_1)) and to the fact that the boundary condition in [1] is an integral equation with suitable support properties.

On the other hand, another result of regional null controllability was also recently obtained in [7, 8] for an other degenerate parabolic equation but with a type of degeneracy that is quite different of the one that we study in this paper. We studied the (strongly) degenerate equation

$$u_t - (a(x)u_x)_x = \chi_{(\alpha,\beta)} f,$$

where a may vanish on $[0, \alpha']$ for $0 \leq \alpha' < \beta$. For all $T > 0$, we proved a result of regional null controllability at time T in the region $(\alpha + \delta, 1)$ (for all $\delta > 0$ such that $\alpha + \delta \leq 1$).

4. For nondegenerate parabolic equations, (*global*) null controllability holds. Then, since the energy of the uncontrolled energy is nonincreasing, once the system has been driven to zero in time T, it remains indefinitely at zero without being controlling anymore.

In our case, the situation is not the same. First, if the boundary condition u_1 is equal to zero and if $T > x_0$, then the solution u^f which satisfies (2.4) is also equal to zero on the domain $\mathcal{R}_1 \cup \mathcal{R}_2$ (see Figure 2.2 below) without being controlling anymore. Now in the general case, given $u_1 \in L^2_{\text{loc}}([0, +\infty); L^2(0, 1))$, due to the transport phenomenon, the solution u remains equal to zero on the domain $\mathcal{R}_2$.

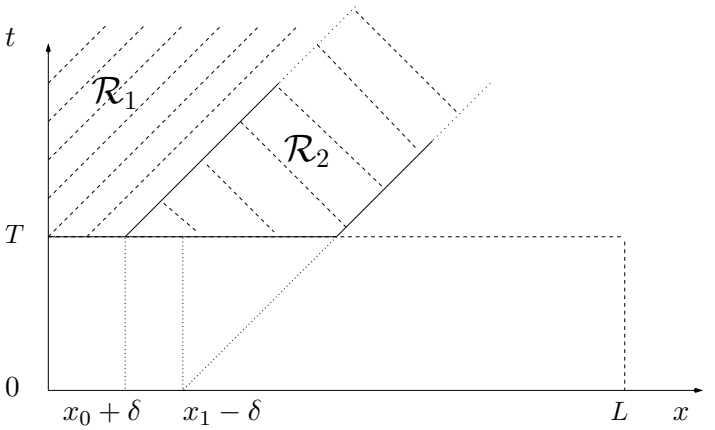


FIG. 2.2.

Since, given $\tau \geq T$, it is possible to find a control f^τ that drives the system to rest at time τ on the set $(x_0 + \delta, x_1 - \delta + \tau) \times (0, 1)$, it would be interesting to know if, given $T' > T$, it is possible to find a control $f^{T,T'}$ that drives the system to rest on the time-space domain $\{(x, y, \tau), \tau \in (T, T'), (x, y, \tau) \in (x_0 + \delta, x_1 - \delta + \tau) \times (0, 1) \times \{\tau\}\}$.

3. The case of a boundary control. Let $0 < x_0 < x_1 < L$ be fixed. For $u_0 \in L^2(\Omega)$, $u_1 \in L^2((0, 1) \times (0, T))$, and $f \in L^2((x_0, x_1) \times (0, T))$, we consider the following control problem:

$$\begin{cases} u_t + u_x - u_{yy} = 0, & (x, y, t) \in \Omega \times (0, T), \\ u(x, 0, t) = 0, & (x, t) \in (0, L) \times (0, T), \\ u(x, 1, t) = \chi_{(x_0,x_1)}(x) f(x, t), & (x, t) \in (0, L) \times (0, T), \\ u(0, y, t) = u_1(y, t), & (y, t) \in (0, 1) \times (0, T), \\ u(x, y, 0) = u_0(x, y), & (x, y) \in \Omega. \end{cases} \tag{3.1}$$

First one can prove that the problem is well-posed: for all $u_0 \in L^2(\Omega)$, $u_1 \in L^2((0,1)\times(0,T))$, and $f \in L^2((x_0,x_1)\times(0,T))$, problem (3.1) has a unique solution that satisfies $u \in \mathcal{C}^0([0,T];L^2(0,L;H^{-1}(0,1))) \cap \mathcal{C}^0([0,L];L^2(0,T;H^{-1}(0,1))) \cap L^2(\Omega\times(0,T))$.

Then we study the problem of regional null controllability. (As in the distributed case, due to the transport phenomenon in the x-direction, the region of influence in x of a control supported in x in (x_0,x_1) is only (x_0,x_1+T) in the case $T<L-x_1$ and is only (x_0,L) in the case $T>L-x_1$.) We assume that $T>0$, and we still define $\Omega_C(T,\delta)$ by (2.1) for all $0<\delta<(x_1-x_0)/2$. Then we prove the following observability estimate.

THEOREM 3.1. *Under the previous assumptions, there exists $C(T,\delta)>0$ such that the solutions v of the adjoint equation (2.2) belonging to $\mathcal{C}^0([0,T];L^2((0,L)\times(0,1))) \cap \mathcal{C}^0([0,L];L^2((0,T)\times(0,1))) \cap L^2((0,L)\times(0,T);H^2\cap H_0^1(0,1))$ satisfy*

$$\begin{aligned}(3.2)\quad &\iint_{(0,L)\times(0,1)} v(x,y,0)^2\,dydx \;+\; \iint_{(0,1)\times(0,T)} v(0,y,t)^2\,dtdy\\ &\le\; C\,(T,\delta)\left(\iint_{(0,T)\times(x_0,x_1)} v_y(x,1,t)^2\,dxdt\right.\\ &\quad\left.+\iint_{\Omega\setminus\Omega_C(T,\delta)} v(x,y,T)^2\,dydx \;+\; \iint_{(0,1)\times(0,T)} v(L,y,t)^2\,dtdy\right).\end{aligned}$$

Then we deduce the following result of null controllability.

THEOREM 3.2. *Under the previous assumptions, for all $u_0 \in L^2(\Omega)$, $u_1 \in L^2((0,1)\times(0,T))$, there exists $f \in L^2((x_0,x_1)\times(0,T))$ such that the solution u of (3.1) satisfies*

$$u(x,y,T)=0 \quad \textit{for } (x,y)\in\Omega_C(T,\delta).$$

4. The case of Neumann or mixed boundary conditions.

4.1. Locally distributed control. It is easy to adapt the proof of Theorem 2.2 to Neumann–Dirichlet or Neumann boundary conditions: let u^f be the solution of

$$(4.1)\qquad \begin{cases} u_t+u_x-u_{yy}=\chi_\omega(x,y)f(x,y,t), & (x,y,t)\in\Omega\times(0,T),\\ u_y(x,0,t)=u(x,1,t)=0, & (x,t)\in(0,L)\times(0,T),\\ u(0,y,t)=u_1(y,t), & (y,t)\in(0,1)\times(0,T),\\ u(x,y,0)=u_0(x,y), & (x,y)\in\Omega, \end{cases}$$

or of

$$(4.2)\qquad \begin{cases} u_t+u_x-u_{yy}=\chi_\omega(x,y)f(x,y,t), & (x,y,t)\in\Omega\times(0,T),\\ u_y(x,0,t)=u_y(x,1,t)=0, & (x,t)\in(0,L)\times(0,T),\\ u(0,y,t)=u_1(y,t), & (y,t)\in(0,1)\times(0,T),\\ u(x,y,0)=u_0(x,y), & (x,y)\in\Omega. \end{cases}$$

Then we have the following result of null controllability.

THEOREM 4.1. *Under the previous assumptions, for all $u_0 \in L^2(\Omega)$, $u_1 \in L^2((0,1)\times(0,T))$, there exists $f\in L^2(\omega\times(0,T))$ such that the solution u^f of (4.1) (or the solution u^f of (4.2)) satisfies*

$$(4.3)\qquad u^f(x,y,T)=0 \quad \textit{for } (x,y)\in\Omega_C(T,\delta).$$

The proof of Theorem 4.1 is very close to the one of Theorem 2.2, and we leave it to the reader.

4.2. Boundary control. It is easy to adapt the proof of Theorem 3.2 to consider Neumann–Dirichlet or Neumann boundary conditions: consider the solution u^f of

$$\begin{cases} u_t + u_x - u_{yy} = 0, & (x,y,t) \in \Omega \times (0,T), \\ u_y(x,0,t) = 0, & (x,t) \in (0,L) \times (0,T), \\ u(x,1,t) = \chi_{(x_0,x_1)}(x) f(x,t), & (x,t) \in (0,L) \times (0,T), \\ u(0,y,t) = u_1(y,t), & (y,t) \in (0,1) \times (0,T), \\ u(x,y,0) = u_0(x,y), & (x,y) \in \Omega, \end{cases} \tag{4.4}$$

or of

$$\begin{cases} u_t + u_x - u_{yy} = 0, & (x,y,t) \in \Omega \times (0,T), \\ u_y(x,0,t) = 0, & (x,t) \in (0,L) \times (0,T), \\ u_y(x,1,t) = \chi_{(x_0,x_1)}(x) f(x,t), & (x,t) \in (0,L) \times (0,T), \\ u(0,y,t) = u_1(y,t), & (y,t) \in (0,1) \times (0,T), \\ u(x,y,0) = u_0(x,y), & (x,y) \in \Omega. \end{cases} \tag{4.5}$$

Then we have the following result of null controllability.

THEOREM 4.2. *Under the previous assumptions, for all $u_0 \in L^2(\Omega)$, $u_1 \in L^2((0,1) \times (0,T))$, there exists $f \in L^2((x_0,x_1) \times (0,T))$ such that the solution u^f of (4.4) (or the solution u^f of (4.5)) satisfies*

$$u^f(x,y,T) = 0 \quad \textit{for } (x,y) \in \Omega_C(T,\delta). \tag{4.6}$$

The proof of Theorem 4.2 is very close to the one of Theorem 3.2. In section 6.3, we indicate how to derive the observability inequality needed for the proof of Theorem 4.2.

5. Proofs in the distributed case.

5.1. Proof of Proposition 2.1. Consider the eigenvalues $\lambda_k = k^2\pi^2$ and the eigenfunctions $\phi_k(y) = \sqrt{2}\sin(k\pi y)$ of the problem

$$-(\phi_k)_{yy} = \lambda_k \phi_k \text{ for } y \in (0,1) \text{ and } \phi_k(0) = \phi_k(1) = 0.$$

Let $u_0 \in L^2(\Omega)$, $u_1 \in L^2((0,1) \times (0,T))$, and $f \in L^2(\omega \times (0,T))$ and consider u the solution of (1.2). Since $(\phi_k)_k$ forms an orthonormal basis of $L^2(0,1)$, we can write

$$u_1(y,t) = \sum_{k=0}^{\infty} u_1^k(t)\phi_k(y), \qquad u_0(x,y) = \sum_{k=0}^{\infty} u_0^k(x)\phi_k(y)$$

$$\text{and } \chi_\omega(x,y) f(x,y,t) = \sum_{k=0}^{\infty} f^k(x,t)\phi_k(y), \qquad u(x,y,t) = \sum_{k=0}^{\infty} u^k(x,t)\phi_k(y).$$

Note that for all k, f^k is supported in x in (x_0,x_1). Then for all $k \in \mathbb{N}$, $u^k(x,t)$ is solution of

$$\begin{cases} u_t^k + u_x^k + \lambda_k u^k = f^k, & (x,t) \in (0,L) \times (0,T), \\ u^k(0,t) = u_1^k(t), & t \in (0,T), \\ u^k(x,0) = u_0^k(x), & x \in (0,L). \end{cases}$$

Consequently, we obtain

$$u^k(x,t) = \begin{cases} e^{-\lambda_k x}\, u_1^k(t-x) + \displaystyle\int_{t-x}^{t} e^{-\lambda_k (t-s)} f^k(x-(t-s),s)\, ds & \text{if } x < t, \\ e^{-\lambda_k t}\, u_0^k(x-t) + \displaystyle\int_{0}^{t} e^{-\lambda_k (t-s)} f^k(x-(t-s),s)\, ds & \text{if } x > t. \end{cases}$$

Since $\mathrm{supp}\,(f^k) \subset (x_0,x_1)\times(0,T)$, we deduce that for all $(x,T) \notin \Omega_C(T)$,

$$u^k(x,T) = \begin{cases} e^{-\lambda_k x}\, u_1^k(T-x) & \text{if } x < T, \\ e^{-\lambda_k T}\, u_0^k(x-T) & \text{if } x > T. \end{cases}$$

For all $\tilde\Omega$ such that $\Omega_C(T) \subset \tilde\Omega \subset \Omega$ and $\Omega_C(T) \neq \tilde\Omega$, this allows us to construct $u_0 \in L^2(\Omega)$ and $u_1 \in L^2((0,1)\times(0,T))$ such that for all $f \in L^2(\omega\times(0,T))$, the solution u of (1.2) is not identically equal to zero in $\tilde\Omega$. □

5.2. Proof of Theorem 2.1.

Step 1. In a first step, we prove that it is sufficient to establish (2.3) for *regular* solutions v of (2.2).

Let $v \in \mathcal{C}^0([0,T];L^2((0,L)\times(0,1))) \cap \mathcal{C}^0([0,L];L^2((0,T)\times(0,1))) \cap L^2((0,L)\times(0,T);H_0^1(0,1))$ be a solution of the adjoint problem (2.2). Let $v_L \in L^2((0,T)\times(0,1))$ and $v_T \in L^2((0,L)\times(0,1))$ be the functions defined by $v_L(y,t) := v(L,y,t)$ and $v_T(x,y) := v(x,y,T)$. Then v is a weak solution of the equation

$$\begin{cases} v_t + v_x + v_{yy} = 0, & (x,y,t) \in \Omega\times(0,T), \\ v(x,0,t) = v(x,1,t) = 0, & (x,t) \in (0,L)\times(0,T), \\ v(L,y,t) = v_L(y,t), & (y,t) \in (0,1)\times(0,T), \\ v(x,y,T) = v_T(x,y), & (x,y) \in \Omega. \end{cases}$$

Let $(v_L^n)_n$ be a sequence in $\mathcal{C}_c^1((0,1)\times[0,T)) \cap \mathcal{C}^2([0,1]\times[0,T])$ converging to v_L in $L^2((0,T)\times(0,1))$, and let $(v_T^n)_n$ be a sequence in $\mathcal{C}_c^1([0,L)\times(0,1)) \cap \mathcal{C}^2([0,L]\times[0,1])$ converging to v_T in $L^2((0,L)\times(0,1))$. Denote by v^n the solution of the equation

$$\begin{cases} v_t^n + v_x^n + v_{yy}^n = 0, & (x,y,t) \in \Omega\times(0,T), \\ v^n(x,0,t) = v^n(x,1,t) = 0, & (x,t) \in (0,L)\times(0,T), \\ v^n(L,y,t) = v_L^n(y,t), & (y,t) \in (0,1)\times(0,T), \\ v^n(x,y,T) = v_T^n(x,y), & (x,y) \in \Omega. \end{cases}$$

Using the explicit expression of v^n, we verify that $v^n \in \mathcal{C}^1([0,L]\times[0,T];L^2(0,1)) \cap \mathcal{C}^0([0,L]\times[0,T];H^2 \cap H_0^1(0,1))$ and that $v^n \to v$ in $\mathcal{C}^0([0,T];L^2((0,L)\times(0,1))) \cap \mathcal{C}^0([0,L];L^2((0,T)\times(0,1))) \cap L^2((0,L)\times(0,T);H_0^1(0,1))$ as $n \to +\infty$. Consequently, it is sufficient to prove (2.3) only for solutions having the regularity of v^n. Then (2.3) follows for v by passing to the limit as $n \to +\infty$.

Step 2. Now we assume $v \in \mathcal{C}^1([0,L]\times[0,T];L^2(0,1)) \cap \mathcal{C}^0([0,L]\times[0,T];H^2 \cap H_0^1(0,1))$ and we prove (2.3). We only treat the case $T < L - x_1 + \delta$ and $T \geq x_0 + \delta$ and we use a decomposition of the domain represented in Figure 5.1. (The other cases can be treated with a slightly different decomposition of the domain.)

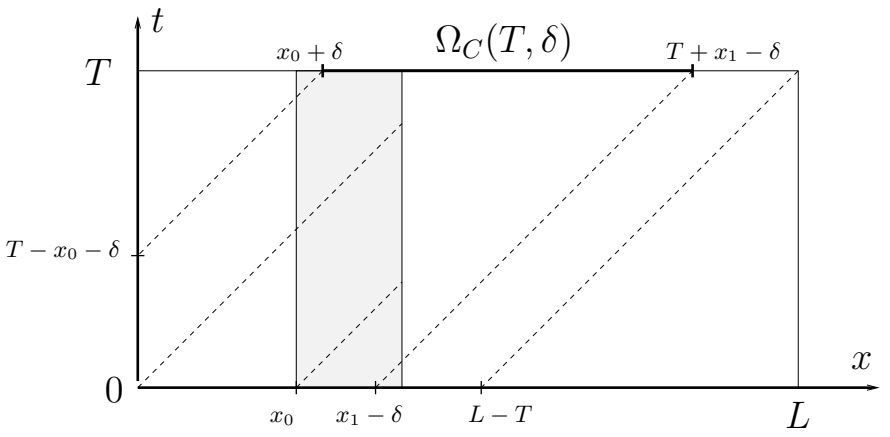


FIG. 5.1.

First we decompose the left-hand side of (2.3) as follows:

$$\begin{aligned}(5.1)\quad &\iint_{(0,L)\times(0,1)} v(x,y,0)^2\,dydx\\ &= \iint_{(0,x_0)\times(0,1)} v(x,y,0)^2\,dydx \;+\; \iint_{(x_0,x_1-\delta)\times(0,1)} v(x,y,0)^2\,dydx\\ &+ \iint_{(x_1-\delta,L-T)\times(0,1)} v(x,y,0)^2\,dydx \;+\; \iint_{(L-T,L)\times(0,1)} v(x,y,0)^2\,dydx\end{aligned}$$

and

$$\begin{aligned}(5.2)\quad &\iint_{(0,1)\times(0,T)} v(0,y,t)^2\,dtdy\\ &= \iint_{(0,1)\times(0,T-x_0-\delta)} v(0,y,t)^2\,dtdy \;+\; \iint_{(0,1)\times(T-x_0-\delta,T)} v(0,y,t)^2\,dtdy.\end{aligned}$$

Then we prove the following inequalities:

$$(5.3)\qquad \iint_{(0,x_0)\times(0,1)} v(x,y,0)^2\,dydx \;\le\; C\int_{x_0}^{x_1}\int_{\omega_y}\int_{x-x_0}^{x} v(x,y,t)^2\,dtdydx,$$

$$(5.4)\qquad \iint_{(x_0,x_1-\delta)\times(0,1)} v(x,y,0)^2\,dydx \;\le\; C\int_{x_0}^{x_1}\int_{\omega_y}\int_{0}^{x-x_0} v(x,y,t)^2\,dtdydx,$$

$$(5.5)\qquad \iint_{(x_1-\delta,L-T)\times(0,1)} v(x,y,0)^2\,dydx \;\le C\int_{x_1-\delta+T}^{L}\int_{0}^{1} v(x,y,T)^2\,dydx,$$

$$(5.6)\qquad \iint_{(L-T,L)\times(0,1)} v(x,y,0)^2\,dydx \;\le C\int_{0}^{1}\int_{0}^{T} v(L,y,t)^2\,dtdy,$$

$$\iint_{(0,1)\times(0,T-x_0-\delta)} v(0,y,t)^2\,dtdy \;\le C\int_{x_0}^{x_1}\int_{\omega_y}\int_x^T v(x,y,t)^2\,dtdydx, \tag{5.7}$$

$$\iint_{(0,1)\times(T-x_0-\delta,T)} v(0,y,t)^2\,dtdy \;\le C\int_0^{x_0+\delta}\int_0^1 v(x,y,T)^2\,dydx. \tag{5.8}$$

Theorem 2.1 follows clearly from (5.1)–(5.8). Note that the proofs of (5.5), (5.6), and (5.8) are similar. Thus we will prove only (5.8). The proofs of (5.3), (5.4), and (5.7) are also similar. And we will prove only (5.3) and (5.4). (We add the proof of (5.4) to show where we use the assumption $\delta > 0$.)

In order to prove these inequalities, we first note that if we consider the solution v of (2.2) along the characteristics, then v is solution of a nondegenerate parabolic equation.

Indeed, let $v \in \mathcal{C}^1([0,L]\times[0,T];L^2(0,1))\cap\mathcal{C}^0([0,L]\times[0,T];H^2\cap H_0^1(0,1))$ be a solution of (2.2). For all $\xi\in(-T,L), t\in(0,T), y\in(0,1)$ with $\xi+t\in(0,L)$, we introduce $w(\xi,y,t) := v(\xi+t,y,t)$. Then we verify that

$$w_t + w_{yy} = v_t + v_x + v_{yy} = 0, \quad (\xi,y,t)\in D,$$

where $D = \{(\xi,y,t)\in(-T,L)\times(0,1)\times(0,T) \;\mid\; \xi+t\in(0,L)\}$.

In particular, for all $\xi\in(-T,L)$ fixed, $w^\xi(y,t) := w(\xi,y,t)$ is the solution of a nondegenerate parabolic equation:

$$\begin{cases} w^\xi_t + w^\xi_{yy} = 0, & (y,t)\in(0,1)\times(t_0^\xi,t_1^\xi),\\ w^\xi(0,t) = w^\xi(1,t) = 0, & t\in(t_0^\xi,t_1^\xi), \end{cases} \tag{5.9}$$

where $t_0^\xi = \max(0,-\xi)$, $t_1^\xi = \min(T,L-\xi)$, and $w^\xi\in\mathcal{C}^0([t_0^\xi,t_1^\xi];H^2\cap H_0^1(0,1))$.

Note that, for all $\xi\in(-T,L)$, the energy of w^ξ is increasing, i.e.,

$$\text{for all } t_0^\xi\le T_0\le T_1\le t_1^\xi, \quad \int_0^1 w^\xi(y,T_0)^2\,dy \le \int_0^1 w^\xi(y,T_1)^2\,dy. \tag{5.10}$$

Then the proof of (5.5), (5.6), and (5.8) will be deduced from the fact that the energy of solutions of (5.9) is increasing. And the proof of (5.3), (5.4), and (5.7) will follow from Carleman estimates for the solutions of (5.9), using also the fact that the energy of solutions of (5.9) is increasing.

Proof of (5.8). From the definition of w, it follows that

$$\int_{T-x_0-\delta}^T\int_0^1 v(0,y,t)^2\,dydt \;=\; \int_{T-x_0-\delta}^T\int_0^1 w(-t,y,t)^2\,dydt.$$

Moreover, using that the energy of w^{-t} is increasing (we apply (5.10) to $\xi = -t\in(-T,-T+x_0+\delta)$, $T_0 = t$ and $T_1 = T$), for all $t\in(T-x_0-\delta,T)$, we have

$$\int_0^1 w(-t,y,t)^2\,dy \;\le\; \int_0^1 w(-t,y,T)^2\,dy.$$

We deduce that

$$\int_{T-x_0-\delta}^T\int_0^1 v(0,y,t)^2\,dydt \;\le\; \int_{T-x_0-\delta}^T\int_0^1 w(-t,y,T)^2\,dydt.$$

Since

$$\int_{T-x_0-\delta}^{T}\int_0^1 w(-t,y,T)^2\,dydt \;=\; \int_0^{x_0+\delta}\int_0^1 v(x,y,T)^2\,dydx,$$

we deduce (5.8). □

Proof of (5.3). We will use the following classical lemma.

LEMMA 5.1. *Let $\tilde T>0$ and let ω_y be an nonempty open set of $(0,1)$. Then there exists $C(\tilde T,\omega_y)>0$ such that the solutions $w\in\mathcal{C}^0([0,\tilde T];H^2(0,1)\cap H_0^1(0,1))$ of*

$$w_t(y,t)+w_{yy}(y,t)=0,\quad (y,t)\in(0,1)\times(0,\tilde T), \tag{5.11}$$

satisfy

$$\int_0^1 w(y,0)^2\,dy \;\le\; C(\tilde T,\omega_y)\int_0^{\tilde T}\int_{\omega_y} w(y,t)^2\,dydt. \tag{5.12}$$

This is by now a well-known observability inequality for the nondegenerate parabolic equation (5.11), and it follows from Carleman's estimates (see, for example, [16, 14, 2]).

In order to prove (5.3), we use w defined as previously. Let $\xi\in(0,x_0)$ be given. Then using that the energy of w^ξ is increasing (we apply (5.10) to $T_0=0$ and $T_1=x_0-\xi$), we have

$$\int_0^1 w(\xi,y,0)^2\,dy\le\int_0^1 w(\xi,y,x_0-\xi)^2\,dy.$$

Thus taking the integral over $\xi\in(0,x_0)$, we have

$$\int_0^{x_0}\int_0^1 w(\xi,y,0)^2\,dyd\xi \;\le\; \int_0^{x_0}\int_0^1 w(\xi,y,x_0-\xi)^2\,dyd\xi.$$

On the other hand, we note that (5.9) is a nondegenerate parabolic equation; we deduce from Lemma 5.1 that for all ξ, the solutions of (5.9) satisfy

$$\text{for all } t_0^\xi\le T_0\le T_1\le t_1^\xi,\quad \int_0^1 w^\xi(y,T_0)^2\,dy \;\le\; C\int_{T_0}^{T_1}\int_{\omega_y} w^\xi(y,t)^2\,dydt, \tag{5.13}$$

where C is a constant independent of ξ that depends on T_1-T_0 and on ω_y. Applying (5.13) for all $\xi\in(0,x_0)$ to $T_0=x_0-\xi$ and $T_1=x_1-\xi$, we deduce that

$$\int_0^1 w(\xi,y,x_0-\xi)^2\,dy \;\le\; C\int_{x_0-\xi}^{x_1-\xi}\int_{\omega_y} w(\xi,y,t)^2\,dydt,$$

where C is independent of ξ (it depends on x_1-x_0 and ω_y). Hence

$$\begin{aligned}\int_0^{x_0}\int_0^1 v(x,y,0)^2\,dydx \;&=\; \int_0^{x_0}\int_0^1 w(\xi,y,0)^2\,dyd\xi\\ &\le\; \int_0^{x_0}\int_0^1 w(\xi,y,x_0-\xi)^2\,dyd\xi\\ &\le\; C\int_0^{x_0}\int_{x_0-\xi}^{x_1-\xi}\int_{\omega_y} w(\xi,y,t)^2\,dydtd\xi.\end{aligned}$$

Making the change of variables $(x,t) := (\xi + t, t)$ and observing that

$$\begin{cases} 0 \le \xi \le x_0, \\ x_0 - \xi \le t \le x_1 - \xi \end{cases} \iff \begin{cases} x_0 \le x \le x_1, \\ x - x_0 \le t \le x, \end{cases}$$

we deduce that

$$\begin{aligned} \int_0^{x_0} \int_0^1 v(x,y,0)^2 \, dydx &\le C \int_{x_0}^{x_1} \int_{x-x_0}^{x} \int_{\omega_y} w(x-t,y,t)^2 \, dydtdx \\ &= C \int_{x_0}^{x_1} \int_{x-x_0}^{x} \int_{\omega_y} v(x,y,t)^2 \, dydtdx. \qquad \square \end{aligned}$$

Proof of (5.4). More precisely, we will prove

$$\iint_{(x_0,x_1-\delta)\times(0,1)} v(x,y,0)^2 \, dydx \le C \int_{x_1-\delta}^{x_1} \int_{\omega_y} \int_{x-(x_1-\delta)}^{x-x_0} v(x,y,t)^2 \, dtdydx.$$

We use w defined as previously. Let $\xi \in (x_0, x_1 - \delta)$ be given. Then using that the energy of w^ξ is increasing (we apply (5.10) to $T_0 = 0$ and $T_1 = x_1 - \delta - \xi$), we have

$$\int_0^1 w(\xi,y,0)^2 \, dy \le \int_0^1 w(\xi,y,x_1-\delta-\xi)^2 \, dy.$$

Thus taking the integral over $\xi \in (x_0, x_1 - \delta)$, we have

$$\int_{x_0}^{x_1-\delta} \int_0^1 w(\xi,y,0)^2 \, dyd\xi \le \int_{x_0}^{x_1-\delta} \int_0^1 w(\xi,y,x_1-\delta-\xi)^2 \, dyd\xi.$$

On the other hand, applying (5.13) for all $\xi \in (x_0, x_1 - \delta)$ to $T_0 = x_1 - \delta - \xi$ and $T_1 = x_1 - \xi$, we deduce that

$$\int_0^1 w(\xi,y,x_1-\delta-\xi)^2 \, dy \le C \int_{x_1-\delta-\xi}^{x_1-\xi} \int_{\omega_y} w(\xi,y,t)^2 \, dydt,$$

where C depends on $T_1 - T_0 = \delta$ (and is independent of ξ). Hence

$$\begin{aligned} \int_{x_0}^{x_1-\delta} \int_0^1 v(x,y,0)^2 \, dydx &= \int_{x_0}^{x_1-\delta} \int_0^1 w(\xi,y,0)^2 \, dyd\xi \\ &\le \int_{x_0}^{x_1-\delta} \int_0^1 w(\xi,y,x_1-\delta-\xi)^2 \, dyd\xi \\ &\le C \int_{x_0}^{x_1-\delta} \int_{x_1-\delta-\xi}^{x_1-\xi} \int_{\omega_y} w(\xi,y,t)^2 \, dydtd\xi \\ &= C \int_{x_1-\delta}^{x_1} \int_{x-(x_1-\delta)}^{x-x_0} \int_{\omega_y} w(x-t,y,t)^2 \, dydtdx \\ &= C \int_{x_1-\delta}^{x_1} \int_{x-(x_1-\delta)}^{x-x_0} \int_{\omega_y} v(x,y,t)^2 \, dydtdx. \qquad \square \end{aligned}$$

5.3. Proof of Theorem 2.2. For all $\varepsilon > 0$, consider the penalized problem

$$\text{Min } \{J_\varepsilon(f) \ \mid \ f \in L^2(\omega \times (0,T))\}, \tag{5.14}$$

where

$$J_\varepsilon(f) \ := \ \frac{1}{2} \iiint_{\omega\times(0,T)} f(x,y,t)^2 \, dtdydx \ + \ \frac{1}{2\varepsilon} \iint_{\Omega_C(T,\delta)} u^f(x,y,T)^2 \, dydx,$$

with u^f the solution of (1.2) associated with f.

Step 1 (characterization of the solution). The functional J_ε is continuous on $L^2(\Omega \times (0,T))$ and strictly convex, and $J_\varepsilon(f) \to \infty$ as $\|f\|_{L^2(\Omega\times(0,T))} \to \infty$. Thus, for all $\varepsilon > 0$, problem (5.14) has a unique solution f^ε. And we can verify that it is characterized by

$$f^\varepsilon(x,y,t) = -v^\varepsilon(x,y,t)\chi_\omega(x,y), \tag{5.15}$$

where v^ε is the solution of the adjoint problem

$$\begin{cases} v^\varepsilon_t + v^\varepsilon_x + v^\varepsilon_{yy} = 0, & (x,y,t) \in \Omega \times (0,T), \\ v^\varepsilon(x,0,t) = v^\varepsilon(x,1,t) = 0, & (x,t) \in (0,L) \times (0,T), \\ v^\varepsilon(L,y,t) = 0, & (y,t) \in (0,1) \times (0,T), \\ v^\varepsilon(x,y,T) = \frac{1}{\varepsilon}\chi_{\Omega_C(T,\delta)} u^{f^\varepsilon}(x,y,T), & (x,y) \in \Omega. \end{cases} \tag{5.16}$$

Indeed, f^ε is characterized by $DJ_\varepsilon(f^\varepsilon) \cdot h = 0$ for all $h \in L^2(\omega \times (0,T))$. (As usual, $DJ_\varepsilon(f^\varepsilon) \cdot h = 0$ denotes the differential of the functional J_ε computed at the point f^ε and applied to the element h.) By classical computations, we obtain

$$DJ_\varepsilon(f) \cdot h \ = \ \iiint_{\omega\times(0,T)} fh \, dtdydx + \ \frac{1}{\varepsilon} \iint_{\Omega_C(T,\delta)} u^f(T) z^h(T) \, dydx,$$

where z^h is the solution of

$$\begin{cases} z^h_t + z^h_x - z^h_{yy} = \chi_\omega h, & (x,y,t) \in \Omega \times (0,T), \\ z^h(x,0,t) = z^h(x,1,t) = 0, & (x,t) \in (0,L) \times (0,T), \\ z^h(0,y,t) = 0, & (y,t) \in (0,1) \times (0,T), \\ z^h(x,y,0) = 0, & (x,y) \in \Omega. \end{cases} \tag{5.17}$$

On the other hand, we multiply (5.16) by z^h and (5.17) by v^ε. We add these two relations and we take the integral over $\Omega \times (0,T)$. This gives

$$\begin{aligned} &\iint_\Omega v^\varepsilon(x,y,T) z^h(x,y,T) \, dydx - \iint_\Omega v^\varepsilon(x,y,0) z^h(x,y,0) \, dydx \\ &\quad + \int_0^T \int_0^1 v^\varepsilon(L,y,t) z^h(L,y,t) \, dydt - \int_0^T \int_0^1 v^\varepsilon(0,y,t) z^h(0,y,t) \, dydt \\ &\quad = \iiint_{\Omega\times(0,T)} \chi_\omega(x,y) h(x,y,t) v^\varepsilon(x,y,t) \, dtdydx. \end{aligned}$$

Since $z^h(0,y,t) = 0$, $z^h(x,y,0) = 0$, $v^\varepsilon(x,y,T) = \frac{1}{\varepsilon}\chi_{\Omega_C(T,\delta)} u^{f^\varepsilon}(x,y,T)$, and $v^\varepsilon(L,y,t) = 0$, we deduce

$$\frac{1}{\varepsilon} \iint_{\Omega_C(T,\delta)} u^{f^\varepsilon}(T) z^h(T) \, dydx = \iiint_{\omega\times(0,T)} h v^\varepsilon \, dtdydx.$$

Thus f^ε is characterized as follows: for all $h \in L^2(\omega \times (0,T))$,

$$\iiint_{\omega\times(0,T)} f^\varepsilon h \, dtdydx + \iiint_{\omega\times(0,T)} v^\varepsilon h \, dtdydx = 0,$$

which gives (5.15).

Step 2 (a priori estimates). Now we need suitable a priori estimates independent of ε to let $\varepsilon \to 0$. We multiply (1.2) by v^ε and (5.16) by u^{f^ε}. Then we add these identities, and we take the integral over $\Omega \times (0,T)$ to obtain

$$\begin{aligned}
&\iint_\Omega u^{f^\varepsilon}(x,y,T)v^\varepsilon(x,y,T)\,dydx \;-\; \iint_\Omega u_0(x,y)v^\varepsilon(x,y,0)\,dydx \\
&+ \iint_{(0,1)\times(0,T)} u^{f^\varepsilon}(L,y,t)v^\varepsilon(L,y,t)\,dtdy \;-\; \iint_{(0,1)\times(0,T)} u_1(y,t)v^\varepsilon(0,y,t)\,dtdy \\
&= \iiint_{\Omega\times(0,T)} \chi_\omega f^\varepsilon(x,y,t)v^\varepsilon(x,y,t)\,dtdydx \;=\; -\iiint_{\omega\times(0,T)} f^\varepsilon(x,y,t)^2\,dtdydx.
\end{aligned}$$

Since $v^\varepsilon(L,y,t) = 0$ and $v^\varepsilon(x,y,T) = \frac{1}{\varepsilon}u^{f^\varepsilon}(x,y,T)\chi_{\Omega_C(T,\delta)}(x,y)$, with Young's inequality, we obtain

$$\begin{aligned}
&\frac{1}{\varepsilon}\iint_{\Omega_C(T,\delta)} u^{f^\varepsilon}(x,y,T)^2\,dydx \;+\; \iiint_{\omega\times(0,T)} f^\varepsilon(x,y,t)^2\,dtdydx \\
&\quad= \iint_\Omega u_0(x,y)v^\varepsilon(x,y,0)\,dydx \;+\; \iint_{(0,1)\times(0,T)} u_1(y,t)v^\varepsilon(0,y,t)\,dtdy \\
&\quad\le \frac{1}{4\gamma}\iint_\Omega u_0(x,y)^2 dydx \;+\; \frac{1}{4\gamma}\iint_{(0,1)\times(0,T)} u_1(y,t)^2\,dtdy \\
&\qquad+ \gamma\iint_\Omega v^\varepsilon(x,y,0)^2\,dydx \;+\; \gamma\iint_{(0,1)\times(0,T)} v^\varepsilon(0,y,t)^2\,dtdy
\end{aligned}$$

for all $\gamma > 0$. From Theorem 2.1, it follows that we can choose $\gamma > 0$ small enough to have

$$\begin{aligned}
&\frac{1}{\varepsilon}\iint_{\Omega_C(T,\delta)} u^{f^\varepsilon}(x,y,T)^2\,dydx \;+\; \iiint_{\omega\times(0,T)} f^\varepsilon(x,y,t)^2\,dtdydx \\
&\quad\le C\left(\|u_0\|^2_{L^2(\Omega)} \;+\; \|u_1\|^2_{L^2((0,1)\times(0,T))}\right) \;+\; \frac{1}{2}\iiint_{\omega\times(0,T)} v^\varepsilon(x,y,t)^2\,dtdydx \\
&\qquad+ C\iint_{\Omega\setminus\Omega_C(T,\delta)} v^\varepsilon(x,y,T)^2\,dydx \;+\; C\iint_{(0,1)\times(0,T)} v^\varepsilon(L,y,t)^2\,dtdy.
\end{aligned}$$

Since $v_\varepsilon(x,y,T) = 0$ on $\Omega\setminus\Omega_C(T,\delta)$, $v_\varepsilon(L,y,t) = 0$ on $(0,1)\times(0,T)$, and $v_\varepsilon = -\chi_\omega f^\varepsilon$, we obtain

$$\begin{aligned}
&\frac{1}{\varepsilon}\iint_{\Omega_C(T,\delta)} u^{f^\varepsilon}(x,y,T)^2\,dydx \;+\; \frac{1}{2}\iiint_{\omega\times(0,T)} f^\varepsilon(x,y,t)^2\,dtdydx \\
&\quad\le C\left(\|u_0\|^2_{L^2(\Omega)} \;+\; \|u_1\|^2_{L^2((0,1)\times(0,T))}\right).
\end{aligned}$$

This gives the a priori estimates that allows us to pass to the limit in (1.2) as $\varepsilon \to 0$, which gives a solution to the null controllability problem. Indeed, since the sequence

$(f^\varepsilon)_\varepsilon$ is bounded in $L^2(\omega\times(0,T))$, there exists a subsequence still denoted by $(f^\varepsilon)_\varepsilon$ and there exists $\tilde f\in L^2(\omega\times(0,T))$ such that

$$f^\varepsilon \rightharpoonup \tilde f \text{ weakly in } L^2(\omega\times(0,T)) \text{ as } \varepsilon\to 0.$$

We denote by $\tilde u$ the solution of (1.2) associated with $\tilde f$. From the convergence of $(f^\varepsilon)_\varepsilon$ to $\tilde f$, it follows that $\left(\chi_{\Omega_C(T,\delta)}u^{f^\varepsilon}(T)\right)_\varepsilon$ converges to $\chi_{\Omega_C(T,\delta)}\tilde u(T)$ weakly in $L^2(\Omega)$. Moreover the sequence $\left(\varepsilon^{-1/2}\chi_{\Omega_C(T,\delta)}u^{f^\varepsilon}(T)\right)_\varepsilon$ is also bounded in $L^2(\Omega)$. Thus there exists a subsequence still indexed by ε and there exists $v_T\in L^2(\Omega)$ such that

$$\varepsilon^{-1/2}\chi_{\Omega_C(T,\delta)}u^{f^\varepsilon}(T)\rightharpoonup v_T \text{ weakly in } L^2(\Omega) \text{ as } \varepsilon\to 0.$$

Thus $\left(\chi_{\Omega_C(T,\delta)}u^{f^\varepsilon}(T)\right)_\varepsilon$ also converges to 0 weakly in $L^2(\Omega)$. Therefore we have $\chi_{\Omega_C(T,\delta)}\tilde u(T)\equiv 0$, which proves Theorem 2.2. ∎

6. Proofs in the boundary case.

6.1. Proof of Theorem 3.1. As in the proof of Theorem 2.1, it is sufficient to prove (3.2) only for *regular* solutions of (2.2). Thus we assume that $v\in\mathcal{C}^1([0,L]\times[0,T];L^2(0,1))\cap\mathcal{C}^0([0,L]\times[0,T];H^2\cap H_0^1(0,1))$.

We only treat the case $T<L-x_1+\delta$ and $T\geq x_0+\delta$ and we still use the decomposition of the domain represented in Figure 5.1. First we decompose the left-hand side of (3.2) as in the proof of Theorem 2.1 (see (5.1) and (5.2)).

Then we remark that (5.5), (5.6), and (5.8) still hold (with the same proof). Next, instead of (5.3), (5.4), and (5.7), it is sufficient to prove, respectively, the following inequalities:

$$\iint_{(0,x_0)\times(0,1)} v(x,y,0)^2\,dydx \;\leq\; C\int_{x_0}^{x_1}\int_{x-x_0}^{x} v_y(x,1,t)^2\,dtdx, \tag{6.1}$$

$$\iint_{(x_0,x_1-\delta)\times(0,1)} v(x,y,0)^2\,dydx \;\leq\; C\int_{x_0}^{x_1}\int_{0}^{x-x_0} v_y(x,1,t)^2\,dtdx, \tag{6.2}$$

$$\iint_{(0,1)\times(0,T-x_0-\delta)} v(0,y,t)^2\,dtdy \;\leq C\int_{x_0}^{x_1}\int_{x}^{T} v_y(x,1,t)^2\,dtdx. \tag{6.3}$$

Since the proofs are similar, we will just completely prove (6.1).

Proof of (6.1). Let v be a regular solution of (2.2). We use w and w^ξ defined in the proof of Theorem 2.1. As in the proof of (5.3), we have

$$\int_0^{x_0}\int_0^1 w(\xi,y,0)^2\,dyd\xi \;\leq\; \int_0^{x_0}\int_0^1 w(\xi,y,x_0-\xi)^2\,dyd\xi.$$

On the other hand, we recall the classical boundary observability inequality for nondegenerate parabolic equations that follows from classical Carleman's estimates (see for example [16, 14, 2]):

LEMMA 6.1. *For all $\tilde T>0$, there exists $C(\tilde T)>0$ such that the solutions $w\in\mathcal{C}^0([0,\tilde T];H^2\cap H_0^1(0,1))$ of*

$$w_t(y,t)+w_{yy}(y,t)=0,\qquad (y,t)\in(0,1)\times(0,\tilde T),$$

satisfy

$$\int_0^1 w(y,0)^2\,dy \;\le\; C(\tilde T)\;\int_0^{\tilde T} w_y(1,t)^2\,dt.$$

From Lemma 6.1, we deduce that for all ξ the solutions of (5.9) satisfy the following *boundary* observability inequality:

$$\text{for all } T_1 > T_0 \ge 0, \qquad \int_0^1 w^\xi(y,T_0)^2\,dy \;\le\; C\;\int_{T_0}^{T_1} w_y^\xi(1,t)^2\,dt, \tag{6.4}$$

where C is a constant independent of ξ that depends on $T_1 - T_0$. Applying (6.4) for all $\xi \in (0,x_0)$ to $T_0 = x_0 - \xi$ and $T_1 = x_1 - \xi$, we deduce that

$$\int_0^1 w(\xi,y,x_0-\xi)^2\,dy \;\le\; C\;\int_{x_0-\xi}^{x_1-\xi} w_y(\xi,1,t)^2\,dt.$$

Hence

$$\begin{aligned}\int_0^{x_0}\int_0^1 v(x,y,0)^2\,dydx \;&=\; \int_0^{x_0}\int_0^1 w(\xi,y,0)^2\,dyd\xi\\ &\le\; \int_0^{x_0}\int_0^1 w(\xi,y,x_0-\xi)^2\,dyd\xi \;\le\; C\int_0^{x_0}\int_{x_0-\xi}^{x_1-\xi} w_y(\xi,1,t)^2\,dtd\xi\\ &=\; C\int_{x_0}^{x_1}\int_{x-x_0}^{x} w_y(x-t,1,t)^2\,dtdx \;=\; C\int_{x_0}^{x_1}\int_{x-x_0}^{x} v_y(x,1,t)^2\,dtdx. \qquad \square\end{aligned}$$

6.2. Proof of Theorem 3.2. First, we note that for all $u_0 \in L^2(\Omega)$, $u_1 \in L^2((0,1)\times(0,T))$ and $f \in L^2((x_0,x_1)\times(0,T))$; then the weak solution u of (3.1), defined by the transposition method (see Lions [21]), belongs to $\mathcal{C}^0([0,T];L^2(0,L;H^{-1}(0,1)))\cap\mathcal{C}^0([0,L];L^2(0,T;H^{-1}(0,1)))\cap L^2(\Omega\times(0,T))$. Since we only know that $u(\cdot,T)$ belongs to $L^2(0,L;H^{-1}(0,1))$, we will introduce a modified penalized problem (so that the solution will be characterized via an adjoint problem with a regular terminal condition).

We denote by $(-\Delta_y)^{-1}$ the inverse of the isomorphism $-\Delta_y : p \in H_0^1(0,1) \mapsto -p_{yy} \in H^{-1}(0,1)$. We also denote by $\langle\cdot,\cdot\rangle$ the duality product $\langle\cdot,\cdot\rangle_{H^{-1}(0,1)\times H_0^1(0,1)}$. We recall that $q \mapsto \langle q,(-\Delta_y)^{-1}q\rangle^{1/2}$ defines a norm that is equivalent to the norm $q \mapsto \|q\|_{H^{-1}(0,1)}$ on $H^{-1}(0,1)$. Indeed, using that $\|(-\Delta_y)^{-1}q\|_{H_0^1(0,1)}$ is a norm equivalent to the norm $\|q\|_{H^{-1}(0,1)}$ on $H^{-1}(0,1)$, we deduce

$$\begin{aligned}C_1\|q\|^2_{H^{-1}(0,1)} \le \|(-\Delta_y)^{-1}q\|^2_{H_0^1(0,1)} &= \langle(-\Delta_y)(-\Delta_y)^{-1}q,(-\Delta_y)^{-1}q\rangle\\ &= \langle q,(-\Delta_y)^{-1}q\rangle \le \|q\|_{H^{-1}(0,1)}\|(-\Delta_y)^{-1}q\|_{H_0^1(0,1)} \le C_2\|q\|^2_{H^{-1}(0,1)}.\end{aligned}$$

Finally, we also set $\mathcal{O} := (x_0+\delta, x_1+T-\delta)$; we recall that $\Omega_C(T,\delta) = \mathcal{O}\times(0,1)$. Thus $\chi_{\Omega_C(T,\delta)}(x,y) = \chi_{\mathcal{O}}(x)$ for all $(x,y) \in (0,L)\times(0,1)$.

Then for all $\varepsilon > 0$, we consider the penalized problem

$$\text{Min } \{J_\varepsilon(f) \;\mid\; f \in L^2((x_0,x_1)\times(0,T))\}, \tag{6.5}$$

where

$$\begin{aligned}J_\varepsilon(f) \;:=\; &\frac12\iint_{(0,T)\times(x_0,x_1)} f(x,t)^2\,dxdt\\ &+\;\frac{1}{2\varepsilon}\int_0^L \chi_{\mathcal{O}}(x)\Big\langle u^f(x,\cdot,T),(-\Delta_y)^{-1}u^f(x,\cdot,T)\Big\rangle\,dx,\end{aligned}$$

with u^f the solution of (3.1) associated with f.

Step 1 (characterization of the solution). This problem has a unique solution f^ε that it is characterized by

$$f^\varepsilon(x,t) = v_y^\varepsilon(x,1,t)\chi_{(x_0,x_1)}(x), \tag{6.6}$$

where v^ε is the solution of the adjoint problem

$$\begin{cases} v_t^\varepsilon + v_x^\varepsilon + v_{yy}^\varepsilon = 0, & (x,y,t) \in \Omega \times (0,T), \\ v^\varepsilon(x,0,t) = v^\varepsilon(x,1,t) = 0, & (x,t) \in (0,L) \times (0,T), \\ v^\varepsilon(L,y,t) = 0, & (y,t) \in (0,1) \times (0,T), \\ v^\varepsilon(x,y,T) = \frac{1}{\varepsilon}\chi_{\mathcal{O}}(x)(-\Delta_y)^{-1}u^{f^\varepsilon}(x,y,T), & (x,y) \in \Omega. \end{cases} \tag{6.7}$$

Indeed, f^ε is characterized by $DJ_\varepsilon(f^\varepsilon)\cdot h = 0$ for all $h \in L^2((x_0,x_1)\times(0,T))$. By classical computations, we obtain

$$\begin{aligned} DJ_\varepsilon(f)\cdot h \;=\; & \int_0^T \int_{x_0}^{x_1} fh\,dxdt \\ & + \frac{1}{2\varepsilon}\int_0^L \chi_{\mathcal{O}}(x)\Big\langle u^f(x,\cdot,T), (-\Delta_y)^{-1}z^h(x,\cdot,T)\Big\rangle\,dx \\ & + \frac{1}{2\varepsilon}\int_0^L \chi_{\mathcal{O}}(x)\Big\langle z^h(x,\cdot,T), (-\Delta_y)^{-1}u^f(x,\cdot,T)\Big\rangle\,dx \\ =\; & \int_0^T \int_{x_0}^{x_1} fh\,dxdt + \frac{1}{\varepsilon}\int_0^L \chi_{\mathcal{O}}(x)\Big\langle z^h(x,\cdot,T), (-\Delta_y)^{-1}u^f(x,\cdot,T)\Big\rangle\,dx, \end{aligned}$$

where z^h is the solution of

$$\begin{cases} z_t^h + z_x^h - z_{yy}^h = 0, & (x,y,t) \in \Omega \times (0,T), \\ z^h(x,0,t) = 0, & (x,t) \in (0,L) \times (0,T), \\ z^h(x,1,t) = \chi_{(x_0,x_1)}(x)h(x,t), & (x,t) \in (0,L) \times (0,T), \\ z^h(0,y,t) = 0, & (y,t) \in (0,1) \times (0,T), \\ z^h(x,y,0) = 0, & (x,y) \in \Omega. \end{cases} \tag{6.8}$$

On the other hand, we multiply (6.7) by z^h and (6.8) by v^ε and we add these two relations. Then using the initial and terminal data of z^h and v^ε, this gives

$$\begin{aligned} & \frac{1}{\varepsilon}\int_0^L \chi_{\mathcal{O}}(x)\Big\langle z^h(x,\cdot,T), (-\Delta_y)^{-1}u^{f^\varepsilon}(x,\cdot,T)\Big\rangle\,dx \\ & = \int_0^L \Big\langle z^h(x,\cdot,T), v^\varepsilon(x,\cdot,T)\Big\rangle\,dx \;=\; -\int_0^T\int_0^L \chi_{(x_0,x_1)}(x)v_y(x,1,t)h(x,t)\,dxdt. \end{aligned}$$

Thus f^ε is characterized as follows: for all $h \in L^2((x_0,x_1)\times(0,T))$,

$$\int_0^T \int_{x_0}^{x_1} f(x,t)h(x,t)\,dxdt - \int_0^T \int_{x_0}^{x_1} v_y(x,1,t)h(x,t)\,dxdt = 0,$$

which gives (6.6).

Step 2 (a priori estimates). Now we need suitable a priori estimates to let $\varepsilon \to 0$. We multiply (3.1) by v^ε and (6.7) by u^{f^ε}. Then we add these identities, and we integrate over $\Omega \times (0,T)$ to obtain

$$\begin{aligned}
&\int_0^L \Big\langle u^{f^\varepsilon}(x,\cdot,T), v^\varepsilon(x,\cdot,T)\Big\rangle\, dx - \int_0^L \Big\langle u^{f^\varepsilon}(x,\cdot,0), v^\varepsilon(x,\cdot,0)\Big\rangle\, dx \\
&\quad + \int_0^T \Big\langle u^{f^\varepsilon}(L,\cdot,t), v^\varepsilon(L,\cdot,t)\Big\rangle\, dt - \int_0^T \Big\langle\ u^{f^\varepsilon}(0,\cdot,t), v^\varepsilon(0,\cdot,t)\Big\rangle\, dt \\
&\quad = -\int_0^T \int_0^L v_y^\varepsilon(x,1,t) u^{f^\varepsilon}(x,1,t)\, dxdt = -\int_0^T \int_0^L \chi_{(x_0,x_1)}(x) f^\varepsilon(x,t)^2\, dxdt.
\end{aligned}$$

Since $v^\varepsilon(L,y,t) = 0$ and $v^\varepsilon(x,y,T) = \frac{1}{\varepsilon}\chi_{\mathcal{O}}(x)(-\Delta_y)^{-1} u^{f^\varepsilon}(x,y,T)$, with Young's inequality we obtain

$$\begin{aligned}
&\frac{1}{\varepsilon}\int_0^L \chi_{\mathcal{O}}(x)\Big\langle u^{f^\varepsilon}(x,\cdot,T), (-\Delta_y)^{-1} u^{f^\varepsilon}(x,\cdot,T)\Big\rangle\, dx \ +\ \iint_{(0,T)\times(x_0,x_1)} f^\varepsilon(x,t)^2\, dxdt \\
&\quad =\ \iint_\Omega u_0(x,y) v^\varepsilon(x,y,0)\, dydx \ +\ \iint_{(0,1)\times(0,T)} u_1(y,t) v^\varepsilon(0,y,t)\, dtdy \\
&\quad \le\ \frac{1}{4\gamma}\left(\|u_0\|^2_{L^2(\Omega)} \ +\ \|u_1\|^2_{L^2((0,1)\times(0,T)}\right) \\
&\qquad +\ \gamma \iint_\Omega v^\varepsilon(x,y,0)^2\, dydx \ +\ \gamma \iint_{(0,1)\times(0,T)} v^\varepsilon(0,y,t)^2\, dtdy
\end{aligned}$$

for all $\gamma > 0$. From Theorem 3.1, it follows that we can choose $\gamma > 0$ small enough to have

$$\begin{aligned}
&\frac{1}{\varepsilon}\int_0^L \chi_{\mathcal{O}}(x)\Big\langle u^{f^\varepsilon}(x,\cdot,T), (-\Delta_y)^{-1} u^{f^\varepsilon}(x,\cdot,T)\Big\rangle\, dx \ +\ \iint_{(0,T)\times(x_0,x_1)} f^\varepsilon(x,t)^2\, dxdt \\
&\quad \le\ C\left(\|u_0\|^2_{L^2(\Omega)} \ +\ \|u_1\|^2_{L^2((0,1)\times(0,T))}\right) \ +\ \frac{1}{2}\iint_{(0,T)\times(x_0,x_1)} v_y^\varepsilon(x,1,t)^2\, dxdt \\
&\qquad +\ C\iint_{\Omega\setminus\Omega_C(T,\delta)} v^\varepsilon(x,y,T)^2\, dydx \ +\ C\iint_{(0,1)\times(0,T)} v^\varepsilon(L,y,t)^2\, dtdy.
\end{aligned}$$

Since $v_\varepsilon(x,y,T) = 0$ on $\Omega \setminus \Omega_C(T,\delta)$, $v_\varepsilon(L,y,t) = 0$ on $(0,1)\times(0,T)$, and $f^\varepsilon(x,t) = v_y^\varepsilon(x,1,t)\chi_{(x_0,x_1)}(x)$, we obtain that

$$\begin{aligned}
&\frac{1}{\varepsilon}\int_0^L \chi_{\mathcal{O}}(x)\Big\langle u^{f^\varepsilon}(x,\cdot,T), (-\Delta_y)^{-1} u^{f^\varepsilon}(x,\cdot,T)\Big\rangle\, dx \ +\ \frac{1}{2}\iint_{(0,T)\times(x_0,x_1)} f^\varepsilon(x,t)^2\, dxdt \\
&\quad \le\ C\left(\|u_0\|^2_{L^2(\Omega)} \ +\ \|u_1\|^2_{L^2((0,1)\times(0,T))}\right).
\end{aligned}$$

Since $\chi_{\mathcal{O}}(x) = \chi_{\Omega_C(T,\delta)}(x,y)$, this implies

$$\begin{aligned}
&\frac{1}{\varepsilon}\int_0^L \|\chi_{\Omega_C(T,\delta)} u^{f^\varepsilon}(x,\cdot,T)\|^2_{H^{-1}(0,1)}\, dx \ +\ \frac{1}{2}\iint_{(0,T)\times(x_0,x_1)} f^\varepsilon(x,t)^2\, dxdt \\
&\qquad \le\ C\left(\|u_0\|^2_{L^2(\Omega)} \ +\ \|u_1\|^2_{L^2((0,1)\times(0,T))}\right).
\end{aligned}$$

Finally we deduce that the sequence $(f^\varepsilon)_\varepsilon$ is bounded in $L^2((x_0,x_1)\times(0,T))$ and that the sequence $\left(\varepsilon^{-1/2}\chi_{\Omega_C(T,\delta)} u^{f^\varepsilon}(T)\right)_\varepsilon$ is also bounded in $L^2(0,L;H^{-1}(0,1))$. This gives a priori estimates that allows us to pass to the limit in (3.1) as $\varepsilon \to 0$, which gives a solution to the null controllability problem. ∎

6.3. Main tool for the proof of Theorem 4.2. To prove Theorem 4.2, we need to combine the previous proof with a suitable observability estimate. In the case of Neumann boundary conditions (4.5), we use the following.

LEMMA 6.2. *For all $\tilde{T} > 0$, there exists $C(\tilde{T}) > 0$ such that the solutions $w \in \mathcal{C}^0([0,\tilde{T}]; H^2(0,1))$ of*

$$\begin{cases} w_t(y,t) + w_{yy}(y,t) = 0, & (y,t) \in (0,1) \times (0,\tilde{T}), \\ w_y(0,t) = 0 = w_y(1,t), & t \in (0,\tilde{T}), \end{cases} \tag{6.9}$$

satisfy

$$\int_0^1 w(y,0)^2 \, dy \;\le\; C(\tilde{T}) \int_0^{\tilde{T}} w(1,t)^2 \, dt. \tag{6.10}$$

This result follows from the well-known equivalence between null controllability and observability (see, for example, [26, Theorem 2.6, p. 213]). For the reader's convenience, we give a short proof of (6.10): first consider the solution u of

$$\begin{cases} u_t - u_{yy} = 0, & (y,t) \in (0,1) \times (0,T), \\ u_y(0,t) = 0, & (y,t) \in (0,L) \times (0,T), \\ u_y(1,t) = h(t), & t \in (0,T), \\ u(0,t) = u_0(y), & y \in (0,1), \\ u(y,T) = 0, & y \in (0,1). \end{cases} \tag{6.11}$$

Using, e.g., [12], given $u_0 \in L^2(0,1)$, there always exists $h \in L^2(0,T)$ such that the problem (6.11) has a solution. Moreover, h can be determined so that

$$\|h\|_{L^2(0,T)} \le C \|u_0\|_{L^2(0,1)}$$

for some positive constant C independent of u_0. Now we multiply (6.9) by u, and after some integrations by parts, we obtain that

$$\int_0^1 w(y,0) u_0(y) \, dy = -\int_0^T w(1,t) h(t) \, dt.$$

Using the Cauchy–Schwarz inequality

$$\left| \int_0^1 w(y,0) u_0(y) \, dy \right| \le \|h\|_{L^2(0,T)} \|w(1,\cdot)\|_{L^2(0,T)} \le C \|u_0\|_{L^2(0,1)} \|w(1,\cdot)\|_{L^2(0,T)}.$$

Since this is true for all $u_0 \in L^2(0,1)$, it implies that

$$\|w(\cdot,0)\|_{L^2(0,1)} \le C \|w(1,\cdot)\|_{L^2(0,T)}. \qquad \square$$

REFERENCES

[1] B. AINSEBA AND S. ANITA, *Local exact controllability of the age-dependent population dynamics with diffusion*, Abstr. Appl. Anal., 6 (2001), pp. 357–368.

[2] P. ALBANO AND P. CANNARSA, *Lectures on Carleman Estimates for Elliptic and Parabolic Operators and Applications*, in preparation.

[3] S. ANIŢA AND V. BARBU, *Null controllability of nonlinear convective heat equations*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 157–173.

[4] J.-M. Buchot, *Stabilisation et contrôle optimal des équations de Prandtl*, Ph.D. thesis, E.N.S.A.E. Toulouse, Toulouse, France, 2002.
[5] J.-M. Buchot and J.-P. Raymond, *A linearized model for boundary layer equations*, in Optimal Control of Complex Structures, Internat. Ser. Numer. Math. 139, Birkhäuser, Basel, 2002, pp. 31–42.
[6] J.-M. Buchot and P. Villedieu, *Construction de modèles pour le contrôle de la position de transition laminaire-turbulent sur une plaque plane*, Technical report, 1/3754.00 DTIMT/T, 1999.
[7] P. Cannarsa, P. Martinez, and J. Vancostenoble, *Nulle contrôlabilité régionale pour des équations de la chaleur dégénérées*, C. R. Mec. Acad. Sci. Paris, 330 (2002), pp. 397–401.
[8] P. Cannarsa, P. Martinez, and J. Vancostenoble, *Persistent Regional Null Controllability for a Class of Degenerate Parabolic Equations*, in preparation.
[9] S. Dolecki and D. L. Russell, *A general theory of observation and control*, SIAM J. Control Optim., 15 (1977), pp. 185–220.
[10] Y. V. Egorov, *Some problems in the theory of optimal control*, Z. Vychisl. Mat. Mat. Fiz., 5 (1963), pp. 887–904.
[11] C. Fabre, J. P. Puel, and E. Zuazua, *Approximate controllability for the semilinear heat equation*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 31–61.
[12] H. O. Fattorini and D. L. Russell, *Exact controllability theorems for linear parabolic equations in one space dimension*, Arch. Ration. Mech. Anal. 4 (1971), pp. 272–292.
[13] H. O. Fattorini and D. L. Russell, *Uniform bounds on biorthogonal functions for real exponentials with an application to the control theory of parabolic equations*, Quart. Appl. Math., 32 (1974), p. 45–69.
[14] E. Fernández-Cara and E. Zuazua, *The cost of approximate controllability for heat equations: The linear case*, Adv. Differential Equations, 5 (2000), pp. 465–514.
[15] E. Fernández-Cara and E. Zuazua, *Null and approximate controllability for weakly blowing-up semilinear heat equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 583–616.
[16] A. V. Fursikov and O. Y. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Ser. 34, Seoul National University, Seoul, Korea, 1996.
[17] O. Yu. Imanuvilov, *Boundary controllability of parabolic equations*, Mat. Sb., 186 (1995), pp. 109–132.
[18] O. Yu. Imanuvilov and M. Yamamoto, *On Carleman Inequalities for Parabolic Equations in Sobolev Spaces of Negative Order and Exact Controllability for Semilinear Parabolic Equations*, preprint 98-46, University of Tokyo, Grade School of Mathematics, Komobo, Tokyo, Japan, 1998.
[19] I. Lasiecka and R. Triggiani, *Carleman estimates and exact boundary controllability for a system of coupled, non conservative second order hyperbolic equations*, in Partial Differential Equations Methods in Control and Shape Analysis, Lecture Notes in Pure and Appl. Math. 188, Marcel Dekker, New York, 1994, pp. 215–243.
[20] G. Lebeau and L. Robbiano, *Contrôle exact de l'équation de la chaleur*, Comm. Partial Differential Equations, 20 (1995), pp. 335–356.
[21] J.-L. Lions, *Controle optimal des systemes gouvernes par des equations aux derivees partielles*, Dunod, Paris, 1969.
[22] P. Martinez, J.-P. Raymond, and J. Vancostenoble, *Nulle contrôlabilité régionale d'une équation de type Crocco linéarisée*, C. R. Math. Acad. Sci. Paris, 334 (2002), pp. 1–4.
[23] O. A. Oleinik and V. N. Samokhin, *Mathematical Models in Boundary Layer Theory*, Appl. Math. Math. Comput. 15, Chapman and Hall/CRC, Boca Raton, London, New York, 1999.
[24] D. L. Russell, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., 52 (1973), pp. 189–221.
[25] D. Tataru, *Carlemann estimates, unique continuation and controllability for anizotropic PDEs*, in Optimization Methods in Partial Differential Equations, Contemp. Math. 209, 1997, pp. 267–279.
[26] J. Zabczyk, *Mathematical Control Theory: An Introduction*, Birkhauser Boston, Boston, 1992.
[27] X. Zhang, *A remark on null exact controllability of the heat equation*, SIAM J. Control Optim., 40 (2001), pp. 39–53.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 729–745

OPTIMALITY CONDITIONS FOR DEGENERATE EXTREMUM PROBLEMS WITH EQUALITY CONSTRAINTS*

OLGA A. BREZHNEVA† AND ALEXEY A. TRET'YAKOV‡

Abstract. In this paper we consider an optimization problem with equality constraints given in operator form as $F(x) = 0$, where $F : X \to Y$ is an operator between Banach spaces. The paper addresses the case when the equality constraints are not regular in the sense that the Fréchet derivative $F'(x^*)$ is not onto. In the first part of the paper, we pursue an approach based on the construction of p-regularity. For p-regular constrained optimization problems, we formulate necessary conditions for optimality and derive sufficient conditions for optimality. In the second part of the paper, we consider a generalization of the concept of p-regularity and derive *generalized* necessary conditions for optimality for an optimization problem that is neither regular nor p-regular. For this problem, we show that the tangent cone to a level surface of F can consist of rays (rather than lines). This is in contrast to the regular and the p-regular cases, for which the tangent cone is always "two-sided." We state that if the gradient of the generalized p-regular problem is nonzero, it can belong to an open set, despite the fact that all constructions are usually closed. Both p-regular and *generalized* conditions for optimality reduce to classical conditions for regular cases, but they give new and nontrivial conditions for nonregular cases. The presented results can be considered as a part of the p-regularity theory.

Key words. Lyusternik theorem, constrained optimization, nonregular problems, optimality conditions

AMS subject classifications. 49K27, 46N10, 49N60, 90C30

DOI. 10.1137/S0363012901388488

1. Introduction.

In this paper we consider the nonlinear optimization problem

$$\begin{array}{ll} \underset{x\in X}{\text{minimize}} & f(x) \\ \text{subject to} & F(x) = 0, \end{array} \tag{1.1}$$

where $f : X \to \mathbb{R}$ is a sufficiently smooth function in a Banach space X, and $F : X \to Y$ is a sufficiently smooth mapping from a Banach space X to a Banach space Y.

We will be interested in the case when the equality constraints are not regular at a solution x^* of (1.1) in the sense that the Fréchet derivative $F'(x^*)$ is not onto. In this case, Euler–Lagrange necessary conditions for optimality,

$$\lambda_0 f'(x^*) + F'(x^*)^* y^* = 0,$$

are trivially satisfied with $\lambda_0 = 0$ and $y^* \in \mathrm{Ker} F'(x^*)$ and provide no additional information about solutions of (1.1).

The development of optimality conditions for nonregular problems has become an active research topic (see [1, 2, 3, 4, 5, 10, 12, 14, 15, 16, 20, 21, 23, 24] and references therein). In this paper we develop an approach based on the construction of p-regularity introduced in [25, 26, 27]. The main idea of this approach is to replace

*Received by the editors April 16, 2001; accepted for publication (in revised form) December 2, 2002; published electronically June 12, 2003.
http://www.siam.org/journals/sicon/42-2/38848.html

†Institute for Mathematics and Its Applications, University of Minnesota, 400 Lind Hall, 207 Church St. SE, Minneapolis, MN 55455 (olga@ima.umn.edu).

‡Computing Center of the Russian Academy of Sciences, Vavilova 40, Moscow, GSP-1, Russia and University of Podlasie in Siedlce, 3 Maja, Siedlce, Poland (tret@ap.siedlce.pl). The research of this author was supported in part by the Russian Foundation for Basic Research grant 99-01-00472.

the operator $F'(x^*)$, which is not onto, with a linear operator $\Psi_p(x^*)$, related to the pth order Taylor polynomial of F at x^*, which is onto. The operator $\Psi_p(x^*)$ contains up to the pth order derivative of F, so in our consideration F is p-times continuously Fréchet differentiable in a neighborhood of x^*. The order p is chosen as the minimum number for which the operator $\Psi_p(x^*)$ is regular.

In [11, 26, 27] a generalization of the Lyusternik theorem for p-regular mappings was first derived and proved. Using this theorem, nontrivial necessary and sufficient p-order optimality conditions for p-regular problems were obtained in [6, 18, 26]. These conditions contain the gradient of the objective function with a nonzero multiplier. Note that in [6, 18, 26], the sufficient conditions for optimality for p-regular problems were proposed in terms of the Euler–Lagrange function,

$$L(x, y) = f(x) + \langle y, F(x)\rangle.$$

In this paper, we obtain new sufficient conditions for optimality in terms of the generalized Lagrange function or p-factor–Lagrange function. This result is complementary to the necessary optimality conditions proposed in [6, 18, 26].

A new class of nonregular problems that satisfy a *generalized* condition of 2-regularity was introduced in [17, 18]. In contrast to regular and p-regular cases, the results obtained for the *generalized* 2-regular problems allow us to analyze nonregular problems by means of curves $x(t) = x^* + th + t^\alpha \tilde{h} + r(t)$ that lie in $M(x^*) = \{x \in X \mid F(x) = F(x^*)\}$, $t > 0$, and $h, \tilde{h} \in X$. Note that α is not necessarily integer. It can be fractional. For example, problem (1.1) with $F(x_1, x_2) = x_1^2 - x_2^3$ is neither 2-regular nor 3-regular, but the mapping F satisfies the *generalized* condition of 2-regularity. In [7, 8], new necessary optimality conditions were derived for optimization problems where the constraint mapping $F(x)$ satisfies the *generalized* condition of 2-regularity at the solution. Here we extend this approach to $p \geq 2$ and derive new necessary optimality conditions for problem (1.1) that satisfies the *generalized* condition of p-regularity.

To compare our approach with others, we can note that Ledzewicz and Schättler [21, 23] use the terminology p-regular, but in a different sense. A mapping is called p-regular at a point x^* along an element h_1 in our sense if it is p-regular in the direction of the sequence $H_{p-1} = (h_1, 0, \ldots, 0)$ in the sense of Ledzewicz and Schättler [21, 23]. But both our definition and the definition from [21, 23] reduce to the same definition of 2-regularity for $p = 2$. Furthermore, some problems that satisfy the *generalized* condition of 2-regularity could not be treated using the approach presented in [21, 23]. For example, problem (1.1) with $F(x_1, x_2) = x_1^2 - x_2^{5/2}$ satisfies the *generalized* condition of 2-regularity, but this problem is not p-regular in any direction in the sense of the definition given in [21, 23].

The organization of the paper is as follows. In section 2, we recall the definition of p-regularity and formulate a generalization of the Lyusternik theorem for p-regular problems. In section 3, we formulate necessary conditions for optimality [18] and derive new sufficient conditions for optimality for p-regular optimization problems (1.1). In section 4, we consider a generalization of the concept of p-regularity [17, 18] and obtain new necessary optimality conditions for problem (1.1) that satisfy the *generalized* condition of p-regularity.

Our results are applicable to a variety of extremum and optimal control problems with nonregular constraints [2, 21, 22]. Moreover, the approach presented in the paper extends to inequality-constraint optimization problems. In the future we are going to

prove conditions for optimality for these problems using techniques proposed in this paper.

Notation. We denote by $\mathcal{L}(X, Y)$, the space of all continuous linear operators from X to Y. Further, $\operatorname{Ker}\Lambda = \{x \in X \mid \Lambda x = 0\}$ denotes the null-space (kernel) of a given linear operator $\Lambda : X \to Y$, and $\operatorname{Im}\Lambda = \{y \in Y \mid y = \Lambda x \text{ for some } x \in X\}$ is its image space. Also, $\Lambda^* : Y^* \to X^*$ denotes the adjoint of Λ , where X^* and Y^* denote the dual spaces of X and Y, respectively. $M^\perp = \{h^* \in X^* \mid \langle h^*, x\rangle = 0 \quad \forall x \in M\}$ denotes the annihilator of the set M. For a subset S of a space X, we denote by $\operatorname{Sp}(S)$, the linear space spanned (generated) by S, and by $\operatorname{cl} S$, the closure of S.

If F is a p-times Fréchet differentiable mapping at x^*, then $F^{(p)}(x^*)$ denotes the pth order derivative that is pth order mapping from $X \times X \times \cdots \times X$ to Y. By definition of p-order mapping, we have

$$F^{(p)}(x^*)[h]^p = F^{(p)}(x^*)[h, \dots, h].$$

In our consideration, $\operatorname{Ker}^p F^{(p)}(x^*) = \{h \in X \mid F^{(p)}(x^*)[h]^p = 0\}$ denotes the *p-kernel* of the p-order mapping $F^{(p)}(x^*)$.

2. *P*-regular mappings and a generalization of the Lyusternik theorem. As is well known, the Lyusternik theorem provides a useful tool for constructive description of the tangent cone to the set $M(x^*) = \{x \in X : F(x) = F(x^*)\}$ at a given point x^* for the regular mapping F. Let us recall some definitions and the Lyusternik theorem [13].

We say a mapping F is *regular* at some point x^* if

$$\operatorname{Im} F'(x^*) = Y. \tag{2.1}$$

The mapping F is called *nonregular (irregular, degenerate, abnormal)* if the regularity condition (2.1) is not satisfied.

Definition 2.1. *Let M be a subset of a Banach space X. A vector $h \in X$ is said to be tangent to the set M at a point x^* if there exists an $\varepsilon > 0$ and a mapping $t \to r(t)$ of the interval $[0, \varepsilon]$ into X such that*

$$x^* + th + r(t) \in M \qquad \forall t \in [0, \varepsilon],$$

$$\lim_{t \to 0} \frac{\|r(t)\|}{t} = 0.$$

A set of vectors tangent to the set M at a point x^* is called the *tangent cone* to the set M at the point x^* and is denoted by $T_1 M$.

Theorem 2.2 (Lyusternik theorem). *Let X and Y be Banach spaces, U a neighborhood of a point $x^* \in X$, and let $F : U \to Y$ be a continuously Fréchet differentiable mapping of U into Y. Assume that F is regular at x^*, and its derivative $F' : U \to \mathcal{L}(X, Y)$ is continuous at x^*.*

Then the tangent cone to the set $M(x^*) = \{x \in U \mid F(x) = F(x^*)\}$ at the point x^* coincides with the kernel of the operator $F'(x^*)$,

$$T_1 M(x^*) = \operatorname{Ker} F'(x^*). \tag{2.2}$$

We consider the case when the regularity condition (2.1) does not hold, but the mapping F is p-regular. For this case a generalization of the Lyusternik theorem was derived in [26, 27]. First let us review the definition of p-regularity and construction

of a p-factor-operator. Throughout the paper we assume that $F : U \to Y$ is a p-times continuously Fréchet differentiable mapping in U, where U is a neighborhood of a point $x^* \in X$.

We construct here a p-factor-operator under an assumption that the space Y is decomposed into a direct sum

$$Y = Y_1 \oplus \cdots \oplus Y_p, \tag{2.3}$$

where $Y_1 = \operatorname{cl} \operatorname{Im} F'(x^*)$, $Y_i = \operatorname{cl} \operatorname{Sp} (\operatorname{Im} P_{Z_i} F^{(i)}(x^*)[\cdot]^i)$, $i = 2, \ldots, p-1$, $Y_p = Z_p$, Z_i is a closed complementary subspace for $(Y_1 \oplus \cdots \oplus Y_{i-1})$ with respect to Y, $i = 2, \ldots, p$, and $P_{Z_i} : Y \to Z_i$ is the projection operator onto Z_i along $(Y_1 \oplus \cdots \oplus Y_{i-1})$ with respect to Y, $i = 2, \ldots, p$.

Define the mappings [18]

$$f_i(x) : U \to Y_i, \quad f_i(x) = P_{Y_i} F(x), \quad i = 1, \ldots, p,$$

where $P_{Y_i} : Y \to Y_i$ is the projection operator onto Y_i along $(Y_1 \oplus \cdots \oplus Y_{i-1} \oplus Y_{i+1} \oplus \cdots \oplus Y_p)$ with respect to Y, $i = 1, \ldots, p$.

The consideration of another case, when (2.3) is not satisfied, is given in detail, for example, in [18].

DEFINITION 2.3. *The linear operator* $\Psi_p(h) \in \mathcal{L}(X, Y_1 \oplus \cdots \oplus Y_p)$, $h \in X$,

$$\Psi_p(h) = f_1'(x^*) + \frac{1}{2!} f_2''(x^*)[h] + \cdots + \frac{1}{p!} f_p^{(p)}(x^*)[h]^{p-1},$$

is called a p-factor-operator.

DEFINITION 2.4. *We say the mapping F is p-regular at x^* along an element h if* $\operatorname{Im} \Psi_p(h) = Y$.

DEFINITION 2.5. *We say the mapping F is p-regular at x^* if it is p-regular along any h from the set*

$$H_p(x^*) = \left\{ \bigcap_{i=1}^{p} \operatorname{Ker} f_i^{(i)}(x^*) \right\} \setminus \{0\}.$$

DEFINITION 2.6. *The mapping F is called strongly p-regular at the point x^* if there exists $\gamma > 0$ such that*

$$\sup_{h \in H_\gamma} \|\{\Psi_p(h)\}^{-1}\| < \infty,$$

where

$$H_\gamma = \{h \in X \mid \|f_i^{(i)}(x^*)[h]^i\|_{Y_i} \le \gamma \quad \forall i = 1, \ldots, p, \quad \|h\|_X = 1\}.$$

Remark. Not only Ψ_p, but also every Y_i, $i = 2, \ldots, p$, in $Y_1 \oplus \cdots \oplus Y_p$ depends on the element h. To simplify our notation, we use Y_i instead of $Y_i(h)$, $i = 2, \ldots, p$.

We are ready to formulate a generalization of the Lyusternik theorem [27] that is applied to prove necessary optimality conditions for p-regular mappings. This theorem differs from the p-order Lyusternik theorem formulated and proven by Ledzewicz and Schättler in [21].

THEOREM 2.7 (generalization of the Lyusternik theorem). *Let X and Y be Banach spaces, U a neighborhood of a point $x^* \in X$, and let $F : U \to Y$ be a p-times*

continuously Fréchet differentiable mapping in U. Assume that F is p-regular at x^*. Then

$$T_1 M(x^*) = H_p(x^*).$$

At about the same time, in the case $F^{(r)}(x^*) \equiv 0$, $r = 1, \dots, p-1$, Theorem 2.7 was proved in [9] for finite dimensional spaces and in [25] for Banach spaces. In a general case, when there exists $F^{(r)}(x^*) \neq 0$, $r < p$, the theorem was presented in [27].

The following theorem [27] will be used in our analysis.

THEOREM 2.8. *Let X and Y be Banach spaces, U a neighborhood of a point $x^* \in X$, and let $F : X \to Y$ be a p-times continuously Fréchet differentiable mapping in U. Assume that F is strongly p-regular at x^*. Then there exists a neighborhood $U' \subseteq U$ of the point x^*, a mapping $\eta \to x(\eta) : U' \to X$, and constants $\delta_1 > 0$ and $\delta_2 > 0$ such that*

$$F(\eta + x(\eta)) = F(x^*) \qquad \forall \eta \in U',$$

$$\|x(\eta)\|_X \leq \delta_1 \sum_{i=1}^{p} \frac{\|f_i(\eta) - f_i(x^*)\|_{Y_i}}{\|\eta - x^*\|^{i-1}} \qquad \forall \eta \in U', \tag{2.4}$$

$$\|x(\eta)\|_X \leq \delta_2 \sum_{i=1}^{p} \|f_i(\eta) - f_i(x^*)\|_{Y_i}^{1/i} \qquad \forall \eta \in U'.$$

3. Optimality conditions for p-regular problems. In this section we formulate necessary optimality conditions [18] and pursue new sufficient optimality conditions for p-regular constrained optimization problems.

We define the *p-factor–Lagrange function*,

$$\mathcal{L}_p(x, h, \lambda_0(h), y(h)) = \lambda_0(h) f(x) + \sum_{i=1}^{p} \left\langle y_i(h),\, f_i^{(i-1)}(x)[h]^{i-1} \right\rangle, \tag{3.1}$$

where $x \in X$, $h \in X$, $\lambda_0(h) \in \mathbb{R}$, $y_i(h) \in Y_i^*$, $i = 1, \dots, p$. Note that the function (3.1) is a generalization of the Lagrange function and it reduces to the Lagrange function for the regular case.

THEOREM 3.1 (necessary conditions for optimality). *Let X and Y be Banach spaces, U a neighborhood of the point $x^* \in X$, $f : U \to \mathbb{R}$ a twice continuously Fréchet differentiable function in U, and let $F : U \to Y$ be a p-times continuously Fréchet differentiable mapping in U. Assume that for an element $h \in H_p(x^*)$ the set $\operatorname{Im} \Psi_p(h)$ is closed in $Y_1 \oplus \cdots \oplus Y_p$.*

If x^* is a local solution to problem (1.1), then there exist $\lambda_0(h) \in \mathbb{R}$ and multipliers $y_i(h) \in Y_i^*$, $i = 1, \dots, p$, such that they do not all vanish, and

$$\mathcal{L}'_{p\,x}(x^*, h, \lambda_0(h), y(h)) = \lambda_0(h)\, f'(x^*) + \sum_{i=1}^{p} (f_i^{(i)}(x^*)[h]^{i-1})^* y_i(h) = 0.$$

If, moreover, $\operatorname{Im} \Psi_p(h) = Y_1 \oplus \cdots \oplus Y_p$, then $\lambda_0(h) \neq 0$.

If $f'(x^*) = 0$, then the necessary conditions for optimality given in Theorem 3.1 are trivially satisfied with $y^* = 0$ and any $\lambda_0 \neq 0$. In the following theorem, we

formulate other informative necessary optimality conditions for p-regular problems, which were derived for $q = 2$ in [6] and for any q in [18].

THEOREM 3.2. *Let X and Y be Banach spaces, U a neighborhood of $x^* \in X$, $f : U \to \mathbb{R}$ a q-times continuously Fréchet differentiable function in U, and let $F : U \to Y$ be a p-times continuously Fréchet differentiable mapping in U. Assume that for $h \in H_p(x^*)$, $\operatorname{Im}\Psi_p(h) = Y_1 \oplus \cdots \oplus Y_p$. Assume also that $f^{(i)}(x^*) = 0$ for all $i = 1, \dots, q-1$.*

If x^* is a local solution to problem (1.1), *then either $f^q(x^*)[h]^q > 0$ or $f^q(x^*)[h]^q = 0$ and, in the last case, there exist multipliers $y_i(h) \in Y_i^*$, $i = 1, \dots, p$, such that*

$$f^q(x^*)[h]^{q-1} + \sum_{i=1}^{p} (f_i^{(i)}(x^*)[h]^{i-1})^* y_i(h) = 0.$$

In the following theorem we present new sufficient optimality conditions for p-regular problems.

THEOREM 3.3 (sufficient conditions for optimality). *Let X and Y be Banach spaces, U be a neighborhood of a point $x^* \in X$, $f : U \to \mathbb{R}$ be a twice continuously Fréchet differentiable function in U, and $F : U \to Y$ be a $(p+1)$-times continuously Fréchet differentiable mapping in U. Assume that the set $\operatorname{Im}\Psi_p(h)$ is closed in $Y_1 \oplus \cdots \oplus Y_p$ for any element $h \in H_p(x^*)$ and $\operatorname{Im}\Psi_p(h) = Y_1 \oplus \cdots \oplus Y_p$. Assume also that F is strongly p-regular at x^*.*

If there exist $\alpha > 0$ and multipliers $y_i(h) \in Y_i^*$, $i = 1, \dots, p$, such that

$$\mathcal{L}'_{p\,x}(x^*, h, 1, y(h)) = 0 \tag{3.2}$$

and

$$\mathcal{L}''_{p\,xx}(x^*, h, 1, y(h))[h_p]^2 \geq \alpha \|h_p\|^2 \qquad \forall h_p \in H_p(x^*), \tag{3.3}$$

then x^* is a strict local minimizer to problem (1.1).

Proof. We consider an element $x \in U$ such that $F(x) = 0$. We must prove that x can be represented as

$$x = x^* + th + \xi, \tag{3.4}$$

where $\|\xi\| \leq C\,t^2$, $|t| = \|x - x^*\| + o(\|x - x^*\|)$, $h \in H_p(x^*)$, $\|h\| = 1$, and $C > 0$.

Since $F(x) = 0$ and $f_i^{(k)}(x^*) = 0$ for all $k < i$, we have

$$0 = f_i(x) = \frac{1}{i!} f_i^{(i)}(x^*)[x - x^*]^i + \omega_i(x) \qquad \forall i = 1, \dots, p,$$

where $\|\omega_i(x)\| \leq \alpha_i \|x - x^*\|^{i+1}$, $\alpha_i \geq 0$. Hence,

$$\|f_i^{(i)}(x^*)[x - x^*]^i\| \leq i!\alpha_i \|x - x^*\|^{i+1} \leq \beta_i \|x - x^*\|^{i+1}, \qquad \beta_i > 0. \tag{3.5}$$

Consider the equation

$$\Psi(z) = \sum_{i=1}^{p} \frac{1}{i!} f_i^{(i)}(x^*)[z]^i = 0.$$

Note that any solution z to the last equation belongs to the set H_p. Apply Theorem 2.8 to $\Psi(\eta + x(\eta)) = \Psi(0)$ with

$$\eta = x - x^*, \qquad x(\eta) = x - x^* - th,$$

where $h \in H_p$, $\|h\| = 1$, $t \geq 0$. By virtue of (2.4) and (3.5), we have

$$\|x - x^* - th\| \leq \delta_1 \sum_{i=1}^{p} \frac{\|f_i^{(i)}(x^*)[x - x^*]^i - 0\|}{\|x - x^*\|^{i-1}}$$
$$\leq \delta_1 \sum_{i=1}^{p} \frac{\beta_i \|x - x^*\|^{i+1}}{\|x - x^*\|^{i-1}} \leq C\|x - x^*\|^2, \qquad C > 0.$$

Hence,

$$\|x - x^* - th\| \leq C\|x - x^*\|^2, \tag{3.6}$$

where h is an element from the set H_p, $\|h\| = 1$, and $t > 0$. By virtue of the last relation, the representation (3.4) is proved.

Let $y_i \in Y_i^*$ be functionals such that (3.2) and (3.3) are satisfied. Using $t \neq 0$ and $F(x) = 0$, represent $f(x)$ as

$$f(x) = f(x) + \langle y_1, F(x)\rangle + \frac{\langle y_2, P_2 F(x)\rangle}{t} + \cdots + \frac{\langle y_p, P_p F(x)\rangle}{t^{p-1}},$$

where P_i is the projection operator onto Y_i, defined at the point x^*. Then

$$f(x) - f(x^*) = f(x) - f(x^*) + \langle y_1, F(x^* + th + \xi)\rangle$$
$$+ \frac{\langle y_2, P_2 F(x^* + th + \xi)\rangle}{t} + \cdots + \frac{\langle y_p, P_p F(x^* + th + \xi)\rangle}{t^{p-1}}.$$

Consider the Taylor expansion of f and F. We have

$$f(x) - f(x^*) = f'(x^*)[th + \xi] + \frac{1}{2} f''(x^*)[th + \xi]^2 + o(t^2)$$
$$+ \left\langle y_1, F(x^*) + F'(x^*)[th + \xi] + \frac{F''(x^*)}{2}[th + \xi]^2 + o(t^2) \right\rangle$$
$$+ \left\langle y_2, P_2 \left(\frac{F'(x^*)[th + \xi]}{t} + \frac{F''(x^*)}{2t}[th + \xi]^2 \right.\right.$$
$$\left.\left. + \frac{F'''(x^*)}{6t}[th + \xi]^3 + o(t^2) \right) \right\rangle + \cdots$$
$$+ \left\langle y_p, P_p \left(\frac{F'(x^*)[th + \xi]}{t^{p-1}} + \frac{F''(x^*)}{2\,t^{p-1}}[th + \xi]^2 + \cdots \right.\right.$$
$$\left.\left. + \frac{F^{(p+1)}(x^*)}{p!\,t^{p-1}}[th + \xi]^{p+1} + o(t^2) \right) \right\rangle.$$

Transform the last relation into a new form, using the property $P_i F^{(j)}(x^*) = 0$ for

$i > j$. We then have an expression of the form

$$\begin{aligned} f(x) - f(x^*) = {} & f'(x^*)[th+\xi] + \frac{1}{2}f''(x^*)[th+\xi]^2 + o(t^2) \\ & + \left\langle y_1, F'(x^*)[th+\xi] + \frac{F''(x^*)}{2}[th+\xi]^2 + o(t^2) \right\rangle \\ & + \left\langle y_2, P_2\frac{F''(x^*)}{2t}[th+\xi]^2 + P_2\frac{F'''(x^*)}{6t}[th+\xi]^3 + o(t^2) \right\rangle \\ & + \cdots \\ & + \left\langle y_p, P_p\frac{F^{(p)}(x^*)}{p!t^{p-1}}[th+\xi]^p + P_p\frac{F^{(p+1)}(x^*)}{(p+1)!t^{p-1}}[th+\xi]^{p+1} \right. \\ & \qquad \left. + o(t^2) \right\rangle. \end{aligned}$$

We can represent the right part of the last equality as $(A_1 + A_2)$ where

$$\begin{aligned} A_1 = {} & f'(x^*)[th+\xi] + \langle y_1, F'(x^*)[th+\xi]\rangle + \left\langle y_2, P_2\frac{F''(x^*)}{2t}[th+\xi]^2 \right\rangle \\ & + \cdots + \left\langle y_p, P_p\frac{F^{(p)}(x^*)}{p!t^{p-1}}[th+\xi]^p \right\rangle + o(t^2) \end{aligned}$$

and

$$\begin{aligned} A_2 = {} & \frac{1}{2}f''(x^*)[th+\xi]^2 + \left\langle y_1, \frac{F''(x^*)}{2}[th+\xi]^2 \right\rangle \\ & + \left\langle y_2, P_2\frac{F'''(x^*)}{6t}[th+\xi]^3 \right\rangle + \cdots + \left\langle y_p, P_p\frac{F^{(p+1)}(x^*)}{(p+1)!t^{p-1}}[th+\xi]^{p+1} \right\rangle. \end{aligned}$$

By definitions of $\mathcal{L}_p(x^*, h, 1, y(h))$ and ξ, we obtain

$$\begin{aligned} A_1 = {} & \left\langle f'(x^*) + (F'(x^*))^* y_1 + \left(P_2\frac{F''(x^*)}{2}[h]\right)^* y_2 + \cdots \right. \\ & \left. + \left(P_p\frac{F^{(p)}(x^*)}{p!}[h]^{(p-1)}\right)^* y_p, th+\xi \right\rangle + o(t^2) \\ = {} & \langle \mathcal{L}'_{p\,x}(x^*, h, 1, y(h)), th+\xi\rangle + o(t^2) = o(t^2), \end{aligned}$$

since $\mathcal{L}'_{p\,x}(x^*, h, 1, y(h)) = 0$. But also

$$A_2 = \mathcal{L}''_{p\,xx}(x^*, h, 1, y(h))[th]^2 + o(t^2) \geq \alpha\|th\|^2 + o(t^2).$$

The last inequality follows from the condition that $h \in H_p(x^*)$ and (3.3). Finally, we thus have

$$f(x) - f(x^*) = A_1 + A_2 \geq \alpha\|th\|^2 + o(t^2) > 0 \qquad \forall x \in U(x^*),$$

and, therefore, x^* is a strict local minimizer to problem (1.1). □

4. Generalization of the concept of *p*-regularity. In this section we consider a generalization of the concept of p-regularity and derive new necessary optimality conditions for problem (1.1) that is neither regular nor p-regular. We are interested

only in the case $f'(x^*) \neq 0$, since otherwise the classical Euler–Lagrange-type necessary conditions are trivially satisfied with $\lambda_0 \neq 0$ and $y = 0$.

We construct necessary optimality conditions for problem (1.1) under an assumption that F does not satisfy the definition of p-regularity and $F^{(r)}(x^*) \equiv 0$, $r = 1, \dots, p-1$, $p \geq 2$. We believe that this case of absolute degeneration is the most important in analysis of properties of degenerate problems. Moreover, it was proved in [18] that the case of general degeneration, when there exists $F^{(r)}(x^*) \neq 0$, $r < p$, can be reduced to the absolute degeneration case.

Assume that, for some element $h \in X$, the space Y can be decomposed into a direct sum of subspaces

$$Y = Y_1(h) \oplus Y_2(h), \tag{4.1}$$

where $Y_1(h) = \mathrm{Im} F^{(p)}(x^*)[h]^{p-1}$, $Y_2(h)$ is a linear subspace that complements $Y_1(h)$ with respect to Y, and both $Y_1(h)$ and $Y_2(h)$ are closed in Y.

The projection operators onto $Y_1(h)$ along $Y_2(h)$ and onto $Y_2(h)$ along $Y_1(h)$ are denoted by $\mathrm{P}_1(h)$ and $\mathrm{P}_2(h)$, respectively.

We introduce

$$\Lambda(F, h, x^*) = \{h^* \in X^* \,|\, \langle h, h^* \rangle \leq 0\}, \qquad h \in \mathrm{Ker}^{\,\mathrm{p}} F^{(p)}(x^*),$$

$$\overset{0}{\Lambda}(F, h, x^*) = \{h^* \in X^* | \langle h, h^* \rangle < 0\} \cup \{0\}, \qquad h \in \mathrm{Ker}^{\,\mathrm{p}} F^{(p)}(x^*).$$

Define

$$Z = \mathrm{Im}(F^{(p)}(x^*)[h]^{p-1} + P_2(h) F^{(p)}(x^*)\,[h]^{p-2}\,[\widetilde{h}])^*, \qquad \widetilde{h} \in X,$$

and assume that, for $\widetilde{h} \in X$, the space Z is decomposed into a direct sum

$$Z = Z_1(\widetilde{h}) \oplus Z_2(\widetilde{h}), \tag{4.2}$$

where

$$Z_1(\widetilde{h}) = \{z \in Z \,|\, \langle z, \widetilde{h} \rangle = 0\}$$

and $Z_2(\widetilde{h})$ is its complementary subspace with respect to Z. Furthemore, assume that both $Z_1(\widetilde{h})$ and $Z_2(\widetilde{h})$ are closed in Z. Denote by $P_1(\widetilde{h})$, the projection operator onto $Z_1(\widetilde{h})$ along $Z_2(\widetilde{h})$.

Note that under our assumptions all necessary decompositions and constructions are possible in Hilbert spaces. Moreover, they take place in Banach spaces using factorization of spaces.

Under these assumptions, the following theorem holds for problem (1.1).

THEOREM 4.1. *Let X and Y be Banach spaces, x^* a solution to problem* (1.1)*, and U a neighborhood of x^* in X. Suppose that the mapping $F : U \to Y$ is p-times continuously Fréchet differentiable and the function $f : U \to \mathbb{R}$ is twice continuously Fréchet differentiable. Suppose that condition* (4.1) *is satisfied for $h \in \mathrm{Ker}^{\,\mathrm{p}} F^{(p)}(x^*)$ and there exists an element $\widetilde{h} \in X$ $(c \leq \|h\|, \|\widetilde{h}\| \leq C)$, $0 < c \leq C < \infty$ and $D > 0$ such that* (4.2) *is fulfilled and*

$$\|F^{(p)}(x^*)\,[h]^{p-1}\,[\widetilde{h}]\,\| = 0, \tag{4.3}$$

$$\|P_2(h)F(x^* + th + t^\alpha \widetilde{h})\| \le t^{\,(p-2)+2\alpha+\varepsilon}, \quad 1 < \alpha \le \frac{3}{2}, \quad \varepsilon \in (0,1), \tag{4.4}$$

$$\|\{F^{(p)}(x^*)\,[h]^{p-1} + P_2(h)F^{(p)}(x^*)\,[h]^{p-2}\,[\widetilde{h}]\}^{-1}\| \le D, \tag{4.5}$$

where $t \in (0, \delta)$ and $\delta > 0$ is sufficiently small.

Then

(1) *we have $h \in T_1 M(x^*)$; moreover,*

$$\alpha^{\,+}(t) = x^* + th + t^{\,\alpha}\widetilde{h} + r^+(t) \in M(x^*),$$

$$\alpha^{\,-}(t) = x^* + th - t^{\,\alpha}\widetilde{h} + r^-(t) \in M(x^*),$$

so that $\| r^{\pm}(t) \| = o(t^{\,\alpha})$;

(2) *the following inclusion holds:*

$$-f'(x^*) \in \left\{ \overset{0}{\Lambda}(F, h, x^*) \cup \operatorname{Im}((F^{(p)}(x^*)\,[h]^{p-1} + P_2(h)F^{(p)}(x^*)\,[h]^{p-2}\,[\widetilde{h}])\,P_1{}^*(\widetilde{h}))^* \right\}. \tag{4.6}$$

Proof. (1) Consider the case $(+\widetilde{h})$. Define the mapping

$$\Phi(x) = x - (F^{(p)}(x^*)\,[th]^{p-1} + P_2(h)F^{(p)}(x^*)\,[th]^{p-2}\,[t^\alpha \widetilde{h}])^{-1} F(x^* + th + t^\alpha \widetilde{h} + x).$$

Taking into account (4.3) and applying the contraction multimapping principle [13] to the mapping Φ under $x_0 = 0$, there exists a function $r^+(t)$ such that

$$F(x^* + th + t^\alpha \widetilde{h} + r^+(t)) = 0$$

and

$$\|r^+(t)\| \le D\left(\frac{\|P_1(h)F(x^* + th + t^\alpha \widetilde{h})\|}{t^{p-1}} + \frac{\|P_2(h)F(x^* + th + t^\alpha \widetilde{h})\|}{t^{p-2}\,t^\alpha} \right) \le D(t^{2\alpha-1} + t^{\alpha+\varepsilon}) = o(t^\alpha).$$

The case $(-\widetilde{h})$ is considered in a similar way.

(2) First let us note that $\langle -f'(x^*), h\rangle \le 0$, since otherwise x^* would not be a local minimizer.

If $\langle -f'(x^*), h\rangle < 0$, then, by the definition of $\overset{0}{\Lambda}(F, h, x^*)$,

$$-f'(x^*) \in \overset{0}{\Lambda}(F, h, x^*), \tag{4.7}$$

which proves inclusion (4.6). It remains to consider the case $\langle f'(x^*), h\rangle = 0$, $f'(x^*) \ne 0$. Let us show that, in this case,

$$f'(x^*) \in \operatorname{Im}((F^{(p)}(x^*)[h]^{(p-1)} + P_2(h)F^{(p)}(x^*)\,[h]^{(p-2)}\,[\widetilde{h}])\,P_1{}^*(\widetilde{h}))^*. \tag{4.8}$$

By assertion (1) of the theorem, the arcs $\alpha^+(t)$ and $\alpha^-(t)$ belong to $M(x^*)$ for the vector $\widetilde{h} \in \operatorname{Ker} F^{(p)}(x^*)\,[h]^{p-1}$. Using the Taylor formula, we obtain $\langle f'(x^*), \widetilde{h}\rangle = 0$,

since otherwise x^* would not be a local minimizer for problem (1.1), because the function $f(x)$ would decrease along either $\alpha^+(t)$ or $\alpha^-(t)$.

It can be proved in a similar way that the equality $\langle f'(x^*), \widehat{h}\rangle = 0$ holds for any $\widehat{h} \in \operatorname{Ker}(F^{(p)}(x^*)[h]^{p-1} + P_2(h)\,F^{(p)}(x^*)[h]^{p-2}[\widetilde{h}])$. This follows from the existence of $\delta \in (0,1)$ and $\alpha < 2-\delta$ such that

$$\beta^{\pm}(t) = x^* + th \pm t^{\alpha}\widetilde{h} \pm t^{\alpha+\delta}\widehat{h} + \omega^{\pm}(t) \in M(x^*),$$

where $\|\omega^{\pm}(t)\| = o(t^{\alpha+\delta})$ (see, for instance, [8]).

We define

$$L = \operatorname{Sp}\{\operatorname{Ker}(F^{(p)}(x^*)[h]^{p-1} + P_2(h)\,F^{(p)}(x^*)[h]^{p-2}[\widetilde{h}]), \widetilde{h}\}.$$

Since $\langle f'(x^*), \widehat{h}\rangle = 0$ and $\langle f'(x^*), \widetilde{h}\rangle = 0$, then, by definitions of the vectors $\widehat{h}$ and $\widetilde{h}$, we have

$$\langle f'(x^*), \xi\rangle = 0 \qquad \forall \xi \in L. \tag{4.9}$$

Let us show that $Z_1^{\perp}(\widetilde{h}) = L$. By virtue of the annihilator lemma [13],

$$\begin{aligned} Z &= \operatorname{Im}(F^{(p)}(x^*)[h]^{p-1} + P_2(h)\,F^{(p)}(x^*)[h]^{p-2}[\widetilde{h}])^* \\ &= \operatorname{Ker}(F^{(p)}(x^*)[h]^{p-1} + P_2(h)\,F^{(p)}(x^*)[h]^{p-2}[\widetilde{h}])^{\perp}. \end{aligned}$$

Hence, $Z_1(\widetilde{h})$ can be represented as

$$\begin{aligned} Z_1(\widetilde{h}) = \{z \in Y^* | \langle z, \widetilde{h}\rangle = 0, \quad \langle z, y\rangle = 0 \\ \forall y \in \operatorname{Ker}(F^{(p)}(x^*)[h]^{p-1} + P_2(h)\,F^{(p)}(x^*)[h]^{p-2}[\widetilde{h}])\}. \end{aligned}$$

By virtue of a corollary from the bipolar theorem [19], the last relation implies that

$$Z_1^{\perp}(\widetilde{h}) = \operatorname{Sp}\{y, \widetilde{h}\}, \tag{4.10}$$

where $y \in \operatorname{Ker}(F^{(p)}(x^*)[h]^{p-1} + P_2(h)\,F^{(p)}(x^*)[h]^{p-2}[\widetilde{h}])$.

By the definition of $Z_1(\widetilde{h})$, we obtain

$$Z_1(\widetilde{h}) = \operatorname{Im}(P_1(\widetilde{h})(F^{(p)}(x^*)[h]^{p-1} + P_2(h)\,F^{(p)}(x^*)[h]^{p-2}[\widetilde{h}])^*);$$

hence, by the annihilator lemma [13],

$$Z_1^{\perp}(\widetilde{h}) = \operatorname{Ker}((F^{(p)}(x^*)[h]^{p-1} + P_2(h)\,F^{(p)}(x^*)[h]^{p-2}[\widetilde{h}])P_1^*(\widetilde{h})).$$

By virtue of (4.9) and (4.10), this yields

$$\begin{aligned} &\langle f'(x^*), \xi\rangle = 0 \\ &\quad \forall \xi \in \operatorname{Ker}((F^{(p)}(x^*)[h]^{p-1} + P_2(h)\,F^{(p)}(x^*)[h]^{p-2}[\widetilde{h}])\,P_1^*(\widetilde{h})), \end{aligned}$$

so, again by the annihilator lemma [13],

$$f'(x^*) \in \operatorname{Im}(P_1(\widetilde{h})\,(F^{(p)}(x^*)[h]^{p-1} + P_2(h)\,F^{(p)}(x^*)[h]^{p-2}[\widetilde{h}])^*),$$

which proves the inclusion (4.8) in the case under consideration. Together with (4.7), this proves the theorem. □

Remark. Note that according to (4.6), the gradient $-f'(x^*)$ belongs to an open set, despite the fact that all constructions are closed. This is related to condition (4.4), which yields a deterministic structure of the arc $\alpha(t)$. It can be proved that the open set becomes closed when condition (4.4) is not satisfied.

Remark. Relation (4.5) is a new condition which is weaker than the p-regularity condition. We say the operator in the left part of (4.5) is *a generalization of a p-factor-operator* for $p \geq 2$ and condition (4.5) is a *a generalization of the concept of p-regularity* for $p \geq 2$.

COROLLARY 4.2. *If there exists an element $h \neq 0$ such that the conditions of Theorem* 4.1 *are satisfied for h and $-h$, then*

(4.11)
$$\begin{aligned} f'(x^*) \in \ & \mathrm{Im}((F^{(p)}(x^*)[-h]^{p-1} + P_2(-h)F^{(p)}(x^*)[-h]^{p-2}[\widetilde{h}^-])P_1{}^*(\widetilde{h}^-))^* \\ & \cap \mathrm{Im}((F^{(p)}(x^*)[h]^{p-1} + P_2(h)F^{(p)}(x^*)[h]^{p-2}[\widetilde{h}^+])P_1{}^*(\widetilde{h}^+))^*. \end{aligned}$$

The approach considered in this section can be extended to the case of general degeneration. Then, the corresponding *generalized* p-factor-operators $\Psi(h)$ can be constructed by analogy with (4.5) and are generated by elements $h = (h_1, h_2, \dots)$ in such a way that $\mathrm{Im}\Psi(h) = Y$.

5. Examples. In this section we give several examples which illustrate how Theorem 3.3 and Theorem 4.1 can be applied to analyze wide classes of problems that could not be investigated before. In these examples $p = 2$.

The first example illustrates the application of Theorem 3.3.

Example 5.1. Consider the problem

$$\begin{aligned} & x_2^2 + x_3 \to \min, \\ & F(x) = \begin{pmatrix} 0,5(x_1^2 - x_2^2 + x_3^2) + x_2^3 \\ 0,5(x_1^2 - x_2^2 + x_3^2) + x_1x_3 + x_2^3 \end{pmatrix} = 0. \end{aligned} \tag{5.1}$$

As is easy to verify, the point $x^* = 0$ is a local minimum to problem (5.1). Let us prove that the sufficient conditions for optimality given in Theorem 3.3 are satisfied at $x^* = 0$.

For $x^* = 0$, we have $F'(0) = 0$,

$$\mathrm{Ker}^2 F''(0) = \mathrm{Sp}\left\{\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}\right\} \bigcup \mathrm{Sp}\left\{\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}\right\}.$$

Consider the element $h = (1, 1, 0)^T$. Since $\mathrm{Im}F''(0)h = \mathbb{R}^2$, the mapping $F(x)$ is 2-regular at $x^* = 0$ along the element h. Consider the 2-factor–Lagrange function with $\alpha_0 = 1$. After some transformations we obtain

$$\begin{aligned} L_2(x, h, y(h)) = x_2^2 + x_3 + \alpha(x_1 - x_2 + 3x_2^2) \\ + \beta(x_1 - x_2 + x_3 + 3x_2^2), \end{aligned} \tag{5.2}$$

where $y(h) = (y_1(h), y_2(h))$, $y_2(h) = (\alpha, \beta)$. Let us calculate the coefficients α and β. Using the equality $L'_{2\,x}(x^*, h, y(h)) = 0$, we obtain $\alpha = 1, \beta = -1$. Putting the coefficients into (5.2), we have

$$L_2(x, h, y(h)) = x_2^2.$$

Therefore, $L''_{2\,xx}(x^*, h, y(h))[h]^2 = 2 \geq 0$.

For the element $h = (1, -1, 0)^T$, all conditions of Theorem 3.3 are verified in the same way.

Hence, we proved that sufficient conditions for optimality are satisfied at the point $x^* = 0$. This means that x^* is a strict local minimizer to (5.1). □

The following examples illustrate the application of Theorem 4.1.

Example 5.2. Consider problem (1.1) with $x^* = 0$ and

$$F(x) = F(x_1, x_2) = x_1^2 - x_2^3 + x_2^{7/2} = 0.$$

We have $F'(x) = (2x_1, -3x_2^2 + \frac{7}{2}x_2^{5/2})$, $F'(x^*) = (0, 0)$,

$$F''(x) = \begin{pmatrix} 2 & 0 \\ 0 & -6x_2 + \frac{35}{4}x_2^{3/2} \end{pmatrix}, \qquad F'''(x^*)\,[h]^2 = (0, -6h_2^2).$$

Therefore,

$$\operatorname{Ker}{}^2 F''(x^*) = \{h \in \mathbb{R}^2 \,|\, h_1 = 0,\ h_2 \in \mathbb{R}\},$$

and we obtain $F''(x^*)\,[h] = (0, 0)$ for all $h \in \operatorname{Ker}{}^2 F''(x^*)$; i.e., the mapping F is not 2-regular along the elements of the kernel of the second derivative, and

$$\operatorname{Im} F''(x^*)[h] = \{0\}.$$

Hence,

$$Y_1(h) = \{0\}, \quad Y_2(h) = \mathbb{R}, \quad P_2(h) = 1.$$

For the element $h = (0, \pm 1) \in \operatorname{Ker}{}^2 F''(x^*)$, there exists an element $\widetilde{h} = (1, 0)$ such that

$$F''(x^*)[h] + F''(x^*)[\widetilde{h}] = (2\widetilde{h}_1, 0)^T$$

and

$$F''(x^*)\,[h, \widetilde{h}] = 0, \qquad F''(x^*)\,[\widetilde{h}]^2 = 2,$$

$$\operatorname{Im}(F''(x^*)\,[h] + P_2(h)\,F''(x^*)\,[\widetilde{h}]) = \mathbb{R},$$

$$\begin{aligned} P_2(h)\,F(x^* + th + t^\alpha \widetilde{h}) = F(t^\alpha \widetilde{h}_1,\, th_2) &= t^{2\alpha}\widetilde{h}_1^2 - t^3 h_2^3 + t^{7/2} h_2^{7/2} \\ &= t^3(\widetilde{h}_1^2 - h_2^3) + t^{7/2} h_2^{7/2}, \qquad \alpha = \frac{3}{2}. \end{aligned}$$

By virtue of the latter relations, for all the conditions of Theorem 4.1 to be satisfied, it is necessary that the equation

$$\widetilde{h}_1^2 - h_2^3 = 0 \tag{5.3}$$

has a nonzero solution. This holds for $h_2 > 0$. However, (5.3) does not have a nonzero solution if $h_2 < 0$. Hence, the arc

$$x^* + th + t^\alpha \widetilde{h} + r_1(t), \qquad \| r_1(t) \| = o(t^\alpha)$$

belongs to $M(x^*)$ only for $h = (0, +1)$. This example illustrates the essential difference from the p-regular case, where a tangent cone is always two-sided. In the case under study, half-spaces may come into play, which will extend the scope of the analysis.

Furthermore, $F''(x^*)[\widetilde{h}] = (2, 0)$,

$$Z_1 = \{z \in \operatorname{Im}(F''(x^*)[\widetilde{h}])^* \mid \langle z, \widetilde{h}\rangle = 0\}$$
$$= \left\{ z = t \begin{pmatrix} 2 \\ 0 \end{pmatrix} \;\middle|\; 2t + 0 = 0, \quad t \in \mathbb{R} \right\} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

i.e.,

$$P_1(\widetilde{h}) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Hence,

$$\operatorname{Im}((F''(x^*)[h] + P_2(h)F''(x^*)[\widetilde{h}])\, P_1{}^*(\widetilde{h}))^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$f'(x^*) \in \overset{0}{\Lambda}(F, h, x^*) \cup \{0\} = \{z \in \mathbb{R}^2 \mid \langle z, h\rangle < 0\} \cup \{0\}.$$

Thus, the gradient $f'(x^*)$ must either be situated strictly in the lower half-plane $X_2 < 0$ or be equal to zero. □

Example 5.3. Consider problem (1.1) with $x^* = 0$ and

$$F(x) = F(x_1, x_2) = \begin{pmatrix} x_1^2 - |x_2|^3 - x_3^4 \\ x_1 x_3 \end{pmatrix} = 0.$$

We have

$$\operatorname{Ker}{}^2 F''(x^*) = \{(0, h_2, h_3) \mid h_2 \in \mathbb{R}, h_3 \in \mathbb{R}\}$$

and

$$F''(x^*)[h] = \begin{pmatrix} 0 & 0 & 0 \\ h_3 & 0 & 0 \end{pmatrix}$$

for all $h \in \operatorname{Ker}{}^2 F''(x^*)$; i.e., the mapping F is not 2-regular along the elements of the kernel of the second derivative. For the element $h = (0, \pm 1, 0) \in \operatorname{Ker}{}^2 F''(x^*)$ we have

$$\operatorname{Im} F''(x^*)[h] = \{0\}.$$

Hence,

$$Y_1(h) = \{0\}, \quad Y_2(h) = \mathbb{R}^2, \quad P_2(h) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

For the element h and $\alpha = \frac{3}{2}$ there exists an element $\widetilde{h} = (1, 0, 0)$ such that

$$F''(x^*)[h] + P_2(h)F''(x^*)[\widetilde{h}] = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$F(x^* + th + t^\alpha \widetilde{h}) = F(t^\alpha \widetilde{h}_1,\, th_2) = \begin{pmatrix} t^3 - t^3|h_2|^3 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad t > 0.$$

Hence, all conditions of Theorem 4.1 are satisfied.

Furthermore, $Z = (0, 0, z),\ z \in \mathbb{R}$, and

$$P_1(\widetilde{h}) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad P_1(\widetilde{h})\,(F''(x^*)[\widetilde{h}])^T = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

On the one hand, by (4.6), we have for $h = (0, 1, 0)$

$$f'(x^*) \in \overset{0}{\Lambda}\,(F, h, x^*) \cup \operatorname{Im}(P_1(\widetilde{h})(F''(x^*)[\widetilde{h}])^T).$$

On the other hand, for the element $-h = (0, -1, 0)$,

$$f'(x^*) \in \overset{0}{\Lambda}\,(F, -h, x^*) \cup \operatorname{Im}(P_1(\widetilde{h})(F''(x^*)[\widetilde{h}])^T).$$

From the definition of $\overset{0}{\Lambda}$ and the last inclusions we obtain

$$f'(x^*) \in \operatorname{Im}(P_1(\widetilde{h})(F''(x^*)[\widetilde{h}])^T) = (0, 0, z)^T, \qquad z \in \mathbb{R}. \qquad \square$$

Let us note, by virtue of Corollary 4.2, if there exists an element h such that the intersection of the sets in (4.11) is empty, then

$$f'(x^*) = 0.$$

We illustrate this by the following example.

Example 5.4. Consider problem (1.1) with $x^* = 0$ and

$$F : \mathbb{R}^3 \to \mathbb{R}, \qquad F(x) = F(x_1, x_2, x_3) = x_1^2 - x_2^3 - x_3^2.$$

For the element $h = (0, \pm 1, 0)^T \in \operatorname{Ker}^2 F''(x^*)$ we obtain

$$\widetilde{h}^+ = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \widetilde{h}^- = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad P_1(\widetilde{h}^+) = 0, \quad P_1(\widetilde{h}^-) = 0.$$

Hence, for any problem (1.1) with

$$F(x) = F(x_1, x_2, x_3) = x_1^2 - x_2^3 - x_3^2 = 0,$$

the following assertion holds:

$$f'(x^*) = 0. \qquad \square$$

6. Conclusions. In this paper we derived new optimality conditions for problems with nonregular equality constraints. Our approach is based on constructions of p-regularity. In the first part of the paper, we derived new sufficient conditions for optimality for p-regular constrained optimization problems. This result complements necessary conditions for optimality which were obtained earlier. In the second part of the paper, we proved new necessary conditions for optimality for problems that satisfy a *generalized* condition of p-regularity. Both p-regular and *generalized* conditions for optimality reduce to classical conditions for regular cases, but they give new and nontrivial conditions for nonregular cases. The presented results can be considered as a part of the p-regularity theory.

Acknowledgments. The authors thank the corresponding editor Professor Heinz Schättler and the anonymous reviewers for their comments and suggestions which helped us improve considerably the content and presentation of this paper.

REFERENCES

[1] A. V. Arutyunov, *Higher-order conditions in abnormal extremal problems with equality-type constraints*, Soviet Math. Dokl., 42 (1991), pp. 799–804.

[2] A. V. Arutyunov, *Extremum Conditions. Abnormal and Degenerate Cases*, Factorial, Moscow, Russia, 1997 (in Russian).

[3] E. R. Avakov, *Extremum conditions for smooth problems with equality-type constraints*, USSR Comput. Math. and Math. Phys., 25 (1985), pp. 24–32.

[4] E. R. Avakov, *Necessary conditions for a minimum for nonregular problems in Banach spaces. The maximum principle for abnormal optimal control problems*, Trudy Mat. Inst. Steklov., 185 (1988), pp. 3–29 (in Russian).

[5] E. R. Avakov, *Necessary conditions for an extremum for smooth abnormal problems with constraints of equality- and inequality type*, Math. Notes, 45 (1989), pp. 431–437.

[6] K. N. Belash and A. A. Tret'yakov, *Methods for solving degenerate problems*, USSR Comput. Math. Math. Phys., 28 (1988), pp. 90–94.

[7] O. A. Brezhneva and A. A. Tret'yakov, *New Methods for Solving Singular Nonlinear Problems*, Computing Center of the Russian Academy of Sciences, Moscow, Russia, 2000 (in Russian).

[8] O. A. Brezhneva, A. A. Tret'yakov, and A. Chmura, *Generalization of the concept of p–regularity and higher order optimality conditions*, Comput. Math. Math. Phys., 41 (2001), pp. 187–196.

[9] M. Buchner, J. Marsden, and S. Schecter, *Applications of the blowing-up construction and algebraic geometry to bifurcation problems*, J. Differential Equations, 48 (1983), pp. 404–433.

[10] R. H. Byrd, D. Feng, and R. B. Schnabel, *On Optimality Conditions for Singular Constrained Optimization*, Tech. report 95.03, Research Institute for Advanced Computer Science, NASA Ames Research Center, Moffett Field, CA, 1995.

[11] D. V. Denisov, V. G. Karmanov, and A. A. Tret'yakov, *The accelerated Newton method for the solution of functional equations*, Dokl. Acad. Nauk. SSSR, 281 (1985), pp. 1293–1297 (in Russian).

[12] A. V. Dmitruk, *Quadratic conditions for a Pontryagin minimum in an optimal control problem linear with respect to the control.* II. *Theorems on the relaxing of constraints on the equality*, Math. USSR-Izv., 31 (1987), pp. 121–141.

[13] A. D. Ioffe and V. M. Tihomirov, *Theory of Extremal Problems,* North–Holland, Amsterdam, The Netherlands, 1979.

[14] A. F. Izmailov, *Optimality conditions for degenerate extremum problems with inequality-type constraints*, Comput. Math. Math. Phys., 34 (1994), pp. 723–736.

[15] A. F. Izmailov and M. V. Solodov, *The theory of 2–regularity for mappings with Lipschitzian derivatives and its applications to optimality conditions*, Math. Oper. Res., 27 (2002), pp. 614–635.

[16] A. F. Izmailov and M. V. Solodov, *Optimality conditions for irregular inequality-constrained problems*, SIAM J. Control Optim., 40 (2001), pp. 1280–1295.

[17] A. F. Izmailov and A. A. Tret'yakov, *On question about inversibility homogeneous p-order polynomial mappings*, Comput. Math. Math. Phys., 33 (1993), pp. 289–299.

[18] A. F. Izmailov and A. A. Tret'yakov, *Factor-Analysis of Nonlinear Mappings*, Nauka, Moscow, 1994 (in Russian).

[19] L. V. Kantorovitch and G. P. Akilov, *Functional Analysis*, Pergamon Press, Oxford, New York, 1982.

[20] U. Ledzewicz and H. Schättler, *Second-order conditions for extremum problems with non-regular equality constraints*, J. Optim. Theory Appl., 86 (1995), pp. 113–144.

[21] U. Ledzewicz and H. Schättler, *A high-order generalization of the Lyusternik theorem*, Nonlinear Anal., 34 (1998), pp. 793–815.

[22] U. Ledzewicz and H. Schättler, *High-order extended maximum principles for optimal control problems with non-regular constraints*, in Optimal Control: Theory, Algorithms, and Applications, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 298–325.

[23] U. Ledzewicz and H. Schättler, *High-order approximations and generalized necessary conditions for optimality*, SIAM J. Control Optim., 37 (1998), pp. 33–53.

[24] A. A. Milyutin, *On quadratic conditions for an extremum in smooth problems with a finite-dimensional image*, in Methods of the Theory of Extremal Problems in Economics, V. L. Lenin, ed., Nauka, Moscow, 1981, pp. 138–177 (in Russian).

[25] A. A. Tret'yakov, *Necessary conditions for optimality of p-th order*, Control and Optimization, MGU, 1983, pp. 28–35 (in Russian).

[26] A. A. Tret'yakov, *Necessary and sufficient conditions for optimality of p-th order*, Comput. Math. Math. Phys., 24 (1984), pp. 123–127.

[27] A. A. Tret'yakov, *The implicit function theorem in degenerate problems*, Russian Math. Surveys, 42 (1987), pp. 179–180.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 2, pp. 746–768

SPILLOVER STABILIZATION IN FINITE-DIMENSIONAL CONTROL AND OBSERVER DESIGN FOR DISSIPATIVE EVOLUTION EQUATIONS*

GREGORY HAGEN† AND IGOR MEZIĆ‡

Abstract. We consider the problem of global stabilization of a semilinear dissipative evolution equation by finite-dimensional control with finite-dimensional outputs. Coupling between the system modes occurs directly through the nonlinearity and also through the control influence functions. Similar modal coupling occurs in the infinite-dimensional error dynamics through the nonlinearity and measurements. For both the control and observer designs, rather than decompose the original system into Fourier modes, we consider Lyapunov functions based on the infinite-dimensional dynamics of the state and error systems, respectively. The inner product terms of the Lyapunov derivative are decomposed into Fourier modes. Upper bounds on the terms representing control and observation spillover are obtained. Linear quadratic regulator (LQR) designs are used to stabilize the state and error systems with these upper bounds. Relations between system and LQR design parameters are given to ensure global stability of the state and error dynamics with robustness with respect to control and observation spillover, respectively. It is shown that the control and observer designs can be combined to yield a globally stabilizing compensator. The control and observer designs are numerically demonstrated on the problem of controlling stall in a model of axial compressors.

Key words. linear-quadratic control, nonlinear reaction-diffusion equations, Lyapunov stability

AMS subject classifications. 49N10, 35K57, 34D20

DOI. 10.1137/S0363012900378942

1. Introduction.

Modeling and control of infinite-dimensional systems is commonly practiced by approximating the underlying system by a finite-dimensional lumped system (via Galerkin projection [33], proper orthogonal decomposition [28, 20, 21, 30], inertial manifolds [45], etc.). This is not necessary when the control is also infinite-dimensional and the system is spatially invariant [4, 25], which simplifies the analysis of the infinite-dimensional system. However, for practical reasons, it is desirable to obtain finite-dimensional controllers that will stabilize the entire system (see, e.g., [3, 26, 34]). The problem inherent in decomposing a nonlinear infinite-dimensional system into Fourier modes is that the nonlinearity couples all of the modes, often making the decomposed system just as intractable as the original infinite-dimensional one. Common to all dissipative systems, the higher modes of the system often exhibit some form of stability which allows for the treatment of a truncated system of a finite number of modes. Control design can then be systematically carried out to stabilize the reduced order system. However, a systematic and quantifiable method of determining the minimum order of truncation remains elusive. The high modes of the system cannot be completely neglected due to the fact that modes are still coupled through the nonlinearity and that the control can inadvertently introduce energy into these high modes. This energy can act to destabilize the high modes or can be transferred to the low modes through nonlinear coupling. This phenomenon is referred to as spillover [2]. Analysis of reduced order systems [10, 17]

*Received by the editors October 2, 2000; accepted for publication (in revised form) December 9, 2002; published electronically June 12, 2003.
http://www.siam.org/journals/sicon/42-2/37894.html
†United Technologies Research Center, 411 Silver Lane, East Hartford, CT 06108 (HagenGS@utrc.utc.com).
‡Department of Mechanical and Environmental Engineering, University of California, Santa Barbara, CA 93106 (mezic@engineering.ucsb.edu).

and control spillover has since become a topic of work dealing with the control of flexible structures and/or robust control (see, e.g., [9, 39, 41]).

In particular, recent work has employed the use of singular perturbation theory [12] to construct globally stabilizing finite-dimensional controllers [13] and locally stabilizing finite-dimensional controller/observers [1] for dissipative parabolic PDEs. However, the use of singular perturbation methods requires a priori knowledge of the initial conditions of the system [31] before such controllers can be constructed. Additionally, the use of approximate inertial manifolds and the nonlinear Galerkin method [33] has been utilized to derive reduced order models. These methods require the eigenspectrum of the linear parts of the system to satisfy a gap condition in which consecutive stable eigenvalues have a sufficiently large difference between them. This often requires the controller to stabilize modes that are already stable (see the appendix of [12]). Furthermore, the gap condition is even more difficult to satisfy when the system has strong convective terms and/or a small diffusion parameter [45].

In the present work, to alleviate some of the problems created by the coupling of modes through the nonlinearity of the system, we avoid directly decomposing the system into Fourier modes and instead consider a control Lyapunov function (CLF) for the original infinite-dimensional system and an observer Lyapunov function (OLF) for the infinite-dimensional error system. Details concerning Lyapunov stability theory in the context of infinite-dimensional systems can be found in [6, 14]. By applying a linear bound of the system nonlinearity, as in [25], we obtain an upper bound on the CLF derivative in terms of the Fourier coefficients of the state. This in turn provides for the direct construction of a linear finite-dimensional controller of the low modes of the system. The CLF derivative explicitly shows the destabilizing effects of the high modal content of the controller. Upper bounds of the high modes of the CLF derivative are obtained to quantify the effects of control spillover. A linear quadratic regulator (LQR) control design is used to make the CLF derivative negative with these upper bounds, and we state relations between the system and LQR parameters that are sufficient for global stability and robustness with respect to control spillover. The infinite-dimensional observer based on finite-dimensional measurements is designed by similar analysis of the OLF derivative. Furthermore, we show that the controller and observer can be combined while maintaining global stability.

The paper is organized as follows. Section 2 contains the problem setup and description. We first consider the control problem with full state information. In section 3, we state the CLF to be used and show that coupling between modes occurs through the nonlinearity and the control terms. The analysis of the nonlinear coupling is eliminated by considering a linear bound of the nonlinearity. This essentially diagonalizes the CLF derivative as in [25]; however, coupling still occurs through the control inputs. In section 4, we provide sufficient conditions on the control design and the system parameters to ensure that the closed loop system is robust to the destabilizing effects of control spillover. In section 5, we consider the estimation problem with finite-dimensional outputs in the absence of control. We consider an OLF, and the analysis is similar to that of sections 3 and 4. In section 6, we show that the closed loop system is globally stable when the controller and observer are combined. The duality principles between the control and observer designs and the respective spillover analyses are illustrated in section 7. In section 8, we apply the control and observer designs and spillover analyses to the problem of stabilizing a nonlocal evolution equation describing the rotating stall phenomena in axial compressors by a finite-dimensional controller with finite-dimensional outputs.

2. Problem formulation. We consider the following type of nonlocal dissipative evolution equation with periodic boundary conditions:

$$\frac{\partial y}{\partial t} = a^2 \frac{\partial^2 y}{\partial \theta^2} - \beta \frac{\partial y}{\partial \theta} + f(y) - \alpha \overline{f(y)} + B\mathcal{V}, \tag{1}$$

$$z = Hy,$$

$$y(t,0) = y(t,2\pi), \quad y(0,\theta) = y_0(\theta), \tag{2}$$

where a^2, β, and α are positive constants, the function $f(y)$ is a polynomial of odd degree with a strictly negative highest order coefficient, and

$$\overline{f(y)} = \frac{1}{2\pi} \int_0^{2\pi} f(y)\, d\theta. \tag{3}$$

This system describes the rotating stall phenomenon in axial compressors [25, 23], phase separation in binary mixtures [40, 36], chemical reaction systems [43, 45], as well as other systems [11, 19]. From now on, we shall take $\alpha = 1$ for convenience. This will restrict the average of $y(t,\theta)$ over θ to be constant for all time. Therefore, without loss of generality, we can assume that $y(t,\theta)$ is zero-average over θ for all time. The analysis in this paper is applicable to cases where $\alpha \neq 1$. In this case, the control design would have to incorporate stabilization of the zeroth Fourier mode of the system. The control $\mathcal{V} \in \mathbb{R}^N$ and the linear operator $B : \mathbb{R}^N \to \bar{L}_2(0,2\pi)$, where $\bar{L}_2(0,2\pi)$ is the space of square integrable functions that are 2π-periodic and have zero-average. We write $B\mathcal{V} = \sum_{i=1}^N v^i b^i(\theta)$, so the controls enter the system additively through N different amplitudes v^i with shape functions $b^i(\theta)$. The output $z \in \mathbb{R}^N$ and the linear operator $H : \bar{L}_2(0,2\pi) \to \mathbb{R}^N$. The required regularity results for the feedback system (1) can be found in [5, 25, 38]. We note that the superscripts i are for notation and do not represent powers. The control objective is to globally stabilize the equilibrium solution $y(\theta) = 0$.

3. Modal decomposition and Lyapunov stability analysis. We first consider the control problem with full state information. Assuming $y(t,\theta)$ is zero-average over θ for all t, we decompose the function $y(t,\theta)$ into Fourier modes,

$$y(t,\theta) = \sum_{n=1}^{\infty} y_n(t)\phi_n(\theta), \tag{4}$$

where $y_n(t) = \langle y(t,\theta), \phi_n(\theta)\rangle$ is the nth Fourier coefficient of $y(t,\theta)$ and $\phi_n(\theta)$ is the nth orthonormal function of the Fourier series satisfying the boundary conditions. The expression $\langle \cdot, \cdot \rangle$ is the inner product on $\bar{L}_2(0,2\pi)$. If we immediately substitute (4) into (1), we obtain for each $n \geq 1$

$$\dot{y}_n = (-a^2 n^2 - \beta j n) y_n + f_n\left(\sum_{n=1}^{\infty} y_n \phi_n\right) + \sum_{i=1}^{N} v^i b_n^i; \quad j = \sqrt{-1}. \tag{5}$$

It is easy to see that the modes are coupled through $f(y)$ and the shape functions $b^i(\theta)$. We consider the CLF

$$V(y) = \frac{1}{2}\|y\|^2 = \frac{1}{2}\langle y, y\rangle, \tag{6}$$

where $\|\cdot\|$ is the usual norm in $\bar{L}_2(0,2\pi)$.

In order to bypass the difficulty caused by nonlinear modal coupling in (5), we apply a linear bound of the system nonlinearity before decomposing into Fourier modes. For a given polynomial $f(x)$ of odd degree with a strictly negative highest order coefficient and $f(0) = 0$, there exists a constant C such that

$$xf(x) \leq Cx^2 \quad \forall x \in \mathbb{R}. \tag{7}$$

Since $f(y)$ is independent of the spatial variable θ, we can extend this inequality to the infinite-dimensional case,

$$\langle y, f(y) \rangle \leq C\|y\|_2^2 \quad \forall y \in \bar{L}_2(0, 2\pi). \tag{8}$$

Now we take the time derivative of the CLF:

$$\begin{aligned}
\dot{V} &= \left\langle y, a^2 \frac{\partial^2 y}{\partial \theta^2} - \beta \frac{\partial y}{\partial \theta} + f(y) - \overline{f(y)} + \sum_{i=1}^{N} v^i b^i(\theta) \right\rangle \\
&= -a^2 \left\| \frac{\partial y}{\partial \theta} \right\|^2 + \langle y, f(y) \rangle - \langle y, \overline{f(y)} \rangle + \left\langle y, \sum_{i=1}^{N} v^i b^i(\theta) \right\rangle \qquad (9) \\
&= -a^2 \left\| \frac{\partial y}{\partial \theta} \right\|^2 + \langle y, f(y) \rangle + \left\langle y, \sum_{i=1}^{N} v^i b^i(\theta) \right\rangle \qquad (10) \\
&\leq -a^2 \left\| \frac{\partial y}{\partial \theta} \right\|^2 + C\|y\|^2 + \left\langle y, \sum_{i=1}^{N} v^i b^i(\theta) \right\rangle, \qquad (11)
\end{aligned}$$

where (9) is obtained from integrating by parts and applying the boundary conditions and (10) is obtained from the fact that $\overline{f(y)}$ is a constant and $y(t, \theta)$ is zero-average over θ for all time. Inequality (11) is obtained by using the linear bound (8).

REMARK 1. *The CLF given with the periodic boundary conditions cancels the term $\beta \frac{\partial y}{\partial \theta}$. Similarly, the presence of the Burgers nonlinearity in the system* (1), (2) *is inconsequential to the forthcoming stability analysis because, by integration by parts, $\langle y, y \frac{\partial y}{\partial \theta} \rangle = 0$.*

REMARK 2. *If $\beta = 0$, then the forthcoming analysis also applies for homogeneous Dirichlet and Neumann boundary conditions.*

Instead of dealing directly with (5), we now decompose $y(t, \theta)$ into Fourier modes and substitute into (6) and (11) to get

$$V(y) = \frac{1}{2} \left\langle \sum_{n=1}^{\infty} y_n \phi_n, \sum_{n=1}^{\infty} y_n \phi_n \right\rangle = \frac{1}{2} \sum_{n=1}^{\infty} (y_n)^2 \tag{12}$$

and

$$\begin{aligned}
\sum_{n=1}^{\infty} (y_n) \dot{y}_n = \dot{V} &\leq \sum_{n=1}^{\infty} -a^2 n^2 y_n^2 + C \sum_{n=1}^{\infty} y_n^2 + \sum_{i=1}^{N} v^i \left(\sum_{n=1}^{\infty} y_n b_n^i \right) \\
&= \sum_{n=1}^{\infty} \left[\left(-a^2 n^2 + C \right) y_n^2 + y_n \sum_{i=1}^{N} v^i b_n^i \right], \qquad (13)
\end{aligned}$$

where we have also decomposed the input shape functions $b^i(\theta)$ into Fourier modes. In (13), we can directly compare the stabilizing $(-a^2 n^2)$ term with the destabilizing (C) term for each n. The last term on the right side of (13) shows how the control can introduce energy into the high modal content $\dot{V}$.

4. Linear control with spillover analysis. We use (13) to design a linear controller. We define $\dot{V}_n$ to be the nth term of the summation in (13),

$$\dot{V}_n = (-a^2n^2 + C)y_n^2 + y_n \sum_{i=1}^{N} v^i \; b_n^i, \tag{14}$$

and note that $\dot{V} \leq \sum_{n=1}^{\infty} \dot{V}_n$. Because the modes are coupled, we stabilize the entire system by stabilizing $\dot{V}_n$ for every n. Consider the uncontrolled system. We have, for sufficiently large values of n,

$$\dot{V}_n = (-a^2n^2 + C)y_n^2 \leq 0. \tag{15}$$

This shows that for the purpose of control design, the high modes exhibit some form of stability. We assume that N is such that

$$-a^2N^2 + C > 0 > -a^2(N+1)^2 + C. \tag{16}$$

Note that we have assumed that there are N actuators. This assumption states that the number of actuators must be at least the number of unstable modes of the operator $(a^2 \frac{\partial^2}{\partial \theta^2} + C)$. We now consider stabilizing the first N terms in the upper estimate of the CLF derivative. We define

$$\mathbf{Y}_N := [y_1 \dots y_N]^T, \;\; \mathcal{V} := \left[v^1 \dots v^N\right]^T, \;\; \mathbf{V}_N := \sum_{n=1}^{N} V_n, \;\; \mathbf{V}_{n>N} := \sum_{n>N} V_n,$$

$$A_N := \begin{bmatrix} -a^2 + C & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -a^2N^2 + C \end{bmatrix}, \tag{17}$$

$$b_n := [b_n^1 \dots b_n^N]^T, \;\; \mathcal{B}_N := \begin{bmatrix} b_1^1 & \dots & b_1^N \\ \vdots & \ddots & \vdots \\ b_N^1 & \dots & b_N^N \end{bmatrix}. \tag{18}$$

Thus we have $\dot{V} \leq \dot{\mathbf{V}}_N + \dot{\mathbf{V}}_{n>N}$. We let $\mathcal{V}$ be a linear controller

$$\mathcal{V} = -K\mathbf{Y}_N, \tag{19}$$

where K is an $N \times N$ gain matrix. Using (17)–(19), the first N terms of the estimate (13) become

$$\dot{\mathbf{V}}_N = \sum_{i=1}^{N} \dot{V}_i \leq \frac{1}{2}\mathbf{Y}_N^T \left([A_N - \mathcal{B}_N K] + [A_N - \mathcal{B}_N K]^T \right) \mathbf{Y}_N := \mathbf{Y}_N^T \Lambda_N \mathbf{Y}_N. \tag{20}$$

Assuming that $\mathcal{B}_N$ is such that the system is controllable, we can design K such that Λ_N is negative definite. This can be achieved by an LQR design on the finite-dimensional system (see, e.g., [4, 25]).

$$\text{Minimize: } J(\mathcal{V}, \mathbf{Y}_N) = \int_0^{\infty} \left(\mathbf{Y}_N^T Q \mathbf{Y}_N + \mathcal{V}^T R \mathcal{V} \right) \, dt$$

$$\text{Subject to: } \dot{\mathbf{Y}}_N = A_N \mathbf{Y}_N + \mathcal{B}_N \mathcal{V}.$$

The LQR problem gives a feedback law

$$\mathcal{V} = -R^{-1}\mathcal{B}_N^T P \mathbf{Y}_N, \tag{21}$$

where P is the solution of the algebraic Riccati equation (ARE),

$$0 = A_N^T P + P A_N + Q - P\mathcal{B}_N R^{-1}\mathcal{B}_N^T P. \tag{22}$$

Substituting $K = R^{-1}\mathcal{B}_N^T P$ into (22), we obtain the Lyapunov equation,

$$(A_N - \mathcal{B}_N K)^T P + P(A_N - \mathcal{B}_N K) = -Q - K^T R K. \tag{23}$$

By applying (23) with $P = \frac{1}{2}I$ to (20), we have

$$\dot{\mathbf{V}}_N = \mathbf{Y}_N^T \left[-Q - K^T R K\right] \mathbf{Y}_N. \tag{24}$$

Now that we have constructed a control law that will stabilize the first N terms of the summation in (13), we analyze the possibly destabilizing effect the control has on the higher modes of the system. We repeat (14) here,

$$\dot{V}_n = (-a^2 n^2 + C) y_n^2 + y_n \sum_{i=1}^{N} v^i \; b_n^i. \tag{25}$$

For $n > N$, the first term is always ≤ 0, and the second term is sign indefinite. This term quantifies the control spillover onto the higher modes of the system. Specifically, if the input shape functions $b^i(\theta)$ have high frequency content, there is the possibility of more control spillover. We define the following scalars:

$$c_n := a^2 n^2 - C, \quad d_n := -b_n^T K \mathbf{Y}_N. \tag{26}$$

Substituting (19) and (26) into (25) results in

$$\dot{V}_n = -c_n y_n^2 + d_n y_n. \tag{27}$$

It is easy to see that this expression reaches its maximum value at $y_n = \frac{1}{2}\frac{d_n}{c_n}$, and so for each $n > N$,

$$\dot{V}_n \leq \frac{1}{4}\frac{d_n^2}{c_n} = \frac{1}{4c_n}\mathbf{Y}_N^T K^T b_n b_n^T K \mathbf{Y}_N. \tag{28}$$

We see from (28) how high frequency content of the input shape functions and high controller gains can destabilize the entire system. Since each $\dot{V}_n$ corresponds to the possible destabilization of a high mode, the stabilization provided by (24) must be such that $\dot{V}$ remains negative with the addition of terms such as (28). We will show that this is guaranteed given certain LQR design parameters and some conditions on the reduced order model $(A_N, \mathcal{B}_N)$.

Recall that $\dot{V}$ is majorized by summing the terms $\dot{V}_n$. The controller (21) stabilizes the first N modes of the system with stability given by (24), while the rest of the terms of the summation of the CLF derivative estimate are each bounded according to (28). The total possible destabilizing effect produced by control spillover is given by summing over the all the terms for $n > N$,

$$\dot{\mathbf{V}}_{n>N} \leq \frac{1}{4}\mathbf{Y}_N^T K^T \left[\sum_{n>N} \frac{1}{c_n} b_n b_n^T\right] K \mathbf{Y}_N. \tag{29}$$

Since the input operators are bounded, the summation in (29) exists, and we define the $N \times N$ matrix

$$\Xi := \sum_{n>N} \frac{1}{c_n} b_n b_n^T. \tag{30}$$

REMARK 3. *Given sufficient regularity of solutions of* (1), (2), *the assumption of bounded input operators may be lifted. For example, since* $c_n = \mathcal{O}(n^2)$, *the summation exists for unbounded "point" operators.*

This gives

$$\dot{\mathbf{V}}_{n>N} \leq \frac{1}{4} \mathbf{Y}_N^T K^T \Xi K \mathbf{Y}_N. \tag{31}$$

The $N \times N$ matrix Ξ quantifies the destabilizing effects of the control spillover. Notice that increasing the control gain K and the high frequency content of the input functions $b^i(\theta)$ both contribute to the destabilizing effect. We can now go back and analyze the original CLF (13). Adding (24) and (31), we get

$$\dot{V} \leq \mathbf{Y}_N^T \left[-Q - K^T R K + \frac{1}{4} K^T \Xi K \right] \mathbf{Y}_N. \tag{32}$$

We now point out that we need to solve the LQR design problem in reverse order. That is, we have specified $P = \frac{1}{2} I$ in deriving (24). We will specify the positive definite matrix R to account for the effects of control spillover and then solve the ARE for an admissible positive definite matrix Q.

Motivated by (32), we specify a positive definite matrix

$$R \geq \frac{1}{4} \Xi \tag{33}$$

to cancel the effects of control spillover. If we can solve the ARE for a positive definite Q with our choices of P and R, then we have

$$\dot{V} \leq -\mathbf{Y}_N^T Q \mathbf{Y}_N \leq 0. \tag{34}$$

When $\mathbf{Y}_N^T Q \mathbf{Y}_N = 0$, the control is zero, and from (15) we see that the high modes decay naturally. Therefore, $\dot{V} = 0$ only if $y_n = 0$ for all n. We now solve the ARE (22) for Q with our choice of $P = \frac{1}{2} I$. Since A_N is diagonal,

$$Q = -A_N + \frac{1}{4} \mathcal{B}_N R^{-1} \mathcal{B}_N^T, \tag{35}$$

which is positive definite if and only if

$$A_N < \frac{1}{4} \mathcal{B}_N R^{-1} \mathcal{B}_N^T. \tag{36}$$

The control design is now reduced to finding a positive definite matrix R that satisfies (33) and (36). Note that A_N is positive definite (see (16), (17)), and therefore $\mathcal{B}_N$ must be invertible. Equation (36) is equivalent to requiring $R < \frac{1}{4} \mathcal{B}_N^T A_N^{-1} \mathcal{B}_N$, and combining this with (33), it is necessary that

$$\mathcal{B}_N^T A_N^{-1} \mathcal{B}_N > \Xi. \tag{37}$$

This inequality can be checked by computing the eigenvalues of $\mathcal{B}_N^T A_N^{-1} \mathcal{B}_N - \Xi$ (see [29]), and the matrix R can be chosen accordingly. We have just shown the following result.

THEOREM 4. *Given the system* (1)*, with* $f(y)$ *satisfying* (8)*, along with the definitions* (17), (18)*, and* (30)*, if the matrix inequality* (37) *is satisfied, then there exists* $R > 0$ *such that the closed loop system* (1) *with feedback control* (21)*, with* $P = \frac{1}{2} I$*, is globally stable.*

Controller synthesis. We outline the synthesis of the controllers resulting from Theorem 4:

1. Given the system nonlinearity $f(x)$, determine C according to (7).
2. Given the operators $\mathcal{A} := a^2 \frac{\partial^2 \cdot}{\partial \theta^2} - \beta \frac{\partial \cdot}{\partial \theta}$ and B, compute the matrices $A_N, \mathcal{B}_N$, and b_n according to (17), (18).
3. Compute Ξ according to (30).
4. Choose $R \geq \Xi$. For example, take $R = \Xi + \delta I$, with $\delta > 0$.
5. According to (37), check if $\mathcal{B}_N^T A_N^{-1} \mathcal{B}_N - R > 0$. If so, then the controller $\mathcal{V} = -\frac{1}{2} R^{-1} \mathcal{B}_N^T \mathbf{Y}_N$ is globally stabilizing. If not, then the construction may not stabilize the control spillover.

5. Linear observation with spillover analysis. In this section, we consider the system (1) with a finite-dimensional output. Our goal is to construct an observer based on the available finite-dimensional measurements that is robust to observation spillover. By the linear principle of duality, we expect to find requirements similar to (33) and (36). We consider (1), (2) with no control and with $\alpha = 1$,

$$\frac{\partial y}{\partial t} = a^2 \frac{\partial^2 y}{\partial \theta^2} - \beta \frac{\partial y}{\partial \theta} + f(y) - \overline{f(y)}, \tag{38}$$

with a finite-dimensional output $z \in \mathbb{R}^N$ given by

$$z = Hy, \tag{39}$$

where $H : \bar{L}_2(0, 2\pi) \to \mathbb{R}^N$ is linear. For the control design of section 4, we used only the Fourier coefficients y_n of $y(t, \theta)$ for feedback. This motivates us to consider directly estimating the finite number of Fourier coefficients. However, just as it was necessary to analyze the stability of the high modes of the state in the control design, in order to achieve global results, we will need to analyze the decay of the high modes of the error. Thus, in order to analyze the observation spillover, we need to construct an estimate of the full state, $\hat{y}(t, \theta)$. We write the estimator system,

$$\frac{\partial \hat{y}}{\partial t} = a^2 \frac{\partial^2 \hat{y}}{\partial \theta^2} - \beta \frac{\partial \hat{y}}{\partial \theta} + f(\hat{y}) - \overline{f(\hat{y})} + LH(y - \hat{y}), \tag{40}$$

$$\hat{y}(t, 0) = \hat{y}(t, 2\pi), \quad \hat{y}(0, \theta) = 0, \tag{41}$$

where L will be the linear observer gain that is to be designed. For reasons that will be evident later, we choose L such that $L : \mathbb{R}^N \to \text{Span}\{\phi_1 \dots \phi_N\}$. If $H : \bar{L}_2(0, 2\pi) \to \mathbb{R}^N$, we can write

$$H := \left[H^1 \dots H^N\right]^T ; H^i : \bar{L}_2(0, 2\pi) \to \mathbb{R}, \tag{42}$$

and L is represented by

$$L := \left[L^1(\theta) \dots L^N(\theta)\right]. \tag{43}$$

We can immediately see the duality between the control and observer designs. In the control design, the input shape functions $b^i(\theta)$ are given, and the design objective is to determine the vector v. In the observer design, the vector output Hy is given, and the objective is to determine the estimation shape functions $L^i(\theta)$. We note that the superscripts in (42) and (43) are for notation and do not represent exponents.

In constructing the observer, we will use an inequality similar to the linear bound given by (7) in order to simplify the spillover analysis. We will stabilize the zero solution of the error state,

$$\tilde{y} := y - \hat{y}, \tag{44}$$

whose dynamics are governed by

$$\frac{\partial \tilde{y}}{\partial t} = a^2 \frac{\partial^2 \tilde{y}}{\partial \theta^2} - \beta \frac{\partial \tilde{y}}{\partial \theta} + f(y) - \overline{f(y)} - (f(\hat{y}) - \overline{f(\hat{y})}) - LH\tilde{y}, \tag{45}$$

$$\tilde{y}(t,0) = \tilde{y}(t,2\pi), \;\; \tilde{y}(0,\theta) = y_0(\theta). \tag{46}$$

Note that we have subtracted the averaged quantities which enforce $\overline{\tilde{y}}(t,\theta) = 0$ for all $t > 0$. Since y is evolving in time, we need to consider the linear bound of $f(y)$ for all y. We assume that the constant C_L is the maximum positive slope of $f(y)$, so

$$(x-y)(f(x) - f(y)) \leq C_L (x-y)^2 \;\; \forall x, y \in \mathbb{R}. \tag{47}$$

This inequality is less conservative than the Lipschitz property. We extend this inequality to the infinite-dimensional case

$$\langle x - y, f(x) - f(y) \rangle \leq C_L \|x - y\|_2^2 \;\; \forall x, y \in \bar{L}_2(0, 2\pi). \tag{48}$$

We will now follow a process similar to that of section 4. We consider an OLF

$$W(\tilde{y}) = \frac{1}{2}\|\tilde{y}\|_2^2 = \frac{1}{2}\langle \tilde{y}, \tilde{y} \rangle \tag{49}$$

and take the time derivative

$$\begin{aligned}
\dot{W} &= \left\langle \tilde{y}, a^2 \frac{\partial^2 \tilde{y}}{\partial \theta^2} - \beta \frac{\partial \tilde{y}}{\partial \theta} + f(y) - \overline{f(y)} - (f(\hat{y}) - \overline{f(\hat{y})}) - LH\tilde{y} \right\rangle \\
&= -a^2 \left\| \frac{\partial \tilde{y}}{\partial \theta} \right\|^2 + \langle \tilde{y}, f(y) - \overline{f(y)} - (f(\hat{y}) - \overline{f(\hat{y})}) \rangle - \langle \tilde{y}, LH\tilde{y} \rangle \qquad (50) \\
&= -a^2 \left\| \frac{\partial \tilde{y}}{\partial \theta} \right\|^2 + \langle \tilde{y}, f(y) - f(\hat{y}) \rangle - \langle \tilde{y}, LH\tilde{y} \rangle \qquad (51) \\
&\leq -a^2 \left\| \frac{\partial \tilde{y}}{\partial \theta} \right\|^2 + C_L \|\tilde{y}\|_2^2 - \langle \tilde{y}, LH\tilde{y} \rangle, \qquad (52)
\end{aligned}$$

where (50) is obtained from integrating by parts and applying the boundary conditions and (51) is obtained from the fact that $\overline{f(y)}$ and $\overline{f(\hat{y})}$ are constants and $\tilde{y}(t,\theta)$ is zero-average over θ for all time. Inequality (52) is obtained by using the inequality (48). Just as in section 4, we now decompose $\tilde{y}(t,\theta)$ into Fourier modes, $\tilde{y}(t,\theta) = \sum_{n=1}^{\infty} \tilde{y}_n(t)\phi_n(\theta)$, and substitute into (49) and (52) to get

$$W(\tilde{y}) = \frac{1}{2}\left\langle \sum_{n=1}^{\infty} \tilde{y}_n \phi_n, \sum_{n=1}^{\infty} \tilde{y}_n \phi_n \right\rangle = \frac{1}{2} \sum_{n=1}^{\infty} (\tilde{y}_n)^2, \tag{53}$$

$$\dot{W} \leq \sum_{n=1}^{\infty} \left[(-a^2 n^2 + C_L)\tilde{y}_n^2 - \tilde{y}_n \left(LH\tilde{y}\right)_n \right], \tag{54}$$

where $(LH\tilde{y})_n = \langle LH\tilde{y}, \phi_n \rangle$. From (42) and (43),

$$(LH\tilde{y})_n = \left\langle \sum_{i=1}^{N} L^i(\theta) H^i \tilde{y}, \phi_n \right\rangle = \sum_{i=1}^{N} L_n^i H^i \tilde{y} = \left[L_n^1 \dots L_n^N \right] H\tilde{y}, \tag{55}$$

where $L_n^i = \langle L^i(\theta), \phi_n \rangle$. Just as in section 4, we take N such that $-a^2N^2 + C_L > 0 > -a^2(N+1)^2 + C_L$. We define $\dot{W}_n$ to be the nth term in the summation of the estimate (54),

$$\dot{W}_n = (-a^2n^2 + C_L)\hat{y}_n^2 - \tilde{y}_n \left(LH\tilde{y}\right)_n, \tag{56}$$

and note that $\dot{W} \leq \sum_{n=1}^{\infty} \dot{W}_n$. We define

$$\tilde{\mathbf{Y}}_N := [\tilde{y}_1 \dots \tilde{y}_N]^T, \quad \mathbf{W}_N := \sum_{n=1}^{N} W_n, \quad \mathbf{W}_{n>N} := \sum_{n>N} W_n,$$

$$A := \begin{bmatrix} -a^2 + C_L & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -a^2N^2 + C_L \end{bmatrix}, \quad \mathbf{L}_N = \begin{bmatrix} L_1^1 & \dots & L_1^N \\ \vdots & \ddots & \vdots \\ L_N^1 & \dots & L_N^N \end{bmatrix}, \tag{57}$$

$$h_n := H\phi_n, \quad \mathbf{H}_N := [h_1 \dots h_N]. \tag{58}$$

Thus we have $\dot{W} \leq \dot{\mathbf{W}}_N + \dot{\mathbf{W}}_{n>N}$. We write $H\tilde{y}$ as

$$H\tilde{y} = LH\left(\sum_{n=1}^{\infty} \tilde{y}_n \phi_n \right) = \sum_{n=1}^{\infty} \tilde{y}_n H\phi_n = \mathbf{H}_N \tilde{\mathbf{Y}}_N + \sum_{n>N} \tilde{y}_n h_n, \tag{59}$$

so

$$\dot{\mathbf{W}}_N = \frac{1}{2}\tilde{\mathbf{Y}}_N^T (A - \mathbf{H}_N^T \mathbf{L}_N^T)\tilde{\mathbf{Y}}_N + \frac{1}{2}\tilde{\mathbf{Y}}_N^T (A - \mathbf{H}_N^T \mathbf{L}_N^T)^T \tilde{\mathbf{Y}}_N - \tilde{\mathbf{Y}}_N^T \mathbf{L}_N \sum_{m>N} \tilde{y}_m h_m, \tag{60}$$

$$\dot{\mathbf{W}}_{n>N} = \sum_{n>N} (-a^2n^2 + C_L)\tilde{y}_n^2. \tag{61}$$

The term $\sum_{m>N} \tilde{y}_m h_m$ in (60) represents the projection of the high modes of the state onto the output operator H, which quantifies the observation spillover. The analysis now follows in a similar fashion as in section 4. However, in this case, the observation spillover affects the stabilization of the first N terms of the OLF derivative, as is shown in (60). Adding (60) and (61) and rearranging the terms give

$$\begin{aligned} \dot{W} \leq \dot{\mathbf{W}}_N + \dot{\mathbf{W}}_{n>N} = {} & \frac{1}{2}\tilde{\mathbf{Y}}_N^T (A - \mathbf{H}_N^T \mathbf{L}_N^T)\tilde{\mathbf{Y}}_N + \frac{1}{2}\tilde{\mathbf{Y}}_N^T (A - \mathbf{H}_N^T \mathbf{L}_N^T)^T \tilde{\mathbf{Y}}_N \\ & + \sum_{n>N} \left[-\tilde{\mathbf{Y}}_N^T \mathbf{L}_N \tilde{y}_n h_n + (-a^2n^2 + C_L)\tilde{y}_n^2 \right]. \end{aligned} \tag{62}$$

We define for $n \geq N$

$$c_n := a^2n^2 - C_L, \quad d_n := -\tilde{\mathbf{Y}}_N^T \mathbf{L}_N h_n \tag{63}$$

and see that the expression $-c_n\tilde{y}_n^2 + d_n\tilde{y}_n$ attains its maximum at $\tilde{y}_n = \frac{d_n}{2c_n}$, so the terms of the summation in (62) are bounded by

$$-(a^2n^2 - C_L)\tilde{y}_n^2 - \tilde{\mathbf{Y}}_N^T\mathbf{L}_N h_n\tilde{y}_n \leq \frac{1}{4c_n}\tilde{\mathbf{Y}}_N^T\mathbf{L}_N h_n h_n^T\mathbf{L}_N^T\tilde{\mathbf{Y}}_N. \tag{64}$$

We define the $N \times N$ matrix

$$\Omega := \sum_{n>N} \frac{1}{c_n} h_n h_n^T, \tag{65}$$

and it follows that

$$\dot{W} \leq \frac{1}{2}\tilde{\mathbf{Y}}_N^T(A - \mathbf{H}_N^T\mathbf{L}_N^T)\tilde{\mathbf{Y}}_N + \frac{1}{2}\tilde{\mathbf{Y}}_N^T(A - \mathbf{H}_N^T\mathbf{L}_N^T)^T\tilde{\mathbf{Y}}_N + \frac{1}{4}\tilde{\mathbf{Y}}_N^T\mathbf{L}_N\Omega\mathbf{L}_N^T\tilde{\mathbf{Y}}_N. \tag{66}$$

We now construct a linear optimal observer based on the following problem:

$$\text{Minimize: } J(\tilde{\mathbf{Y}}_N) = \int_0^\infty (\tilde{\mathbf{Y}}_N^T Q\tilde{\mathbf{Y}}_N + \tilde{\mathbf{Y}}_N\mathbf{L}_N R\mathbf{L}_N^T\tilde{\mathbf{Y}}_N)\, dt$$
$$\text{Subject to: } \dot{\tilde{\mathbf{Y}}}_N = A\tilde{\mathbf{Y}}_N - \mathbf{H}_N^T\mathbf{L}_N^T\tilde{\mathbf{Y}}_N.$$

The resulting observer gain is

$$\mathbf{L}_N^T = R^{-1}\mathbf{H}_N P, \tag{67}$$

where $P > 0$ solves the ARE

$$0 = A^T P + PA + Q - P\mathbf{H}_N^T R^{-1}\mathbf{H}_N P. \tag{68}$$

We note here that (68) is the control version of the ARE and is not the dual (or observer) form of the ARE (see, e.g., [44]). If the underlying system were linear, then the two would be equivalent. In this case, we are using the solution of the ARE as a tool to cancel the observation spillover. We discuss the duality between control and observation spillover further in section 7. Substituting (67) into (68), we obtain the closed loop Lyapunov equation,

$$(A - \mathbf{H}_N^T\mathbf{L}_N^T)^T P + P(A - \mathbf{H}_N^T\mathbf{L}_N^T) = -Q - \mathbf{L}_N R\mathbf{L}_N^T. \tag{69}$$

We take $P = \frac{1}{2}I$ and substitute (69) into (66) to obtain

$$\dot{W} \leq \tilde{\mathbf{Y}}_N^T\left[-Q - \mathbf{L}_N R\mathbf{L}_N^T + \frac{1}{4}\mathbf{L}_N\Omega\mathbf{L}_N^T\right]\tilde{\mathbf{Y}}_N. \tag{70}$$

We now obtain the requirements on R as in (33) and (36),

$$R \geq \frac{1}{4}\Omega, \tag{71}$$

$$A < \frac{1}{4}\mathbf{H}_N^T R^{-1}\mathbf{H}_N, \tag{72}$$

where (72) is to ensure that $Q > 0$. As in the previous section, if A is positive definite, then H_N must be invertible, and hence it is necessary that

$$\mathbf{H}_N A^{-1}\mathbf{H}_N^T > \Omega. \tag{73}$$

This inequality can be checked by computing the eigenvalues of $\mathbf{H}_N A^{-1}\mathbf{H}_N^T - \Omega$ (see [29]), and the matrix R can be chosen accordingly. We have just shown the following result.

THEOREM 5. *Given the estimation system* (38)–(46)*, where* $f(y)$ *satisfies* (48)*, along with the definitions* (57)*,* (58)*,* (65)*, if the matrix inequality* (73) *is satisfied, then there exists* $R > 0$ *such that the closed loop system* (45) *with* L *parameterized by* (58)*,* (67)*, with* $P = \frac{1}{2}I$*, is globally stable.*

Observer synthesis. We outline the synthesis of the observers resulting from Theorem 5:

1. Given the system nonlinearity $f(x)$, determine C_L according to (48).
2. Given the operators $\mathcal{A} := a^2 \frac{\partial^2 \cdot}{\partial \theta^2} - \beta \frac{\partial \cdot}{\partial \theta}$ and H, compute the matrices $A, \mathbf{H}_N$, and h_n according to (57), (58).
3. Compute Ω according to (65).
4. Choose $R \geq \Omega$. For example, take $R = \Omega + \epsilon I$.
5. According to (73), check if $\mathbf{H}_N A^{-1}\mathbf{H}_N^T - R > 0$. If so, then the observer gain $\mathbf{L}_N^T = \frac{1}{2}R^{-1}\mathbf{H}_N$ globally stabilizes the error system (45). If not, then the observer construction may not stabilize the observation spillover.

6. Combined control and observation. In this section, we combine the controller derived in section 4 and the observer derived in section 5. For a linear system, we can simply combine the controller with the observer with no additional analysis. Since the system we are dealing with is nonlinear, we need to check the stability of the closed loop system. The state and error dynamics are

$$\frac{\partial y}{\partial t} = a^2 \frac{\partial^2 y}{\partial \theta^2} - \beta \frac{\partial y}{\partial \theta} + f(y) - \overline{f(y)} + Bv(y) - Bv(\tilde{y}), \tag{74}$$

$$\frac{\partial \tilde{y}}{\partial t} = a^2 \frac{\partial^2 \tilde{y}}{\partial \theta^2} - \beta \frac{\partial \tilde{y}}{\partial \theta} + f(y) - \overline{f(y)} - (f(\hat{y}) - \overline{f(\hat{y})}) - LH\tilde{y}, \tag{75}$$

with periodic boundary conditions. We will use the combined Lyapunov function

$$V(y, \tilde{y}) = \frac{1}{2}\|y\|_2^2 + \frac{\gamma}{2}\|\tilde{y}\|_2^2 = \frac{1}{2}\sum_{n=1}^{\infty} y_n^2 + \frac{\gamma}{2}\sum_{n=1}^{\infty} \tilde{y}_n^2,$$

where γ is a positive constant, and we will take the time derivative

$$\dot{V} = \frac{1}{2}\sum_{n=1}^{\infty} y_n \dot{y}_n + \frac{1}{2}\sum_{n=1}^{\infty} \dot{y}_n y_n + \frac{\gamma}{2}\sum_{n=1}^{\infty} \dot{\tilde{y}}_n \tilde{y}_n + \frac{\gamma}{2}\sum_{n=1}^{\infty} \tilde{y}_n \dot{\tilde{y}}_n.$$

For ease of notation, we will assume that A_N and A are $N \times N$ matrices. In general these matrices can be of different dimensions because $C_L \geq C$. The forthcoming stability analysis can be easily modified for this case. Using the notation of the previous sections, we have

$$\dot{V} \leq \mathbf{Y}_N^T \frac{1}{2}\left([A_N - \mathcal{B}_N K] + [A_N - \mathcal{B}_N K]^T\right)\mathbf{Y}_N \tag{76}$$

$$+ \sum_{n>N} -(a^2 n^2 - C) y_n^2 - y_n b_n^T K \mathbf{Y}_N + \sum_{n>N} y_n b_n^T K \tilde{\mathbf{Y}}_N \tag{77}$$

$$+ \frac{1}{2}\mathbf{Y}_N^T \mathcal{B}_N K \tilde{\mathbf{Y}}_N + \frac{1}{2}\tilde{\mathbf{Y}}_N^T K^T \mathcal{B}_N^T \mathbf{Y}_N \tag{78}$$

$$+ \tilde{\mathbf{Y}}_N^T \frac{\gamma}{2}\left([A - \mathbf{H}_N^T \mathbf{L}_N^T] + [A - \mathbf{H}_N^T \mathbf{L}_N^T]^T\right)\tilde{\mathbf{Y}}_N + \frac{\gamma}{4}\tilde{\mathbf{Y}}_N^T \mathbf{L}_N \Omega \mathbf{L}_N^T \tilde{\mathbf{Y}}_N. \tag{79}$$

We see that the terms in (76) and (79) are exactly the terms from the previous sections. However, there are extra dynamics due to the fact that the controller v is now a function of the estimate $\hat{y}$. These effects are quantified by the last term in (77) and the terms in (78). Following the reasoning of (25)–(31), we see that

$$\sum_{n>N} -(a^2n^2 - C)y_n^2 - \sum_{n>N} y_n b_n^T K(\mathbf{Y}_N - \tilde{\mathbf{Y}}_N) \leq \frac{1}{4}(\mathbf{Y}_N - \tilde{\mathbf{Y}}_N)^T K^T \Xi K(\mathbf{Y}_N - \tilde{\mathbf{Y}}_N).$$

We will use the control and observer designs of the previous sections. With $P = \frac{1}{2}I$, we have from (23)

$$\frac{1}{2}\left[A_N - \mathcal{B}_N K\right]^T + \frac{1}{2}\left[A_N - \mathcal{B}_N K\right] = -Q_c - K^T R_c K < -K^T R_c K,$$

and from (69) we have

$$\frac{\gamma}{2}\left[A - \mathbf{H}_N^T \mathbf{L}_N^T\right]^T + \frac{\gamma}{2}\left[A - \mathbf{H}_N^T \mathbf{L}_N^T\right] = -\gamma Q_o - \gamma \mathbf{L}_N R_o \mathbf{L}_N^T < -\gamma \mathbf{L}_N R_o \mathbf{L}_N^T.$$

Now, (76)–(79) becomes

$$\dot{V} \leq -\begin{bmatrix} \mathbf{Y}_N^T & \tilde{\mathbf{Y}}_N^T \end{bmatrix} M \begin{bmatrix} \mathbf{Y}_N \\ \tilde{\mathbf{Y}}_N \end{bmatrix},$$

$$M := \begin{bmatrix} K^T R_c K - \frac{1}{4}K^T \Xi K & -\frac{1}{2}\mathcal{B}_N K + \frac{1}{4}K^T \Xi K \\ -\frac{1}{2}K^T \mathcal{B}_N^T + \frac{1}{4}K^T \Xi K & \gamma \mathbf{L}_N R_o \mathbf{L}_N^T - \frac{\gamma}{4}\mathbf{L}_N \Omega \mathbf{L}_N^T - \frac{1}{4}K^T \Xi K \end{bmatrix}.$$

Just as in the previous designs, we will choose R_c and R_o to make M positive semidefinite and then determine the further requirements to find positive definite matrices Q_c and Q_o. We state a useful theorem from [29].

THEOREM 6. *If M is a Hermitian matrix partitioned as*

$$M = \begin{bmatrix} A & B \\ B^* & C \end{bmatrix}$$

such that A and C are square and A is invertible, then M is positive semidefinite if and only if A is positive semidefinite and $C \geq B^* A^{-1} B$.

Note that the condition that A be invertible can be lifted by using the pseudoinverse of A (see [32]). However, for the purpose of this analysis, this restriction is minor. In order to satisfy the first part of the theorem, we must choose R_c such that

$$S_c := R_c - \frac{1}{4}\Xi \geq 0, \tag{80}$$

which is exactly the requirement (33). We similarly define

$$S_o := R_o - \frac{1}{4}\Omega \geq 0, \tag{81}$$

where we note that at this point R_o is unspecified. Also note that (81) is exactly the requirement (71). We note that $K = \frac{1}{2}R_c^{-1}\mathcal{B}_N^T$, so $\mathcal{B}_N = 2K^T R_c$ so that the cross terms of the matrix M can be written as

$$-\frac{1}{2}\mathcal{B}_N K + \frac{1}{4}K^T \Xi K = -K^T R_c K + \frac{1}{4}K^T \Xi K = -K^T S_c K.$$

Similarly, $\left(K^T S_c K\right)^{-1} = K^{-1} S_c^{-1} K^{-T}$. In order to satisfy the second part of the theorem, we must choose R_o such that

$$\begin{aligned}\gamma \mathbf{L}_N S_o \mathbf{L}_N^T &\geq K^T \left[\frac{1}{4}\Xi + S_c K \left(K^T S_c K\right)^{-1} K^T S_c\right] K \\ &= K^T \left[\frac{1}{4}\Xi + S_c\right] K \\ &= K^T R_c K = \frac{1}{4}\mathcal{B}_N R_c^{-1} \mathcal{B}_N^T, \qquad (82)\end{aligned}$$

where (82) is obtained by substituting (80). If (81) is satisfied, then we can always find a $\gamma > 0$ such that (82) is satisfied. We must now solve the AREs (22) and (68) with the choices for R_c and R_o, respectively. In solving (22), we have the same requirements (80) and

$$4\mathcal{B}_N^{-1} A_N \mathcal{B}_N^{-T} < R_c^{-1}.$$

This requirement is exactly (36), which shows that we can directly apply the control design of section 4 with no changes. In solving (68) with (81), we must have

$$A < \frac{1}{4}\mathbf{H}_N^T R_o^{-1} \mathbf{H}_N, \qquad (83)$$

and, since A is positive definite, (83) with (82) implies that we need

$$\mathbf{H}_N A^{-1} \mathbf{H}_N^T > \Omega + \frac{1}{\gamma}\mathbf{L}_N^{-1} \mathcal{B}_N R_c^{-1} \mathcal{B}_N^T \mathbf{L}_N^{-T}. \qquad (84)$$

We also see that if (73) is satisfied, then we can find a $\gamma > 0$ such that (84) is satisfied. Therefore, we have shown the following corollary to Theorems 4 and 5.

COROLLARY 7. *If the conditions of Theorems* 4 *and* 5 *are satisfied, then the resulting controller and observer can be combined with no alterations, and the closed loop system with dynamic compensator* (74), (75) *is globally stable.*

7. Duality of control and observation spillover. In this section, we briefly discuss the apparent duality between control and observation spillover. In the construction of finite-dimensional controllers and observers, there are two obstacles in each of the problems. The first obstacle is the nonlinearity, and the second obstacle is spillover. We point out here that in [42], finite-dimensional compensators were constructed for linear systems with no spillover. We have constructed a compensator with finite-dimensional inputs and outputs with infinite-dimensional internal dynamics (the infinite-dimensional estimator) that globally stabilizes systems with linearly bounded nonlinearities with the possibility of spillover destabilization.

In both designs, we consider a Lyapunov function that is the squared $\bar{L}_2(0, 2\pi)$-norm of the state or error. The given Lyapunov functions provide the uniform decay of the Fourier modes. This allows the use of the linear bounds on $f(y)$, so the controller and observer can be constructed to stabilize the respective systems governed by the linear equations. In both cases, the gains were determined by solving the control version of the ARE. These gains were determined to stabilize the low modes of the respective Lyapunov derivatives, with robustness to spillover.

The control design can be stated as follows. Given a linear operator $B : \mathbb{R}^N \to \bar{L}_2(0, 2\pi)$, determine the vector $\mathcal{V} \in \mathbb{R}^N$, where $\mathcal{V} = \mathcal{V}(\mathbf{Y}_N)$, to stabilize the solution

$y(\theta) = 0$. The observer design can be stated as follows: Given a vector $Hy \in \mathbb{R}^N$, determine the linear operator $L : \mathbb{R}^N \to \bar{L}_2(0, 2\pi)$, which depends on the projection of the low modes of $y(\theta)$ onto H, that stabilizes the error solution $\hat{y}(\theta) = 0$.

The control $\mathcal{V} \in \mathbb{R}^N$ and the linear operator $B : \mathbb{R}^N \to \bar{L}_2(0, 2\pi)$, while the output $z \in \mathbb{R}^N$ and the linear operator $H : \bar{L}_2(0, 2\pi) \to \mathbb{R}^N$. The control spillover is given by the high modal content of $B\mathcal{V}$ for a given control $\mathcal{V}$. In our construction, $\mathcal{V} = \mathcal{V}(\mathbf{Y}_N)$, and hence the spillover was quantified by the high modal content of the influence functions $b^i(\theta)$. The observation spillover is given by the projection of the high modes of $y(\theta)$ onto the output operator H. The observer was constructed to stabilize the low modes of the error $\tilde{y}(\theta)$, and hence the observation spillover is given by the vectors $H\phi_n$ for $n > N$. These definitions of control and observation spillover are in accordance with [3]. In both cases, an upper bound on the possible destabilizing effects of spillover was quantified by summing over the high modal content of the Lyapunov derivatives. The ARE was used to determine lower bounds on the positive definite matrix R to cancel the effects of the spillover. Upper bounds on R were determined to ensure the existence of a positive definite matrix Q such that $\dot{V} \leq \mathbf{Y}_N^T Q \mathbf{Y}_N$. These upper bounds relate the unstable nature of the plant matrices A_N and A to the low modal content of the operators $\mathcal{B}_N$ and $\mathbf{H}_N$, respectively. The control and observer designs were combined by using the combined Lyapunov function $V(y, \tilde{y}) = \frac{1}{2}\|y\|_2^2 + \frac{\gamma}{2}\|\tilde{y}\|_2^2$. This discussion is summarized in Table 1.

TABLE 1
Duality between control and observation spillover analysis.

	Control	Observation
Lyapunov function	$V(y) = \frac{1}{2}\|y\|_2^2$	$V(\tilde{y}) = \frac{1}{2}\|\tilde{y}\|_2^2$
Linear bound	$\langle y, f(y)\rangle \leq C\|y\|_2^2$	$\langle \tilde{y}, f(y) - f(\hat{y})\rangle \leq C_L\|\tilde{y}\|_2^2$
Given	$B : \mathbb{R}^N \to \bar{L}_2(0, 2\pi)$	$H : \bar{L}_2(0, 2\pi) \to \mathbb{R}^N$
Design	$\mathcal{V} : \bar{L}_2(0, 2\pi) \to \mathbb{R}^N$	$L : \mathbb{R}^N \to \bar{L}_2(0, 2\pi)$
Gain	$K = R^{-1}\mathcal{B}_N^T P$	$\mathbf{L}_N^T = R^{-1}\mathbf{H}_N P$
Spillover matrix	$\Xi := \sum_{n>N} \frac{1}{c_n} b_n b_n^T$	$\Omega := \sum_{n>N} \frac{1}{c_n} h_n h_n^T$
R Lower bound	$R \geq \frac{1}{4}\Xi$	$R \geq \frac{1}{4}\Omega$
R Upper bound	$A_N < \frac{1}{4}\mathcal{B}_N R^{-1}\mathcal{B}_N^T$	$A < \frac{1}{4}\mathbf{H}_N^T R^{-1}\mathbf{H}_N$
Combined Lyapunov function	$V(y, \tilde{y}) = \frac{1}{2}\|y\|_2^2 + \frac{\gamma}{2}\|\tilde{y}\|_2^2$	

8. Application: Control of axial compressor stall with air injection actuation. In this section, we apply the control and observer designs with their respective spillover analyses to the problem of controlling a model describing rotating stall in axial compressors. We use a model of axial compressors with air injectors derived in [25, 23]. The model bears close resemblance to the familiar Moore–Greitzer model [35] with air injection actuation [8, 15]. A description of the physical setup of compression system control with air injection actuation and flow sensors is given in [27]. Continuous air injection was employed by Day [16] in one of the earliest studies of compressor stall control, where a finite number of air injectors were used. Further experimental investigations of stall control appear in [7, 15, 22]. These works cite a number of physical limitations involved with the implementation of stall control systems. In particular, limitations due to finite-dimensional implementation are

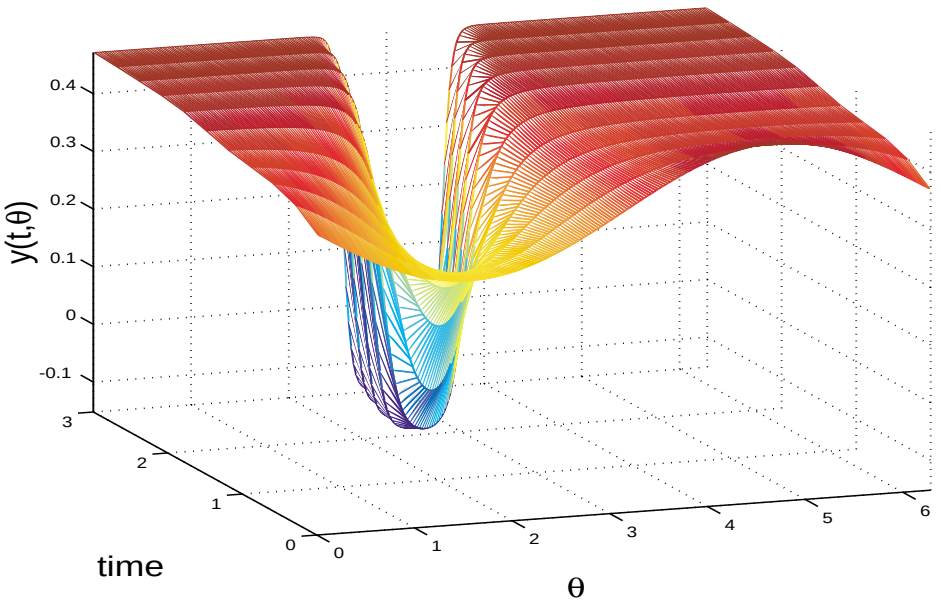


FIG. 1. *An initial sinusoidal disturbance develops into a stable piecewise constant solution, representing stalled flow.*

described in [37]. Other limiting factors such as actuator rate limits and noise are discussed in [18, 47, 46], and for further discussion of some of these issues, the reader is referred to [25]. In this work, we address only the issue of finite-dimensional implementation as it relates to the present theoretical results.

In this model, $y(t,\theta)$ denotes the axial flow through the compressor, and the constants appearing in (1) are

$$a^2 = 10^{-2}, \quad \beta = 0.5, \quad \alpha = 1.$$

The polynomial $f(y)$ is referred to as the compressor characteristic function (see [35]) and is given by

$$f(y) = 0.86 - 0.3y - 15.64y^2 - 27.43y^3, \tag{85}$$

where the numerical values are the same as those used in [5]. This system exhibits stable steady state (in a rotating reference frame) piecewise constant solutions which signify the stalling phenomena (see [25]). The evolution of an initial sinusoidal disturbance to stalled flow is shown in Figure 1. The goal is to control the axial flow such that the uniform flow $y(t,\theta) = M$ is stable. We take note of two things. First, since $\alpha = 1$, the first mode of the system remains constant, and hence the control design is directly applicable to this problem. Second, the polynomial given by (85) does not exhibit the property of $f(0) = 0$. However, if we look at our original system (1), we see that its average over θ is subtracted. Therefore, for practical purposes, we can add or subtract any constant to the polynomial, and the dynamics of the system would not change.

The linear bound (48) will be more restrictive than (8). In this model, the compressor characteristic satisfies the inequality (48) with $C_L = 2.7$. Taking $N = 16$

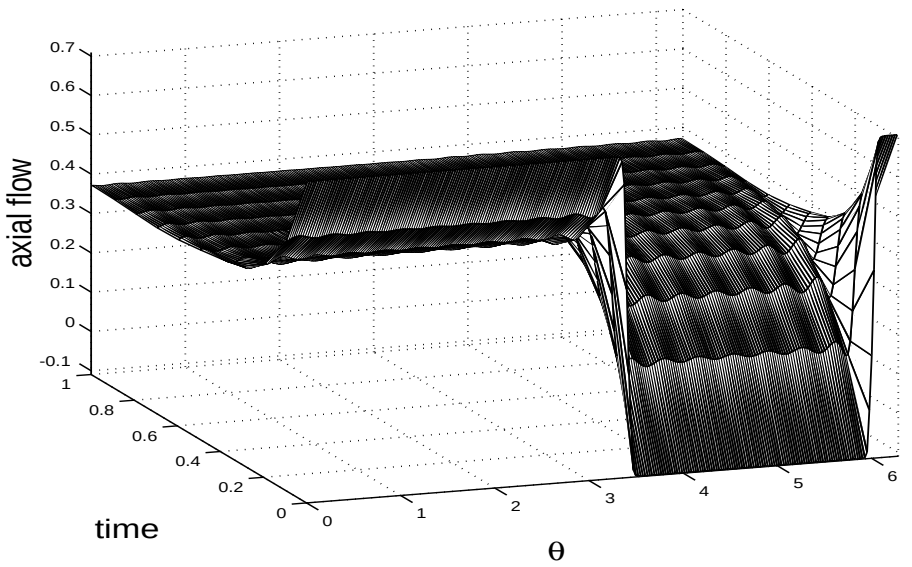


FIG. 2. *Stall cell decay for truncated LQR design. The initial condition is set to a piecewise constant function (stalled flow). The finite-dimensional controller constructed from the eigenmodes of the diffusion operator stabilizes the flow to a uniform value.*

satisfies $-a^2N^2 + C_L > 0 > -a^2(N+1)^2 + C_L$. We will first present some results of the controlled system with full state information to illustrate the role of the influence functions $b^i(\theta)$. Then we will present simulation results of the observer system in the absence of control. Finally, we will present results of the combined controller with the observer.

We present two cases with different input functions $b^i(\theta)$. For the first case, we simply take the first N eigenmodes of the diffusion operator as the input functions. In this trivial case, the high frequency content of the input functions is zero, i.e., $\Xi = 0$, and thus it is impossible for the control to destabilize the higher modes of the system. As stated before, we are not restricted to choosing P and R as in section 4. In [25], an infinite-dimensional LQR controller was constructed to stabilize the linear part of this system. It was shown that the sufficient conditions for this LQR control design to globally stabilize the system are equivalent to optimally stabilizing the linear system governed by the sector condition of the nonlinear function. This is exactly how we constructed the feedback controller in section 4, so we can simply truncate the LQR design derived in [25] after the 16th mode and be guaranteed global stability. The response of the system with this truncated controller is shown in Figure 2. The initial condition is set at a piecewise constant solution (stalled flow). We see that controller stabilizes the flow, eliminating the stall cell.

For the second case, we consider input shape functions with high frequency content. We consider an array of 16 injectors that are evenly spaced about the circumference of the compressor. Corresponding with this physical setup, the shape functions will be given by

$$b^i(\theta) = \begin{cases} 1, & \theta_i - \frac{\Delta}{2} < \theta < \theta_i + \frac{\Delta}{2}, \\ 0 & \text{otherwise}, \end{cases} \tag{86}$$

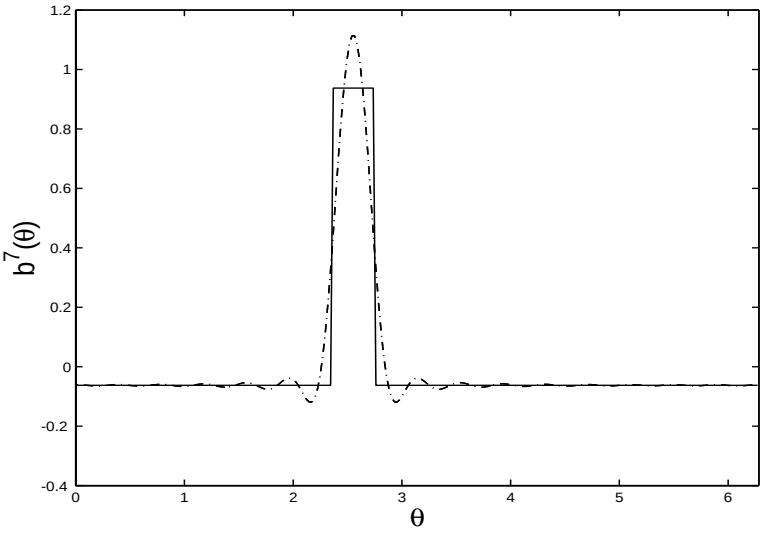


FIG. 3. *Shape function for* $b^7(\theta)$ *(shown as solid) and its reconstruction from its first* 16 *Fourier modes (shown as dotted).*

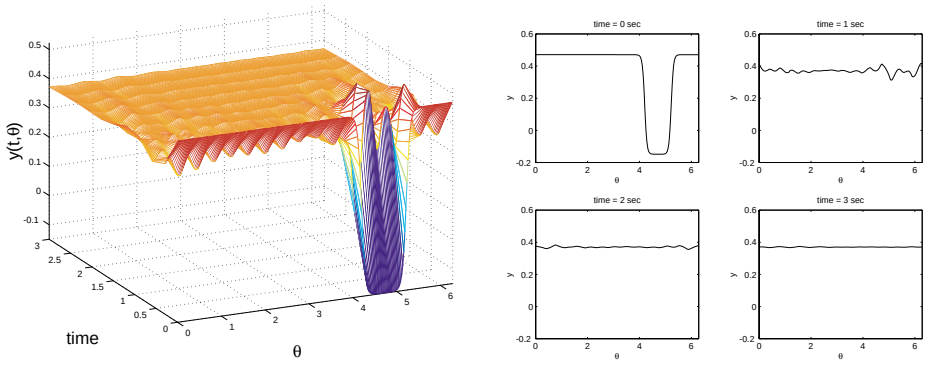


FIG. 4. *Left: Stall cell decay with controller constructed from* 16 *injectors spanning the circumference of the compressor. The initial condition is set to a piecewise constant function representing stalled flow. The control design eliminates control spillover from the high frequency content of the input shape functions, stabilizing the flow to a uniform value. Right: Stall cell decay of the left figure at different times.*

where the width of each injector is given by $\Delta \approx \frac{1}{18.25} 2\pi$. We point out here that the input shape functions given in (86) are not zero-average functions. In this case, it would be necessary to have a separate control to cancel out the zeroth mode of these shape functions. This is a common practice in the theoretical control of PDEs [1, 13], and for control of compression systems, this can be implemented by a throttle control [5]. Figure 3 shows a zero-average version of the input function for $i = 7$ along with its reconstruction from the first 16 Fourier modes. We can see that these shape functions have high frequency content. We apply the synthesis algorithm of section 4 for the linear control design. Figure 4 shows the closed loop system response with the linear controller. The initial condition is set at the open loop steady state piecewise constant solution (stalled flow). We see that the controller stabilizes the

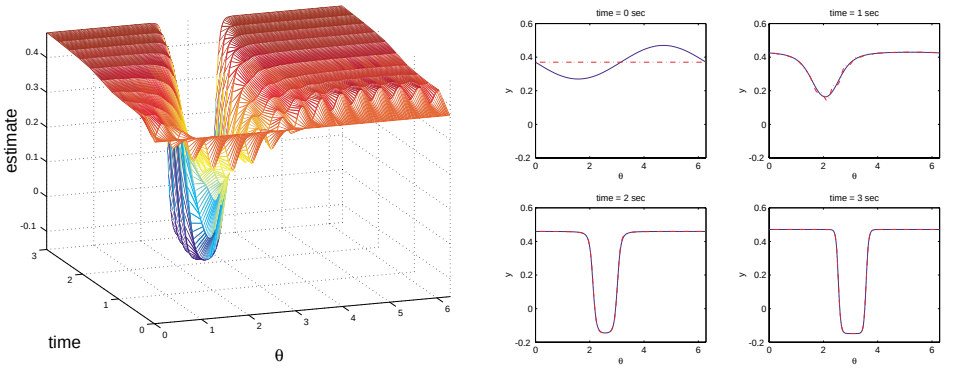


FIG. 5. *Left: Estimate evolution for the uncontrolled system. The estimate quickly matches the actual state shown in Figure* 1 *which is evolving into stall. Right: State (solid, from Figure* 1*) and estimate (dashed, from left figure) shown at different times. The estimate is very close to the state at time* = 3 *and* 4.

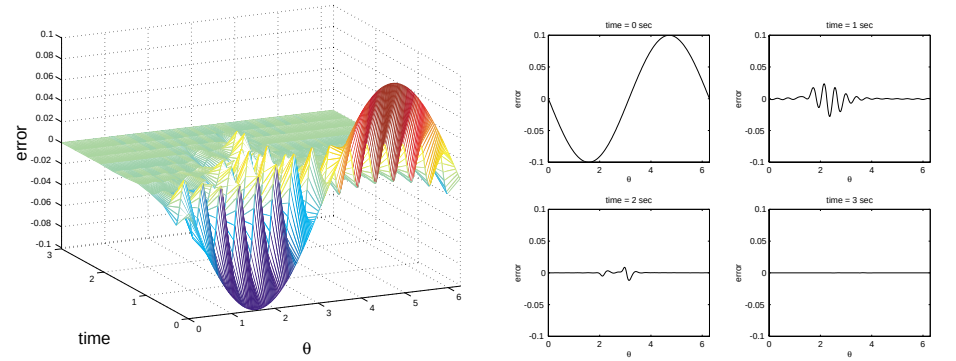


FIG. 6. *Left: Error evolution for the uncontrolled observer system. Right: Error evolution from the left figure shown at different times. The error decays to zero as the estimate matches the state.*

flow, eliminating the stall cell.

We now discuss the results of the observer without control. We consider the physical setup of an array of 16 flow sensors that span the circumference. Corresponding to the physical setup, each element of the output z will be the spatial average of $y(t,\theta)$ over a finite width Δ, and the output kernels $h^i(\theta)$ are given by

$$h^i(\theta) = \begin{cases} \frac{1}{\Delta}, & \theta_i - \frac{\Delta}{2} < \theta < \theta_i + \frac{\Delta}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Just as in the control design, we used the ARE (68) with $R > 0$ satisfying (71) and (72). The results of the numerical simulations are shown in Figure 1 (showing the uncontrolled state evolution) and Figures 5 and 6 (showing the estimation and error evolutions, respectively). The evolution of the estimate is shown in Figure 5 with the initial condition $\hat{y}(0,\theta) = 0$. We see that the estimate quickly matches the original state shown in Figure 1. This is also shown in Figure 6, which shows the evolution of the error $\tilde{y}(t,\theta)$. We see that the error quickly goes to zero uniformly in θ.

We now present simulation results for the system with finite-dimensional controls and observation. The B and H operators are the same as those given before. By the

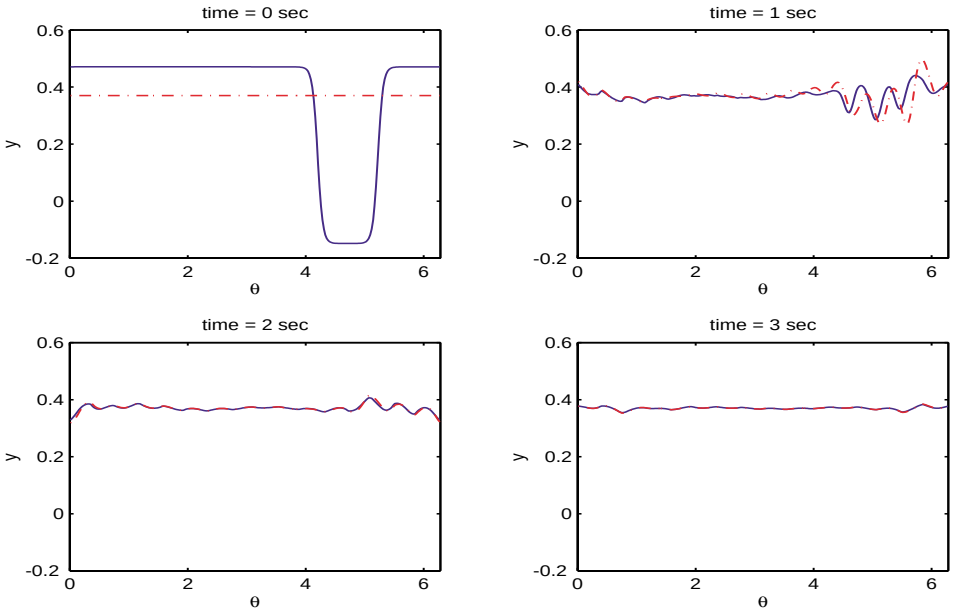


FIG. 7. *State and estimate evolution at different times. The state is shown as a solid line, and the estimate is shown as a dashed line. The estimate is very close to the state at time = 2 and 3 seconds.*

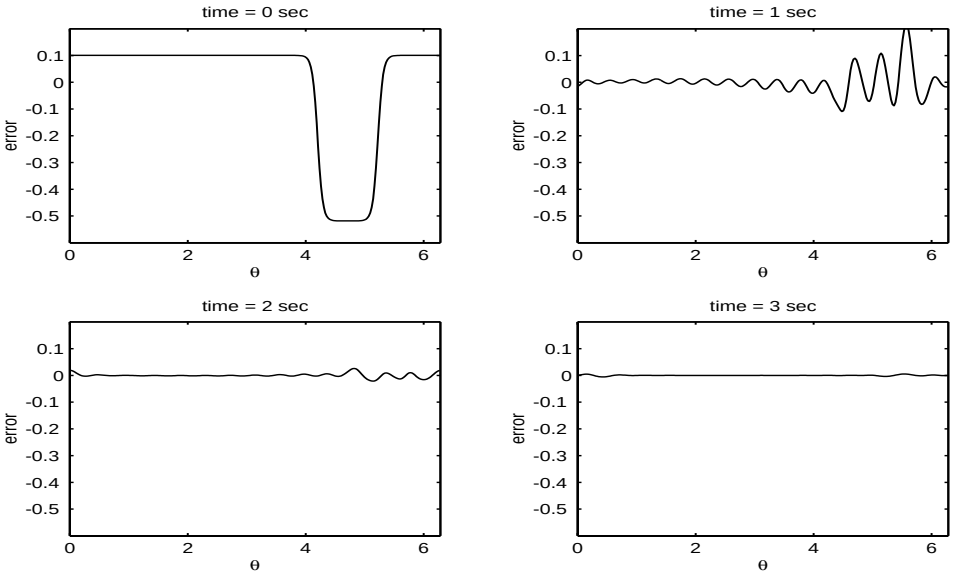


FIG. 8. *Error between state and estimate at different times.*

analysis given in section 6, we can directly combine the two designs. Figures 7 and 8 show $y(t,\theta)$, $\hat{y}(t,\theta)$, and $\tilde{y}(t,\theta)$ at different times. We see that stability of the state and error are maintained when the two designs are combined. The estimate $\hat{y}(t,\theta)$ converges to the state $y(t,\theta)$, and the controller stabilizes $y(t,\theta) = 0$.

9. Conclusion. We have considered the problem of controlling semilinear dissipative evolution equations with additive control input of finite dimension and the dual problem of observation with finite-dimensional output. The control input enters the equation through a finite number of shape functions and their respective amplitude values, while the output is given by a finite-dimensional vector. For the control design, we used a CLF based on the infinite-dimensional dynamics. The derivative of the CLF was decomposed into Fourier modes and was diagonalized by application of the linear bound of the nonlinearity. Since the system is coupled through the terms representing the control input, there is a possibility of destabilizing control spillover. The modal CLF derivative was split into low and high modes. An LQR control design technique was employed to stabilize the low modal content of the CLF derivative. An upper bound on the destabilizing terms representing control spillover was obtained. Sufficient conditions on the system and LQR design parameters were stated to ensure global stability and robustness with respect to control spillover. Unlike reduced order model control formulations, which employ Galerkin truncation, the control design of this paper is based on the global stabilization of the infinite-dimensional dynamics of the full system. A similar design and analysis were conducted for the finite-dimensional observation problem. It was shown that the closed loop system with the controller and observer was globally stable and furthermore that the control and observer designs could be combined with no changes in the separate designs. Duality properties between the control and observer designs and respective spillover analyses were discussed. The finite-dimensional controller and observer were demonstrated on a model describing rotating stall in axial compressors. The control methodology outlined in this paper is extended to decentralized and nonlinear controllers in [24].

REFERENCES

[1] A. Armaou and P. D. Christofides, *Global stabilization of the Kuramoto-Sivashinsky equation via distributed output feedback control*, Phys. D, 137 (2000), pp. 49–61.

[2] M. J. Balas, *Active control of flexible systems*, J. Optim. Theory Appl., 25 (1978), pp. 415–436.

[3] M. J. Balas, *Distributed parameter control of nonlinear flexible structures with linear finite-dimensional controllers*, J. Math. Anal. Appl., 108 (1985), pp. 528–545.

[4] B. Bamieh, F. Paganini, and M. A. Dahleh, *Distributed control of spatially-invariant systems*, IEEE Trans. Automat. Control, 47 (2002), pp. 1091–1107.

[5] A. Banaszuk, H. A. Hauksson, and I. Mezić, *A backstepping controller for a nonlinear partial differential equation model of compression system instabilities*, SIAM J. Control Optim., 37 (1999), pp. 1503–1537.

[6] S. P. Banks, *State-Space and Frequency-Domain Methods in the Control of Distributed Parameter Systems*, Peter Peregrinus, Ltd., London, 1983.

[7] R. L. Behnken, *Nonlinear Control and Modeling of Rotating Stall in an Axial Flow Compressor*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 1997.

[8] R. L. Behnken and R. M. Murray, *Combined air injection control of rotating stall and bleed valve control of surge*, in Proceedings of the American Control Conference, Albuquerque, NM, 1997, pp. 987–992.

[9] J. Bontsema and R. F. Curtain, *A note on spillover and robustness for flexible systems*, IEEE Trans. Automat. Control, 33 (1988), pp. 567–569.

[10] J. A. Burns and B. B. King, *A reduced basis approach to the design of low order feedback controllers for nonlinear continuous systems*, J. Vib. Control, 4 (1998), pp. 297–323.

[11] N. Chafee, *The electric balast resistor: Homogeneous and nonhomogeneous equilibria*, in Nonlinear Differential Equations, Academic Press, New York, 1981, pp. 97–127.

[12] P. D. Christofides, *Robust control of parabolic PDE systems*, Chemical Engineering Science, 57 (1998), pp. 2949–2965.

[13] P. D. Christofides and A. Armaou, *Feedback control of the Kuramoto-Sivashinsky equation*, Systems Control Lett., 39 (2000), pp. 283–294.

[14] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.
[15] R. D'Andrea, R. L. Behnken, and R. M. Murray, *Rotating stall control of an axial flow compressor using pulsed air injection*, Transactions ASME Journal of Turbomachinery, 119 (1997), pp. 742–752.
[16] I. J. Day, *Active suppression of rotating stall in axial compressors*, Transactions ASME Journal of Turbomachinery, 115 (1993), pp. 40–47.
[17] A. L. Faulds and B. B. King, *Sensor location for feedback control of partial differential equation systems*, in Proceedings of the IEEE Conference on Control Applications, IEEE Control Systems Society, Piscataway, NJ, 2000, pp. 536–541.
[18] C. Freeman, A. G. Wilson, I. J. Day, and M. A. Swinbanks, *Experiments in active control of stall on an aeroengine gas turbine*, Transactions ASME Journal of Turbomachinery (1997).
[19] J. Furter and M. Grinfeld, *Local vs. non-local interactions in population dynamics*, J. Math. Biol., 27 (1990), pp. 65–80.
[20] W. R. Graham, J. Peraire, and K. Y. Tang, *Optimal control of vortex shedding using low-order models.* I. *Open-loop model development*, Internat. J. Numer. Methods Engrg., 44 (1999), pp. 945–972.
[21] W. R. Graham, J. Peraire, and K. Y. Tang, *Optimal control of vortex shedding using low-order models.* II. *Model-based control*, Internat. J. Numer. Methods Engrg., 44 (1999), pp. 973–990.
[22] D. L. Gysling and E. M. Greitzer, *Dynamic control of rotating stall in axial compressors using aeromechanical feedback*, Transactions ASME Journal of Turbomachinery (1995), pp. 307–319.
[23] G. Hagen, *Large Scale Flow Phenomena in Axial Compressors: Modeling, Analysis, and Control with Air Injectors*, Ph.D. thesis, University of California, Santa Barbara, CA, 2001.
[24] G. Hagen, *Finite-dimensional decentralized and nonlinear control of semilinear dissipative evolution equations*, Systems Control Lett., submitted.
[25] G. Hagen, I. Mezić, and B. Bamieh, *Distributed control design for parabolic evolution equations; application to compressor stall control*, IEEE Trans. Automat. Control, to appear.
[26] Y. Harn and E. Polak, *On the design of finite-dimensional stabilizing compensators for infinite-dimensional feedback-systems via semiinfinite optimization*, IEEE Trans. Automat. Control, 35 (1990), pp. 1135–1140.
[27] G. J. Hendricks and D. L. Gysling, *Theoretical study of sensor-actuator schemes for rotating stall control*, J. Propulsion and Power, 10 (1994), pp. 101–109.
[28] P. Holmes, J. L. Lumley, and G. Berkooz, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge University Press, Cambridge, UK, 1998.
[29] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1996.
[30] G. Kepler, H. Tran, and H. Banks, *Compensator Control for Chemical Vapor Deposition Film Growth Using Reduced Order Design Models*, Report CRSC-TR99-41, North Carolina State University, Raleigh, NC, 1999; IEEE Trans. Semiconductor Manufacturing, 14 (2001), pp. 231–241.
[31] H. K. Khalil, *Nonlinear Systems*, 2nd ed., Prentice–Hall, Upper Saddle River, NJ, 1996.
[32] A. J. Laub, *Computational Matrix Analysis*, Course Notes for ECE 234, University of California, Santa Barbara, CA, 1995.
[33] M. Marion and R. Temam, *Nonlinear Galerkin methods*, SIAM J. Numer. Anal., 26 (1989), pp. 1139–1157.
[34] L. Meirovitch and H. Baruh, *On the problem of observation spillover in self-adjoint distributed parameter systems*, J. Optim. Theory Appl., 39 (1983), pp. 269–291.
[35] F. K. Moore and E. M. Greitzer, *A theory of post-stall transients in axial compression systems.* I. *Development of equations*, Transactions ASME Journal of Engineering for Gas Turbines and Power, 108 (1986), pp. 68–76.
[36] N. C. Owen and P. Sternberg, *Gradient flow and front propagation with boundary contact energy*, Proc. Roy. Soc. London Ser. A, 437 (1992), pp. 715–728.
[37] J. D. Paduano, E. M. Greitzer, and A. H. Epstein, *Compression system stability and active control*, in Annu. Rev. Fluid Mech. 33, Annual Reviews, Palo Alto, CA, 2001, pp. 491–517.
[38] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[39] A. Preumont, *Spillover alleviation for nonlinear active control of vibration*, J. Guidance Control Dynam., 11 (1988), pp. 124–130.

[40] J. Rubinstein and P. Sternberg, *Nonlocal reaction diffusion equations and nucleation*, IMA J. Appl. Math., 48 (1992), pp. 249–264.
[41] H. Sano and Y. Sakawa, *$\mathcal{H}_\infty$ control of diffusion systems by using a finite-dimensional controller*, SIAM J. Control Optim., 37 (1998), pp. 409–428.
[42] J. M. Schumacher, *Dynamic Feedback in Finite and Infinite-Dimensional Linear Systems*, Mathematisch Centrum, Amsterdam, 1981.
[43] S. Y. Shvartsman, C. Theodoropoulos, R. Rico-Martínez, I. G. Kevrekidis, E. S. Titi, and T. J. Mountziaris, *Order reduction for nonlinear dynamic models of distributed reacting systems*, J. Process Control, 10 (2000), pp. 177–184.
[44] R. F. Stengel, *Optimal Control and Estimation*, Dover, New York, 1994.
[45] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Springer-Verlag, New York, 1997.
[46] Y. Wang and R. M. Murray, *Effects of noise and actuator limits on active control of rotating stall and surge*, in Proceedings of the IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1997, pp. 4602–4607.
[47] S. Yeung and R. M. Murray, *Reduction of bleed valve rate requirements for control of rotating stall using continuous air injection*, in Proceedings of the IEEE International Conference on Control Applications, IEEE Control Systems Society, Piscataway, NJ, 1997, pp. 683–690.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 769–788

SOLVING SEMI-INFINITE OPTIMIZATION PROBLEMS WITH INTERIOR POINT TECHNIQUES*

OLIVER STEIN† AND GEORG STILL‡

Abstract. We introduce a new numerical solution method for semi-infinite optimization problems with convex lower level problems. The method is based on a reformulation of the semi-infinite problem as a Stackelberg game and the use of regularized nonlinear complementarity problem functions. This approach leads to central path conditions for the lower level problems, where for a given path parameter a smooth nonlinear finite optimization problem has to be solved. The solution of the semi-infinite optimization problem then amounts to driving the path parameter to zero.

We show convergence properties of the method and give a number of numerical examples from design centering and from robust optimization, where actually so-called generalized semi-infinite optimization problems are solved. The presented method is easy to implement, and in our examples it works well for dimensions of the semi-infinite index set at least up to 150.

Key words. generalized semi-infinite optimization, convexity, Stackelberg game, nonlinear complementarity problem function, smoothing, optimality condition

AMS subject classifications. 90C34, 90C25, 49M37, 65K05

DOI. 10.1137/S0363012901398393

1. Introduction.

In this article we introduce a bi-level solution method for so-called *generalized semi-infinite optimization problems*. These problems have the form

$$GSIP: \qquad \text{minimize} \quad f(x) \quad \text{subject to} \quad x \in M$$

with

$$M \;=\; \{x \in \mathbb{R}^n |\; g_j(x,y) \le 0,\; y \in Y(x),\; j \in J\}$$

and

$$Y(x) \;=\; \{y \in \mathbb{R}^m |\; v_\ell(x,y) \le 0,\; \ell \in L\}\,.$$

All defining functions $f, g_j\,,\; j \in J = \{1,\dots,p\},\; v_\ell\,,\; \ell \in L = \{1,\dots,s\}\,$, are assumed to be real-valued and d times continuously differentiable on their respective domains with $d \ge 2$.

As opposed to a standard semi-infinite optimization problem *SIP*, the possibly infinite index set $Y(x)$ of inequality constraints is x-dependent in a *GSIP*. For surveys about *standard* semi-infinite optimization we refer to [10, 18, 37].

Engineering applications that give rise to generalized semi-infinite optimization problems include robot design [12, 19], reverse Chebyshev approximation [29], time-optimal control [31], and design centering [35, 36]. In finite optimization with uncertainty about parameters y from a fixed set Y, the robust (i.e., worst-case) formulation of inequality constraints gives rise to a standard semi-infinite problem [1]. If the set

*Received by the editors November 19, 2001; accepted for publication (in revised form) January 22, 2003; published electronically June 18, 2003.
http://www.siam.org/journals/sicon/42-3/39839.html

†Department of Mathematics – C, Aachen University, Aachen, Germany (stein@mathC.rwth-aachen.de).

‡Faculty of Mathematical Sciences, University of Twente, Enschede, The Netherlands (g.still@math.utwente.nl).

of uncertain parameters is state-dependent, then the worst-case formulation takes the form of *GSIP*. Furthermore, min-max problems can be reformulated as either standard or generalized semi-infinite programs, depending on whether the feasible set of the maximization (inner) problem depends on the minimization (outer) variable.

The growing interest in *GSIP* over recent years has resulted in various contributions on the structure of the feasible set M [25, 41, 48, 49, 50, 57] and on first and second order optimality conditions [19, 25, 29, 40, 51, 53, 57]. The articles [55] and [57] investigate how the known methods from *SIP* have to be modified in order to cover the more general situation of *GSIP*. The Newton-SQP approach, which works well in standard semi-infinite programming (see, e.g., [13]), can be transferred to *GSIP* if the so-called reduction ansatz holds. In [16] such a Newton-type method is applied to the terminal variational problems from [29]. Since the reduction ansatz is of local nature, also generalizations of the discretization and exchange methods from *SIP* are desired. In [55] it is shown that discretization methods converge if the x-dependent grid points are chosen such that they depend continuously on x. Moreover, [56] studies how for a discretization method the rate of convergence depends on a consistent treatment of the boundary points of $Y(x)$.

As these generalized discretization methods are not easy to implement, in the present article we concentrate on the case of convex lower level problems. Based upon the observation that, under natural assumptions, *GSIP* can be reformulated as a special Stackelberg game (cf. also [54]), we design a numerical solution method, which exploits the lower level convexity. As opposed to the exchange and discretization methods presented in [55], this approach is not a generalization of known methods from standard semi-infinite programming, but it provides a new and different way of numerical treatment. Moreover, as standard semi-infinite programming is a special case of *GSIP*, as a by-product we obtain a new solution method for standard semi-infinite optimization problems, too. Our point of view also was implicitly taken in [33], where a branch and bound method is developed for generalized semi-infinite optimization problems with linear-quadratic lower level problems and additional convexity in the upper level. An approach using exact penalization to transform *GSIP* to *SIP* is given in [32].

We remark that the inclusion of equality constraints in the definitions of M and $Y(x)$, as well as a j-dependence of the index set $Y(x)$, is straightforward and will not be considered here for ease of presentation.

The article is organized as follows. In section 2 we recall the so-called reduction ansatz and the resulting first order optimality condition for *GSIP*. Furthermore, we give analogous results under a convexity assumption and recall the concept of nonlinear complementarity problem (NCP) functions. Section 3 presents the numerical method, and in section 4 we study its convergence properties. A number of numerical results conclude the article in section 5.

2. Preliminaries.

2.1. The reduction ansatz. In this section we briefly recall the reduction ansatz and explain the concept of Fritz John points for *GSIP*. Since optimality conditions are well known for points from the topological interior of M, in the following we focus our attention on a given feasible boundary point of M, i.e., a point $\bar{x} \in M \cap \partial M$, where ∂M denotes the topological boundary of M.

Recall that the set-valued mapping Y is called locally bounded around $\bar{x}$ if there exists a neighborhood U of $\bar{x}$ such that the set $\cup_{x \in U} Y(x)$ is bounded.

Assumption 1 (local boundedness). The set-valued mapping Y is locally bounded around $\bar{x}$.

Let Assumption 1 hold throughout this article, and fix U to be some corresponding bounded open neighborhood of $\bar{x}$.

The n-parametric so-called lower level problems of *GSIP* are given by

$$Q^j(x): \qquad \text{maximize} \quad g_j(x,y) \quad \text{subject to} \quad y \in Y(x)$$

with $j \in J$. Associated with $Q^j(x)$ are its optimal value function

$$\varphi_j(x) \;=\; \begin{cases} \max\limits_{y \in Y(x)} g_j(x,y) & \text{if} \quad Y(x) \neq \emptyset, \\ -\infty & \text{else} \end{cases}$$

and, in case of solvability, its solution set

$$Y_\star^j(x) \;=\; \{y \in Y(x) |\; g_j(x,y) = \varphi_j(x)\}.$$

It is easily seen that M and the set $\{x \in \mathbb{R}^n |\; \varphi_j(x) \leq 0\,,\; j \in J\}$ coincide.

Since the defining functions of $Y(x)$ are continuous, the set-valued mapping Y is closed. Together with Assumption 1 this means that Y is upper semicontinuous in the sense of Berge [2] around $\bar{x}$. As a consequence (cf., e.g., [21]), the $\varphi_j\,,\; j \in J$, are upper semicontinuous on U. From this it is not hard to derive that the sets $M \cap U$ and $\{x \in U |\; \varphi_j(x) \leq 0\,,\; j \in J_0(\bar{x})\}$ coincide (possibly after shrinking U), where $J_0(\bar{x}) = \{j \in J |\; \varphi_j(\bar{x}) = 0\}$ denotes the set of active indices at $\bar{x}$. Since for $j \in J_0(\bar{x})$ the problem $Q^j(\bar{x})$ has vanishing optimal value, the set of its solution points can be described as

$$Y_\star^j(\bar{x}) \;=\; Y_0^j(\bar{x}) \;=\; \{y \in Y(\bar{x}) |\; g_j(\bar{x},y) = 0\},$$

and we have $j \in J_0(\bar{x})$ if and only if $Y_0^j(\bar{x}) \neq \emptyset$.

Next, we give a local description of M by finitely many *smooth* constraints for the case when certain regularity assumptions hold in the lower level problems.

The *linear independence constraint qualification (LICQ)* is said to hold at a point $\bar{y} \in Y(\bar{x})$ if the family of vectors $D_y v_\ell(\bar{x},\bar{y})\,,\; \ell \in L_0(\bar{x},\bar{y})$, is linearly independent, and the weaker *Mangasarian–Fromovitz constraint qualification (MFCQ)* holds if there exists some vector $\eta \in \mathbb{R}^m$ such that $D_y v_\ell(\bar{x},\bar{y})\,\eta < 0\,,\; \ell \in L_0(\bar{x},\bar{y})$. Here $L_0(\bar{x},\bar{y}) = \{\ell \in L |\; v_\ell(\bar{x},\bar{y}) = 0\}$ denotes the set of lower level active indices, and $D_y v_\ell$ stands for the row vector of partial derivatives of v_ℓ with respect to y.

In what follows, let v be the column vector of the functions $v_\ell\,,\; \ell \in L = \{1,\ldots,s\}$, let $\operatorname{diag}(\gamma)$ stand for the (s,s)-diagonal matrix with diagonal vector $\gamma \in \mathbb{R}^s$, and let $j \in J_0(\bar{x})$. Since each $\bar{y} \in Y_0^j(\bar{x})$ is a solution of $Q^j(\bar{x})$, upon definition of the lower level Lagrange function

$$\mathcal{L}_j(x,y,\gamma) \;=\; g_j(x,y) - \gamma^\top v(x,y)$$

the Karush–Kuhn–Tucker theorem states that the following system of equalities and inequalities has a solution γ if the MFCQ holds at such a $\bar{y}$:

$$D_y^\top \mathcal{L}_j(\bar{x},\bar{y},\gamma) = 0, \tag{2.1}$$

$$-\operatorname{diag}(\gamma)\, v(\bar{x},\bar{y}) = 0, \tag{2.2}$$

$$\gamma \geq 0, \tag{2.3}$$

$$-v(\bar{x},\bar{y}) \geq 0. \tag{2.4}$$

Note that γ is uniquely determined under the LICQ. We denote the set of Kuhn–Tucker multipliers corresponding to $\bar{x}$ and $\bar{y} \in Y_0^j(\bar{x})$ by

$$KT^j(\bar{x}, \bar{y}) \;=\; \{\gamma \in \mathbb{R}^s |\ \gamma \text{ satisfies (2.1)–(2.4)}\}.$$

The point $\bar{y}$ is said to satisfy the *strict complementary slackness (SCS)* condition if $\gamma_\ell > 0\,,\ \ell \in L_0(\bar{x}, \bar{y})$.

Under the LICQ the tangent space to $Y(\bar{x})$ at $\bar{y}$ can be described as $T_{\bar{y}}Y(\bar{x}) = \{\eta \in \mathbb{R}^m |\ D_y v_\ell(\bar{x}, \bar{y})\,\eta = 0\,,\ \ell \in L_0(\bar{x}, \bar{y})\}$. Let $\bar{y} \in Y_0^j(\bar{x})$ and let $\bar{\gamma}$ be the corresponding solution of (2.1)–(2.4). The point $\bar{y}$ is said to satisfy the *second order sufficiency condition (SOSC)* if the matrix $D_y^2 \mathcal{L}_j(\bar{x}, \bar{y}, \bar{\gamma})|_{T_{\bar{y}}Y(\bar{x})}$ possesses only negative eigenvalues. Here, $D_y^2 \mathcal{L}_j = D_y D_y^\top \mathcal{L}_j$ denotes the Hessian matrix of $\mathcal{L}_j$ with respect to y, and $D_y^2 \mathcal{L}_j(\bar{x}, \bar{y}, \bar{\gamma})|_{T_{\bar{y}}Y(\bar{x})} = V^\top D_y^2 \mathcal{L}_j(\bar{x}, \bar{y}, \bar{\gamma}) V$ for any matrix V of m-vectors which form a basis of the tangent space $T_{\bar{y}}Y(\bar{x})$.

Definition 2.1. *Let $\bar{x} \in \partial M \cap M$ and $j \in J_0(\bar{x})$. A point $\bar{y} \in Y_0^j(\bar{x})$ is called the* nondegenerate global maximizer *of $Q^j(\bar{x})$ if the LICQ holds at $\bar{y}$ and if SCS and the SOSC are valid with the vector γ from* (2.1)–(2.4).

Assumption 2 (reduction ansatz). For each $j \in J_0(\bar{x})$ all global maximizers of $Q^j(\bar{x})$ are nondegenerate.

The reduction ansatz was originally formulated for standard semi-infinite optimization problems in [58, 17] under weaker regularity assumptions and was transferred to generalized semi-infinite optimization problems in [19]. For standard semi-infinite optimization problems the reduction ansatz is a natural assumption in the sense that for problems with defining functions in general position *it holds at each local minimizer* (cf. [59, 47]). For *GSIP* this result could be transferred to local minimizers $\bar{x}$ with $\sum_{j \in J_0(\bar{x})} |Y_0^j(\bar{x})| \geq n$ in [49]. Moreover, in [54] it is shown that it holds in the "completely linear" case, i.e., when all defining functions f, $g_j\,,\ j \in J$, $v_\ell\,,\ \ell \in L$, of *GSIP* are affine linear on their respective domains.

As under Assumption 2 the global maximizers of $Q^j(\bar{x})$ are isolated points in $Y(\bar{x})$, and the latter set is compact, there are only finitely many global maximizers, say

$$Y_0^j(\bar{x}) \;=\; \{\bar{y}^{j,k},\ k \in J_0^j(\bar{x})\}$$

with $|J_0^j(\bar{x})| < \infty$. An application of the implicit function theorem (cf. [7]) shows that for each $\bar{y}^{j,k}$ with $k \in J_0^j(\bar{x})$ and corresponding multiplier vector $\bar{\gamma}^{j,k}$ there are locally defined C^{d-1}-functions $y^{j,k}$ and $\gamma^{j,k}$ with $y^{j,k}(\bar{x}) = \bar{y}^{j,k}$ and $\gamma^{j,k}(\bar{x}) = \bar{\gamma}^{j,k}$ such that $y^{j,k}(x)$ is the locally unique local maximizer of $Q^j(x)$ with multiplier $\gamma^{j,k}(x)$. Hence, we may introduce the locally defined optimal value functions

$$\varphi_{j,k}(x) \;=\; g_j(x, y^{j,k}(x))\,, \quad k \in J_0^j(\bar{x})\,,\ j \in J_0(\bar{x}).$$

Lemma 2.2 (cf., e.g., [24]). *The functions $\varphi_{j,k}$ are of differentiability class C^d, and their gradients satisfy*

$$D\varphi_{j,k}(\bar{x}) = D_x \mathcal{L}_j(\bar{x}, \bar{y}^{j,k}, \bar{\gamma}^{j,k})\,.$$

The next result follows from a more general reduction lemma that we shall prove in section 4 (cf. Lemma 4.2).

Theorem 2.3 (reduction lemma; cf. [19, 47]). *Let Assumption* 2 *be satisfied at $\bar{x}$. Then the sets M and*

$$M_{\bar{x}} \;=\; \{x \in U |\ \varphi_{j,k}(x) \leq 0,\ k \in J_0^j(\bar{x})\,,\ j \in J_0(\bar{x})\}$$

coincide locally around $\bar{x}$.

Theorem 2.3 shows that under the reduction ansatz the original problem *GSIP* is locally equivalent to the reduced problem $\min f|_{M_{\bar{x}}}$. Hence, local optimality conditions from finite optimization may be applied to yield results for the semi-infinite case. In particular, we obtain a Fritz John–type first order necessary optimality condition (cf. [23]).

THEOREM 2.4. *Let* $\bar{x}$ *be a local minimizer of GSIP, and let Assumption* 2 *hold. Then there exist multipliers* $\kappa \geq 0,\ \lambda_{j,k} \geq 0,\ k \in J_0^j(\bar{x})\,,\ j \in J_0(\bar{x})$*, not all vanishing, such that*

$$\kappa\, Df(\bar{x}) \ + \sum_{j \in J_0(\bar{x})} \sum_{k \in J_0^j(\bar{x})} \lambda_{j,k}\, D_x \mathcal{L}_j(\bar{x}, \bar{y}^{j,k}, \bar{\gamma}^{j,k}) \ = \ 0. \tag{2.5}$$

Note that in Theorem 2.4 we do not need to impose a complementarity condition since all appearing constraints are active by definition. Moreover, recall that due to Carathéodory's theorem at most $n+1$ nonvanishing multipliers $\lambda_{j,k}$ are required in (2.5). In what follows we will call each point $\bar{x}$ that satisfies the reduction ansatz and the necessary optimality condition from Theorem 2.4 a *Fritz John point* for *GSIP*.

2.2. Convex lower level problems. We call a problem $Q^j(x),\ j \in J$, convex if the functions $-g_j(x,\cdot)\,,\ v_\ell(x,\cdot),\ \ell \in L$, are convex on $\mathbb{R}^m$. The main assumption of the present article is the following.

Assumption 3. The lower level problems $Q^j(x),\ j \in J$, are convex for all $x \in \mathbb{R}^n$.

Under Assumption 3 a set $Y(x)$ with $x \in \mathbb{R}^n$ is said to satisfy the Slater condition if there exists $y^\star$ such that $v_\ell(x, y^\star) < 0$ for all $\ell \in L$.

Assumption 4. The sets $Y(x)$ are bounded and satisfy the Slater condition for all $x \in \mathbb{R}^n$.

Under Assumptions 3 and 4 the sets $Y_\star^j(x)$ are nonempty and locally bounded around each $\bar{x} \in \mathbb{R}^n$ (cf. [22, Lemma 2]), so that the optimal value functions $\varphi_j(x) = \max_{y \in Y(x)} g_j(x,y)\,,\ j \in J,$ are well defined and continuous on $\mathbb{R}^n$ [22]. Hence, the feasible set M is closed, and as in section 2.1 we have that M and the set $\{x \in \mathbb{R}^n |\ \varphi_j(x) \leq 0\,,\ j \in J_0(\bar{x})\}$ coincide around each boundary point $\bar{x}$.

For the next theorem, which is a slight generalization of [42, Theorem 4.2], recall that the sets $KT^j(\bar{x}),\ j \in J_0(\bar{x})$, do not depend on the variable y in the convex case (cf., e.g., [11]). Without loss of generality we set $J_0(\bar{x}) = \{1, \ldots, p_0\}$.

THEOREM 2.5. *Let* $\bar{x}$ *be a local minimizer of GSIP and let Assumptions* 3 *and* 4 *be satisfied. Then for each selection* $(\bar{\gamma}^1, \ldots, \bar{\gamma}^{p_0}) \in KT^1(\bar{x}) \times \cdots \times KT^{p_0}(\bar{x})$ *there exist* $y^{j,k} \in Y_0^j(\bar{x}),\ k = 1, \ldots, p_j\,,\ j \in J_0(\bar{x}),\ \sum_{j \in J_0(\bar{x})} p_j \leq n+1$*, and multipliers* $\kappa \geq 0\,,\ \lambda_{j,k} \geq 0\,,$ *not all equal to zero, such that*

$$\kappa\, Df(\bar{x}) \ + \sum_{j \in J_0(\bar{x})} \sum_{k=1}^{p_j} \lambda_{j,k}\, D_x \mathcal{L}^j(\bar{x}, y^{j,k}, \bar{\gamma}^j) \ = \ 0\,.$$

If, in addition to Assumptions 3 and 4, the reduction ansatz (Assumption 2) also holds, then the sets $KT^j(\bar{x}) = \{\bar{\gamma}^j\}$ and $Y_0^j(\bar{x}) = \{\bar{y}^j\}$, $j \in J_0(\bar{x})$, are singletons. In this case, Theorems 2.4 and 2.5 obviously simplify to the following result.

COROLLARY 2.6. *Let* $\bar{x}$ *be a local minimizer of GSIP and let Assumptions* 2, 3, *and* 4 *be satisfied. Then there exist multipliers* $\kappa \geq 0\,,\ \lambda_j \geq 0\,,\ j \in J_0(\bar{x})$*, not all equal to zero, such that*

$$\kappa\, Df(\bar{x}) \ + \sum_{j \in J_0(\bar{x})} \lambda_j\, D_x \mathcal{L}^j(\bar{x}, \bar{y}^j, \bar{\gamma}^j) \ = \ 0\,.$$

The following lemma is well known. A short proof can be found in [52].

LEMMA 2.7. *Let Assumptions* 3 *and* 4 *be satisfied. Then a point* $\bar{y}^j$ *is a nondegenerate global maximizer of* $Q^j(\bar{x})$ *with corresponding multiplier vector* $\bar{\gamma}^j$ *if and only if* (2.1)–(2.4) *hold and if the Jacobian of* (2.1), (2.2) *with respect to* (y^j, γ^j),

$$A^j \;=\; A^j(\bar{x}, \bar{y}^j, \bar{\gamma}^j) \;=\; \begin{pmatrix} D_y^2 \mathcal{L}_j(\bar{x}, \bar{y}^j, \bar{\gamma}^j) & -D_y^\top v(\bar{x}, \bar{y}^j) \\ -\operatorname{diag}(\bar{\gamma}^j) D_y v(\bar{x}, \bar{y}^j) & -\operatorname{diag}(v(\bar{x}, \bar{y}^j)) \end{pmatrix},$$

is nonsingular.

2.3. NCP functions. A function $\psi : \mathbb{R}^2 \to \mathbb{R}$ with

$$\psi(a,b) \;=\; 0 \quad \text{if and only if} \quad a \geq 0\,,\; b \geq 0\,,\; ab \;=\; 0$$

is called an *NCP function*. Let us remark that the existence of a C^∞-NCP function is clear from a theorem by Whitney [3]. However, as smooth NCP functions are degenerate at the origin, in the following we will work with the nonsmooth NCP functions

$$\psi^{NR}(a,b) \;=\; \frac{1}{2}\left(a + b - \sqrt{(a-b)^2}\right)$$

and

$$\psi^{FB}(a,b) \;=\; a + b - \sqrt{a^2 + b^2}\;.$$

The function ψ^{NR} is the so-called natural residual or min-function since it coincides with $\min(a,b)$, and ψ^{FB} is known as the Fischer–Burmeister function [8].

For numerical purposes one can regularize these nondifferentiable NCP functions. The so-called Chen–Harker–Kanzow–Smale function [4, 27, 46] is given by

$$\psi_\tau^{NR}(a,b) \;=\; \frac{1}{2}\left(a + b - \sqrt{(a-b)^2 + 4\tau^2}\right),$$

whereas the so-called smoothed Fischer–Burmeister function is

$$\psi_\tau^{FB}(a,b) \;=\; a + b - \sqrt{a^2 + b^2 + 2\tau^2}\;.$$

Obviously, ψ_τ^{NR} and ψ_τ^{FB} are continuously differentiable for all $\tau \neq 0$, and for $\tau = 0$ they coincide with ψ^{NR} and ψ^{FB}, respectively. Moreover, both functions share the following important properties.

LEMMA 2.8. *Let* $\tau \neq 0$, *and let* ψ_τ *denote one of the functions* ψ_τ^{NR} *and* ψ_τ^{FB}. *Then the following assertions hold:*

(i) *We have* $\psi_\tau(a,b) \;=\; 0$ *if and only if* $a > 0\,,\; b > 0\,,\; ab \;=\; \tau^2$.

(ii) *For a zero* (a,b) *of* ψ_τ *the gradient* $D\psi_\tau(a,b)$ *does not explicitly depend on* τ *and is given by* $(a+b)^{-1}\,(b,a)$.

Proof. Part (i) was observed in [27], and part (ii) is easily verified. □

In what follows we mainly need the results of Lemma 2.8, so we will not distinguish between ψ_τ^{NR} and ψ_τ^{FB} but simply write ψ_τ.

3. The numerical approach. The aim of our numerical method is to replace *GSIP* by a sequence of finite nonlinear programming problems which are numerically tractable and whose solutions or stationary points converge to a solution or a stationary point of *GSIP*, respectively. Unlike other numerical methods for semi-infinite

programming, our approach does not discretize the index set $Y(x)$, but we take advantage of the fact that the solution set of a regular convex lower level problem is characterized by its first order optimality condition. Thus, let Assumptions 3 and 4 hold throughout this section.

In a first step we reformulate *GSIP* as a special Stackelberg game:

$$SG: \quad \min_{x,y^1,\ldots,y^p} f(x) \text{ subject to (s.t.) } g_j(x,y^j) \le 0\,, \text{ and } y^j \text{ solves } Q^j(x),\ j \in J.$$

Note that the *SG* possesses two special features: its objective function f does not depend on the variables $y^j,\ j \in J$, and its upper level inequality constraint functions coincide with its lower level objective functions. In [54] it is shown that *GSIP* and *SG* are equivalent problems whenever the index set $Y(x)$ is nonempty for all $x \in \mathbb{R}^n$. The latter is the case under Assumption 4. We point out that for $Y(x) = \emptyset$ the point x would be feasible for *GSIP* but infeasible for *SG*.

Next, since the problems $Q^j(x)$ are convex, we may replace the restrictions "y^j solves $Q^j(x)$" in *SG* equivalently by their first order optimality conditions: for each $j \in J$ there is a solution γ^j of (2.1)–(2.4). The latter statement is true under Assumption 4, since Slater's condition guarantees the existence of Kuhn–Tucker multipliers. However, unlike in the case of the LICQ, these multipliers are not necessarily uniquely determined. By this reformulation, *SG* is equivalent to the following mathematical programming problem with equilibrium constraints:

$$\begin{aligned} MPEC: \quad \min_{x,y^1,\gamma^1,\ldots,y^p,\gamma^p} f(x) \text{ s.t. } \quad & g_j(x,y^j) && \le && 0, \\ & D_y^\top \mathcal{L}_j(x,y^j,\gamma^j) && = && 0, \\ & -\operatorname{diag}(\gamma^j)\, v(x,y^j) && = && 0, \\ & \gamma^j && \ge && 0, \\ & -v(x,y^j) && \ge && 0,\ j \in J. \end{aligned}$$

At this point, *GSIP* has been replaced by an equivalent finite nonlinear programming problem. However, numerical standard software cannot be expected to solve this problem since due to the appearance of complementarity conditions the MFCQ is violated at all points of the feasible set of *MPEC* (cf. [43]). In [26, 38] it is shown that the MFCQ is a necessary condition for the stability of smooth nonlinear programs under data perturbations and thus for the stability of numerical methods in the presence of round-off errors.

Given an NCP function ψ and $a, b \in \mathbb{R}^s$ we define the vectorization

$$\Psi(a,b) \;=\; (\psi(a_1\,,\,b_1), \ldots, \psi(a_s\,,\,b_s))^\top$$

so that *MPEC* can be equivalently rewritten as

$$\begin{aligned} P: \quad \min_{x,y^1,\gamma^1,\ldots,y^p,\gamma^p} f(x) \text{ s.t. } \quad & g_j(x,y^j) && \le && 0, \\ & D_y^\top \mathcal{L}_j(x,y^j,\gamma^j) && = && 0, \\ & \Psi(\gamma^j, -v(x,y^j)) && = && 0,\ j \in J. \end{aligned}$$

We now apply an interior point approach to the lower level problems $Q^j(x)$. For $j \in J$ we replace the Karush–Kuhn–Tucker system (2.1)–(2.4) at y^j and its corresponding multiplier vector γ^j by the perturbed system

$$D_y^\top \mathcal{L}_j(x,y^j,\gamma^j) = 0, \tag{3.1}$$

$$-\operatorname{diag}(\gamma^j)\, v(x,y^j) = \tau^2\, e_s, \tag{3.2}$$

$$\gamma^j \geq 0, \tag{3.3}$$

$$-v(x,y^j) \geq 0, \tag{3.4}$$

depending on $\tau \in \mathbb{R}$ (and on x). Here we set $e_s = (1,\ldots,1)^\top \in \mathbb{R}^s$. With one of the regularized NCP functions Ψ_τ in vector form, P is thus embedded into the parameterized family of optimization problems

$$\left\{ \begin{array}{lll} P_\tau : & \min\limits_{x,y^1,\gamma^1,\ldots,y^p,\gamma^p} f(x) \ \text{ s.t. } & \begin{array}{rcl} g_j(x,y^j) & \leq & 0, \\ D_y^\top \mathcal{L}_j(x,y^j,\gamma^j) & = & 0, \\ \Psi_\tau(\gamma^j, -v(x,y^j)) & = & 0,\ j \in J, \end{array} \end{array} \right. \tag{3.5}$$

with $P_0 = P$. We note that a similar approach for the solution of *MPEC*s is presented in [6].

The following proposition shows that problem P_τ is numerically tractable in the sense that the inherent singularity in the equality constraints of problem P has now been removed. Its proof follows straightforwardly from Lemma 2.8(ii). We remark that more detailed proofs of this and the following results can be found in [52].

PROPOSITION 3.1. *Let $\tau \neq 0$ and let $(x,y^1,\gamma^1,\ldots,y^p,\gamma^p)$ be a feasible point of P_τ such that for each $j \in J$ the matrix*

$$A^j \;=\; A^j(x,y^j,\gamma^j) \;=\; \begin{pmatrix} D_y^2 \mathcal{L}_j(x,y^j,\gamma^j) & -D_y^\top v(x,y^j) \\ -\operatorname{diag}(\gamma^j) D_y v(x,y^j) & -\operatorname{diag}(v(x,y^j)) \end{pmatrix}$$

is nonsingular. Then the gradients of the equality constraints of P_τ are linearly independent in $(x,y^1,\gamma^1,\ldots,y^p,\gamma^p)$.

We now recall the connection of the perturbed Karush–Kuhn–Tucker systems (3.1)–(3.4) with the barrier problems

$$Q_\tau^j(x): \qquad \max_y \; b_\tau^j(x,y) \;:=\; g_j(x,y) + \tau^2 \sum_{\ell \in L} \ln(-v_\ell(x,y))$$

(depending on x and τ) for $j \in J$. A necessary and in the convex case also sufficient optimality condition for $Q_\tau^j(x)$ is

$$0 \;=\; D_y b_\tau^j(x,y) \;=\; D_y g_j(x,y) + \sum_{\ell \in L} \frac{\tau^2}{v_\ell(x,y)} D_y v_\ell(x,y)\ .$$

Furthermore, the Hessian of $b_\tau^j(x,y)$ with respect to y reads as

$$\begin{aligned} D_y^2 b_\tau^j(x,y) = D_y^2 g_j(x,y) &+ \sum_{\ell \in L} \frac{\tau^2}{v_\ell(x,y)} D_y^2 v_\ell(x,y) \\ &- \sum_{\ell \in L} \frac{\tau^2}{[v_\ell(x,y)]^2} D_y^\top v_\ell(x,y) D_y v_\ell(x,y)\ . \end{aligned}$$

LEMMA 3.2. *Let $j \in J$ and $\tau \neq 0$.*

(i) *The point y^j is a solution of $Q^j_\tau(x)$ if and only if (y^j, γ^j) with $\gamma^j_\ell = -\tau^2/v_\ell(x, y^j)$, $\ell \in L$, is a solution of (3.1)–(3.4). Moreover, for the latter solutions $D^2_y b^j_\tau(x, y^j)$ is nonsingular if and only if $A^j(x, y^j, \gamma^j)$ is nonsingular.*

(ii) *If at least one of the matrices $D^2_y g_j(x, y^j)$, $D^2_y v_\ell(x, y^j)$, $\ell \in L$, is nonsingular, then $A^j(x, y^j, \gamma^j)$ is nonsingular, too.*

Proof. The first part of (i) is evident from a comparison between the relations $D_y \mathcal{L}_j(x, y^j, \gamma^j) = 0$ and $D_y b^j_\tau(x, y^j) = 0$. For the second part note that in view of Lemma 2.8(i) the matrix $\operatorname{diag}(v(x, y^j))$ is nonsingular so that $A^j = A^j(x, y^j, \gamma^j)$ is nonsingular if and only if the Schur complement S^j of $\operatorname{diag}(v(x, y^j))$ in A^j is nonsingular. This Schur complement is just $S^j = D^2_y b^j_\tau(x, y^j)$.

Due to our convexity assumptions (Assumption 3) it is easily seen that S^j is the sum of negative semidefinite matrices. Under the assumption of part (ii), at least one of the matrices $D^2_y g_j(x, y^j)$, $-D^2_y v_\ell(x, y^j)$, $\ell \in L$, is actually negative definite and, since all numbers $\tau^2/v_\ell(x, y^j)$, $\ell \in L$, are negative, S^j is negative definite, too. Together with part (i) this shows the assertion of part (ii). ∎

A different proof for a weaker result related to part (ii) of the preceding lemma can be found in [28]. Let us point out that Lemma 3.2 provides sufficient conditions for the assumption of nonsingular matrices $A^j(x, y^j, \gamma^j)$, $j \in J$, in Proposition 3.1. From Lemma 2.7 and a simple continuity argument it follows that $A^j(x, y^j, \gamma^j)$ is also nonsingular if (x, y^j, γ^j) is sufficiently close to a point $(\bar{x}, \bar{y}^j, \bar{\gamma}^j)$ such that $\bar{y}^j$ is a nondegenerate solution of $Q^j(\bar{x})$ with corresponding multiplier vector $\bar{\gamma}^j$.

The ideas presented so far lead to a simple continuation method for the numerical solution of *GSIP* which is easy to implement and can be given conceptually in the following form.

Numerical method.

Step 1. Choose a sequence $\{\tau_\nu\}$ of nonzero reals with $\lim_{\nu\to\infty} \tau_\nu = 0$ and a starting point $x^0 \in \mathbb{R}^n$.

Step 2. Compute a starting point $(x^{0,0}, y^{1,0,0}, \gamma^{1,0,0}, \ldots, y^{p,0,0}, \gamma^{p,0,0})$ of P_{τ_0} and set $\nu = 0$.

Step 3. Find a solution $(x^{\nu,\star}, y^{1,\nu,\star}, \ldots, \gamma^{p,\nu,\star})$ of P_{τ_ν}.

Step 4. Set $(x^{\nu+1,0}, y^{1,\nu+1,0}, \ldots, \gamma^{p,\nu+1,0}) = (x^{\nu,\star}, y^{1,\nu,\star}, \ldots, \gamma^{p,\nu,\star})$, $\nu := \nu + 1$, and go to Step 3.

In Step 2, we clearly choose $x^{0,0} = x^0$. In order to obtain the corresponding values $(y^{1,0,0}, \ldots, \gamma^{p,0,0})$ numerically, one might try to find a zero of

$$\begin{pmatrix} D_y^\top \mathcal{L}_j(x^0, y^j, \gamma^j) \\ \Psi_{\tau_0}(\gamma^j, -v(x^0, y^j)) \end{pmatrix}$$

for each $j \in J$. Another method will be given below.

Step 3 is a "black box" which stands for any standard solution method for nonlinear finite optimization problems. In view of Steps 2 and 4, a minimal requirement is that the method should be able to process infeasible starting points.

Conceptually, termination criteria might be the relative error of optimal points or of optimal values, as well as the error in the first order optimality condition for *GSIP* (cf. Corollary 2.6) and combinations thereof. We emphasize that the availability of an easily checkable first order optimality condition is crucial for the numerical performance of the method.

4. Convergence results. Recall that in section 3 we reformulated *GSIP* equivalently first as an *SG* and then as a finite optimization problem *P*. Then we embedded *P* in the parametric family P_τ with $\tau \in \mathbb{R}$. Let us now clarify what the equivalent embedding for *GSIP* is. Throughout this section, let Assumptions 3 and 4 hold.

With the observations in section 3 the problem P_τ with $\tau \neq 0$ can equivalently be written in the form of an *SG*:

$$SG_\tau: \quad \min_{x,y^1,\dots,y^p} f(x) \ \text{ s.t. } \ g_j(x,y^j) \le 0\,, \ \text{ and } \ y^j \ \text{ solves } \ Q_\tau^j(x),\ j \in J.$$

Under the reduction ansatz, locally near a feasible point $\bar{x}$ of *GSIP* (i.e., for $\bar{\tau} = 0$) there exists a unique solution of (3.1)–(3.4) with $\tau \neq 0$.

PROPOSITION 4.1. *Let $\bar{x} \in M$ and $j \in J$ be given. Assume that $\bar{y}^j$ is a solution of $Q^j(\bar{x})$ with corresponding multiplier vector $\bar{\gamma}^j$ such that $A^j(\bar{x},\bar{y}^j,\bar{\gamma}^j)$ is nonsingular. Then there exist neighborhoods U of $\bar{x}$ and T of $\bar{\tau} = 0$ as well as C^{d-1}-functions $y^j : V \to \mathbb{R}^m$, $\gamma^j : V \to \mathbb{R}^s$ ($V := U \times T$) such that $y^j(\bar{x},0) = \bar{y}^j$, $\gamma^j(\bar{x},0) = \bar{\gamma}^j$ and such that for all $(x,\tau) \in V$ $(y^j(x,\tau),\gamma^j(x,\tau))$ is the unique solution of* (3.1)–(3.4).

Proof. The proof follows directly by applying the implicit function theorem to the system of equations (3.1), (3.2), and using that $A^j(\bar{x},\bar{y}^j,\bar{\gamma}^j)$, its Jacobian with respect to (y,γ) at $(\bar{x},\bar{y}^j,\bar{\gamma}^j,\bar{\tau})$, is nonsingular. Note that (3.3) and (3.4) hold thanks to continuity arguments. ∎

Under the reduction ansatz (Assumption 2) the assumptions of Proposition 4.1 are satisfied for $j \in J_0(\bar{x})$ due to Lemma 2.7. Because of Lemma 3.2(i) for all $x \in U$, $\tau \in T$ the condition

$$g_j(x,y^j) \le 0 \quad \text{and} \quad y^j \quad \text{solves} \quad Q_\tau^j(x)$$

is equivalent with

$$g_j(x,y^j(x,\tau)) \le 0.$$

Note that under Assumptions 3 and 4 for $j \in J \setminus J_0(\bar{x})$ the value function $\tilde{\varphi}_j(x,\tau)$ of $Q_\tau^j(x)$ is continuous in the neighborhood $U \times T$ of $(\bar{x},0)$. So for all $\tau \in T$ the problem P_τ, locally restricted to $x \in U$, is equivalent to the reduced problem

$$GSIP_\tau(\bar{x}): \quad \min_{x \in U} f(x) \ \text{ s.t. } \ g_j(x,y^j(x,\tau)) \le 0 \text{ for all } j \in J_0(\bar{x}).$$

Summarizing, we have obtained the following reduction lemma, which provides the basis for our convergence analysis.

LEMMA 4.2 (parametric reduction lemma). *Let the reduction ansatz hold at a point $\bar{x} \in M$. Then locally in a neighborhood $U \times T$ of $(\bar{x},0)$ (cf. Proposition* 4.1*) the problems P_τ and $GSIP_\tau(\bar{x})$ are equivalent in the sense that for all $\tau \in T$ the vector $x_\tau \in U$ is a solution of $GSIP_\tau(\bar{x})$ if and only if $(x_\tau,y^1,\dots,y^p)$ with $x_\tau \in U$ solves SG_τ. In particular, problem $GSIP_0(\bar{x})$ is locally in U equivalent with $SG_0 = SG$ and, hence, with GSIP.*

The above lemma yields local reductions for all problems P_τ when (x,τ) is sufficiently close to $(\bar{x},0)$, which shows that the parametric reduction lemma implies the reduction lemma (Theorem 2.3).

The following theorem is related to a result by Shimizu and Aiyoshi [45] about the convergence behavior of the solutions of SG_τ for $\tau \to 0$. However, our proof relies only on the parametric reduction lemma and some well-known results from parametric optimization.

THEOREM 4.3. *Let* $(\tau_\nu)_{\nu\in\mathbb{N}}$ *be a sequence with* $\lim_{\nu\to\infty}\tau_\nu = 0$, *and let* $(x^\nu, y^{1,\nu}, \gamma^{1,\nu}, \ldots, y^{p,\nu}, \gamma^{p,\nu})_{\nu\in\mathbb{N}}$ *be a sequence of global solutions of* P_{τ_ν}, $\nu\in\mathbb{N}$ *(cf.* (3.5)*). If* $x^\star$ *is an accumulation point of the sequence* $(x^\nu)_{\nu\in\mathbb{N}}$ *such that Assumption* 2 *holds at* $x^\star$ *and such that the MFCQ holds at some solution of* $GSIP_0(x^*)$, *then* $x^\star$ *is a global solution of GSIP.*

Proof. Without loss of generality, let (x^ν, τ_ν) converge to $(x^\star, 0)$ for $\nu\to\infty$. By Proposition 4.1 for the solutions $y^{j,\nu}$, $j\in J_0(x^*)$, we have (for sufficiently large ν) $y^{j,\nu} = y(x^\nu, \tau_\nu)$, and x^ν is a solution of $GSIP_{\tau_\nu}(x^*)$. By continuity x^* is feasible for $GSIP_0(x^*) = GSIP$ (locally in U). As the MFCQ holds at some solution of $GSIP_0(x^\star)$, by a result due to Gauvin and Dubeau [9] the value function $\omega(\tau)$ of $GSIP_\tau(x^\star)$ is continuous in τ. Consequently $\omega(\tau_\nu) = f(x^\nu)\to\omega(0)$ and $x^\star$ is a solution of *GSIP*. □

Theorem 4.3 is primarily of theoretical interest, as numerical standard software may usually not find global solution points of the problems P_{τ_ν}, $\nu\in\mathbb{N}$. One can at most expect a point which satisfies a first order optimality condition like the one of Fritz John. Consequently, a numerical solution method for *GSIP* can also only be expected to find Fritz John points in the sense of section 2.

In the following we study how Fritz John points of the finite problems P_τ $(\tau\neq 0)$ are related to the Fritz John points of $GSIP_\tau(\bar{x})$ and *GSIP*.

LEMMA 4.4. *Let the reduction ansatz hold at a point* $\bar{x}\in M$, *and let* $(x, y^1, \gamma^1, \ldots, y^p, \gamma^p)$ *be a Fritz John point of* P_τ *(cf.* (3.5)*) with* (x,τ) *sufficiently close to* $(\bar{x}, 0)$. *Moreover, let the matrices*

$$A^j = \begin{pmatrix} D_y^2\mathcal{L}_j(x, y^j, \gamma^j) & -D_y^\top v(x, y^j) \\ -\operatorname{diag}(\gamma^j) D_y v(x, y^j) & -\operatorname{diag}(v(x, y^j)) \end{pmatrix}, \quad j\in J\setminus J_0(\bar{x}),$$

be nonsingular. Then x *is a Fritz John point of* $GSIP_\tau(\bar{x})$.

Proof. It is not hard to see that x is feasible for $GSIP_\tau(\bar{x})$. The feasibility of $(x, y^1, \gamma^1, \ldots, y^p, \gamma^p)$ for P_τ implies particularly that y^j solves $Q_\tau^j(x)$. Since the matrices A^j are nonsingular for all $j\in J$, for (x,τ) sufficiently close to $(\bar{x}, 0)$ the point y^j coincides with the unique solution $y^j(x,\tau)$ of $Q_\tau^j(x)$ (cf. Proposition 4.1). These observations, together with some simple continuity arguments, yield the reduction of the Fritz John condition of P_τ to the one of $GSIP_\tau(\bar{x})$. □

Now we can prove the main result of this section.

THEOREM 4.5. *Let* $(\tau_\nu)_{\nu\in\mathbb{N}}$ *be a sequence with* $\lim_{\nu\to\infty}\tau_\nu = 0$, *and let* $(x^\nu, y^{1,\nu}, \gamma^{1,\nu}, \ldots, y^{p,\nu}, \gamma^{p,\nu})$ *be Fritz John points of* P_{τ_ν}, $\nu\in\mathbb{N}$ *(cf.* (3.5)*), with an accumulation point* $(x^\star, y^{1,\star}, \gamma^{1,\star}, \ldots, y^{p,\star}, \gamma^{p,\star})$. *Let the reduction ansatz (Assumption* 2*) hold at* $x^\star$, *and let the matrices*

$$A^j = \begin{pmatrix} D_y^2\mathcal{L}_j(x^\star, y^{j,\star}, \gamma^{j,\star}) & -D_y^\top v(x^\star, y^{j,\star}) \\ -\operatorname{diag}(\gamma^{j,\star}) D_y v(x^\star, y^{j,\star}) & -\operatorname{diag}(v(x^\star, y^{j,\star})) \end{pmatrix}, \quad j\in J\setminus J_0(x^\star),$$

be nonsingular. Then $x^\star$ *is a Fritz John point of GSIP.*

Proof. For sufficiently large $\nu\in\mathbb{N}$ all assumptions of Lemma 4.4 are satisfied so that x^ν is a Fritz John point of $GSIP_{\tau_\nu}(x^\star)$. By a continuity argument and Lemma 2.2 $x^\star$ is thus a Fritz John point for *GSIP* in the sense of Corollary 2.6. □

The following proposition gives a criterion for the existence of an accumulation point in the assumption of Theorem 4.5.

PROPOSITION 4.6. *Let* $(\tau_\nu)_{\nu\in\mathbb{N}}$ *be a sequence with* $\lim_{\nu\to\infty}\tau_\nu = 0$, *and let* $(x^\nu, y^{1,\nu}, \gamma^{1,\nu}, \ldots, y^{p,\nu}, \gamma^{p,\nu})$ *be feasible points of* P_{τ_ν}, $\nu\in\mathbb{N}$ *(cf.* (3.5)*). Moreover, let* $x^\star$ *be an accumulation point of the sequence* $(x^\nu)_{\nu\in\mathbb{N}}$ *such that the LICQ holds*

everywhere in $Y(x^\star)$. *Then the sequence* $(x^\nu, y^{1,\nu}, \gamma^{1,\nu}, \ldots, y^{p,\nu}, \gamma^{p,\nu})_{\nu\in\mathbb{N}}$ *possesses an accumulation point* $(x^\star, y^{1,\star}, \gamma^{1,\star}, \ldots, y^{p,\star}, \gamma^{p,\star})$.

Proof. After taking a subsequence, let $x^\nu \to x^\star$ for $\nu \to \infty$. For $\nu \in \mathbb{N}$ the feasibility of $(x^\nu, y^{1,\nu}, \gamma^{1,\nu}, \ldots, y^{p,\nu}, \gamma^{p,\nu})$ for P_{τ_ν} implies $\psi_{\tau_\nu}(\gamma_\ell^{j,\nu}, -v_\ell(x^\nu, y^{j,\nu})) = 0$ and thus, by Lemma 2.8(i), $v_\ell(x^\nu, y^{j,\nu}) < 0$ for all $\ell \in L,\ j \in J$. Hence, $y^{j,\nu} \in Y(x^\nu)$, $j \in J$, and the upper semicontinuity of Y yields that an accumulation point $y^{j,\star}$ of $y^{j,\nu}, \nu \in \mathbb{N}$, exists and is contained in $Y(x^\star)$, $j \in J$.

Next assume that for some $j \in J$ the sequence $(\gamma^{j,\nu})_{\nu\in\mathbb{N}}$ is unbounded. Then for sufficiently large ν we can consider the vectors $\|\gamma^{j,\nu}\|^{-1}\gamma^{j,\nu}$, which converge to a vector η^j with $\|\eta^j\| = 1$, possibly after taking a subsequence.

The feasibility of $(x^\nu, y^{1,\nu}, \gamma^{1,\nu}, \ldots, y^{p,\nu}, \gamma^{p,\nu})$ for P_{τ_ν} yields

$$D_y g_j(x^\nu, y^{j,\nu}) - \gamma^{j,\nu} D_y v(x^\nu, y^{j,\nu}) \;=\; 0 \tag{4.1}$$

and, by Lemma 2.8(i),

$$-\gamma_\ell^{j,\nu} \cdot v_\ell(x^\nu, y^{j,\nu}) \;=\; \tau_\nu^2\ , \quad \ell \in L\ . \tag{4.2}$$

Division of (4.2) by $\|\gamma^{j,\nu}\|$ and taking the limit for $\nu \to \infty$ yields $\eta_\ell^j = 0$ for $\ell \in L \setminus L_0(x^\star, y^{j,\star})$. However, in the same way (4.1) then implies that the LICQ is violated at $y^{j,\star}$ in $Y(x^\star)$, contradicting our assumptions. Consequently, for each $j \in J$ the sequence $(\gamma^{j,\nu})_{\nu\in\mathbb{N}}$ is bounded and, thus, has an accumulation point. ∎

The existence of some accumulation point $x^\star$ in the assumption of Proposition 4.6 can of course be guaranteed if an additional restriction $x \in X$ is incorporated into *GSIP*, with X nonempty and compact.

We end this section with a result on the rate of convergence of the method.

LEMMA 4.7. *For* $\bar{x} \in M$ *and* $j \in J$ *let* y^j *be a nondegenerate solution of* $Q^j(\bar{x})$, *and let* $(y^j(x,\tau), \gamma^j(x,\tau))$ *denote the locally unique solution of* (3.1)–(3.4) *around* $(\bar{x}, 0)$. *Then we have*

$$\begin{pmatrix} D_\tau y^j(x,0) \\ D_\tau \gamma^j(x,0) \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

for x *sufficiently close to* $\bar{x}$.

In the following proposition we call $\bar{x} \in M$ a nondegenerate solution of *GSIP* if the reduction ansatz holds at $\bar{x}$ and if $\bar{x}$ is a nondegenerate local minimizer for the locally reduced problem $GSIP_0(\bar{x})$.

PROPOSITION 4.8. *Let the assumptions of Theorem* 4.3 *hold, and let the solution* $x^\star$ *of GSIP be nondegenerate. Then for each subsequence of* $(x^\nu)_{\nu\in\mathbb{N}}$ *that converges to* $x^\star$, *the optimal values of* P_{τ_ν} *satisfy*

$$f(x^\nu) - f(x^\star) \;=\; O(\tau_\nu^2).$$

Proof. Let $\omega(\tau)$ denote the optimal value of P_τ. In the proof of Theorem 4.3 we have seen that $\omega(\tau)$ coincides with the optimal value of $GSIP_\tau(x^\star)$ if τ is sufficiently close to zero. Now Lemma 2.2, Lemma 4.7, and a Taylor expansion of ω around 0 yield the assertion. ∎

5. Numerical examples. For the numerical illustrations we implemented the method from section 3 in *Matlab* 5.3 and used the routine *fmincon* from its *Optimization Toolbox* 2.0, i.e., an SQP method with BFGS updates for the Hessian of the

Lagrangian, to replace the "black box" in Step 3 of the method. All examples were run on an 800 MHz Linux PC.

In Step 2 we do *not* solve the nonlinear problems

$$\begin{pmatrix} D_y^\top \mathcal{L}_j(x^0, y^j, \gamma^j) \\ \Psi_{\tau_0}(\gamma^j, -v(x^0, y^j)) \end{pmatrix} = 0\,, \quad j \in J\,,$$

since an appropriate starting point for an iteration procedure is not at hand. Instead we solve the unconstrained, concave problems

$$Q^j_{\tau_0}(x^0): \qquad \max_y \; g_j(x^0, y) + \tau_0^2 \sum_{\ell \in L} \ln(-v_\ell(x^0, y))\,, \quad j \in J\,,$$

by the routine *fminunc* to obtain $y^{j,0,0},\ j \in J$, and put

$$\gamma_\ell^{j,0,0} \;=\; -\frac{\tau_0^2}{v_\ell(x^0, y^{j,0,0})}\,,\ \ell \in L\,,\ j \in J\,.$$

Here the problem of finding an appropriate starting point is solved easily by the determination of some Slater point of the set $Y(x^0)$. The latter task can be fulfilled by solving the convex problem

$$\min_{y,\eta} \; \eta \quad \text{s.t.} \quad v_\ell(x^0, y) - \eta \;\le\; 0\,, \quad \ell \in L\,,$$

where the choice of a starting point is obvious.

As an a priori τ-sequence we use $\tau_\nu = 10 \cdot 100^{-\nu},\ \nu \in \mathbb{N}$. The iteration terminates if either the relative error in the optimal point or in the optimal value is less than 10^{-6}. If the method is provided with the gradients of f, g, and v with respect to x, then also the first order condition for *GSIP* from Corollary 2.6 is checked (in the Euclidean norm).

5.1. Design centering in two dimensions. The general design centering problem (see also [14]) consists of maximizing some measure, e.g., the volume, of a parameterized body $B(x)$ which is contained in a second body G:

$$\max_{x \in \mathbb{R}^n} \; \mathit{Vol}(B(x)) \;\; \text{s.t.} \;\; B(x) \subset G\,.$$

In the first examples we let $G = \{y \in \mathbb{R}^2 |\ g(y) \le 0\}$ with

$$g(y) = \begin{pmatrix} -y_1 - y_2^2 \\ y_1/4 + y_2 - 3/4 \\ -y_2 - 1 \end{pmatrix}\,.$$

The two-dimensional volume of the resulting body is easily calculated to be 20/3. An equivalent formulation of the general design centering problem as *GSIP* is

$$\max_{x \in \mathbb{R}^n} \; \mathit{Vol}(B(x)) \;\; \text{s.t.} \;\; g(y) \le 0 \;\; \text{for all} \;\; y \in B(x)\,.$$

Problem 1. We look for the largest ball with free center and radius that is contained in G. Thus, we have $n = 3$ and

$$B(x) \;=\; \{y \in \mathbb{R}^2 |\ (y_1 - x_1)^2 + (y_2 - x_2)^2 - x_3^2 \le 0\}\,, \quad \mathit{Vol}(B(x)) \;=\; \pi\, x_3^2\,.$$

TABLE 1
Approximation of G with ψ^{NR}.

Problem	ov	$\varepsilon_{\mathtt{ov}}$	$\varepsilon_{\mathtt{op}}$	$\varepsilon_{\mathtt{FOC}}$	$\mathtt{CPU}_{\mathtt{init}}$	$\mathtt{CPU}_{\mathtt{iter}}$	$\#_{\mathtt{iter}}$
1	1.8606	2.5596e-06	$\star$	$\star$	2.94	1.04	3
2	3.4838	$\star$	$\star$	4.8781e-06	2.96	3.55	4
3	3.7234	$\star$	$\star$	n.a.	3.31	7.50	4
4	3.0792	3.9278e-06	$\star$	2.1812e-06	3.76	1.85	3

As an initial point we use the infeasible point $x^0 = (0, 0, 1)^\top$.

Problem 2. We search the largest ellipsoid with free center and axis lengths that is contained in G. The axes are supposed to be parallel to the coordinate axes. We have $n = 4$ and

$$B(x) \;=\; \left\{ y \in \mathbb{R}^2 \;\middle|\; \frac{(y_1 - x_1)^2}{x_3^2} + \frac{(y_2 - x_2)^2}{x_4^2} - 1 \le 0 \right\}, \;\; \mathit{Vol}(B(x)) \;=\; \pi\, x_3 x_4 \;.$$

The initial point is $x^0 = (0, 0, 1, 1)^\top$.

Problem 3. Now the ellipsoid from Problem 2 is allowed to have axes in arbitrary position, and with $n = 6$ we can set

$$B(x) \;=\; \left\{ y \in \mathbb{R}^2 \;\middle|\; \left(y - \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right)^\top \left(\begin{pmatrix} x_3 & x_4 \\ x_5 & x_6 \end{pmatrix} \begin{pmatrix} x_3 & x_5 \\ x_4 & x_6 \end{pmatrix} \right)^{-1} \left(y - \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) - 1 \;\le\; 0 \right\}.$$

Then we have

$$\mathit{Vol}(B(x)) \;=\; \pi \left| \det \begin{pmatrix} x_3 & x_4 \\ x_5 & x_6 \end{pmatrix} \right|,$$

and we choose the initial point $x^0 = (0, 0, 1, 0, 0, 1)^\top$.

Problem 4. In this problem we inscribe the largest box with sides parallel to the coordinate axes into G. For $n = 4$ we have

$$B(x) \;=\; \{ y \in \mathbb{R}^2 |\; y_1 - x_1 \;\le\; 0, \quad y_2 - x_2 \;\le\; 0, \; -y_1 + x_3 \;\le\; 0, \; -y_2 + x_4 \;\le\; 0 \}$$

with

$$\mathit{Vol}(B(x)) \;=\; (x_1 - x_3) \cdot (x_2 - x_4)\,,$$

and we choose the infeasible initial point $x^0 = (1, 1, -1, -1)^\top$.

The columns of the following tables are labeled as follows: ov, optimal value; $\varepsilon_{\mathtt{ov}}$, relative error in optimal value; $\varepsilon_{\mathtt{op}}$, relative error in optimal point; $\varepsilon_{\mathtt{FOC}}$, error in first order optimality condition; $\mathtt{CPU}_{\mathtt{init}}$, CPU time for initialization step in seconds; $\mathtt{CPU}_{\mathtt{iter}}$, CPU time for iterations in seconds; $\#_{\mathtt{iter}}$, number of outer iterations. If (relative) errors are below 10^{-6}, we replace the actual number by the symbol "$\star$". In Problem 3 the gradients entering the first order optimality condition are not available for the method, so this criterion is not checked.

For the results in Table 1 we used the natural residual function ψ^{NR} as the NCP function. In this example the performance of the method does not change significantly if the NCP function ψ^{NR} is replaced by ψ^{FB} (see [52] for more details).

Concerning the solution of Problem 4 it is worth mentioning, as expected, we have $x_4^\star = -1$. This means that the method converges although $Q^3(x^\star)$ clearly is a *degenerate* problem. In fact, the computed "maximal box" in G is the set $B(x^\star) = [-0.024, 3.619] \times [-1, -0.155]$, so that the solution set of $Q^3(x^\star)$ coincides with the facet $[-0.024, 3.619] \times \{-1\}$ of $B(x^\star)$. From all these optimal points the method chooses their "midpoint" as $y^{3,\star} = (1.7975, -1)$. In fact, in our approach the lower level linear problems are solved by the central path method. It is well known in linear programming that under weak assumptions the interior point sequence converges to the so-called analytic center of the optimal facet (cf. [39]).

5.2. Robust optimization. Robustness questions arise when an optimization problem is subject to uncertain data. If an inequality constraint function $g(x, y)$ depends on some uncertain parameter vector $y \in Y \subset \mathbb{R}^m$, then the "most cautious" way to deal with this constraint is to use its worst-case reformulation

$$g(x, y) \ \leq 0 \ \text{ for all } \ y \in Y\,,$$

which is clearly of semi-infinite type. When the uncertainty set Y also depends on the state variable x, we arrive at a generalized semi-infinite constraint.

The following robust optimization problem is studied in [1] for elliptic uncertainty sets. In the case of ellipticity the lower level optimal value functions can be computed explicitly in such a way that the semi-infinite problem is reduced to a nonsmooth finite problem which can be tackled by SDP methods (cf. [1] for details). We will show that our numerical method solves not only this specially structured problem but also two nontrivial generalizations.

Let 1 euro be invested in a portfolio comprised of N shares. At the end of a given period the return per 1 euro invested in share i is $y_i > 0$. The goal is to determine the amount x_i to be invested in share i, $i = 1, \ldots, N$, so as to maximize the end-of-period portfolio value $y^\top x$.

If the vector y was certain, the solution of this optimization problem would be evident. A more realistic assumption is that y varies in some nonempty compact set $Y \subset \mathbb{R}^N$. Upon moving the objective function to the constraint set we obtain the following standard semi-infinite optimization problem with $n = N + 1$ and $m = N$:

$$\max_{x, x_{N+1}} \ x_{N+1} \quad \text{s.t.} \quad x_{N+1} - y^\top x \leq 0\,,\ y \in Y\,,\ \sum_{i=1}^{N} x_i = 1\,,\ x \geq 0.$$

Apart from its special structure used in [1] for the case of an ellipsoidal set Y, this is also a linear semi-infinite optimization problem, meaning that the semi-infinite constraint function is linear in the variable (x, x_{N+1}). Solution methods for these types of problems are described, e.g., in [10, 20]. Note, however, that the index set of the semi-infinite constraint is N-dimensional, where N might be a large number.

Problem 5. In [1] the set Y has the form

$$Y \ = \ \left\{ y \in \mathbb{R}^N \,\middle|\, \sum_{i=1}^{N} \frac{(y_i - \bar{y}_i)^2}{\sigma_i^2} \ \leq \ \theta^2 \right\},$$

where $\bar{y}_i$ is some "nominal" value of y_i, σ_i is a scaling parameter, $i = 1, \ldots, N$, and θ measures the risk aversion of the decision maker. With the particular choices

$$\bar{y}_i = 1.15 + i \cdot \frac{0.05}{N}\,, \quad i = 1, \ldots, N,$$

TABLE 2
Optimal portfolio with ellipsoidal uncertainty and ψ^{NR}.

N	ov	$\varepsilon_{\texttt{ov}}$	$\varepsilon_{\texttt{op}}$	$\varepsilon_{\texttt{FOC}}$	$\texttt{CPU}_{\texttt{init}}$	$\texttt{CPU}_{\texttt{iter}}$	$\#_{\texttt{iter}}$
10	1.15	$\star$	1.3693e-03	3.0860e-05	1.55	3.71	3
50	1.15	$\star$	5.4195e-05	6.6652e-05	7.24	27.23	4
100	1.15	$\star$	3.3458e-05	7.5987e-05	66.27	241.6	4
150	1.15	$\star$	1.9149e-05	4.1678e-05	272.94	884.22	4

TABLE 3
Optimal portfolio with ellipsoidal uncertainty and ψ^{FB}.

N	ov	$\varepsilon_{\texttt{ov}}$	$\varepsilon_{\texttt{op}}$	$\varepsilon_{\texttt{FOC}}$	$\texttt{CPU}_{\texttt{init}}$	$\texttt{CPU}_{\texttt{iter}}$	$\#_{\texttt{iter}}$
10	1.15	$\star$	7.7231e-04	2.8254e-05	1.53	3.68	4
50	1.15	$\star$	5.4165e-05	5.5333e-05	7.21	34.52	4
100	1.15	$\star$	4.7199e-05	1.5974e-04	65.17	333.8	5
150	1.15	$\star$	4.7277e-05	3.7021e-04	271.49	912.72	4

$$\sigma_i = \frac{0.05}{3\,N}\sqrt{2N(N+1)i}\,, \quad i = 1, \ldots, N,$$
$$\theta = 1.5,$$

one can show that the optimal policy is to invest equally in all shares, i.e., $x_i = 1/N\,,\ i = 1, \ldots, N$, with optimal value $1.15\,$. We use the starting point $x^0 = (1, 0, \ldots, 0)$ in $\mathbb{R}^{N+1}$.

Having the large dimensions of Y in mind, the method performs very well when ψ^{NR} is used (cf. Table 2). In the case of ψ^{FB} as the NCP function, the "black box" part of the method (i.e., the *Matlab* routine *fmincon*) does not converge for $N = 100$ (cf. Table 3), so that we solved this particular problem with the a priori sequence $\tau_\nu = 10^{-\nu},\ \nu \in \mathbb{N}$.

Problem 6. A more general choice of Y is

$$Y_\delta \;=\; \{\; y \in \mathbb{R}^N \,|\; \|\operatorname{diag}(\sigma)^{-1}(y - \bar{y})\|_\delta \;\leq\; \theta \;\}$$

with $\delta \in [1, \infty]$. Whereas Y_2 is the ellipsoid from Problem 5, the sets Y_1 and Y_∞ are polytopes. For all other choices of δ we still obtain a nonempty compact convex set $Y_\delta\,$. As polytopes can be considered as ellipsoidal sets in the sense of [1], let us use our method for a *nonellipsoidal* set like $Y_{10}\,$. Tables 4 and 5 show the results for the starting point $x^0 = 1/N \cdot (1, \ldots, 1, 0)$. The method performs well for dimensions up to $N = 150$. In this example the initialization phase takes about as long as the main iterations. Note that for increasing dimensions the attainment of the first order condition becomes worse when the method terminates because of a small relative error in the optimal value.

Problem 7. Finally, since our method works for x-dependent sets Y, we can also consider the case in which the risk aversion of the decision maker depends on the point x. If for instance his risk aversion increases when the values x_i deviate from $1/N$, $i = 1, \ldots, N$, we can replace θ by the expression

$$\Theta(x) \;=\; \theta \cdot \left(1 + \sum_{i=1}^{N} \left(x_i - \frac{1}{N}\right)^2\right)$$

and obtain the *generalized* semi-infinite optimization problem

$$\max_{x, x_{N+1}} \; x_{N+1} \quad \text{s.t.} \quad x_{N+1} - y^\top x \leq 0\,,\ y \in Y(x)\,,\ \sum_{i=1}^{N} x_i = 1\,,\ x \geq 0,$$

TABLE 4
Optimal portfolio with nonellipsoidal uncertainty and ψ^{NR}.

N	ov	$\varepsilon_{\mathtt{ov}}$	$\varepsilon_{\mathtt{op}}$	$\varepsilon_{\mathtt{FOC}}$	$\mathtt{CPU}_{\mathtt{init}}$	$\mathtt{CPU}_{\mathtt{iter}}$	$\#_{\mathtt{iter}}$
10	1.1190	$\star$	7.8075e-05	1.5139e-03	2.82	5.63	3
50	1.1155	$\star$	1.5402e-05	6.1081e-02	21.36	41.14	3
100	1.1151	$\star$	9.0467e-06	6.0789e-02	246.35	231.22	3
150	1.1150	$\star$	2.4410e-05	2.0368e-01	809.94	708.99	3

TABLE 5
Optimal portfolio with nonellipsoidal uncertainty and ψ^{FB}.

N	ov	$\varepsilon_{\mathtt{ov}}$	$\varepsilon_{\mathtt{op}}$	$\varepsilon_{\mathtt{FOC}}$	$\mathtt{CPU}_{\mathtt{init}}$	$\mathtt{CPU}_{\mathtt{iter}}$	$\#_{\mathtt{iter}}$
10	1.1190	$\star$	5.4529e-04	1.1783e-03	2.81	5.58	3
50	1.1155	$\star$	4.3946e-04	6.0965e-02	21.26	41.71	3
100	1.1151	$\star$	1.5527e-05	5.7950e-02	241.95	228.89	3
150	1.1150	$\star$	1.0349e-05	2.1809e-01	811.68	703.06	3

TABLE 6
Optimal portfolio with state-dependent uncertainty and ψ^{NR}.

N	ov	$\varepsilon_{\mathtt{ov}}$	$\varepsilon_{\mathtt{op}}$	$\varepsilon_{\mathtt{FOC}}$	$\mathtt{CPU}_{\mathtt{init}}$	$\mathtt{CPU}_{\mathtt{iter}}$	$\#_{\mathtt{iter}}$
10	0.7033	$\star$	$\star$	6.9574e-06	2.81	1.45	4
50	0.9638	$\star$	1.1912e-05	1.1828e-03	14.78	5.49	4
100	1.0259	$\star$	2.5937e-05	5.5778e-03	132.54	37.05	3
150	1.0535	$\star$	7.0987e-06	2.3336e-03	643.39	73.96	3

TABLE 7
Optimal portfolio with state-dependent uncertainty and ψ^{FB}.

N	ov	$\varepsilon_{\mathtt{ov}}$	$\varepsilon_{\mathtt{op}}$	$\varepsilon_{\mathtt{FOC}}$	$\mathtt{CPU}_{\mathtt{init}}$	$\mathtt{CPU}_{\mathtt{iter}}$	$\#_{\mathtt{iter}}$
10	0.7033	$\star$	1.3743e-05	1.8857e-04	2.77	1.53	4
50	0.9638	$\star$	3.7616e-05	3.7318e-03	14.76	4.84	3
100	1.0259	$\star$	4.5000e-05	9.6580e-03	133.00	20.48	3
150	1.0535	$\star$	3.6030e-05	1.1934e-02	642.96	58.05	3

with

$$Y(x) \;=\; \left\{ y \in \mathbb{R}^N \;\middle|\; \sum_{i=1}^{N} \frac{(y_i - \bar{y}_i)^2}{\sigma_i^2} \;\leq\; \Theta(x)^2 \right\} .$$

The choice $\sigma_i = 1,\ i = 1, \ldots, N$, prevents the solution of the original Problem 5 to be $x_i = 1/N\,,\ i = 1, \ldots, N$, so that the modified term $\Theta(x)$ can take effect, and we observe a good performance of our method in Tables 6 and 7 up to dimension $N = 150$. Note, however, that in this example the initialization phase takes considerably longer than the main iterations.

6. Final remarks. The essential idea behind the numerical method presented in section 3 is to reformulate the generalized semi-infinite optimization problem as a Stackelberg game and to make use of the convexity in the lower level problems. Starting at this point, there are several possible routes to the design of a numerical method.

First, there are other ways to treat the convex lower level problems than the one used in this article, e.g., penalty, barrier, or cutting plane methods. Furthermore, if one decides to replace the convex optimization problems by their first order optimality conditions and obtain a mathematical program with equilibrium constraints, this

program does not necessarily have to be solved with regularized NCP functions. Instead, the exact penalization approaches from [34, 44] or nonsmooth Newton methods (see, e.g., [30]) could be promising alternatives. Finally, in the case when one uses NCP functions, there are a multitude of functions other than the natural residual and the Fischer–Burmeister function to choose from. For a survey see, e.g., [5].

Further questions concern the required accuracy for the solution of the auxiliary problems by the "black box" method, the implementation of an active set strategy, and the design of a pathfollowing method (cf. [15]) to solve the finite parametric optimization problems P_τ for parameter values tending to $\tau = 0$. Let us point out, however, that the resulting numerical method will then not just solve a sequence of finite dimensional optimization problems that are easily constructed from the problem data, so that the implementation effort for the user increases drastically.

Moreover, generalizations of the convergence proofs from section 4 to cases such as convex lower level problems in which the reduction ansatz is not necessarily satisfied (recall Problem 4 in section 5.1) will be the subject of future research. We finally remark that a direct application of the presented ideas to semi-infinite optimization problems with *nonconvex* lower level problems results, in general, only in a *relaxation* of the original problem, since lower level optimality can then not be replaced equivalently with a first order optimality condition. However, our method can then still be used to obtain lower bounds for the optimal value of *GSIP*.

Acknowledgment. We express our thanks to the referee, whose precise and substantial remarks led to an improved version of the article.

REFERENCES

[1] A. Ben-Tal and A. Nemirovski, *Robust solutions of uncertain linear programs*, Oper. Res. Lett., 25 (1999), pp. 1–13.
[2] C. Berge, *Topological Spaces*, Oliver and Boyd, Edinburgh, 1963.
[3] T. Bröcker and L. Lander, *Differentiable Germs and Catastrophes*, London Math. Soc. Lecture Notes Ser. 17, Cambridge University Press, Cambridge, UK, 1975.
[4] B. Chen and P. T. Harker, *A non-interior-point continuation method for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1168–1190.
[5] B. Chen, X. Chen, and C. Kanzow, *A penalized Fischer-Burmeister NCP-function*, Math. Program., 88 (2000), pp. 211–216.
[6] F. Facchinei, H. Jiang, and L. Qi, *A smoothing method for mathematical programs with equilibrium constraints*, Math. Program., 85 (1999), pp. 107–134.
[7] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.
[8] A. Fischer, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.
[9] J. Gauvin and F. Dubeau, *Differential properties of the marginal function in mathematical programming*, Math. Program. Study, 19 (1982), pp. 101–119.
[10] M. A. Goberna and M. A. López, *Linear Semi-Infinite Optimization*, John Wiley, Chichester, 1998.
[11] E. G. Gol'stein, *Theory of Convex Programming*, Transl. Math. Monogr. 36, AMS, Providence, RI, 1972.
[12] T. J. Graettinger and B. H. Krogh, *The acceleration radius: A global performance measure for robotic manipulators*, IEEE J. Robotics Automat., 4 (1988), pp. 60–69.
[13] G. Gramlich, R. Hettich, and E. W. Sachs, *Local convergence of SQP methods in semi-infinite programming*, SIAM J. Optim., 5 (1995), pp. 641–658.
[14] P. Gritzmann and V. Klee, *On the complexity of some basic problems in computational convexity.* I. *Containment problems*, Discrete Math., 136 (1994), pp. 129–174.
[15] J. Guddat, F. Guerra Vasquez, and H. Th. Jongen, *Parametric Optimization: Singularities, Pathfollowing and Jumps*, John Wiley, Chichester, Teubner, Stuttgart, 1990.
[16] M. Gugat, *Semi-infinite terminal problems: A Newton-type method*, Optimization, 44 (1998), pp. 25–48.

[17] R. Hettich and H. Th. Jongen, *Semi-infinite programming: Conditions of optimality and applications*, in Optimization Techniques, Part 2, Lecture Notes in Control and Inform. Sci. 7, J. Stoer, ed., Springer-Verlag, Berlin, 1978, pp. 1–11.

[18] R. Hettich and K. O. Kortanek, *Semi-infinite programming: Theory, methods, and applications*, SIAM Rev., 35 (1993), pp. 380–429.

[19] R. Hettich and G. Still, *Second order optimality conditions for generalized semi-infinite programming problems*, Optimization, 34 (1995), pp. 195–211.

[20] R. Hettich and P. Zencke, *Numerische Methoden der Approximation und semi-infiniten Optimierung*, Teubner Studienbücher, Stuttgart, 1982.

[21] W. W. Hogan, *Point-to-set maps in mathematical programming*, SIAM Rev., 15 (1973), pp. 591–603.

[22] W. W. Hogan, *Directional derivatives for extremal value functions with applications to the completely convex case*, Oper. Res., 21 (1973), pp. 188–209.

[23] F. John, *Extremum problems with inequalities as subsidiary conditions*, in Studies and Essays, R. Courant Anniversary Volume, Interscience Publishers, New York, 1948, pp. 187–204.

[24] H. Th. Jongen, P. Jonker, and F. Twilt, *Nonlinear Optimization in Finite Dimensions*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.

[25] H. Th. Jongen, J.-J. Rückmann, and O. Stein, *Generalized semi-infinite optimization: A first order optimality condition and examples*, Math. Program., 83 (1998), pp. 145–158.

[26] H. Th. Jongen and G. W. Weber, *Nonlinear optimization: Characterization of structural stability*, J. Global Optim., 1 (1991), pp. 47–64.

[27] C. Kanzow, *Some noninterior continuation methods for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 851–868.

[28] C. Kanzow and H. Jiang, *A continuation method for (strongly) monotone variational inequalities*, Math. Program., 81 (1998), pp. 103–125.

[29] A. Kaplan and R. Tichatschke, *On a class of terminal variational problems*, in Parametric Optimization and Related Topics IV, J. Guddat, H. Th. Jongen, F. Nožička, G. Still, F. Twilt, eds., Peter Lang, Frankfurt, Germany, 1997, pp. 185–199.

[30] M. Kočvara, J. Outrata, and J. Zowe, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints: Theory, Applications and Numerical Results*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[31] W. Krabs, *On time-minimal heating or cooling of a ball*, in Numerical Methods of Approximation Theory, Vol. 8, Internat. Schriftenreihe. Numer. Math. 81, Birkhäuser, Basel, 1987, pp. 121–131.

[32] E. Levitin, *Reduction of Generalized Semi-Infinite Programming Problems to Semi-Infinite or Piece-Wise Smooth Programming Problems*, Preprint no. 8-2001, University of Trier, Trier, Germany, 2000.

[33] E. Levitin and R. Tichatschke, *A branch-and-bound approach for solving a class of generalized semi-infinite programming problems*, J. Global Optim., 13 (1998), pp. 299–315.

[34] Z. Luo, J. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, UK, 1996.

[35] V. H. Nguyen and J. J. Strodiot, *Computing a global optimal solution to a design centering problem*, Math. Program., 53 (1992), pp. 111–123.

[36] E. Polak, *An implementable algorithm for the optimal design centering, tolerancing and tuning problem*, J. Optim. Theory Appl., 37 (1982), pp. 45–67.

[37] R. Reemtsen and J.-J. Rückmann, eds., *Semi-Infinite Programming*, Kluwer Academic Publishers, Boston, 1998.

[38] S. M. Robinson, *Stability theory for systems of inequalities, part* II*: Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.

[39] C. Roos, T. Terlaky, and J.-P. Vial, *Theory and Algorithms for Linear Optimization*, John Wiley, Chichester, UK, 1997.

[40] J.-J. Rückmann and A. Shapiro, *First-order optimality conditions in generalized semi-infinite programming*, J. Optim. Theory Appl., 101 (1999), pp. 677–691.

[41] J.-J. Rückmann and O. Stein, *On linear and linearized generalized semi-infinite optimization problems*, Ann. Oper. Res., 101 (2001), pp. 191–208.

[42] J.-J. Rückmann and O. Stein, *On convex lower level problems in generalized semi-infinite optimization*, in Semi-Infinite Programming – Recent Advances, M. A. Goberna, M. A. Lopez, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 121–134.

[43] H. Scheel and S. Scholtes, *Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity*, Math. Oper. Res., 25 (2000), pp. 1–22.

[44] S. Scholtes and M. Stöhr, *Exact penalization of mathematical programs with equilibrium constraints*, SIAM J. Control Optim., 37 (1999), pp. 617–652.

[45] K. Shimizu and E. Aiyoshi, *A new computational method for Stackelberg and min-max problems by use of a penalty method*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 460–466.
[46] S. Smale, *Algorithms for solving equations*, in Proceedings of the International Congress of Mathematicians, AMS, Providence, RI, 1987, pp. 172–195.
[47] O. Stein, *On Parametric Semi-Infinite Optimization*, Shaker-Verlag, Aachen, Germany, 1997.
[48] O. Stein, *The reduction ansatz in absence of lower semi-continuity*, in Parametric Optimization and Related Topics V, J. Guddat, R. Hirabayashi, H. T. Jongen, and F. Twilt, eds., Peter Lang, Frankfurt, Germany, 2000, pp. 165–178.
[49] O. Stein, *On level sets of marginal functions*, Optimization, 48 (2000), pp. 43–67.
[50] O. Stein, *The feasible set in generalized semi-infinite programming*, in Approximation, Optimization and Mathematical Economics, M. Lassonde, ed., Physica-Verlag, Heidelberg, 2001, pp. 313–331.
[51] O. Stein, *First order optimality conditions for degenerate index sets in generalized semi-infinite programming*, Math. Oper. Res., 26 (2001), pp. 565–582.
[52] O. Stein, *Bi-level Strategies in Semi-infinite Programming*, Kluwer Academic Publishers, Boston, 2003.
[53] O. Stein and G. Still, *On optimality conditions for generalized semi-infinite programming problems*, J. Optim. Theory Appl., 104 (2000), pp. 443–458.
[54] O. Stein and G. Still, *On generalized semi-infinite optimization and bilevel optimization*, European J. Oper. Res., 142 (2002), pp. 444–462.
[55] G. Still, *Generalized semi-infinite programming: Numerical aspects*, Optimization, 49 (2001), pp. 223-242.
[56] G. Still, *Discretization in semi-infinite programming: The rate of convergence*, Math. Program., 91 (2001), pp. 53–69.
[57] G.-W. Weber, *Generalized Semi-Infinite Optimization and Related Topics*, Habilitation Thesis, Darmstadt University of Technology, Darmstadt, Germany, 1999.
[58] W. Wetterling, *Definitheitsbedingungen für relative Extrema bei Optimierungs- und Approximationsaufgaben*, Numer. Math., 15 (1970), pp. 122–136.
[59] G. Zwier, *Structural Analysis in Semi-Infinite Programming*, Thesis, University of Twente, Enschede, The Netherlands, 1987.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 789–809

MINIMUM TIME WITH BOUNDED ENERGY, MINIMUM ENERGY WITH BOUNDED TIME*

MONICA MOTTA† AND CATERINA SARTORI‡

Abstract. Necessary and sufficient conditions for the regularity of the minimum time function and minimum energy function for a control system with controls in $L^p([0,+\infty[,\mathbb{R}^m)$ and $p \geq 1$ are given in terms of topological properties of the reachable sets. In particular, standard local controllability assumptions are sufficient to yield the continuity of both value functions for linear systems and $p \geq 1$ and the Hölder continuity of the minimum time function for nonlinear systems and $p > 1$.

Key words. nonlinear control systems, reachable sets, minimum time, minimum energy

AMS subject classifications. 93B05, 93B06, 49K15, 49L20

DOI. 10.1137/S0363012902385284

1. Introduction.

In this paper we consider the system

$$(\hat{S})_p \qquad \dot{y}(t) = f(y(t)) + \sum_{i=1}^{m} g_i(y(t))u_i(t), \quad t > 0, \quad u \in L^p([0,+\infty[,\mathbb{R}^m)$$

for $p \geq 1$ and give results on the regularity of the functions $\hat{T}_p(x,K)$ and $\hat{E}_p(x,T)$, which are, respectively, the minimum time needed to steer a point $x \in \mathbb{R}^n$ to the origin, along the trajectories of $(\hat{S})_p$, under the constraint $\int_0^{+\infty} |u(s)|^p\, ds \leq K^p$ and the minimum of the needed energy, defined as $(\int_0^T |u(s)|^p\, ds)^{1/p}$, under the constraint $t \leq T$ ($T, K > 0$ given). Strictly related to the regularity of such value functions are the topological properties of the reachable sets defined as

$$\hat{\mathcal{R}}_p(T,K) \doteq \left\{ x \in \mathbb{R}^n : \ \exists u \text{ such that (s.t.) } \int_0^T |u(s)|^p\, ds \leq K^p \text{ and } y_x(T,u) = 0 \right\}.$$

We point out that the cases $p > 1$ and $p = 1$ are very different. In fact, for $p > 1$, the following three properties are obtained among the results of section 2: the sets $\hat{\mathcal{R}}_p(T,K)$ are compact, an optimal control exists for the above minimization problems, and $\hat{T}_p$ and $\hat{E}_p$ are lower semicontinuous. On the other hand, for $p = 1$, we give an example (Example 2.1) in which the sets $\hat{R}_1(T,K)$ are not closed, an optimal control for the minimum time problem does not exist, and $\hat{T}_1$ is not lower semicontinuous. This difference is mainly due to the fact that for $p = 1$ the limit of minimizing sequences of trajectories can be a discontinuous function. However, following [3] all the results obtained for $p > 1$ can be proven also in the case $p = 1$ by considering an *extended system* $(S)_1$ whose trajectories are graphs or limits of graphs of solutions to $(\hat{S})_1$. We then embed the two original minimization problems into two *extended minimization problems* related to the system $(S)_1$ which in general are not equivalent to the original problems. Indeed, in section 2 we show that the *extended reachable sets*

*Received by the editors May 2, 2002; accepted for publication (in revised form) February 24, 2003; published electronically June 18, 2003. This research was partially supported by the M.U.R.S.T. project "Analysis and control of deterministic and stochastic evolution equations."
http://www.siam.org/journals/sicon/42-3/38528.html

†Dipartimento di Matematica Pura e Applicata, Università di Padova, via Belzoni 7, 35131 Padova, Italy (motta@math.unipd.it).

‡Dipartimento di Metodi e Modelli Matematici per le Scienze Applicate, Università di Padova, via Belzoni 7, 35131 Padova, Italy (sartori@dmsa.unipd.it).

are the closure of the original reachable sets and that the *extended minimum time and minimum energy functions*, denoted by $T_1(x,K)$ and $E_1(x,T)$, respectively, are the lower semicontinuous envelopes of $\hat{T}_1$ and $\hat{E}_1$, respectively. The extended problems are in fact equivalent to the original problems if some controllability around the origin is assumed.

In section 3 we begin by giving necessary and/or sufficient conditions for the upper semicontinuity, Lipschitz continuity, and Hölder continuity of the maps $\hat{T}_p$ and $\hat{E}_p$ for $p>1$, and of T_1 and E_1 in terms of *global topological properties* of the reachable sets and of the extended reachable sets, respectively. In subsection 3.2 we show that, for $p=1$, assuming in addition a controllability condition of the original system around the target, many of the previous properties hold also for the original functions $\hat{T}_1$ and $\hat{E}_1$ in the interior of their domains. In subsection 3.3 we show via a dynamic programming approach that assuming just local controllability around the origin is sufficient to yield the local Hölder continuity of $\hat{E}_p$ and $\hat{T}_p$ in the state variable x for $p>1$, while an example (Example 3.3) shows that this is not possible in the case $p=1$ neither for the function $\hat{T}_1$ nor for T_1.

In section 4 we show that controllable linear systems have reachable sets that verify all the global topological properties introduced in the previous section. In particular, this yields that $\hat{T}_1$ and $\hat{E}_1$ are at least continuous in the interior of their domains. In the nonlinear case we show that a classical local controllability condition used for systems with compact valued controls (see, e.g., [8]) implies the local Hölder continuity of $\hat{T}_p$ in the state variable x for $p>1$.

A huge literature treats the regularity of the minimum time and minimum energy functions, mainly under the assumption that the admissible controls are compact valued. To our knowledge, there are results on the regularity of the value functions $\hat{T}_p$ and $\hat{E}_p$ only for linear systems (also in infinite dimension) and for $p>1$ (see, e.g., [4], [6] and the references therein). In fact, in the case $p=1$, the Lipschitz continuity of T_1 has been proved by Rampazzo and Sartori [14] but under assumptions not verified by system $(S)_1$ if the target is a point. The bibliography that we give does not intend to be complete. Besides the articles to which we referred above, we mention here just those papers most related to our point of view. For nonlinear systems Petrov [13] gives the Lipschitz continuity of the minimum time function. For linear systems and for symmetric polysystems the Hölder continuity can be found in Liverovskii [9]. For nonlinear systems the problem is treated in the framework of more general issues on controllability by Bianchini and Stefani [2], Sussmann [18], and many others. All of these last results concern the case of compact valued controls. For linear systems and L^p-constraints on the controls, very sharp estimates on the energy needed to reach the origin as time approaches zero are given by Seidman [16] and by Seidman and Yong [17].

Notation. In what follows p' will denote the integer such that $\frac{1}{p}+\frac{1}{p'}=1$, with the usual convention that $p'=\infty$ and $\frac{1}{p'}=0$ if $p=1$; A° will denote the interior of a given subset $A\subset\mathbb{R}^n$ and $\bar{A}$ its closure; moreover, given a function $u: X\to[-\infty,+\infty]$, $X\subseteq\mathbb{R}^N$, u_* and u^* will denote, respectively, the lower and the upper semicontinuous envelopes.

2. Reachable sets. Minimum time and minimum energy functions.

2.1. Statement of the problems. For any integer $p\geq 1$ we consider the affine control system given by

$$(\hat{S})_p \qquad \dot{y}(t)=f(y(t))+\sum_{i=1}^{m} g_i(y(t))u_i(t), \quad t>0, \quad u\in L^p([0,+\infty[,\mathbb{R}^m),$$

where $f, g_1, \dots, g_m : \mathbb{R}^n \to \mathbb{R}^n$. Throughout the paper we assume that $f, g_1, \dots, g_m$ are locally Lipschitz continuous, sublinear functions. More precisely, if $\varphi \doteq f$ or $\varphi \doteq g_1, \dots, \varphi \doteq g_m$ and $N > 0$, there are some constants $L_\varphi \equiv L_{\varphi,N}$ and M_φ such that

$$\begin{aligned} &|\varphi(x_1) - \varphi(x_2)| \le L_\varphi |x_1 - x_2| \qquad \forall x_1, x_2 \text{ s.t. } |x_1|, |x_2| \le N \\ &|\varphi(x)| \le M_\varphi(1 + |x|) \qquad \forall x \in \mathbb{R}^n. \end{aligned} \tag{2.1}$$

Hence for any $x \in \mathbb{R}^n$ and any control u, we will denote by $y_x(\cdot, u)$ the unique solution to $(\hat{S})_p$ corresponding to u such that $y(0) = x$.

For any $p \ge 1$, $T \ge 0$, and $K \ge 0$ we denote by $\hat{\mathcal{U}}_p(T, K)$ the *set of admissible controls* given by

$$\hat{\mathcal{U}}_p(T, K) \doteq \left\{ u \in L^p([0, T], \mathbb{R}^m) : \int_0^T |u(t)|^p \, dt \le K^p \right\}$$

and define the *reachable set in time T and with energy K* as the subset of $\mathbb{R}^n$ given by

$$\hat{\mathcal{R}}_p(T, K) \doteq \left\{ x \in \mathbb{R}^n : \exists u \in \hat{\mathcal{U}}_p(T, K) \text{ s.t. } y_x(T, u) = 0 \right\}.$$

We also define the *minimum time function with p-energy K* and the *minimum p-energy function in time T* as

$$\hat{T}_p(x, K) \doteq \inf\{T > 0 : \ x \in \hat{\mathcal{R}}_p(T, K)\}, \quad \hat{E}_p(x, T) \doteq \inf\{K > 0 : \ x \in \hat{\mathcal{R}}_p(T, K)\},$$

respectively. For $p > 1$ we will prove that the reachable sets $\hat{\mathcal{R}}_p(T, K)$ are compact and an optimal control for the minimum time and minimum energy problems always exists. For $p = 1$ instead, the following simple example shows that even for linear systems, the reachable sets $\hat{\mathcal{R}}_1(T, K)$ might not be closed and also that minimizing sequences of trajectories can converge to a discontinuous function.

Example 2.1. Consider the system

$$\begin{cases} \dot{y}_1 = -y_2, \\ \dot{y}_2 = -u \end{cases}$$

with scalar control $u \in \hat{\mathcal{U}}_1(T, K)$ for $T, K > 0$. For any u, the solution is given by $(y_1, y_2)_{(x_1,x_2)}(t, u) = (x_1 - \int_0^t (t - s)u(s)\, ds, x_2 - \int_0^t u(s)\, ds)$. As shown in [5, Chap. III, Ex. 3], one has

$$\hat{\mathcal{R}}_1(T, K) = \left\{ (x_1, x_2) \in \mathbb{R}^2 : |Tx_2 - 2x_1| < TK, |x_2| \le K \right\}.$$

Therefore $\hat{\mathcal{R}}_1(T, K)$ is not closed. Moreover, there exists a minimizing sequence for the minimum time problem which does not converge to a solution of the system. Indeed, fix $P = (x_1, x_2) = (\hat{t}, 1)$, $\hat{t} > 0$. We have that $P \in \hat{\mathcal{R}}_1(T, 1)$ for every $T > \hat{t}$ in that the control

$$\hat{u}_T(t) = \begin{cases} (T - \hat{t})^{-1} - (T)^{-1} & \text{for } t \in [0, T - \hat{t}[\\ (\hat{t})^{-1} - T^{-1} & \text{for } t \in [T - \hat{t}, T[\end{cases}$$

belongs to $\hat{\mathcal{U}}_1(T, 1)$ and is such that $y_P(T, \hat{u}_T) = (0, 0)$, but P does not belong to $\hat{\mathcal{R}}_1(\hat{t}, 1)$. Consider now the sequence of controls $(u_n)_{n \in \mathbb{N}}$ where $u_n \doteq \hat{u}_{\hat{t} + \frac{1}{n}}$. It is clear

that $y_P(\hat{t}+\frac{1}{n},u_n)=(0,0)$. Notice though that $\lim_{n\to+\infty}(y_2)_P(\frac{1}{n},u_n)=0$ while for every n one has $(y_2)_P(0,u_n)=1$. Thus the limit function of our minimizing sequence is discontinuous.

Moreover, $\hat{T}_1$ is not lower semicontinuous on the closure of its domain, where the domain is given by

$$\cup_{K>0}\left(\left(\{(x_1,x_2)\in\mathbb{R}^2:\ |x_2|<K\}\cup\{(x_1,sgn(x_1)K):x_1\neq 0\}\right)\times\{K\}\right).$$

In fact, $\hat{T}_1((0,K),K)=+\infty$ while clearly $(\hat{T}_1)_*((0,K),K)=0$.

2.2. Extended system and extended problems. The facts addressed in Example 2.1 lead us to introduce for $p=1$ an *extended system,* whose corresponding *extended reachable sets* coincide with the closure of the original reachable sets and whose trajectories allow us to represent the (eventually discontinuous) limit function of sequences of solutions to $(\hat{S})_1$. In fact, in order to unify the proofs relative to the two cases $p>1$ and $p=1$, let us introduce the following *extended system* for any $p\geq 1$ (see also [15] and Remark 2.1 below):

$$(S)_p\quad\begin{cases}t'(s)=w_0^p(s),\\ k'(s)=|w(s)|^p,\\ y'(s)=f(y(s))w_0^p(s)+\displaystyle\sum_{i=1}^m g_i(y(s))w_i(s)w_0^{p-1}(s),\quad s\in[0,1],\end{cases}$$

where the controls $(w_0,w):[0,1]\to[0,+\infty[\times\mathbb{R}^m$ are measurable functions. For any control (w_0,w) and any $x\in\mathbb{R}^n$ we will denote by $(t(s),k(s),y_x(s))$ (or by $(t(s,w_0,w),$ $k(s,w_0,w),$ $y_x(s,w_0,w))$ if we want to specify the control) the solution to $(S)_p$ corresponding to (w_0,w) such that $(t(0),k(0),y(0))=(0,0,x)$. We will sometimes refer to such a solution as *forward solution* to $(S)_p$. The solution to $(S)_p$ where the third equation is replaced by $y'(s)=-f(y(s))w_0^p(s)-\sum_{i=1}^m g_i(y(s))w_i(s)w_0^{p-1}(s)$ such that $(t(0),k(0),y(0))=(0,0,x)$ will be denoted by $(t(s),k(s),$ $y_x^-(s))$, and we will refer to it as *backward solution* to $(S)_p$.

For any $p\geq 1$, $T\geq 0$, and $K\geq 0$, we denote by $\mathcal{U}_p(T,K)$ the *set of extended admissible controls* given by

$$\mathcal{U}_p(T,K)\doteq\left\{(w_0,w)\in L^p([0,1],[0,+\infty[\times\mathbb{R}^m):\ \int_0^1 w_0^p\,ds\leq T,\quad \int_0^1|w|^p\,ds\leq K^p\right\}$$

and define the *extended reachable set in time T and with energy K* as the subset of $\mathbb{R}^n$ given by

$$\mathcal{R}_p(T,K)\doteq\{x\in\mathbb{R}^n:\ \exists(w_0,w)\in\mathcal{U}_p(T,K)\text{ s.t. }y_x(1,w_0,w)=0\}.$$

We define also the *extended minimum time function with p-energy K* and the *extended minimum p-energy function in time T* as

$$T_p(x,K)\doteq\inf\{T>0:\ x\in\mathcal{R}_p(T,K)\},\quad E_p(x,T)\doteq\inf\{K>0:\ x\in\mathcal{R}_p(T,K)\},$$

respectively. We refer to the appendix for the technical propositions that relate the solution to $(\hat{S})_p$ to the solution of $(S)_p$.

Remark 2.1. In view of Proposition A.1 in the appendix, if $p > 1$, the (t, y)-components of the trajectories of $(S)_p$ are substantially only time reparametrizations of graphs of trajectories of $(\hat{S})_p$, in the sense that when $(w_0, w) = (0, w)$ on some set $[s_1, s_2]$ one has $y_x(\cdot, 0, w) =$ constant on $[s_1, s_2]$. Hence for any $T > 0$, $K > 0$ the reachable set $\mathcal{R}_p(T, K)$ coincides with $\hat{\mathcal{R}}_p(T, K)$, and $T_p(x, K)$ and $E_p(x, T)$ coincide with $\hat{T}_p(x, K)$ and $\hat{E}_p(x, T)$, respectively. In the case $p = 1$ instead, one has $y'(s) = \sum_{i=1}^m g_i(y(s))w_i(s)$ $\forall s \in [s_1, s_2]$. Hence the set of the extended trajectories is larger than the set of the graphs reparametrizations of trajectories of $(\hat{S})_1$. Notice that such extension of $(\hat{S})_1$ is equivalent to an *extension in measure* only in the special case of commutative control systems, i.e., when the Lie brackets $[g_i, g_j] \equiv 0 \ \ \forall i \neq j$. (See, e.g., [8] for an extension in measure in the special case of linear systems; see [3] and [10] for an approach to the general case which agrees with the one followed here.) We point out that system $(S)_p$ is introduced even in the case $p > 1$, not only to give the same proof for several results which are valid for any $p \geq 1$, but also because in the extended problems we can consider extended controls belonging to a compact set, as it follows from Proposition A.2 in the appendix.

2.3. New results. As anticipated before, in this subsection we prove that the reachable sets (the extended reachable sets in the case $p = 1$) are compact; that a bounded optimal control for the extended minimum time and minimum energy problems does always exist; and that the minimum time and the minimum energy functions (the extended functions in the case $p = 1$) are lower semicontinuous. Similar results were already proven in [6] only for $p > 1$ and (infinite dimensional) linear systems. Moreover, for $p = 1$ we show that the extended reachable sets coincide with the closure of the original sets and that the extended functions turn out to be the lower semicontinuous envelopes of the original functions.

PROPOSITION 2.1. *Let $p \geq 1$. For any T, $K \geq 0$ the set $\mathcal{R}_p(T, K)$ is compact. Furthermore, if $p = 1$, one has $\mathcal{R}_1(T, K) = \overline{\cap_{S>T} \hat{\mathcal{R}}_1(S, K)}$. Moreover, if $T > 0$, one has $\mathcal{R}_1(T, K) = \overline{\hat{\mathcal{R}}_1(T, K)}$.*

Proof. The assumptions on f and g_i, $i = 1, \ldots, m$, imply easily that $\mathcal{R}_p(T, K)$ is bounded. To prove that $\mathcal{R}_p(T, K)$ is closed, let us consider a sequence $(x_n)_n \subset \mathcal{R}_p(T, K)$ such that $\lim_n x_n = x$. For any x_n, let $(w_{0n}, w_n) \in \mathcal{U}_p(T, K)$ be a control such that $y_{x_n}(1, w_{0n}, w_n) = 0$. In view of Proposition A.2 in the appendix, we can assume that $|(w_{0n}, w_n)|^p \leq 2^p(K^p + T)$ a.e. Hence the sequence of extended trajectories $((t_n, k_n, y_n))_n$ (where $t_n \doteq t(\cdot, w_{0n}, w_n)$, $k_n \doteq k(\cdot, w_{0n}, w_n)$, $y_n \doteq y_{x_n}(\cdot, w_{0n}, w_n)$ $\forall n$) is equibounded and equi-Lipschitz. By the Ascoli–Arzelà theorem it has a subsequence uniformly converging to a function (t, k, y) such that $(t(0), k(0), y(0)) = (0, 0, x)$, $t(1) \leq T$, $k(1) \leq K^p$, and $y(1) = 0$. Moreover, by a well-known result (see, e.g., [8, Chap. IV]) (t, k, y) is in fact a trajectory of $(S)_p$ since for all $z \in \mathbb{R}^n$ the set $\{(w_0, f(z)w_0 + \sum_{i=1}^m g_i(z)w_i, |w|) : \ w_0 \geq 0, \ w \in \mathbb{R}^m, |(w_0, w)|^p \leq 2^p(K^p + T)\}$ is convex and compact. Then $x \in \mathcal{R}_p(T, K)$. (In the case of linear systems and for $p = 1$ a proof of the above result in terms of an approach in measure can be found in [8].)

The fact that $\mathcal{R}_1(T, K) = \overline{\cap_{S>T} \hat{\mathcal{R}}_1(S, K)}$ can be shown using the same arguments as in [14, Theorem 3.1]. In order to prove the last statement, it suffices to prove that the inclusion $\mathcal{R}_1(T, K) \subset \overline{\hat{\mathcal{R}}_1(T, K)}$ holds for any $T > 0$. Let $x \in \mathcal{R}_1(T, K)$, let $(w_0, w) \in \mathcal{U}_1(T, K)$ be a control such that $|(w_0, w)|^p \leq 2^p(T + K^p)$, and $y_0^-(1, w_0, w) = x$. For any n let us define $w_{0n} \doteq \left(w_0^p + \frac{1}{n}\right)^{1/p}$, and let $\sigma_n \doteq \sup\{\sigma \in [0, 1] : \ \int_0^\sigma w_{0n}^p(s)\, ds \leq T\}$. Hence the backward trajectories $(t_n, k, y_n^-)(\cdot) = (t, k, y_0^-)(\cdot, w_{0n}, w)$ and $(t_{\sigma_n}, k, y_{\sigma_n}^-)(\cdot) = (t, k, y_0^-)(\cdot, w_{0n}\chi_{[0,\sigma_n]}, w)$ of $(S)_1$ satisfy the

estimates

$$|t_{\sigma_n}(s) - t(s)| \le |t_{\sigma_n}(s) - t_n(s)| + |t_n(s) - t(s)| \le \int_{\sigma_n}^1 w_{0n}^p(s)\,ds + \frac{1}{n} \le \frac{2}{n},$$
$$|y^-_{\sigma_n}(1) - x| \le |y^-_{\sigma_n}(1) - y^-_n(1)| + |y^-_n(1) - x| \le \omega\left(\frac{1}{n}\right) \quad \forall s \in [0,1],$$

where $\omega : [0,+\infty[\to [0,+\infty[$ is an increasing function, continuous at 0, such that $\omega(0) = 0$. This concludes the proof in that $x_n \doteq y^-_{\sigma_n}(1) \in \hat{\mathcal{R}}_1(T,K)$ by definition and $\lim_{n\to+\infty} x_n = x$. □

PROPOSITION 2.2. *Let $T \ge 0$, $K \ge 0$, and $p \ge 1$. Then for any $x \in \mathcal{R}_p(T,K)$ there exists a* bounded *optimal control (w_0, w) for the* extended minimum time problem *and a* bounded *optimal control $(\tilde{w}_0, \tilde{w})$ for the* extended minimum energy problem.

Proof. We show the existence of a bounded optimal control only for the extended minimum time problem, the proof for the extended minimum energy problem being analogous. Let $p \ge 1$, and for $x \in \mathcal{R}_p(T,K)$ let $((w_{0n}, w_n))_n$ be a minimizing sequence of controls, i.e., assume that the backward trajectories $(t_n, k_n, y_n^-)(\cdot) \doteq (t,k,y_0^-)(\cdot, w_{0n}, w_n)$ of $(S)_p$ satisfy

$$\lim_n t_n(1) = T_p(x,K), \qquad k_n(1) \le K^p, \quad y_n^-(1) = x \quad \forall n.$$

On the basis of Proposition A.2 in the appendix, we can suppose that $|(w_{0n}, w_n)|^p \le 2^p(T_p(x,K) + K^p) + 1$. At this point, the same arguments used in the proof of Proposition 2.1 allow us to conclude that there exists a subsequence of $(t_n, k_n, y_n^-)(\cdot)$ which converges uniformly to a backward solution (t,k,y^-) of $(S)_p$ associated to a bounded admissible control $(w_0, w) \in \mathcal{U}_p(T_p(x,K),K)$, optimal in that $\int_0^1 w_0^p(s)\,ds = T_p(x,K)$ (and $y^-(1) = x$, $\int_0^1 |w|^p\,ds \le K^p$). □

Remark 2.2. In Proposition 2.2 we proved the existence of an extended optimal control (w_0, w). If $p > 1$, in fact, one could also prove the existence of an optimal control in the original setting (either directly or using the arguments of Remark 2.1). If $p = 1$ instead, as already shown in Example 2.1, an optimal control for the original problem might not exist.

For any $p \ge 1$, $T, K > 0$ we define the sets

$$\mathcal{R}_p(K) \doteq \cup_{T\ge 0}\mathcal{R}_p(T,K), \qquad \mathcal{S}_p(T) \doteq \cup_{K\ge 0}\mathcal{R}_p(T,K), \tag{2.2}$$
$$\hat{\mathcal{R}}_1(K) \doteq \cup_{T\ge 0}\hat{\mathcal{R}}_1(T,K), \qquad \hat{\mathcal{S}}_1(T) \doteq \cup_{K\ge 0}\hat{\mathcal{R}}_1(T,K).$$

Hence the domains of the functions T_p, E_p, $\hat{T}_1$, and $\hat{E}_1$ are given, respectively, by

$$Dom(T_p) = \cup_{K>0}(\mathcal{R}_p(K) \times \{K\}), \quad Dom(E_p) = \cup_{T>0}(\mathcal{S}_p(T) \times \{T\}),$$
$$Dom(\hat{T}_1) = \cup_{K>0}(\hat{\mathcal{R}}_1(K) \times \{K\}), \quad Dom(\hat{E}_1) = \cup_{T>0}(\hat{\mathcal{S}}_1(T) \times \{T\}).$$

As a consequence of Propositions 2.1 and 2.2 we will prove that the functions T_p and E_p are lower semicontinuous. We remark that this holds not only in the state variable x but in their whole domains.

THEOREM 2.1. *For any $p \ge 1$, the functions $T_p : \overline{Dom(T_p)} \to [0,+\infty]$ and $E_p : \overline{Dom(E_p)} \to [0,+\infty]$ are lower semicontinuous. Furthermore, in the case $p = 1$ one has that $(\hat{E}_1)_* = E_1$ and $(\hat{T}_1)_* = T_1$.*

Proof. Let $p \geq 1$. We prove only the statements for T_p, the proofs for E_p being analogous. In order to show that T_p is lower semicontinuous, let us fix $(x, K) \in \overline{Dom(T_p)}$. We argue by contradiction and suppose that there are $T < T_p(x, K)$ and $(x_n, K_n) \in \overline{Dom(T_p)}$ such that $T_p(x_n, K_n) < T$, and $\lim_n (x_n, K_n) = (x, K)$. Hence for all n sufficiently large one has $K_n \leq K + 1$ and $x_n \in \mathcal{R}_p(T, K_n) \subset \mathcal{R}_p(T, K+1)$. By Proposition 2.2, there exist optimal controls $(w_{0n}, w_n) \in \mathcal{U}_p(T, K_n)$ uniformly bounded, e.g., by $2^p(T + (K+1)^p)$, such that the backward trajectories $(t_n, k_n, y_n^-)(\cdot) = (t, k, y_0^-)(\cdot, w_{0n}, w_n)$ to $(S)_p$ satisfy

$$t_n(1) \leq T, \qquad k_n(1) \leq K_n, \qquad y_n^-(1) = x_n.$$

As in Proposition 2.1, known theorems imply that there is a subsequence of $(t_n, k_n, y_n^-)(\cdot)$ uniformly converging to a backward trajectory of $(S)_p$ steering 0 to x in time not greater than T and with energy not greater than K, in contradiction with the hypothesis that $T < T_p(x, K)$.

In order to prove that $(\hat{T}_1)_* = T_1$, we observe that $T_1 \leq (\hat{T}_1)_*$ follows from the inequality $T_1 \leq \hat{T}_1$ and from the lower semicontinuity of T_1. The reverse inequality, instead, is an easy consequence of the fact that $\mathcal{R}_1(T, K) = \overline{\cap_{S>T} \hat{\mathcal{R}}_1(S, K)}$ for each $T \geq 0, K \geq 0$. □

Remark 2.3. The fact that T_p and E_p are lower semicontinuous functions for any $p \geq 1$ allowed us to characterize them together with their domains as the unique lower semicontinuous solutions, in the viscosity sense, of suitable boundary value problems (see [12]). Incidentally, the equalities $(\hat{T}_1)_* = T_1$ and $(\hat{E}_1)_* = E_1$ follow also as a by-product of the results in [12].

We end this section by stating the following last remarkable property of the reachable sets, which is well known if $p = +\infty$.

PROPOSITION 2.3. *For any $p \geq 1$, the set valued map $(T, K) \mapsto \mathcal{R}_p(T, K)$ is a continuous map from $[0, +\infty[\times[0, +\infty[$ to the space of compact subsets of $\mathbb{R}^n$, endowed with the Hausdorff distance.*

Proof. Let (T_0, K_0), $(T, K) \in [0, +\infty[\times[0, +\infty[$. If $T_0 \leq T$ and $K_0 \leq K$, one has $\mathcal{R}_p(T_0, K_0) \subset \mathcal{R}_p(T, K)$ and $\mathcal{R}_p(T_0, K_0) \subset B(\mathcal{R}_p(T, K), \varepsilon)$ $\forall \varepsilon > 0$. If $T_0 > T$ or $K_0 > K$, for any $x \in \mathcal{R}_p(T_0, K_0)$ let $(w_0, w) \in \mathcal{U}_p(T_0, K_0)$ be a control such that $|(w_0, w)|^p \leq 2^p(T_0 + K_0^p)$, and $y_0^-(1, w_0, w) = x$. Let us define the values $\sigma_1 \doteq \sup\{\sigma \in [0,1] : \int_0^\sigma w_0^p(s)\, ds \leq T\}$ and $\sigma_2 \doteq \sup\{\sigma \in [0,1] : \int_0^\sigma |w(s)|^p\, ds \leq K^p\}$. Hence the backward trajectories $(t, k, y^-)(\cdot) = (t, k, y_0^-)(\cdot, w_0, w)$ and $(t_{\sigma_1}, k_{\sigma_2}, y^-_{\sigma_1,\sigma_2})(\cdot) = (t, k, y_0^-)(\cdot, w_0\chi_{[0,\sigma_1]}, w\chi_{[0,\sigma_2]})$ of $(S)_p$ satisfy the estimates

$$|t_{\sigma_1}(s) - t(s)| \leq \int_{\sigma_1}^1 w_0^p(s)\, ds \leq |T_0 - T|,$$

$$|k_{\sigma_2}(s) - k(s)| \leq \int_{\sigma_2}^1 |w(s)|^p\, ds \leq |K_0^p - K^p| \quad \forall s \in [0,1],$$

$$|y^-_{\sigma_1,\sigma_2}(1) - x| \leq \omega(|K_0^p - K^p| + |T_0 - T|),$$

where $\omega : [0, +\infty[\to [0, +\infty[$ is an increasing function, continuous at 0, such that $\omega(0) = 0$. This concludes the proof in that $\bar{x} \doteq y^-_{\sigma_1,\sigma_2}(1) \in \mathcal{R}_p(T, K)$ by definition and $\mathcal{R}_p(T_0, K_0) \subset B(\mathcal{R}_p(T, K), \omega([K_0^p - K^p] + [T_0 - T]))$. The proof is completed by switching (T_0, K_0) with (T, K). □

3. Main results. We split this section into three subsections. In subsection 3.1 we begin by showing that the upper semicontinuity and the Hölder continuity of $x \mapsto$

$T_p(x,K)$ and of $x \mapsto E_p(x,T)$ are equivalent to certain global topological properties of the reachable sets (see Theorems 3.1, 3.2). Furthermore, we give sufficient conditions for the upper semicontinuity of $T_p(x,K)$ and $E_p(x,T)$ in the pair of variables (x,K) and (x,T), respectively (see Theorem 3.3). After that we characterize the reachable sets and their boundaries by means of T_p and E_p (see Propositions 3.1, 3.2), and we get also a maximality property for T_p and E_p (see Proposition 3.3).

In subsection 3.2 we deal with the critical case $p = 1$. Here we prove that the original functions coincide with the extended functions under a (very natural) local controllability assumption (see Lemma 3.1 and Theorem 3.4). Therefore under such an assumption all the regularity results obtained in subsection 3.1 in the extended setting hold also for $\hat{E}_1$, $\hat{T}_1$, and $\hat{\mathcal{R}}_1(T,K)$ (see Corollaries 3.1, 3.2).

In subsection 3.3 we consider only the case $p > 1$, and we show that local controllability assumptions are sufficient for the local Hölder continuity of $x \mapsto E_p(x,T)$ and $x \mapsto T_p(x,K)$ (see Theorems 3.5, 3.6).

3.1. Global topological properties and regularity results for $p \geq 1$. Let us introduce and briefly comment on the *global topological properties* of the reachable sets that we will use in what follows.

(C.1) Fix $p \geq 1$ and $T > 0$. Then

$$\mathcal{R}_p(T,K) \subset \mathcal{R}_p^\circ(T, K+H) \qquad \forall K \geq 0 \ \forall H > 0.$$

(C.2) Fix $p \geq 1$ and $T > 0$. Then there exist $C_2(T)$ and $\bar{\delta} > 0$ such that

$$B(\mathcal{R}_p(T,K), C_2(T)H) \subset \mathcal{R}_p(T, K+H) \quad \forall K \geq 0,\ 0 \leq H \leq \bar{\delta}.$$

(C.3) Fix $p \geq 1$ and $K > 0$. Then

$$\mathcal{R}_p(T,K) \subset \mathcal{R}_p^\circ(T+S, K) \qquad \forall T \geq 0 \ \forall S > 0.$$

(C.4) Fix $p \geq 1$ and $K > 0$. Then there exist $\alpha \geq 1$ (*independent of* K), $C_4(K)$, and $\bar{\delta} > 0$ such that

$$B(\mathcal{R}_p(T,K), C_4(K)S^\alpha) \subset \mathcal{R}_p(T+S, K) \quad \forall T \geq 0,\ 0 \leq S \leq \bar{\delta}.$$

(C.5) Fix $p \geq 1$ and $K > 0$. Then

$$\mathcal{R}_p(T,K) \cap \mathcal{R}_p^\circ(K) \subset \mathcal{R}_p^\circ(T+S, K) \qquad \forall T \geq 0 \ \forall S > 0,$$

where $\mathcal{R}_p(K)$ is defined as in (2.2).

(C.6) Fix $p \geq 1$. Then for any T, $K > 0$ one has that

$$x \in \mathcal{R}_p^\circ(T,K) \Longrightarrow \exists \varepsilon > 0 \ \text{ s.t. } \ x \in \mathcal{R}_p^\circ(T-\varepsilon, K-\varepsilon).$$

Taking into account that the reachable sets depend here on *two* variables, conditions (C.1) and (C.3) are the natural generalization of the classical "expansion property" of the reachable sets defined, e.g., in [7]. Loosely speaking, they say that $\mathcal{R}_p(T,K)$ expands "well" if one increases either the variable K or the variable T at disposal. Conditions (C.2) and (C.4) are stronger than (C.1) and (C.3), respectively, giving also an estimate on the rate of such an expansion.

Condition (C.5) is a weaker version of (C.3), coinciding with it when the set $\mathcal{R}_p(K)$ is open. We are led to introduce it by the fact that in the case $p = 1$ condition (C.3) may be too strong a requirement (see Example 3.1 below). Condition (C.5)

instead is fulfilled, for instance, as soon as the reachable sets are convex and (C.1) holds (see Proposition 3.2). Hence in particular it always holds for linear controllable systems (see section 4). Incidentally, Example 3.2 shows that (C.5) can hold even if the reachable sets are not convex (and (C.3) does not hold).

In the classical minimum time problem for linear systems with compact valued controls, the convexity of the reachable sets yields the so-called maximality property, that is, for all points belonging to the boundary of the reachable set at time T the minimum time turns out to be equal to T (see, e.g., [7]). In what follows we will prove that, assuming (C.6), similar maximality properties for E_p and T_p hold also for our system. Notice that when the reachable sets $\mathcal{R}_p(T,K)$ are convex, condition (C.6) turns out to be verified in view of Proposition 2.3. This fact can be proved exactly as for $p=+\infty$ (see, e.g., [7]). Hence in particular (C.6) is always fulfilled if the control system is linear. However, Example 3.2 again shows that it can be fulfilled even if the reachable sets are not convex.

Example 3.1. Let us consider the (controllable) linear system $x'=\lambda x+u$, where $\lambda\in\mathbb{R}$, x, $u\in\mathbb{R}^n$.

(a) Let $p>1$ and consider $u\in\hat{\mathcal{U}}_p(T,K)$ for some T, $K>0$. It is not difficult to show that if $\lambda\neq 0$, one has

$$\mathcal{R}_p(T,K)=\hat{\mathcal{R}}_p(T,K)=\left\{x\in\mathbb{R}^n:\ |x|\le K\left(\frac{1-e^{-\lambda T p'}}{\lambda p'}\right)^{\frac{1}{p'}}\right\},$$

while if $\lambda=0$, one gets

$$\mathcal{R}_p(T,K)=\hat{\mathcal{R}}_p(T,K)=\left\{x\in\mathbb{R}^n:\ |x|\le KT^{\frac{1}{p'}}\right\}.$$

Therefore conditions (C.2) and (C.3) turn out to be always verified, while (C.4) is in force only in the case $\lambda<0$.

(b) Let $p=1$ and $u\in\hat{\mathcal{U}}_1(T,K)$ for some T, $K>0$. In this case one recovers that

$$\mathcal{R}_1(T,K)=\overline{\hat{\mathcal{R}}_1(T,K)}=\begin{cases}\{x\in\mathbb{R}^n:\ |x|\le e^{-\lambda T}K\} & \text{if } \lambda<0,\\ \{x\in\mathbb{R}^n:\ |x|\le K\} & \text{if } \lambda\ge 0.\end{cases}$$

Hence conditions (C.1) and (C.2) are always verified, while conditions (C.3) and (C.4) hold only in the case $\lambda<0$.

This example suggests that, at least for linear controllable systems, conditions (C.1), (C.2) for $p\ge 1$, and condition (C.3) in the case $p>1$, should be verified (see also section 4), while (C.3) for $p=1$ and condition (C.4) for all $p\ge 1$ are in fact very strong.

Example 3.2. Let us consider in $\mathbb{R}^2$ the system

$$\begin{cases}\dot{x}=-yu+(x+1)v,\\ \dot{y}=(x+1)u+yv\end{cases}$$

with $(u,v)\in\mathcal{U}_p(T,K)$ for T, $K>0$, and $p\ge 1$. With an obvious change of coordinates one can study the system

$$\begin{cases}\dot{x}=-yu+xv,\\ \dot{y}=xu+yv\end{cases}$$

with target $(-1,0)$, which in polar coordinates is given by $\dot\rho = \rho v$, $\dot\theta = u$. In these coordinates for each $T>0$ and $K>0$ the reachable set is given by

$$\mathcal{R}_p(T,K) - \{(-1,0)\} = \bigcup_{0\le k\le K^p} \left\{ (\rho,\theta) : |\theta| \le kT^{\frac{1}{p'}}, e^{-(K^p-k)^{\frac{1}{p}}T^{\frac{1}{p'}}} \le \rho \le e^{(K^p-k)^{\frac{1}{p}}T^{\frac{1}{p'}}} \right\}.$$

Therefore $\mathcal{R}_p(T,K)$ is not convex for every $T,K\ge 0$, but still condition (C.6) is verified for all $p\ge 1$. If $p>1$, condition (C.3) is also verified, while if $p=1$, only the weaker condition (C.5) is fulfilled. Incidentally, notice that in the case of controls (u,v) such that $|u|\le 1, |v|\le 1$, and without L^p-constraints, (C.6) is not verified (see, e.g., [1]).

THEOREM 3.1. *Fix $T>0$. For any $p\ge 1$, the function $E_p(\cdot,T)$ is upper semicontinuous in the set $\mathcal{S}_p(T)$ (defined as in* (2.2)*) and the set $\mathcal{S}_p(T)$ is open if and only if condition* (C.1) *is verified.*

Furthermore, condition (C.2) *is a necessary and sufficient condition for E_p to verify the inequality*

$$|E_p(x_1,T) - E_p(x_2,T)| \le |x_1-x_2|/C_2(T) \qquad \forall x_1,x_2\in\mathbb{R}^n,$$

where $C_2(T)$ is the same as in (C.2).

Proof. Let $x\in\mathcal{S}_p(T)$. In view of the existence of an optimal control for the minimum energy problem stated in Proposition 2.2, $x\in\mathcal{R}_p(T,E_p(x,T))$. Condition (C.1) easily implies that $\mathcal{S}_p(T)$ is open and it is verified if and only if for any $\varepsilon>0$ there is some $\delta>0$ such that $B(x,\delta)\subset\mathcal{R}_p(T,E_p(x,T)+\varepsilon)$ or equivalently if and only if $E_p(y,T)\le E_p(x,T)+\varepsilon \ \ \forall y\in B(x,\delta)$, that is, $E_p(\cdot,T)$ is upper semicontinuous in $\mathcal{S}_p(T)$. Notice that (C.2) implies $\mathcal{S}_p(T)=\mathbb{R}^n$. Furthermore, let $x_1,x_2\in\mathbb{R}^n$ be such that $|x_2-x_1|\le C_2(T)\bar\delta$, where $C_2(T)$ and $\bar\delta$ are the same as in (C.2), let $K \doteq E_p(x_1,T)$, and suppose that $E_p(x_2,T)>K$. In view of Proposition 2.2, $x_1\in\mathcal{R}_p(T,K)$ and, if (C.2) is verified, setting $H\doteq |x_1-x_2|/C_2(T)$ one has that $x_2\in\mathcal{R}_p(T,K+H)$. Hence $E_p(x_2,T)\le E_p(x_1,T)+|x_1-x_2|/C_2(T)$ and the statement of the second sufficient condition holds. The proof of the necessity can be obtained by reversing the previous arguments.

Finally, it is easy to extend these results to all $x_1,x_2\in\mathbb{R}^n$. □

THEOREM 3.2. *Fix $K>0$. For any $p\ge 1$ the function $T_p(\cdot,K)$ is upper semicontinuous in the set $\mathcal{R}_p(K)$ (defined as in* (2.2)*) and the set $\mathcal{R}_p(K)$ is open if and only if condition* (C.3) *is verified.*

Furthermore, condition (C.4) *is a necessary and sufficient condition for T_p to verify the inequality*

$$|T_p(x_1,K) - T_p(x_2,K)| \le \left(\frac{|x_1-x_2|}{C_4(K)}\right)^{1/\alpha}$$

$\forall x_1,x_2\in\mathbb{R}^n$ such that $|x_1-x_2|\le C_4(K)\bar\delta^\alpha$, where $C_4(K)$, $\bar\delta$, and α are the same as in (C.4).

We omit the proof, since it is completely analogous to the proof of Theorem 3.1.

PROPOSITION 3.1. *Let $p\ge 1$ and $T,\ K>0$.*

(a) *One has*

$$\mathcal{R}_p(T,K) = \{x\in\mathbb{R}^n :\ T_p(x,K)\le T\} = \{x\in\mathbb{R}^n :\ E_p(x,T)\le K\}.$$

(b) (Characterization by means of E_p.) *If one assumes* (C.1) *and* (C.6), *one has*

$$\begin{aligned}\mathcal{R}_p^\circ(T,K) &= \{x\in\mathbb{R}^n:\ E_p(x,T)<K\},\\ \partial\mathcal{R}_p(T,K) &= \{x\in\mathbb{R}^n:\ E_p(x,T)=K\},\\ \mathcal{S}_p^\circ(T) &= \mathcal{S}_p(T)=\{x\in\mathbb{R}^n:\ E_p(x,T)<+\infty\}.\end{aligned}$$

(c) (Characterization by means of T_p.) *If one assumes* (C.3) *and* (C.6), *one has*

$$\begin{aligned}\mathcal{R}_p^\circ(T,K) &= \{x\in\mathbb{R}^n:\ T_p(x,K)<T\},\\ \partial\mathcal{R}_p(T,K) &= \{x\in\mathbb{R}^n:\ T_p(x,K)=T\},\\ \mathcal{R}_p^\circ(K) &= \mathcal{R}_p(K)=\{x\in\mathbb{R}^n:\ T_p(x,K)<+\infty\}.\end{aligned}$$

(d) *If* (C.1) *is assumed and the reachable sets are convex, then*

$$\mathcal{R}_p^\circ(K)=\{x\in\mathbb{R}^n : T_p(x,K)<+\infty \text{ and } E_p(x,T)<K \quad \forall T>T_p(x,K)\}. \tag{3.1}$$

(e) *If* (C.6) *is assumed, the relation* (3.1) *is equivalent to* (C.5).

Proof. Statement (a) is immediate because of the existence of extended optimal controls. The inclusions $\mathcal{R}_p^\circ(T,K)\subset\{x\in\mathbb{R}^n:\ E_p(x,T)<K\}$ and $\mathcal{R}_p^\circ(T,K)\subset\{x\in\mathbb{R}^n:\ T_p(x,K)<T\}$ follow by (C.6). To prove the first equality in (b) let us observe that (C.1) implies that

$$\mathcal{R}_p(S,H)\subset\mathcal{R}_p(T,H)\subset\mathcal{R}_p^\circ(T,K)\quad\forall 0<S<T\ \text{ and }\ \forall 0<H<K.$$

Let $x\in\mathbb{R}^n$ be such that $K'\doteq E_p(x,T)<K$. In view of the above inclusions, it suffices to show that

$$x\in\mathcal{R}_p(S,H)\qquad\text{for some } S<T \text{ and } H<K. \tag{3.2}$$

If the optimal control (w_0,w) associated with $E_p(x,T)$ is such that $T'\doteq\int_0^1 w_0^p\,ds<T$, (3.2) is verified for $S=T'$, and $H=K'$. Otherwise, i.e., in the case $T'=T$, then $x\in\mathcal{R}_p(T,K')$ and by (C.1) $x\in\mathcal{R}_p^\circ(T,H)$ for any $K'<H<K$. Hence by (C.6) there exists some $\varepsilon>0$ such that $x\in\mathcal{R}_p(T-\varepsilon,H-\varepsilon)$ and (3.2) is verified for $S=T-\varepsilon$ and $H=H$. The second statement of (b) follows from (a) and from the first part of (b), in view of the fact that the sets $\mathcal{R}_p(T,K)$ are closed. The third statement is a straightforward consequence of (C.1). All the equalities in (c) can be proved in a similar way. To prove (3.1), notice that the inclusion $\mathcal{R}_p^\circ(K)\supset\cup_{T>0}\mathcal{R}_p^\circ(T,K)$ is always verified. If the sets $\mathcal{R}_p(T,K)$ are convex, the converse inclusion is a consequence of the fact that they are closed. Otherwise, since from the previous characterization of $\mathcal{R}_p^\circ(T,K)$ it follows that $\cup_{T>0}\mathcal{R}_p^\circ(T,K)=\{x\in\mathbb{R}^n : T_p(x,K)<+\infty \text{ and } E_p(x,T)<K \text{ for some } T>T_p(x,K)\}$, thus, to conclude, it remains to show that $T_p(x,K)=\inf\{T>0:\ E_p(x,T)<K\}$. Suppose that $\bar T\doteq\inf\{T>0:\ E_p(x,T)<K\}>T_p(x,K)$. Since the function E_p is lower semicontinuous and decreasing in T, $\bar T$ is in fact a minimum and $K'\doteq E_p(x,\bar T)<K$. Thus $x\in\mathcal{R}_p(\bar T,K')$ and (C.1) implies that $x\in\mathcal{R}_p^\circ(\bar T,K)$. By (C.6), $x\in\mathcal{R}_p(\bar T-\varepsilon,K-\varepsilon)$ for some $\varepsilon>0$, so that $E_p(x,S)\le K-\varepsilon$ for all $S\in[\bar T-\varepsilon,\bar T[$, in contradiction with the definition of $\bar T$.

The implication (3.1) $\Longrightarrow$ (C.5) is clear. Conversely, condition (C.5) means that for all $x\in\mathcal{R}_p^\circ(K)$ one has $x\in\mathcal{R}_p^\circ(T,K)\ \ \forall T>T_p(x,K)$, and (C.6) yields that $x\in\mathcal{R}_p(T-\varepsilon,K-\varepsilon)$ for some $\varepsilon>0$. Hence $E_p(x,T)<K\ \forall T>T_p(x,K)$ and the equivalence is proved. ∎

PROPOSITION 3.2. *Let $p \geq 1$.*

(a) *If* (C.1) *and* (C.6) *are assumed, $Dom(E_p)$ is an open set.*

(b) *If* (C.3) *and* (C.6) *are assumed, $Dom(T_p)$ is an open set.*

(c) *If* (C.1) *is assumed and either the reachable sets are convex or* (C.6) *and* (C.5) *are assumed, then $Dom(T_p)$ is not necessarily open but one has that*

$$Dom(T_p)^\circ = \cup_{K>0} \mathcal{R}_p^\circ(K) \times \{K\},$$

where $\mathcal{R}_p^\circ(K)$ is given in (3.1).

Proof. We prove only (c), the proofs of (a) and (b) being similar and, in fact, easier. We begin by showing that for all $p \geq 1$, $\cup_{K>0} \mathcal{R}_p^\circ(K) \times \{K\}$ is an open set. Indeed, given $x \in \mathcal{R}_p^\circ(K)$, by (C.6) and Proposition 3.1 it follows that there exist ε and $\delta > 0$ such that $B(x,\delta) \subset \mathcal{R}_p^\circ(K-\varepsilon)$. Hence $(y,H) \in R_p^\circ(H) \times \{H\}$ $\forall (y,H) \in B(x,\delta) \times]K-\varepsilon, +\infty[$. This concludes the proof if the sets $\mathcal{R}_p(H)$ are open; otherwise, it remains to show that $Dom(T_p)^\circ \subset \cup_{K>0} \mathcal{R}_p^\circ(K) \times \{K\}$. Let $(x,K) \in Dom(T_p)^\circ$. Since $B((x,K),\delta) \subset Dom(T_p)$ for some $\delta > 0$, we have, in particular, that $x \in \mathcal{R}_p(K-\delta)$, so that $E_p(x,T) < K$ for some T. In view of Proposition 3.1, this is equivalent to claim that $x \in \mathcal{R}_1^\circ(K)$. ∎

As already remarked, the controllability assumptions (C.1)–(C.4) yield continuity results only for the maps $x \mapsto E_p(x,T)$ and $x \mapsto T_p(x,K)$. However, the dependence of $E_p(x,T)$ and $T_p(x,K)$ on the scalar variables T and K, respectively, is not trivial. For instance, the Hamilton–Jacobi–Bellman equations associated with E_p and T_p involve the derivatives $\partial E_p/\partial T$ and $\partial T_p/\partial K$, respectively, as suggested in the case $p > 1$ by the dynamic programming principles (TDPP) and (EDPP) stated in Proposition 3.4 below (see also Remark 2.3). Together with condition (C.5) and (C.1), condition (C.6) yields the continuity of the minimum time and of the minimum energy function on its whole domain, respectively, as shown in Theorem 3.3, and also the maximality properties stated in Proposition 3.3.

THEOREM 3.3. *Let $p \geq 1$.*

(a) *Assume* (C.5) *and* (C.6). *Then the minimum time function $T_p : Dom(T_p)^\circ \to [0, +\infty[$ is upper semicontinuous.*

(b) *Assume* (C.1) *and* (C.6). *Then the minimum energy function $E_p : Dom(E_p) \to [0, +\infty[$ is upper semicontinuous.*

Proof. Let $(x,K) \in Dom(T_p)^\circ$. Propositions 3.1 and 3.2 imply that $x \in \mathcal{R}_p^\circ(T_p(x,K) + \varepsilon, K)$ for any $\varepsilon > 0$, and by (C.6) it follows that there exists $\varepsilon' > 0$ such that $x \in \mathcal{R}_p^\circ(T_p(x,K) + \varepsilon - \varepsilon', K - \varepsilon')$, so that $B(x,\delta) \subset \mathcal{R}_p(T_p(x,K) + \varepsilon - \varepsilon', K - \varepsilon')$ for some $\delta > 0$. Hence $T_p(y,H) \leq T_p(x,K) + \varepsilon$ $\forall y \in B(x,\delta)$ $\forall H > K - \varepsilon'$ and this concludes the proof. The proof concerning E_p follows the same lines. ∎

PROPOSITION 3.3. *Let $p \geq 1$, and assume* (C.1), (C.6).

(a) *Fix $K > 0$. Then*

$$E_p(x, T_p(x,K)) = K \qquad \forall x \in \mathcal{R}_p(K) \setminus \mathcal{R}_p(0,K).$$

(b) *Assume* (C.3) *and fix $T > 0$. Then*

$$T_p(x, E_p(x,T)) = T \qquad \forall x \in \mathcal{S}_p(T) \setminus \mathcal{R}_p(T,0).$$

(c) *Assume* (C.5) *and fix $T > 0$. Then*

$$T_p(x, E_p(x,T)) = T \qquad \forall x \in \mathcal{R}_p^\circ(K) \setminus \mathcal{R}_p(T,0), \quad \text{where } K \doteq E_p(x,T).$$

Proof. Let $p \geq 1$, and let (x,K) be such that $T_p(x,K) > 0$. By the existence of the optimal control for the minimum time problem, we can assume that $E_p(x, T_p(x,K)) \leq$

K. Suppose that $E_p(x, T_p(x,K)) < K$. Since by definition one has that $E_p(x,T) \geq K$ $\forall T \in]0, T_p(x,K)[$, in view of Theorem 3.3 we find a contradiction with the fact that E_p is upper semicontinuous and decreasing in T. This yields statement (a). Since the function T_p is decreasing in K, the proof of (b) and (c) follows in an analogous way from Proposition 3.1(c) and Theorem 3.3. □

By the results in the appendix it follows that $\mathcal{R}_p(0,K) = \{0\}$ $\forall p > 1$. For $p = 1$ instead, the set $\mathcal{R}_1(0,K)$ is in general nontrivial, as shown, e.g., by Example 3.1(b). In this case the maximality property (a) above fails if $T_1(x,K) = 0$, that is, for $x \in \mathcal{R}_1(0,K)$. Indeed, $T_1(x,K') = 0$ $\forall K' \geq K$, so that $E_1(x, T_1(x,K')) = K < K'$ $\forall K' > K$. Analogous remarks hold for (b) and (c). Notice that under the assumptions made on the drift f in section 2, for any $p \geq 1$ the set $\mathcal{R}_p(T,0) = \{0\}$ if $f(0) = 0$.

3.2. Regularity results for $\hat{E}_1$ and $\hat{T}_1$. Our goal in this subsection is to prove that the original minimum time and minimum energy functions coincide, in fact, with the extended functions under the following *local controllability condition*:

(C.7) $\exists \hat{\varepsilon} > 0$ such that $\forall \varepsilon < \hat{\varepsilon}:$ $B(0,\delta) \subset \hat{\mathcal{R}}_1(\varepsilon,\varepsilon)$ for some $\delta > 0$.

LEMMA 3.1. *Let $p = 1$ and assume* (C.6) *and* (C.7). *Then*

$$\hat{\mathcal{R}}_1^\circ(T,K) = \mathcal{R}_1^\circ(T,K) \qquad \forall T, K > 0.$$

Proof. The inclusion $\hat{\mathcal{R}}_1^\circ(T,K) \subset \mathcal{R}_1^\circ(T,K)$ is trivial. To prove the converse inclusion, fix $x \in \mathcal{R}_1^\circ(T,K)$. Then (C.6) implies that $x \in \mathcal{R}_1^\circ(T-2\varepsilon, K-2\varepsilon)$ for some positive $\varepsilon < \hat{\varepsilon}$, where $\hat{\varepsilon}$ is the same as in (C.7). Fix $z \in B(x,\mu) \subset \mathcal{R}_1(T-2\varepsilon, K-2\varepsilon)$. Let (w_0, w) be a control such that $\int_0^1 w_0(s)\,ds \leq T-2\varepsilon$, $\int_0^1 |w(s)|\,ds \leq K-2\varepsilon$, and $y_z(1, w_0, w) = 0$. Consider then the control $(w_0 + \frac{1}{n}, w)$ and denote by (t_n, k_n, y_n) the corresponding solution to $(S)_1$. In view of (C.7), let δ be such that $B(0,\delta) \subset \hat{\mathcal{R}}_1(\varepsilon,\varepsilon)$. By standard estimates it follows that $|y_n(1)| < \delta$ and $1/n \leq \varepsilon$ for n large enough, so that $y_n(1) \in \hat{\mathcal{R}}_1(\varepsilon,\varepsilon)$ and hence $z \in \hat{\mathcal{R}}_1(T, K-\varepsilon)$ $\forall z \in B(x,\mu)$. Hence $x \in \hat{\mathcal{R}}_1^\circ(T,K)$. □

THEOREM 3.4. *Let $p = 1$ and assume* (C.6), (C.7).

(a) *If* (C.1) *is verified, then $\hat{E}_1 \equiv E_1$ in $\mathbb{R}^n \times]0, +\infty[$.*

(b) *If* (C.5) *is verified, then $\hat{T}_1 \equiv T_1$ in $Dom(\hat{T}_1)^\circ$. Moreover, $Dom(\hat{T}_1)^\circ = Dom(T_1)^\circ$.*

Proof. (a) Let $(x,T) \in \mathbb{R}^n \times]0, +\infty[$ and set $K \doteq E_1(x,T)$. If $K = +\infty$, then $\hat{E}_1(x,T) = +\infty$. Let $K < +\infty$. Since $x \in \mathcal{R}_1(T,K)$, in view of (C.1) one has that $x \in \mathcal{R}_1^\circ(T, K+\varepsilon)$ $\forall \varepsilon > 0$, and by Lemma 3.1 it follows also that $x \in \hat{\mathcal{R}}_1^\circ(T, K+\varepsilon)$ $\forall \varepsilon > 0$. Hence $\hat{E}_1(x,T) \leq E_1(x,T) + \varepsilon$ $\forall \varepsilon > 0$, and since ε is arbitrary we get $\hat{E}_1(x,T) = E_1(x,T)$.

(b) Let $x \in \mathcal{R}_1^\circ(K)$ and assume (C.5). By Proposition 3.2(e), it follows that $E_1(x,T) < K$ $\forall T > T_1(x,K)$, or equivalently that x belongs to $\mathcal{R}_1^\circ(T_1(x,K) + \varepsilon, K)$ for any $\varepsilon > 0$. In view of Lemma 3.1, this implies that $x \in \hat{\mathcal{R}}_1^\circ(T_1(x,K) + \varepsilon, K)$. Hence $\hat{T}_1(x,K) \leq T_1(x,K) + \varepsilon$ for any $\varepsilon > 0$, which yields $\hat{T}_1(x,K) = T_1(x,K)$ $\forall (x,K) \in Dom(T_1)^\circ$ and also $Dom(\hat{T}_1)^\circ = Dom(T_1)^\circ$, in that $T_1(x,K) = +\infty$ implies $\hat{T}_1(x,K) = +\infty$. □

Taking into account Lemma 3.1 and Theorem 3.4, the following results for $\hat{E}_1$ and $\hat{T}_1$ are straightforward consequences of Propositions 3.1 and 3.3 and Theorems 2.1 and 3.3.

COROLLARY 3.1. *Let $p = 1$ and assume* (C.1), (C.6), *and* (C.7). *Then*

(a) *$Dom(\hat{E}_1)$ is an open set and $\hat{E}_1$ is upper semicontinuous in $Dom(\hat{E}_1)$ and lower semicontinuous in $\overline{Dom(\hat{E}_1)}$;*

(b) *if* (C.5) *holds, then for any $K > 0$ one has $\hat{E}_1(x, \hat{T}_1(x, K)) = K$ $\forall x \in \hat{\mathcal{R}}_1^\circ(K) \setminus \mathcal{R}_1(0, K)$.*

(c) *If* (C.2) *holds, then for any $T > 0$ one has $\hat{\mathcal{S}}_1(T) = \mathbb{R}^n$ and there exists $L_1 > 0$ such that for all $x_1, x_2 \in \mathbb{R}^n$ one has*

$$|\hat{E}_1(x_2, T) - \hat{E}_1(x_1, T)| \le L_1 |x_2 - x_1|.$$

COROLLARY 3.2. *Let $p = 1$ and assume* (C.5), (C.6), *and* (C.7). *Then we have the following:*

(a) *$Dom(\hat{T}_1)^\circ = \cup_{K>0}(\hat{\mathcal{R}}_1^\circ(K) \times \{K\})$, and $\hat{T}_1$ is upper semicontinuous in $Dom(\hat{T}_1)^\circ$ and lower semicontinuous in $\overline{Dom(\hat{T}_1)}$. If* (C.3) *holds, then $Dom(\hat{T}_1)$ is open.*

(b) *If* (C.1) *holds, then $\hat{T}_1(x, \hat{E}_1(x, T)) = T$ $\forall x \in \hat{\mathcal{R}}_1^\circ(K) \setminus \hat{\mathcal{R}}_1(T, 0)$, where $K = \hat{E}_1(x, T)$. If* (C.3) *holds, then $\hat{T}_1(x, \hat{E}_1(x, T)) = T$ $\forall x \in \hat{\mathcal{S}}_1(T) \setminus \hat{\mathcal{R}}_1(T, 0)$.*

(c) *If* (C.4) *holds, then for any $K \ge 0$ one has $\hat{\mathcal{R}}_1(K) = \mathbb{R}^n$ and there exists $L_2 > 0$ such that*

$$|\hat{T}_1(x_1, K) - \hat{T}_1(x_2, K)| \le L_2 |x_1 - x_2|^{1/\alpha}$$

$\forall x_1, x_2 \in \mathbb{R}^n$ such that $|x_1 - x_2|$ is small enough and where α is the same as in (C.4).

3.3. Local controllability conditions and regularity results for $p > 1$. In line with what has already been done for linear systems in the case $p > 1$ and for nonlinear systems for $p = \infty$, we prove *local* versions of Theorems 3.1 and 3.2 for $p > 1$ (see Theorems 3.5 and 3.6, respectively) using the following *local controllability conditions:*

(C.8) Fix $p > 1$. Assume that there are a constant $\bar{\varepsilon} > 0$ and an increasing function τ with $\tau(0) = 0$ such that for all $x_0 \in \mathcal{R}_p(T, K) \cap \{x : \ |x| \le \bar{\varepsilon}\}$ for some T, $K > 0$, one has

$$B(x_0, \tau(T)H) \subset \mathcal{R}_p(T, K + H) \qquad \forall \tau(T)H \le \bar{\varepsilon}.$$

(C.9) Fix $p > 1$. Assume that there are some σ, $\bar{\varepsilon} > 0$, and $\alpha \ge 1$ such that

$$B(0, \sigma K S^\alpha) \subset \mathcal{R}_p(S, K) \quad \forall S, K \ge 0 \quad \text{such that} \quad \sigma K S^\alpha \le \bar{\varepsilon}.$$

Notice that, even if (C.8) is a local condition, it differs essentially from assumption (C.9) and, more generally, from the usual local controllability conditions, where one assumes that there exists a ball centered at the origin contained in any reachable set in small time (and with small energy, in our case). Indeed, condition (C.8) requires that around the origin the reachable sets display a "good expandability" property in the K-variable. More precisely, one has to have that for any x_0 near the origin and belonging to some $\mathcal{R}_p(T, K)$, there exists some $\sigma > 0$ such that all the points in $B(x_0, \sigma H)$ can reach the origin using controls with energy less than or equal to $K + H$. Finally, we refer to Remark 3.1 and Example 3.3 below for some considerations about the local controllability conditions and the regularity for $p = 1$.

The proofs of Theorems 3.5 and 3.6 below are based on the following dynamic programming principles.

PROPOSITION 3.4. *Let $p > 1$. For every $(x, T) \in (\mathbb{R}^n \setminus \{0\}) \times [0, +\infty[$ and every $S \le T$ one has*

$$\text{(EDPP)} \quad E_p^p(x,T) = \inf \left\{ \int_0^S |u|^p \, dt + E_p^p\big(y_x(S,u), T-S\big) : \ u \in L^p([0,S], \mathbb{R}^m) \right\}.$$

For every $(x, K) \in (\mathbb{R}^n \setminus \{0\}) \times [0, +\infty[$ and every $T \le T_p(x, K)$ one has

$$T_p(x,K) = \inf \left\{ T + T_p \left(y_x(T,u), \left(K^p - \int_0^T |u|^p \, dt \right)^{\frac{1}{p}} \right) : \right.$$

$$\text{(TDPP)} \qquad \left. u \in L^p([0,T], \mathbb{R}^m), \quad \int_0^T |u|^p \, dt \le K^p \right\}.$$

THEOREM 3.5. *Let $p > 1$, and suppose that condition* (C.8) *is verified. Then for fixed T, $K > 0$, and $N > 0$ there exists some $L_2 > 0$ such that for all $x_1, x_2 \in R_p(T,K) \cap \{x : \ |x| \le N\}$ one has*

$$|E_p(x_2,T) - E_p(x_1,T)| \le L_2 |x_2 - x_1|^{\frac{1}{p}}.$$

Moreover, $\mathcal{S}_p(T)$ is an open set.

Proof.

Step 1. Let $x_1, x_2 \in S_p(T) \cap \{x : \ |x| \le N\}$, let $K_1 \doteq E_p(x_1, T)$, and suppose that $E_p(x_2, T) > K_1$. In view of Proposition 2.2 and Remark 2.2, there exists an optimal control u such that $E_p(x_1, T) = (\int_0^T |u|^p \, dt)^{1/p}$. Let $\bar{t} < T$ be the first time such that $y_{x_1}(\bar{t}, u) \in \partial B(0, \bar{\varepsilon})$. Since $p > 1$, by the estimates

$$\sup_{t \in [0,T]} |y_{x_1}(t,u)| \le \bar{C}_1 \doteq \left(M_f + \sum_{i=1}^m M_{g_i} \right) (1+N) e^{M_f T + \sum_{i=1}^m M_{g_i} K_1 T^{1/p'}} \tag{3.3}$$

and

$$\begin{aligned} \bar{\varepsilon} = |y_{x_1}(\bar{t},u)| &= \left| \int_{\bar{t}}^T \left[f(y_{x_1}(t,u)) + \sum_{i=1}^m g_i(y_{x_1}(t,u)) u_i(t) \right] dt \right| \\ &\le (1 + \bar{C}_1) \left[M_f (T - \bar{t}) + K_1 \sum_{i=1}^m M_{g_i} (T - \bar{t})^{1/p'} \right] \end{aligned}$$

it follows that there is some positive constant $\bar{C}_2$ such that

$$T - \bar{t} \ge \left[\frac{\bar{\varepsilon}}{\bar{C}_2} \right]^{p'}. \tag{3.4}$$

Moreover, similar standard estimates yield that

$$|y_{x_2}(\bar{t},u) - y_{x_1}(\bar{t},u)| \le |x_2 - x_1| e^{L_f \bar{t} + \sum_{i=1}^m L_{g_i} K_1 \bar{t}^{1/p'}},$$

where L_f and L_{g_i} $(i = 1, \dots, n)$ depend on the compact set to which $y_{x_1}(t,u)$ and $y_{x_2}(t,u)$ belong for all $t \in [0,T]$.

Step 2. Let us first consider only x_1, x_2 such that

$$|x_2 - x_1| \le \rho_{x_1} \doteq \bar\varepsilon / e^{L_f T + \sum_{i=1}^m L_{g_i} K_1 T^{1/p'}},$$

and let

$$H \doteq \frac{|x_2 - x_1| e^{L_f T + \sum_{i=1}^m L_{g_i} K_1 T^{1/p'}}}{\tau\left(\left[\frac{\bar\varepsilon}{\bar C_2}\right]^{p'}\right)}.$$

Hence by applying (C.8) to $x_0 \doteq y_{x_1}(\bar t, u) \in \mathcal{R}_p(T - \bar t, (K^p - \int_0^T |u|^p\, dt)^{1/p})$ we have that $y_{x_2}(\bar t, u) \in \mathcal{R}_p(T - \bar t, (K_1^p - \int_0^T |u|^p\, dt)^{1/p} + H)$, which, in view of (EDPP), yields

$$E_p(x_2, T) - E_p(x_1, T) \le \left(\int_0^T |u|^p\, dt + \left[\left(K_1^p - \int_0^T |u|^p\, dt\right)^{1/p} + H\right]^p\right)^{1/p} - K_1.$$

By assuming $H \le 1$, straightforward calculations lead to the local Hölder continuity estimate of the statement for some $L_2 > 0$.

Step 3. By a standard compactness argument, the above estimate on the local Lipschitz continuity can be easily extended to the whole set $\mathcal{R}_p(T, K) \cap \{x : |x| \le N\}$. Moreover, since for any $x_1 \in \mathcal{S}_p(T)$ there are some $K > 0$ and some $N > 0$ such that $x_1 \in \mathcal{R}_p(T, K) \subset \{x : |x| \le N\}$, by the previous steps it follows that there is some $\rho_{x_1} > 0$ such that $B(x_1, \rho_{x_1}) \subset \mathcal{S}_p(T)$. Hence $\mathcal{S}_p(T)$ turns out to be open. ∎

THEOREM 3.6. *Let $p > 1$, and suppose that condition* (C.9) *is verified. Then, for fixed T, K, and $N > 0$ there exists some $L_4 > 0$ such that for all $x_1, x_2 \in \mathcal{R}_p(T, K) \cap \{x : |x| \le N\}$ one has*

$$|T_p(x_1, K) - T_p(x_2, K)| \le L_4 |x_1 - x_2|^{\frac{1}{\alpha p'}}.$$

Moreover, $\mathcal{R}_p(K)$ is an open set.

Proof.

Step 1. Let $x_1, x_2 \in \mathcal{R}_p(K) \cap \{x : |x| \le N\}$ such that $|x_1 - x_2| \le 1$, let $T_1 \doteq T_p(x_1, K)$, and suppose that $T_p(x_2, K) > T_1$. In view of Proposition 2.2 and Remark 2.2, there exists an optimal control u such that $(\int_0^{T_1} |u|^p\, dt)^{1/p} \le K$ and $y_{x_1}(T_1, u) = 0$. Following [6], let $\lambda \doteq 1 - |x_2 - x_1|$ and consider the trajectory $y_{x_2}(\cdot, \lambda u)$. Since $(K^p - \lambda^p K^p)^{1/p} > K(1 - \lambda)^{1/p}$, by (TDPP) it follows that

$$T_p(x_2, K) \le T_1 + T_p\left(y_{x_2}(T_1, \lambda u), K(1 - \lambda)^{1/p}\right).$$

Standard estimates yield that

$$|y_{x_2}(T_1, \lambda u)| \le (1 + \bar C_1)\left[K T_1^{1/p'} \sum_{i=1}^m M_{g_i} e^{L_f + \sum_{i=1}^m L_{g_i} K T_1^{1/p'}}\right] |x_2 - x_1|$$

for some constant $\bar C_1 > 0$, where L_f and L_{g_i} $(i = 1, \dots, n)$ depend on the compact set to which $y_{x_2}(t, \lambda u)$ and $y_{x_1}(t, u)$ belong for all $t \in [0, T]$.

Step 2. Let us first consider only x_1, x_2 such that

$$|x_2 - x_1| \le \rho_{x_1} \doteq \bar\varepsilon / (1 + \bar C_1)\left[K T_1^{1/p'} \sum_{i=1}^m M_{g_i} e^{L_f + \sum_{i=1}^m L_{g_i} K T_1^{1/p'}}\right],$$

so that by (C.9) and by the definition of λ it follows that

$$T_p\left(y_{x_2}(T_1,\lambda u),K(1-\lambda)^{1/p}\right)\le\left[\frac{|y_{x_2}(T_1,\lambda u)|}{\sigma K(1-\lambda)^{1/p}}\right]^{1/\alpha}$$

$$\le\left[\frac{(1+\bar C_1)\left[T_1^{1/p'}\sum_{i=1}^m M_{g_i}e^{L_f+\sum_{i=1}^m L_{g_i}KT_1^{1/p'}}\right]}{\sigma}|x_2-x_1|^{1/p'}\right]^{1/\alpha}. \tag{3.5}$$

At this point (TDPP) leads to the local Hölder continuity estimate

$$T_p(x_2,K)-T_p(x_1,K)\le L_4|x_2-x_1|^{\frac{1}{\alpha p'}},$$

where L_4 denotes the constant written above.

Step 3. In the same way as in Theorem 3.5, this result can be extended to the whole set $\mathcal{R}_p(T,K)\cap\{x:\ |x|\le N\}$, and $\mathcal{R}_p(K)$ turns out to be open. □

We point out that under condition (C.4) the exponent of Hölder continuity of the minimum time function $T_p(\cdot,K)$ obtained in Theorem 3.2 is $1/\alpha$, which is larger than the exponent $1/\alpha p'$ given in Theorem 3.6 under the weaker condition (C.9). For instance, in Example 3.1 and for $\lambda<0$, Theorem 3.2 yields the local Lipschitz continuity of $T_p(\cdot,K)$ for every $p>1$. Notice though that if only (C.9) is in force, the exponent $1/\alpha p'$ cannot be, in general, improved (see [6]). Analogously, condition (C.2) yields the Lipschitz continuity of E_p (this is the case of controllable linear systems) while condition (C.8) yields only its Hölder continuity.

As straightforward consequences of Theorems 3.1, 3.5 and of Theorems 3.2, 3.6, respectively, one has the following results.

COROLLARY 3.3. *Let $p>1$ and assume* (C.8). *Then the global topological property* (C.1) *turns out to be verified.*

COROLLARY 3.4. *Let $p>1$ and assume* (C.9). *Then the global topological property* (C.3) *turns out to be verified.*

Remark 3.1. If $p=1$, the arguments used in the proofs of Theorems 3.5 and 3.6 do not work, even if we consider the extended minimum time and minimum energy functions (and the corresponding dynamic programming principles). More precisely, both the crucial estimates (3.4) and (3.5) are in force thanks only to the fact that $1-\frac{1}{p}>0$. In fact, no regularity property of $T_1(\cdot,K)$ can be propagated in the whole set $\mathcal{R}_1(K)$ from properties of the system in a neighborhood of the target, as shown by the following example.

Example 3.3. Let us consider the (controllable) linear control system introduced in Example 3.1(b). In the case $\lambda\ge 0$ one has that $\mathcal{R}_1(K)=\mathcal{R}_1(T,K)=\overline{B(0,K)}$. Hence the set $\mathcal{R}_1(K)$ turns out to be closed even if (C.9) is verified. On the contrary, a regularity result for $\hat T_1$ similar to the one obtained in Theorem 3.6 for $p>1$ would actually imply that $\mathcal{R}_1(K)$ is open.

4. Sufficient controllability conditions. In this section we prove that for linear systems the classical Kalman condition implies the topological properties and the local controllability conditions introduced in the previous sections. In the general case of nonlinear systems, we show how a well-known controllability condition around the target yields some of the local controllability assumptions introduced in section 3.

We start by considering a linear control system of the form

$$\dot{y} = Ay + Bu, \tag{L}$$

where A is an $n \times n$ and B is an $n \times m$-real matrix. Let us introduce the Kalman condition

$$\{i : \ \mathrm{rank}[B, AB, \dots, A^i B] = n\} \neq \emptyset, \tag{$\mathcal{K}$}$$

and let $r \doteq \min\{i : \ \mathrm{rank}[B, AB, \dots, A^i B] = n\}$. If $p = +\infty$, it is well known that ($\mathcal{K}$) is necessary and sufficient for the continuity, in fact, for the Hölder continuity, of the minimum time function—see, e.g., [9]. Results on the continuity of $(x, T) \mapsto E_p(x, T)$ for $T > 0$ and on the Hölder continuity of $x \mapsto T_p(x, K)$ for $K > 0$ (for linear control systems) can already be found in [4] and [6] but only in the case $p > 1$.

LEMMA 4.1. *Consider system* (L).

(a) *For every $p \geq 1$, $T > 0$, and $K > 0$, the set $\hat{\mathcal{R}}_p(T, K)$ is convex and*

$$\mathcal{R}_p(T, K) = K\mathcal{R}_p(T, 1). \tag{4.1}$$

(b) *Assume* ($\mathcal{K}$). *Then conditions* (C.1), (C.2), (C.5), *and* (C.6) *are verified for all $p \geq 1$. If $p = 1$, condition* (C.7) *also holds.*

(c) *Assume* ($\mathcal{K}$), *and let $p > 1$. Then condition* (C.9), *with $\alpha \doteq \frac{1}{p'} + r$, and condition* (C.3) *are verified.*

Proof. The homogeneity property (4.1) is proved for $p = 1$ in [8], and one can easily extend the proof to the case $p > 1$. By ($\mathcal{K}$) the dimension of $\mathcal{R}_p(T, K)$ is n, and this together with (4.1) implies (C.1) for $p \geq 1$. In order to prove (C.2), fix $T > 0$ and let $x \in \mathcal{R}_p(T, K)$ for some $K > 0$. By ($\mathcal{K}$), for any $H > 0$ it is possible to find (see, e.g., [8]) $n+1$ controls $u_1, \dots, u_{n+1}$ such that $\int_0^T |u_i(t)|^p \, dt \leq H^p$, with $|u_i(t)| = \frac{H}{T^{1/p}}$ for $t \in [0, T]$, $i = 1, \dots, n+1$, and such that, denoting by $y_i \doteq \int_0^T e^{(T-t)A} B u_i(t) \, dt$, the convex hull generated by $\{x + y_i, i = 1, \dots, n+1\}$ contains a ball $B(x, \delta)$. Moreover it is also easy to show following [9] that there exist some constants $\bar{\delta} > 0$ and $C_0 > 0$ such that for any $T > 0$ one has $\delta \geq C_0 \frac{H}{T^{1/p}} T^r$ for all $H \leq \bar{\delta}$. Hence (C.2) turns out to be verified by setting, e.g., $C_2(T) = C_0 \frac{T^r}{T^{1/p}}$. Since the (original) reachable sets are convex and $\overline{\hat{\mathcal{R}}_1(T, K)} = \mathcal{R}_1(T, K)$ the previous result yields (C.7) for $p = 1$, and by Proposition 3.2 it also follows that (C.6) is verified and (C.1) implies (C.5). The proof of (C.9) follows from [17] (see also [6]). Finally, by Corollary 3.4 it follows that (C.9) implies (C.3). □

Owing to Lemma 4.1, the following results on the minimum time and the minimum energy functions are straightforward consequences of the propositions and the theorems in section 3.

COROLLARY 4.1. *Consider system* (L), *assume* ($\mathcal{K}$), *and let $p > 1$. Then we have the following:*

(a) *$Dom(E_p) = \mathbb{R}^n \times]0, +\infty[$, and the map E_p is continuous on it and lower semicontinuous on $\mathbb{R}^n \times [0, +\infty[$.*

(b) *For any $K > 0$, $E_p(x, T_p(x, K)) = K \ \ \forall x \in \mathcal{R}_p(K) \setminus \{0\}$, and $\mathcal{R}_p(K)$ is an open set.*

(c) *For any fixed $T > 0$ there exists $L_1 > 0$ such that for all $x_1, x_2 \in \mathbb{R}^n$ one has*

$$|E_p(x_2, T) - E_p(x_1, T)| \leq L_1 |x_2 - x_1|.$$

(d) *$Dom(T_p)$ is an open set and the map T_p is continuous on it and lower semicontinuous on $\overline{Dom(T_p)}$.*

(e) *For any* $T > 0$, $T_p(x, E_p(x,T)) = T \quad \forall x \in \mathbb{R}^n \setminus \mathcal{R}_p(T,0)$.

(f) *For any fixed* T, K, *and* $N > 0$ *there exists* $L_2 > 0$ *such that for every* $x_1, x_2 \in \mathcal{R}_p(T,K) \cap \{x : |x| \le N\}$ *one has*

$$|T_p(x_1,K) - T_p(x_2,K)| \le L_2 |x_1 - x_2|^{\frac{1}{\alpha p'}}.$$

COROLLARY 4.2. *Consider system* (L), *assume* ($\mathcal{K}$), *and let* $p = 1$. *Then we have the following:*

(a) $Dom(\hat{E}_1) = \mathbb{R}^n \times]0, +\infty[$, $\hat{E}_1$ *is continuous on* $\mathbb{R}^n \times]0,+\infty[$ *and lower semicontinuous on* $\mathbb{R}^n \times [0, +\infty[$.

(b) *For any* $K > 0$ $\hat{E}_1(x, \hat{T}_1(x,K)) = K \quad \forall x \in \hat{\mathcal{R}}_1^\circ(K) \setminus \mathcal{R}_1(0,K)$.

(c) *For any* $T > 0$ *there exists* $L_1 > 0$ *such that for all* $x_1, x_2 \in \mathbb{R}^n$ *one has*

$$|\hat{E}_1(x_2,T) - \hat{E}_1(x_1,T)| \le L_1 |x_2 - x_1|.$$

(d) $Dom(\hat{T}_1)^\circ = \cup_{K>0}(\hat{\mathcal{R}}_1^\circ(K) \times \{K\})$ *and* $\hat{T}_1$ *is continuous in* $Dom(\hat{T}_1)^\circ$ *and lower semicontinuous in* $\overline{Dom(\hat{T}_1)}$.

(e) *For any* $T > 0$ $\hat{T}_1(x, \hat{E}_1(x,T)) = T \quad \forall x \in \hat{\mathcal{R}}_1^\circ(K) \setminus \hat{\mathcal{R}}_1(T,0)$, *where* $K = \hat{E}_1(x,T)$. *If* (C.3) *holds, then* $\hat{T}_1(x, \hat{E}_1(x,T)) = T \quad \forall x \in \mathbb{R}^n \setminus \hat{\mathcal{R}}_1(T,0)$.

(f) *If* (C.4) *holds, then for any* $K \ge 0$ *the set* $\hat{\mathcal{R}}_1(K) = \mathbb{R}^n$ *and there exists* $L_2 > 0$ *such that*

$$|\hat{T}_1(x_1,K) - \hat{T}_1(x_2,K)| \le L_2 |x_1 - x_2|^{1/\alpha}$$

$\forall x_1, x_2 \in \mathbb{R}^n$ *such that* $|x_1 - x_2|$ *is small enough, where* α *is the same as in* (C.4).

In the framework of nonlinear control systems we prove that the following well-known assumption ($\mathcal{H}$) implies some of the *local* controllability conditions introduced in subsections 3.2 and 3.3 (see, e.g., [8]).

($\mathcal{H}$) $f(0) = 0$ and f is continuously differentiable in a neighborhood of the origin. Let $A \doteq \partial_x f(0)$ ($\partial_x f(0)$ denotes the Jacobian matrix of f in the origin) and $B \doteq (g_1(0), \dots, g_m(0))$; A and B verify ($\mathcal{K}$).

Let us recall that, as shown in section 3, local conditions alone are sufficient in order to obtain some partial regularity results for T_p and E_p only in the case $p > 1$ (see Theorems 3.5, 3.6). Any result for $p = 1$, instead, requires us to assume also some global topological properties of the reachable sets.

LEMMA 4.2. *Consider system* $(\hat{S})_p$ *and assume* ($\mathcal{H}$). *Then conditions* (C.9) *and* (C.3) *hold for* $p > 1$; *condition* (C.7) *holds for* $p = 1$.

Proof. The proof is based on an analogous result proved by Bianchini and Stefani in [1] for compact valued controls. In fact, it is possible to deduce from [1] that for any T and $K > 0$, denoting by $\mathcal{U}_\infty(T) \doteq \{u \in L^\infty([0,T], \mathbb{R}^m) : \ |u_i| \le K,\ i = 1, \dots, m\}$ and by $\mathcal{R}_\infty(T)$ the corresponding reachable set, for sufficiently small ε one has $B(0, \sigma K \varepsilon^\alpha) \subset \mathcal{R}_\infty(\varepsilon)$, where $\alpha \doteq 2r + 1 + \rho$. It is clear that if $u \in \mathcal{U}_\infty(T)$, then for $p \ge 1$ one has that $\frac{u}{m} \in \hat{\mathcal{U}}_p(T,K)$ if $T \le 1$. Therefore for sufficiently small ε, there exists a constant σ' such that $B(0, \sigma' K \varepsilon^\alpha) \subset \hat{\mathcal{R}}_p(\varepsilon, K)$. This implies (C.9) for $p > 1$ and (C.7) for $p = 1$; (C.3) follows by Corollary 3.2. □

As a consequence of this lemma, one has that under the hypotheses of Theorem 3.4 the extended problems are equivalent to the original ones for $p = 1$. Moreover, in view of Theorem 3.6, ($\mathcal{H}$) yields the regularity of $T_p(\cdot, K)$ for $p > 1$ and for any $K > 0$.

COROLLARY 4.3. *Assume* ($\mathcal{H}$). *Then for* $p > 1$, *fixed* T, K, *and* $N > 0$ *there exists* $L_4 > 0$ *such that for every* $x_1, x_2 \in \mathcal{R}_p(T,K) \cap \{x : |x| \le N\}$ *one has*

$$|T_p(x_1,K) - T_p(x_2,K)| \le L_4 |x_1 - x_2|^{\frac{1}{\alpha p'}},$$

with $\alpha = 2r + 1 + \rho$ $\forall \rho > 0$. *Moreover,* $\mathcal{R}_p(K)$ *is an open set.*

Remark 4.1. Due to Theorem 3.5, in order to check the local Hölder continuity of $E_p(\cdot, T)$, one should prove directly condition (C.8). As already remarked at the beginning of subsection 3.3, this condition is essentially different from usual local controllability conditions (for bounded valued control systems), and hence it cannot be easily deduced from them. We just mention that in Example 3.2 conditions (C.8) for $p > 1$ (in fact, also the stronger condition (C.2)) and (C.7) for $p = 1$ turn out to be verified. Moreover, for any control system which is linear just in a neighborhood of the origin and here verifies the Kalman condition, (C.8) for $p > 1$ holds.

Appendix. The following propositions clarify the relation between $(\hat{S})_p$ and $(S)_p$. We omit the proofs in the case $p > 1$, in that they are completely similar to the proofs given for more general nonlinear systems and for $p = 1$ in [11].

PROPOSITION A.1. *Fix* $p \geq 1$. *For every* $y(\cdot) = y(\cdot, u)$ *solution to* $(\hat{S})_p$ *in* $[0, T]$ *and for every increasing and surjective absolutely continuous map* $t : [0, 1] \to [0, T]$, *the graph parametrization* $(t, y \circ t)$ *of* y *is the* (t, y)*-component of a solution of* $(S)_p$ *associated with the control* $(w_0(s), w(s)) \doteq (\sqrt[p]{t'(s)}, \sqrt[p]{t'(s)}u(t(s)))$ *for a.e.* $s \in [0, 1]$.[1]

Moreover, if $s : [0, 1] \to [0, 1]$, $\sigma \mapsto s(\sigma)$, *is a nondecreasing and surjective absolutely continuous map, for every trajectory* $(t, k, y)(s) = (t, k, y)(s, w_0, w)$ *of* $(S)_p$ *the map* $(\hat{t}, \hat{k}, \hat{y})(\sigma) \doteq (t, k, y)(s(\sigma))$ $\forall \sigma \in [0, 1]$ *is still a solution to* $(S)_p$, *corresponding to the control* $(\hat{w}_0, \hat{w})$ *defined by* $\hat{w}_0(\sigma) \doteq w_0(s(\sigma))\frac{ds}{d\sigma}(\sigma)$ *and* $\hat{w}(\sigma) \doteq w(s(\sigma))\frac{ds}{d\sigma}(\sigma)$.

Due to the first part of Proposition A.1, the set of graphs of trajectories of $(\hat{S})_p$ can be identified with the subset of (t, y)-components of trajectories of $(S)_p$ with the corresponding control (w_0, w) such that $w_0 > 0$ a.e. In this sense $(S)_p$ can be considered as an extension of $(\hat{S})_p$.

For any $p \geq 1$, let $(w_0, w) \in L^p([0, 1], [0, +\infty[\times\mathbb{R}^m)$. If $(w_0, w) = 0$ a.e. in $[0, 1]$, we set

$$(w_0^c(s), w^c(s)) = (w_0(s), w(s)) \qquad \text{for a.e. } s \in [0, 1];$$

otherwise let $\sigma : [0, 1] \to [0, 1]$ be defined by

$$\sigma(s) \doteq \frac{\int_0^s |(w_0, w)(s')|^p \, ds'}{\int_0^1 |(w_0, w)(s')|^p \, ds'} \qquad \forall s \in [0, 1].$$

We set

$$\left(w_0^c(\sigma(s))\frac{d\sigma}{ds}(s), w^c(\sigma(s))\frac{d\sigma}{ds}(s)\right) = (w_0(s), w(s)) \qquad \text{for a.e. } s \in [0, 1]. \tag{A.1}$$

In principle (A.1) defines a multivalued control map. Yet (w_0^c, w^c) turns out to be uniquely determined a.e.

PROPOSITION A.2. *Fix* $p \geq 1$. *Given a control* $(w_0, w) \in L^p([0, 1], [0, +\infty[\times\mathbb{R}^m)$, *the expression* (2.2) *defines a measurable map* (w_0^c, w^c) *a.e. on* $[0, 1]$ *and* $|(w_0^c, w^c)|^p(s) = \int_0^1 |(w_0, w)|^p(s)\, ds$ *for a.e.* $s \in [0, 1]$. *The control* (w_0^c, w^c) *and the corresponding solution* $(t^c, k^c, y^c)(\cdot) \doteq (t, k, y)(\cdot, w_0^c, w^c)$ *to* $(S)_p$ *will be called the* canonical representatives *of* (w_0, w) *and of* $(t, k, y)(\cdot, w_0, w)$, *respectively. Moreover, the relation*

$$(t, k, y)\left(\sigma^{-1}(\{\xi\})\right) = (t^c, k^c, y^c)(\xi)$$

[1] If a control z is Lebesgue measurable, here and in what follows we assume to replace it with a Borel measurable control ζ such that $\zeta = z$ a.e., so that the composition with $t(s)$ is still measurable.

holds true for all $\xi \in [0,1]$.

Remark A.1. Due to Proposition A.2, the canonical representative of any control $(w_0, w) \in \mathcal{U}_p(T,K)$ is bounded, in that $|(w_0^c, w^c)|^p(s) \leq 2^p(T + K^p)$ for a.e. $s \in [0,1]$. Moreover, the reachable set $\mathcal{R}_p(T,K)$ and the minimum time and the minimum energy functions do not change if one considers only canonical representatives of controls. Hence in the extended problems one deals in fact with *bounded valued* controls.

Acknowledgment. We would like to sincerely thank Rosa Maria Bianchini for giving us very useful information about the controllability conditions in nonlinear systems and for her kind help.

REFERENCES

[1] R. M. Bianchini and G. Stefani, *Time optimal problem and time optimal map*, Rend. Sem. Mat. Univ. Politec. Torino, 48 (1990), pp. 401–429.

[2] R. M. Bianchini and G. Stefani, *Controllability along a trajectory: A variational approach*, SIAM J. Control Optim., 31 (1993), pp. 900–927.

[3] A. Bressan and F. Rampazzo, *On differential systems with vector-valued impulsive controls*, Boll. Un. Mat. Ital. B (7), 2 (1988), pp. 641–656.

[4] O. Cârjă, *The minimal time function in infinite dimensions*, SIAM J. Control Optim., 31 (1993), pp. 1103–1114.

[5] R. Conti, *Problemi di controllo e di controllo ottimale*, UTET, Torino, Italy, 1974.

[6] F. Gozzi and P. Loreti, *Regularity of the minimum time function and minimum energy problems: The linear case*, SIAM J. Control Optim., 37 (1999), pp. 1195–1221.

[7] H. Hermes and J. P. La Salle, *Functional Analysis and Time Optimal Control*, Academic Press, New York, London, 1969.

[8] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[9] A. Liverovskii, *Some properties of the Bellman function for linear and symmetric polysystems*, Differential Equations, 16 (1980), pp. 414–423.

[10] B. M. Miller *The generalized solutions of nonlinear optimization problems with impulse control*, SIAM J. Control Optim., 34 (1996), pp. 1420–1440.

[11] M. Motta and F. Rampazzo, *Nonlinear systems with unbounded controls and state constraints: A problem of proper extension*, NoDEA Nonlinear Differential Equations Appl., 3 (1996), pp. 191–216.

[12] M. Motta and C. Sartori, *Semicontinuous viscosity solutions to mixed boundary value problems with degenerate convex Hamiltonians*, Nonlinear Anal. Ser. A: Theory Methods, 49 (2002), pp. 905–927.

[13] N. N. Petrov, *On the Bellman function for the time-optimal process problem*, J. Appl. Math. Mech., 34 (1970), pp. 785–791.

[14] F. Rampazzo and C. Sartori, *The minimum time function with unbounded controls*, J. Math. Systems Estim. Control, 8 (1998), pp. 1–34.

[15] F. Rampazzo and C. Sartori, *Hamilton-Jacobi-Bellman equations with fast gradient-dependence*, Indiana Univ. Math. J., 49 (2000), pp. 1043–1077.

[16] T. I. Seidman, *How violent are fast controls?*, Math. Control Signals Systems, 1 (1988), pp. 89–95.

[17] T. I. Seidman and J. Yong, *How violent are fast controls?* II, Math. Control Signals Systems, 9 (1996), pp. 327–340.

[18] H. J. Sussmann, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 810–838

LARGE DEVIATIONS FOR A RANDOM WALK MODEL WITH STATE-DEPENDENT NOISE*

MICHELLE BOUÉ†, DANIEL HERNÁNDEZ-HERNÁNDEZ‡, AND RICHARD S. ELLIS§

Abstract. In this paper we prove the large deviation principle for a class of random walks with state-dependent noise. This type of model has important applications in queueing and communication theory and in the area of stochastic approximation.

Key words. stochastic algorithms, large deviations, Laplace principle, weak convergence

AMS subject classifications. 60F10, 60K30

DOI. 10.1137/S0363012901396618

1. Introduction.

This paper is concerned with proving a large deviation principle for a certain class of random walks, where the evolution of the noise process depends on the state of the random walk. This type of model arises in a natural way in the study of recursive algorithms, which have important applications in queueing and communication theory and in the area of stochastic approximation. In fact, our main motivation for the study of these models is their application to the state-dependent stochastic approximation algorithms presented in [10]. The convergence and rate of convergence analysis of algorithms is, in general, difficult, and it is associated with the solution of a deterministic differential equation. This approach is developed in detail in [10], where a number of different models are analyzed, including classical models like Robbins–Monro and ARMAX. Further examples arising from nonlinear filtering and off-line identification can be found in [12] and from parameter tracking in [9].

A general recursive algorithm has the form

$$\theta_{k+1} = \theta_k + \gamma_k F(\theta_k, \eta_{k+1}), \quad k = 0, 1, \ldots, \tag{1.1}$$

with $\theta_k \in \mathbb{R}^d$, $\eta_k \in \mathbb{R}^m$, and $\{\gamma_k, k = 0, 1, \ldots\}$ is a sequence of positive numbers such that $\gamma_k \to 0$ as $k \to \infty$. As in the Robbins–Monro algorithm, θ_k represents an "estimate" of an object of interest, while η_k is a random variable (or observation) with distribution function possibly depending on previous estimates and observations. In our model η_k will represent the noise entering the system (1.1) and its distribution may be affected by θ_{k-1} and η_{k-1}. The sequence $\{\eta_k, k \in \mathbb{N}\}$ has its own structure, depending on the type of application at hand. It can have a linear structure, as in the identification problem described in [11], or nonlinear, as in the random direction problem in [10, p. 16]. Let us suppose that observations are given by the recursive formula

$$\eta_{k+1} = G(\theta_k, \eta_k, \nu_{k+1}), \quad k = 0, 1, \ldots, \tag{1.2}$$

*Received by the editors October 15, 2001; accepted for publication (in revised form) January 17, 2003; published electronically June 18, 2003.
http://www.siam.org/journals/sicon/42-3/39661.html

†Department of Mathematics, Trent University, Peterborough, Ontario Canada K9L 1Z6 (michelleboue@trentu.ca).

‡Centro de Investigacioń en Matemáticas, Apartado Postal 402, Guanajuato, Gto. 36000, Mexico (dher@cimat.mx). The research of this author was supported by Conacyt grant 37643-E.

§Department of Mathematics, University of Massachusetts, Amherst, MA 01003-4515 (rsellis @math.umass.edu). The research of this author was supported by National Science Foundation grant NSF-DMS-0202309.

where ν_k are $\mathbb{R}^d$ valued independent and identically distributed random variables with strictly positive density g and, for each k, ν_{k+1} is independent of $\theta_j,\ \eta_j,\ \ j \le k$. Assume that, given θ and η, $G(\theta,\eta,\cdot)$ is a diffeomorphism on $\mathbb{R}^m$ with inverse $H(\theta,\eta,\cdot)$. Then, given A, a Borel set in $\mathbb{R}^d$,

$$\begin{aligned}
&\text{Prob}[\eta_{k+1} \in A | \theta_0, \dots, \theta_k,\ \eta_0, \dots, \eta_k] \\
&= \text{Prob}[\nu_{k+1} \in H(\theta_k, \eta_k, A) | \theta_0, \dots, \theta_k,\ \eta_0, \dots, \eta_k] \\
&= \int_{H(\theta_k,\eta_k,A)} g(y)dy \\
&= \int_A g(H(\theta_k,\eta_k,y))|J(\theta_k,\eta_k,y)|dy,
\end{aligned}$$

where J is the Jacobian of H and $|J|$ denotes its determinant. From this argument it can be seen that, under broad general conditions on F, G, and g, the algorithm (1.1)–(1.2) satisfies Hypothesis H.1 below, so that the conclusions of our main theorem apply.

The model. Let $\mathcal{S}$ be a Polish space, and let $p(d\zeta|x,\xi)$ be a stochastic kernel on $\mathcal{S}$ given $\mathbb{R}^d \times \mathcal{S}$. For each $n \in \mathbb{N}$, we consider a sequence of random variables $\{(X_j^n, Z_j^n), j = 0, \dots, n\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ and taking values in $\mathbb{R}^d \times \mathcal{S}$. For $x \in \mathbb{R}^d$, $\xi \in \mathcal{S}$, and b a function mapping $\mathbb{R}^d \times \mathcal{S}$ into $\mathbb{R}^d$, this sequence is defined by setting $X_0^n \doteq x$, $Z_0^n \doteq \xi$, and letting

$$X_{j+1}^n \doteq X_j^n + \frac{1}{n} b(X_j^n, Z_{j+1}^n),$$

where for $j \in \{0, 1, \dots, n-1\}$ the conditional distribution of Z_{j+1}^n given the past is given by

$$P_{x,\xi}\left\{Z_{j+1}^n \in d\zeta | (X_i^n, Z_i^n), i = 0, \dots, j\right\} = p(d\zeta | X_j^n, Z_j^n). \tag{1.3}$$

Here $P_{x,\xi}$ denotes probability conditioned on $X_0^n = x$, $Z_0^n = \xi$. We assume that the stochastic kernel p and the function b satisfy the following hypothesis.

Hypothesis H.1.

(a) $b(x,\xi)$ is bounded, continuous in ξ, and Lipschitz continuous with constant K in x, uniformly in ξ.

(b) $p(d\zeta|x,\xi)$ is weakly continuous in (x,ξ), and there exist a probability measure ϑ on $\mathcal{S}$ and a measurable function $\tilde{p}^x(\xi,\zeta)$ on $\mathcal{S} \times \mathcal{S}$ such that

$$p(d\zeta|x,\xi) = \tilde{p}^x(\xi,\zeta)\vartheta(d\zeta).$$

(c) Given any compact set $\Delta \subset \mathbb{R}^d$, there exist constants $0 < a \le A < \infty$ such that

$$a \le \tilde{p}^x(\xi,\zeta) \le A$$

for all $x \in \Delta$. Moreover, $\tilde{p}^x(\xi,\zeta)$ is continuous in x uniformly in ξ and ζ, for $x \in \Delta$.

Let $X^n = \{X^n(t), 0 \le t \le 1\}$ be the piecewise linear interpolation on $[0,1]$ of $\{X_j^n, j = 0, \dots, n\}$. More precisely, for $t \in [j/n, (j+1)/n]$ and $j = 0, \dots, n-1$,

$$X^n(t) \doteq X_j^n + \left(t - \frac{j}{n}\right) b(X_j^n, Z_{j+1}^n). \tag{1.4}$$

The main result of the paper, Theorem 2.1, states the large deviation principle for the sequence $\{X^n, n \in \mathbb{N}\}$. Although our main theorem, Theorem 2.1, is closely related to

the results in [3], the proof there relies on technical assumptions for the function Λ (see (2.1) and (2.3) below), which is assumed to exist. Under Hypothesis H.1, the function Λ indeed exists and satisfies the technical assumptions required there (see section 4.3 in [3]), thereby implying the large deviation principle. However, our aim is to establish a more direct connection with the applications. The proof presented here depends on assumptions made on the evolution of the process itself (the transition kernels and the function b). This has several advantages. First, for the purposes of using the results in applications, assumptions must be made on the processes, since these are the type of assumptions that can be used there. Moreover, knowledge about the process provides a lot of intuition concerning the averaging procedure required for the proof. This intuition has been heavily exploited by some of the proofs of convergence of state-dependent stochastic algorithms (see [8, 10]), and we have incorporated some of their underlying ideas into the proof. Finally, seeing where each one of the properties of the process is needed in the proof has enabled us to understand the ergodicity properties required to extend our results to more general state-dependent processes. Extensions will be dealt with elsewhere.

2. The main theorem.

THEOREM 2.1. *Let $\mathcal{S}$ be compact. Under Hypothesis* H.1 *the sequence $\{X^n, n \in \mathbb{N}\}$ defined in* (1.4) *satisfies a large deviation principle with rate function $I_x(\cdot)$, where*

$$I_x(\phi) \doteq \begin{cases} \int_0^1 L(\phi, \dot{\phi})dt & \text{if } \phi \text{ is absolutely continuous and } \phi(0) = x, \\ \infty & \text{otherwise.} \end{cases}$$

Here $L(x, \cdot)$ is the Legendre–Fenchel transform with respect to the second variable of the function $\Lambda(x, \cdot)$, which is solution to the eigenvalue problem given by

$$e^{\Lambda(x,\alpha)+\Psi(\xi)} = \int_{\mathcal{S}} e^{\langle \alpha, b(x,\zeta)\rangle + \Psi(\zeta)} p(d\zeta|x, \xi). \tag{2.1}$$

That is, for x and β in $\mathbb{R}^d$,

$$L(x, \beta) \doteq \sup_{\alpha \in \mathbb{R}^d} \{\langle \alpha, \beta\rangle - \Lambda(x, \alpha)\}. \tag{2.2}$$

Remarks.

(1) We have made the strong assumption of compactness of the state space $\mathcal{S}$ in order to guarantee tightness of the measures involved in the proof (see part (a) of Theorem C.1). If the state space is not compact, further assumptions are required. These are discussed in section 5.

(2) $\mathcal{C}([0, 1] : \mathbb{R}^d)$ maps this space into $[0, \infty]$ and has compact level sets.

(3) For a fixed $x \in \mathbb{R}^d$, define the operator T on the set of bounded and measurable functions $\psi : \mathcal{S} \to \mathbb{R}$ as

$$T\psi(\xi) = \int_{\mathcal{S}} e^{\langle \alpha, b(x,\zeta)\rangle} \psi(\zeta) p(d\zeta|x, \xi).$$

The eigenvalue problem mentioned in (2.1) consists in finding the largest eigenvalue of this operator. Under Hypothesis H.1, Theorem 10.1 in [6] guarantees the existence and uniqueness of a solution to this problem, with a bounded and uniformly positive associated eigenfunction, corresponding to e^{Ψ} in (2.1). In fact, we can identify the solution function $\Lambda(x, \alpha)$ in a very explicit manner. Given $x \in \mathbb{R}^d$ and $\xi \in \mathcal{S}$, set

$\xi_0^x = \xi$ and let $\{\xi_j^x, j \geq 0\}$ be a Markov process with transition kernel $p(\cdot|x, \xi_j^x)$. Then the function $\Lambda(x, \alpha)$ satisfies

$$\Lambda(x,\alpha) = \lim_{N\to\infty} \frac{1}{N} \log E_\xi \left\{ \exp\left\langle \alpha, \sum_{j=1}^{N} b(x, \xi_j^x) \right\rangle \right\}, \tag{2.3}$$

where E_ξ denotes expectation conditioned on $\xi_0^x = \xi$. We refer to the process $\{\xi_j^x, j \geq 0\}$ as the "fixed x" process. As can be seen, it is the Markov chain that results if the parameter X_j^n in (1.3) is held constant at value x. This process is intimately connected with the process $\{X_j^n\}$. Indeed, if n is large, then X_j^n varies slowly and thus the "local" evolution of $b(X_j^n, Z_{j+1}^n)$ is very similar to the evolution of the same quantity but with X_j^n taken to be constant (see [10, sections 2.5 and 8.4]). This idea will be exploited heavily throughout the paper; we especially refer the reader to the proof of part (e) of Theorem C.1.

Let $W^n(x,\xi) \doteq -1/n \log E_{x,\xi} \{\exp[-nh(X^n)]\}$, with h in $\mathcal{C}([0,1] : \mathbb{R}^d)$. The proof of Theorem 2.1 is done in two parts. We start by proving an upper bound of the form

$$\liminf_{n\to\infty} W^n(x,\xi) \geq \inf_{\phi\in\mathcal{C}([0,1]:\mathbb{R}^d)} \{I_x(\phi) + h(\phi)\}. \tag{2.4}$$

This is the content of section 3. The lower bound

$$\limsup_{n\to\infty} W^n(x,\xi) \leq \inf_{\phi\in\mathcal{C}([0,1]:\mathbb{R}^d)} \{I_x(\phi) + h(\phi)\} \tag{2.5}$$

is then proved in section 4. These two inequalities are equivalent to a large deviation principle, as is proved in Theorems 2.2.1 and 2.2.3 in [4]. In both cases, a key step in the proof is based on studying (via weak convergence arguments) the limit properties of a sequence of associated stochastic control problems. The underlying simplicity of the basic arguments will be made clear below.

3. Proof of the upper bound. This section is devoted to the proof of (2.4). The proof can be summarized simply as follows. Based on the variational representation given in the next theorem, we associate with $W^n(x,\xi)$ an appropriate sequence of controlled processes and of control measures. The limit properties of this sequence, derived in Theorem C.1, will yield (2.4).

Let us start by introducing all the relevant quantities appearing in the representation for $W^n(x,\xi)$ (obtained in Theorem 3.1 below). The representation can be derived easily by following the same steps as those given in [4, section 4.4].

We define a discrete-time controlled process taking values in $\mathbb{R}^d \times \mathcal{S}$ denoted by $\{(\bar{X}_j^n, \bar{Z}_j^n), j = 0, \ldots, n\}$. The control at time j is the distribution of the controlled random variable $\bar{Z}_j^n$. It is given by a stochastic kernel $\nu_j^n(d\zeta|\bar{X}_0^n, \ldots, \bar{X}_j^n, \bar{Z}_j^n)$ on $\mathcal{S}$ given $(\mathbb{R}^d)^{j+1} \times \mathcal{S}$. That is, ν_j^n is a random variable mapping $(\mathbb{R}^d)^{j+1} \times \mathcal{S}$ into $\mathcal{P}(\mathcal{S})$.[1] A sequence of controls $\{\nu_j^n, j = 0, \ldots, n-1\}$ is what we refer to as an admissible control sequence. Now, setting $\bar{Z}_0^n = \xi$ and $\bar{X}_0^n = x$, the evolution of the controlled process is through the relation

$$\bar{X}_{j+1}^n = \bar{X}_j^n + \frac{1}{n} b(\bar{X}_j^n, \bar{Z}_{j+1}^n),$$

[1] For ease of presentation, the dependence on the underlying probability space of all stochastic kernels appearing in this paper is not made explicit in the notation.

where the conditional distribution of $\bar{Z}^n_{j+1}$ is given by

$$\bar{P}_{x,\xi}\left\{\bar{Z}^n_{j+1} \in d\zeta | \bar{X}^n_0, \ldots, \bar{X}^n_j, \bar{Z}^n_0, \ldots, \bar{Z}^n_j\right\} = \nu^n_j(d\zeta | \bar{X}^n_0, \ldots, \bar{X}^n_j, \bar{Z}^n_j).$$

Finally, we let $\bar{X}^n = \{\bar{X}^n(t), t \in [0,1]\}$ be the piecewise linear interpolation of $\{\bar{X}^n_j, j = 0, \ldots, n\}$.

THEOREM 3.1. *Let h be a bounded measurable function mapping $\mathcal{C}([0,1] : \mathbb{R}^d) \mapsto \mathbb{R}$. Then for all $n \in \mathbb{N}$, $x \in \mathbb{R}^d$, and $\xi \in \mathcal{S}$ we have the representation*

$$W^n(x,\xi) = \inf_{\{\nu^n_j\}} \bar{E}_{x,\xi}\left\{\frac{1}{n}\sum_{j=0}^{n-1} R(\nu^n_j(\cdot)\|p(\cdot|\bar{X}^n_j, \bar{Z}^n_j)) + h(\bar{X}^n)\right\}. \tag{3.1}$$

Here R is the relative entropy function; $\nu^n_j(\cdot) = \nu^n_j(\cdot|\bar{X}^n_0, \ldots, \bar{X}^n_j, \bar{Z}^n_j)$; the infimum is taken over all admissible control sequences $\{\nu^n_j, j = 0, \ldots, n-1\}$; $\bar{E}_{x,\xi}$ denotes expectation conditioned on $\bar{X}^n_0 = x$ and $\bar{Z}^n_0 = \xi$; and $\{(\bar{X}^n_j, \bar{Z}^n_j), j = 0, \ldots, n\}$ is the controlled process associated with a particular control sequence $\{\nu^n_j\}$.

Let $\varepsilon > 0$ be given. For each $n \in \mathbb{N}$, let $\{\nu^n_j, j = 0, \ldots, n-1\}$ be a sequence of nearly optimal admissible controls for the variational problem in (3.1), so that

$$W^n(x,\xi) + \varepsilon \geq \bar{E}_{x,\xi}\left\{\frac{1}{n}\sum_{j=0}^{n-1} R(\nu^n_j(\cdot)\|p(\cdot|\bar{X}^n_j, \bar{Z}^n_j)) + h(\bar{X}^n)\right\}. \tag{3.2}$$

Here $\{(\bar{X}^n_j, \bar{Z}^n_j), j = 0, \ldots, n\}$ is the controlled process associated with the nearly optimal sequence of controls.

We will obtain the limit inferior of the right-hand side of (3.2) by rewriting it in terms of a new sequence of control measures. These are defined as conveniently averaged controls in a space that is independent of n. For that purpose, let $\{m_n, n \in \mathbb{N}\}$ be a sequence of real numbers satisfying $m_n \to \infty$ as $n \to \infty$ and such that if $k_n \doteq m_n/n$, then $\lim_{n\to\infty} k_n = 0$. Also, suppose that 1 is an integer multiple of k_n. Given $\xi \in \mathcal{S}$, let δ_ξ denote the unit point measure at ξ. For $l = 0, \ldots, 1/k_n - 1$, and Borel subsets B_1 and B_2 of $\mathcal{S}$, let

$$\tilde{\nu}^n_l(B_1 \times B_2) \doteq \frac{1}{m_n}\sum_{j=lm_n}^{(l+1)m_n-1} \delta_{\bar{Z}^n_j}(B_1) \times \nu^n_j(B_2|\bar{X}^n_j, \bar{Z}^n_j).$$

The quantity $\tilde{\nu}^n_l$ is a stochastic kernel on $\mathcal{S} \times \mathcal{S}$ with marginals

$$(\nu^n_l)_1(B_1) = \frac{1}{m_n}\sum_{j=lm_n}^{(l+1)m_n-1} \delta_{\bar{Z}^n_j}(B_1) \quad \text{and} \quad (\nu^n_l)_2(B_2) = \frac{1}{m_n}\sum_{j=lm_n}^{(l+1)m_n-1} \nu^n_j(B_2|\bar{X}^n_j, \bar{Z}^n_j).$$

These definitions naturally result when one thinks of collecting terms of the sum appearing in (3.2) in groups of size m_n for the purposes of averaging. As was mentioned earlier, this technique is common in the proofs of convergence of state-dependent stochastic algorithms (see [8, 10]).

Now for each $n \in \mathbb{N}$ and $t \in [0,1]$ define

$$\nu^n(B_1 \times B_2|t) \doteq \begin{cases} \tilde{\nu}^n_l(B_1 \times B_2) \text{ if } t \in [lk_n, (l+1)k_n) \text{ for } l = 0, \ldots, 1/k_n - 2, \\ \tilde{\nu}^n_{(\frac{1}{k_n}-1)}(B_1 \times B_2) \text{ if } t \in [1-k_n, 1]. \end{cases}$$

Finally define the admissible control measure ν^n to be the random probability measure defined for Borel subsets B_1, B_2 of $\mathcal{S}$ and C of $[0,1]$ through

$$\nu^n(B_1 \times B_2 \times C) \doteq \int_C \nu^n(B_1 \times B_2|t)dt. \tag{3.3}$$

If for $B_1 \in \mathcal{B}(\mathcal{S})$ we define the first marginal $\hat{\nu}_1^n(d\zeta|t)$ of $\nu^n(d\zeta \times dy|t)$ through $\hat{\nu}_1^n(B_1|t) \doteq \nu^n(B_1 \times \mathcal{S}|t)$, then for Borel subsets B_1, B_2 of $\mathcal{S}$ and C of $[0,1]$, Theorem A.5.6 in [4] gives the decomposition

$$\nu^n(B_1 \times B_2 \times C) = \int_C \int_{B_1 \times B_2} \hat{\nu}_1^n(d\zeta|t)\hat{\nu}_2^n(dy|\zeta,t)dt = \int_C \int_{B_1} \hat{\nu}_2^n(B_2|\zeta,t)\hat{\nu}_1^n(d\zeta|t)dt, \tag{3.4}$$

where $\hat{\nu}_2^n(dy|\zeta,t)$ is a stochastic kernel on $\mathcal{S}$ given $\mathcal{S} \times [0,1]$. Following the notation in [4], we summarize this decomposition as $\nu^n(d\zeta \times dy \times dt) = \hat{\nu}_1^n(d\zeta|t) \otimes \hat{\nu}_2^n(dy|\zeta,t) \otimes \lambda$, where λ is Lebesgue measure on $[0,1]$.

We can now rewrite the right-hand side of (3.2) in terms of the control measures ν^n. We first use the fact (see [4, Lemma 1.4.3(f)]) that $R(\beta\|\gamma) = R(\alpha \times \beta\|\alpha \times \gamma)$ for any probability measures α, β, and γ on $\mathcal{S}$. This formula applied term by term enables us to write

$$\begin{aligned}
&\bar{E}_{x,\xi}\left\{\frac{1}{n}\sum_{j=0}^{n-1} R(\nu_j^n(\cdot)\|p(\cdot|\bar{X}_j^n, \bar{Z}_j^n))\right\} \\
&= \bar{E}_{x,\xi}\left\{\sum_{l=0}^{\frac{1}{k_n}-1} k_n \frac{1}{m_n} \sum_{j=lm_n}^{(l+1)m_n-1} R\left(\delta_{\bar{Z}_j^n}(\cdot) \times \nu_j^n(\cdot) \,\|\, \delta_{\bar{Z}_j^n}(d\zeta) \otimes p(\cdot|\bar{X}_j^n, \zeta)\right)\right\},
\end{aligned}$$

where we have used the notation $\delta_{\bar{Z}_j^n}(\cdot) \times p(\cdot|\bar{X}_j^n, \bar{Z}_j^n) = \delta_{\bar{Z}_j^n}(d\zeta) \otimes p(\cdot|\bar{X}_j^n, \zeta)$. Applying Jensen's inequality to the convex function $R(\cdot\|\cdot)$, the right-hand side of the preceding display is no less than

$$\bar{E}_{x,\xi}\left\{\sum_{l=0}^{\frac{1}{k_n}-1} k_n R\left(\frac{1}{m_n}\sum_{j=lm_n}^{(l+1)m_n-1} \delta_{\bar{Z}_j^n}(d\zeta) \times \nu_j^n(\cdot|\bar{X}_j^n, \zeta) \middle\| \frac{1}{m_n}\sum_{j=lm_n}^{(l+1)m_n-1} \delta_{\bar{Z}_j^n}(d\zeta) \otimes p\left(\cdot|\bar{X}_j^n, \zeta\right)\right)\right\},$$

which is clearly equal to

$$\bar{E}_{x,\xi}\left\{\int_0^1 R(\nu^n(\cdot|t)\|\gamma^n(\cdot|t))dt\right\}. \tag{3.5}$$

In (3.5), for each $n \in \mathbb{N}$ and $t \in [0,1]$, we have defined

$$\gamma^n(B_1 \times B_2|t) \doteq \frac{1}{m_n}\sum_{j=lm_n}^{(l+1)m_n-1} \delta_{\bar{Z}_j^n}(B_1) \otimes p(B_2|\bar{X}_j^n, \zeta) \text{ if } t \in [lk_n, (l+1)k_n),$$

for $l = 0, \ldots, 1/k_n - 1$. Now define the measure γ^n on $\mathcal{S} \times \mathcal{S} \times [0,1]$ through $\gamma^n(B_1 \times B_2 \times C) \doteq \int_C \gamma^n(B_1 \times B_2|t)dt$. Since for all stochastic kernels α and β on $\mathcal{S}$ given $[0,1]$ and probability measures γ on $[0,1]$, we have (see [4, Lemma 1.4.3(f)])

$$\int_{[0,1]} R(\alpha(\cdot|x)\|\beta(\cdot|x))\gamma(dx) = R(\alpha \otimes \gamma\|\beta \otimes \gamma); \tag{3.6}$$

(3.5) can be rewritten as $\bar{E}_{x,\xi}\{R(\nu^n\|\gamma^n)\}$. (Recall the definition of ν^n given in (3.3).)

Combining this series of inequalities with (3.2), we obtain

$$W^n(x,\xi)+\varepsilon \geq \bar{E}_{x,\xi}\left\{R(\nu^n\|\gamma^n)+h(\bar{X}^n)\right\}.$$

We now wish to take the limit inferior as $n\to\infty$ of both terms in the last inequality. The asymptotic properties of the sequence $\{(\nu^n,\gamma^n,\bar{X}^n),n\in\mathbb{N}\}$ required to do this are proved in Theorem C.1. According to that theorem, there exists a probability space on which a subsequence of $\{(\nu^n,\gamma^n,\bar{X}^n),n\in\mathbb{N}\}$ converges in distribution to some limit $(\nu,\gamma,\bar{X})$. The stochastic kernels ν and γ and the random variable $\bar{X}$ satisfy all the conclusions stated in Theorem C.1. Thanks to the Skorohod representation theorem [5, p. 102] we can assume that convergence takes place with probability 1 (w.p.1). Along the convergent subsequence we thus have that

$$\begin{aligned}
&\liminf_{n\to\infty} W^n(x,\xi)+\varepsilon\\
&\geq \bar{E}_{x,\xi}\left\{R(\nu\|\gamma)+h(\bar{X})\right\}\\
&=\bar{E}_{x,\xi}\left\{R(\hat{\nu}_1(d\zeta|t)\otimes\hat{\nu}_2(dy|\zeta,t)\otimes dt\|\hat{\nu}_1(d\zeta|t)\otimes p(dy|\bar{X}(t),\zeta)\otimes dt)+h(\bar{X})\right\}\\
&=\bar{E}_{x,\xi}\left\{\int_0^1\int_{\mathcal{S}} R(\hat{\nu}_2(dy|\zeta,t)\|p(dy|\bar{X}(t),\zeta))\hat{\nu}_1(d\zeta|t)dt+h(\bar{X})\right\}\\
&\geq \bar{E}_{x,\xi}\left\{\int_0^1 L\left(\bar{X}(t),\int_{\mathcal{S}} b(\bar{X}(t),\zeta)\hat{\nu}_1(d\zeta|t)\right)dt+h(\bar{X})\right\}\\
&=\bar{E}_{x,\xi}\left\{I_x(\bar{X})+h(\bar{X})\right\}\geq \inf_{\phi\in\mathcal{C}([0,1]:\mathbb{R}^d)}\left\{I_x(\phi)+h(\phi)\right\}.
\end{aligned}$$

Lower semicontinuity of $R(\cdot\|\cdot)$, Fatou's lemma, and continuity of h yield the second line of the above display. The third line uses parts (b) and (f) of Theorem C.1 and the fourth uses (3.6). Finally, part (b) of Lemma B.2 and part (e) of Theorem C.1 give the fifth and sixth lines, respectively. Since the above inequality is valid for all $\varepsilon>0$, (2.4) follows, concluding the proof of the upper bound. □

4. Proof of the lower bound. This section is devoted to showing that (2.5) holds for all h in $\mathcal{C}([0,1]:\mathbb{R}^d)$ that are Lipschitz continuous. Thanks to [4, Corollary 1.2.5(b)], this is enough to show that (2.5) holds for all h in $\mathcal{C}([0,1]:\mathbb{R}^d)$. As in the proof of Proposition 6.6.1 in [4], the proof of (2.5) is done by introducing a perturbation to the original random walk by means of a random walk with Gaussian noise. This allows one to obtain necessary smoothness properties for a function L_σ, which is the analogue of the function L defined in (2.2) but for the perturbed process. Weak convergence arguments make use of these continuity properties, implying the desired lower bound when taking the perturbation to be sufficiently small.

Let us first focus on the perturbed problem; the connection with (2.5) will be clear after (4.3). Given $\sigma>0$, let $\{G_{j,\sigma},j\in\mathbb{N}_0\}$ be a sequence of independent and identically distributed random variables on $\mathbb{R}^d$ with common Gaussian distribution ρ_σ, with mean zero and variance σI. We assume them to be independent of $\{\xi_j^x,x\in\mathbb{R}^d,j\in\mathbb{N}_0\}$, where ξ_j^x is the "fixed x" Markov process with transition kernel $p(\cdot|x,\xi_j^x)$. Given $n\in\mathbb{N}$ and $j\in\{0,1,\dots,n-1\}$, let X_j^n and Z_j^n be as before, and define

$$U_{0,\sigma}^n\doteq 0,\qquad U_{j+1,\sigma}^n\doteq U_{j,\sigma}^n+\frac{1}{n}\,G_{j,\sigma}.$$

Denote by $X^n(t)$ and $U_\sigma^n(t)$ the piecewise linear interpolations of $\{X_j^n,j=1,\dots,n\}$ and $\{U_{j,\sigma}^n,j=0,\dots,n\}$ on $[0,1]$, respectively (see (1.4)). Also, define

$$Y_\sigma^n(t)\doteq X^n(t)+U_\sigma^n(t), \tag{4.1}$$

which is the piecewise linear interpolation of $\{X_j^n + U_{j,\sigma}^n\}$. As was mentioned earlier, the point of introducing a perturbation is to replace the function L by a continuous function L_σ. This latter function is defined as the Legendre–Fenchel transform of some convex function Λ_σ. Once again, the function Λ_σ is identified via an eigenvalue problem, which we now describe.

For fixed $x \in \mathcal{S}$, we can identify an additive component of the process (see [7, p. 376]), namely, $b(x, \xi_j^x) + G_{j,\sigma}$. Here ξ_j^x is the "fixed x" Markov process described earlier. Let Q_σ^x be the stochastic kernel on $\mathcal{S} \times \mathbb{R}^d$ given $\xi \in \mathcal{S}$ defined by

$$Q_\sigma^x(B_1 \times B_2|\xi) = \int_{B_1} \int_{\mathbb{R}^d} 1_{B_2}(b(x,\zeta) + y)\rho_\sigma(dy)p(d\zeta|x,\xi),$$

where $B_1 \in \mathcal{B}(\mathcal{S})$ and $B_2 \in \mathcal{B}(\mathbb{R}^d)$. Then, letting

$$\upsilon_\sigma(B_1 \times B_2|x) \doteq \int_{B_1} \int_{\mathbb{R}^d} 1_{B_2}(b(x,\zeta) + y)\rho_\sigma(dy)\vartheta(d\zeta),$$

with ϑ as in Hypothesis H.1, $B_1 \in \mathcal{B}(\mathcal{S})$, and $B_2 \in \mathcal{B}(\mathbb{R}^d)$, we have

$$a\upsilon_\sigma(B_1 \times B_2|x) \le Q_\sigma^x(B_1 \times B_2|\xi) \le A\upsilon_\sigma(B_1 \times B_2|x).$$

These bounds on $Q_\sigma^x(\cdot,\cdot|\xi)$ and the fact that the convex hull of the support of $\upsilon_\sigma(\mathcal{S} \times \cdot|x)$ is $\mathbb{R}^d$ guarantee the existence of a solution to the eigenvalue problem for each $x, \alpha \in \mathbb{R}^d$ [7, Lemma 3.1]. That is, for each $x, \alpha \in \mathbb{R}^d$, there exist a unique $\Lambda_\sigma(x,\alpha) \in \mathbb{R}$ and a bounded function $\Psi_\sigma(x;\alpha,\cdot) : \mathcal{S} \mapsto \mathbb{R}$ such that

$$e^{\Lambda_\sigma(x,\alpha)+\Psi_\sigma(x;\alpha,\xi)} = \int_{\mathcal{S}} \int_{\mathbb{R}^d} e^{\langle \alpha, b(x,\zeta)+y\rangle + \Psi_\sigma(x;\alpha,\zeta)}\rho_\sigma(dy)p(d\zeta|x,\xi).$$

Furthermore, $\Lambda_\sigma(x,\alpha) = \Lambda(x,\alpha) + \frac{\sigma^2}{2}||\alpha||^2$, and $\Psi_\sigma(x;\alpha,\xi) = \Psi(x;\alpha,\xi)$, where $\Psi(x;\alpha,\xi)$ is the eigenfunction associated with $\Lambda(x,\alpha)$ (see (2.1)). The Legendre–Fenchel transform of Λ_σ is given by

$$L_\sigma(x,\beta) \doteq \sup_{\alpha \in \mathbb{R}^d} \{\langle \alpha, \beta\rangle - \Lambda_\sigma(x,\alpha)\}. \tag{4.2}$$

Having introduced the necessary definitions, we now proceed to relate the original and the perturbed processes. Let K_1 be the Lipschitz constant of h and define $B \doteq 2||h||_\infty$. Then

$$h(Y_\sigma^n) = h(X^n + U_\sigma^n) \ge h(X^n) - (K_1||U_\sigma^n||_\infty \wedge B),$$

and, because of independence,

$$\begin{aligned}\frac{1}{n}\log E_{x,\xi}\{\exp[-nh(Y_\sigma^n)]\} &\le \frac{1}{n}\log E_{x,\xi}\{\exp[-nh(X^n)]\cdot\exp[n(K_1||U_\sigma^n||_\infty \wedge B)]\} \\ &= -W^n(x,\xi) + \frac{1}{n}\log E_{x,\xi}\{\exp[n(K_1||U_\sigma^n||_\infty \wedge B)]\}.\end{aligned}$$

Hence

$$\limsup_{n\to\infty} W^n(x,\xi) \le \limsup_{n\to\infty}\left(-\frac{1}{n}\log E_{x,\xi}\{\exp[-nh(Y_\sigma^n)]\}\right) + \frac{K_1^2\sigma^2}{2},$$

where the second term of the inequality follows from [4, p. 189]. This implies that (2.5) holds as long as we can show that

$$\limsup_{n\to\infty} W_\sigma^n(x,\xi) \le \inf_{\varphi\in C([0,1]:\mathbb{R}^d)}\{I_x(\varphi)+h(\varphi)\}+\theta(\sigma), \tag{4.3}$$

with $W_\sigma^n \doteq -\frac{1}{n}\log E_{x,\xi}\exp[-nh(Y_\sigma^n)]$ and $\theta(\sigma)\to 0$ when $\sigma\to 0$. What we will show in fact is that, given $\varepsilon>0$ and $\psi\in\mathcal{C}([0,1]:\mathbb{R}^d)$ satisfying

$$I_x(\psi)+h(\psi) \le \inf_{\varphi\in C([0,1]:\mathbb{R}^d)}\{I_x(\varphi)+h(\varphi)\}+\varepsilon<\infty, \tag{4.4}$$

we have

$$\limsup_{n\to\infty} W_\sigma^n(x,\xi) \le \int_0^1 L_\sigma(\psi(t),\dot\psi(t))dt + h(\psi)+\theta(\sigma). \tag{4.5}$$

Since $L_\sigma(x,\beta)\le L(x,\beta)$ for all x and $\beta\in\mathbb{R}^d$ (part (a) of Lemma B.3), (4.3) will follow after that.

The steps in the proof of (4.5) can be described in simple terms. Starting with the nearly optimal function ψ in (4.4), we construct a sequence of nearly optimal admissible controls for the stochastic control problem that is associated with $W_\sigma^n(x,\xi)$ through the representation in Theorem A.1. The limit properties of this sequence, as well as estimates on the associated sequence of running costs (where continuity of L_σ is required), will lead directly to (4.5).

Let ψ satisfy (4.4), and let ψ^* be as in part (e) of Lemma B.3. The admissible control sequence that we define based on ψ^* (see (4.20) below) has the following properties: the running costs are nearly optimal in (A.3), and, with probability converging to 1, the associated controlled process $\bar Y^n \doteq \bar X^n + \bar U^n$ (see (A.1)) enters a small neighborhood of ψ^* as $n\to\infty$. The construction is given in the following paragraphs.

Define the compact set

$$\Delta \equiv \cup_{t\in[0,1]}\{y\in\mathbb{R}^d : ||y-\psi^*(t)||\le 1\}.$$

Let $\eta=\eta(\Delta,\sigma)\in(0,1)$ satisfy the conclusions of part (d) of Lemma B.3 when taking $\varepsilon=\sigma$. Also, let $\{x_j, j=1,\dots,n\}$ be a sequence in Δ satisfying $\|\psi^*(j/n)-x_j\|<\eta$. For every $n\in\mathbb{N}$, $j=1,\dots,n$, and with $x=\psi^*(j/n)$, $y=x_j$, and $\beta=\dot\psi^*(j/n)$, part (d) of that lemma implies that there exists $\bar\beta_j^n\in\mathbb{R}^d$ such that

$$L_\sigma(x_j,\bar\beta_j^n) - L_\sigma(\psi^*(j/n),\dot\psi^*(j/n)) \le \sigma \tag{4.6}$$

and

$$||\bar\beta_j^n - \dot\psi^*(j/n)|| \le K\|\psi^*(j/n)-x_j\|.$$

Further, $\bar\beta_j^n = \bar\beta_j^{1,n}+\bar\beta_j^{2,n}$, with

$$\bar\beta_j^{1,n} = \int_{\mathcal{S}} b(x_j,\xi)\mu_{j,n}^*(d\xi) \quad \text{and} \quad \bar\beta_j^{2,n} = \int_{\mathbb{R}^d} y\nu_{j,n}^*(dy).$$

Here $\mu_{j,n}^*$ is the invariant measure corresponding to the kernel $\gamma_{j,n}^*$ defined for $B_1\in\mathcal{B}(\mathcal{S})$ as

$$\gamma_{j,n}^*(B_1|\psi^*(j/n),\xi) = \int_{B_1}\exp\{\langle\alpha,b(\psi^*(j/n),\zeta)\rangle - \Lambda(\psi^*(j/n),\alpha) + \Psi_\sigma(\psi^*(j/n);\alpha,\zeta)-\Psi_\sigma(\psi^*(j/n);\alpha,\xi)\}p(d\zeta|\psi^*(j/n),\xi),$$

and for $B_2 \in \mathcal{B}(\mathbb{R}^d)$ as

$$\nu^*_{j,n}(B_2) = \int_{B_2} \exp\left\{ \langle \alpha, y\rangle - \frac{\sigma^2||\alpha||^2}{2} \right\} \rho_\sigma(dy).$$

Note that $\alpha = \alpha(\psi^*(j/n), \dot\psi^*(j/n))$ and $\dot\psi^*(j/n) = \int_{\mathcal{S}} b(\psi^*(j/n),\xi)\mu^*_{j,n}(d\xi) + \int_{\mathbb{R}^d} y\nu^*_{j,n}(dy)$. We observe that, from part (b) of Lemma B.2 in Appendix B,

$$\begin{aligned} L(x_j, \bar\beta^{1,n}_j) &\le \int_{\mathcal{S}} R(\gamma^*_{j,n}(\cdot|\psi^*(j/n),\xi)\|p(\cdot|x_j,\xi))\mu^*_{j,n}(d\xi) \\ &= \langle \alpha(\psi^*(j/n), \dot\psi^*(j/n)), \bar\beta^n_j - \bar\beta^{2,n}_j\rangle - \Lambda(\psi^*(j/n), \alpha(\psi^*(j/n), \dot\psi^*(j/n))) \\ &\le L(\psi^*(j/n), \bar\beta^n_j - \bar\beta^{2,n}_j). \end{aligned}$$

Now, from part (c) of Lemma B.3 and for $\bar\beta^n_j$ as in (4.6), the stochastic kernel $\gamma^{1,n}_j(\cdot|x_j,\xi)$ on $\mathcal{S}$ given $\mathcal{S}\times\mathbb{R}^d$ (with invariant measure μ^n_j) and the measure $\gamma^{2,n}_j$ on $\mathbb{R}^d$ given by

$$\gamma^{1,n}_j(B_1|x_j,\xi) = \int_{B_1} e^{\langle \alpha, b(x_j,\zeta)\rangle + \Psi_\sigma(x_j;\alpha,\zeta) - \Psi_\sigma(x_j;\alpha,\xi) - \Lambda(x_j,\alpha)} p(d\zeta|x_j,\xi) \tag{4.7}$$

and

$$\gamma^{2,n}_j(B_2) = \int_{B_2} e^{\langle \alpha, y\rangle - \frac{\sigma^2}{2}\|\alpha\|^2} \rho_\sigma(dy)$$

achieve the infimum in the representation for L_σ. That is,

$$L_\sigma(x_j, \bar\beta^n_j) = \int_{\mathcal{S}} R(\gamma^{1,n}_j(\cdot|x_j,\xi)||p(\cdot|x_j,\xi))\mu^n_j(d\xi) + R(\gamma^{2,n}_j(\cdot)||\rho_\sigma(\cdot)), \tag{4.8}$$

and

$$\bar\beta^n_j = \int_{\mathcal{S}} b(x_j,\xi)\mu^n_j(d\xi) + \int_{\mathbb{R}^d} y\gamma^{2,n}_j(dy) = \bar\beta^{1,n}_j + \bar\beta^{2,n}_j,$$

where we have used the fact that [4, Corollary C.3.3] for any probability measures γ and θ on $\mathcal{S}$, and λ and μ on $\mathbb{R}^d$,

$$R(\gamma\times\lambda\|\theta\times\mu) = R(\gamma\|\theta) + R(\lambda\|\mu). \tag{4.9}$$

Note that $\alpha = \alpha(x_j, \bar\beta^n_j) = \alpha(x_j, \psi^*(j/n), \dot\psi^*(j/n))$ in both $\gamma^{1,n}_j$ and $\gamma^{2,n}_j$. Hence $\gamma^{2,n}_j$ depends implicitly on x_j (through α), but we do not write this dependence explicitly for ease of notation.

We now use the kernels $\gamma^{1,n}_j$ and $\gamma^{2,n}_j$ to finish the definition of the required sequence of admissible controls. As was the case in section 3, grouping for the purposes of averaging motivates part of the construction. Let $\{m_n, n\in\mathbb{N}\}$ be a sequence as the one used there, so that $m_n\to\infty$ and $k_n = m_n/n \to 0$ as $n\to\infty$. Let $l \in \{0,\dots,\frac{1}{k_n} - 1\}$. Then for $lm_n \le j < (l+1)m_n - 1$ we define

$$\nu^{1,n}_j(d\zeta|x_0,\dots,x_j,u_0,\dots,u_j,\xi_j) \doteq \begin{cases} \gamma^{1,n}_{lm_n}(d\zeta|x_{lm_n},\xi_j) & \text{if } \max\limits_{0\le i\le j}\|x_i - \psi^*(i/n)\| \le \eta, \\ p(d\zeta|x_j,\xi_j) & \text{if } \max\limits_{0\le i\le j}\|x_i - \psi^*(i/n)\| > \eta \end{cases} \tag{4.10}$$

and

$$\nu_j^{2,n}(dy|x_0,\dots,x_j,u_0,\dots,u_j,\xi_j) \doteq \begin{cases} \gamma_{lm_n}^{2,n}(dy) & \text{if} \quad \max_{0\le i\le j}\|x_i-\psi^*(i/n)\|\le\eta, \\ \rho_\sigma(dy) & \text{if} \quad \max_{0\le i\le j}|x_i-\psi^*(i/n)|>\eta. \end{cases} \tag{4.11}$$

To simplify notation we have not made explicit the dependence on σ of $\nu_j^{1,n}$ and $\nu_j^{2,n}$. Finally, we define the required admissible control sequence $\{\nu_{j,prod}^n, j=0,\dots,n-1\}$ on $\mathcal{S}\times\mathbb{R}^d$ as

$$\nu_{j,prod}^n(d\zeta\times dy) = \nu_j^{1,n}(d\zeta)\times\nu_j^{2,n}(dy). \tag{4.12}$$

To show that the control sequence just constructed is nearly optimal in (A.3), we compute the associated running cost directly. In what follows, $\bar X_j^n$, $\bar Z_j^n$, $j=0,\dots,n$, are controlled random variables associated with the sequence $\{\nu_j^n\}$ through definitions (A.1) and (A.2).

Let $\tau^n \doteq \frac{1}{n}(\min\{j\in\{0,1,\dots,n\} : \|\bar X_j^n-\psi^*(j/n)\|>\eta\}\wedge n)$. Then (4.9) and the definition of τ^n give

$$\begin{aligned}
&\bar E_{x,\xi}\left\{\frac{1}{n}\sum_{j=0}^{n-1} R(\nu_{j,prod}^n(\cdot)||(p\times\rho_\sigma)(\cdot|\bar X_j^n,\bar Z_j^n))\right\} \\
&= \bar E_{x,\xi}\left\{\frac{1}{n}\sum_{j=0}^{n\tau^n-1}\left[R(\nu_j^{1,n}(\cdot)\|p(\cdot|\bar X_j^n,\bar Z_j^n)) + R(\nu_j^{2,n}(\cdot)\|\rho_\sigma(\cdot))\right]\right\} \\
&= \bar E_{x,\xi}\left\{\sum_{l=0}^{q_n-1} k_n\left[\frac{1}{m_n}\sum_{j=lm_n}^{(l+1)m_n-1} R(\gamma_{lm_n}^{1,n}(\cdot|\bar X_{lm_n}^n,\bar Z_j^n)\|p(\cdot|\bar X_j^n,\bar Z_j^n)) + R(\gamma_{lm_n}^{2,n}(\cdot)\|\rho_\sigma(\cdot))\right]\right. \\
&\qquad \left. + \frac{1}{n}\sum_{j=q_nm_n}^{n\tau^n-1}\left[R(\gamma_{q_nm_n}^{1,n}(\cdot|\bar X_{q_nm_n}^n,\bar Z_j^n)\|p(\cdot|\bar X_j^n,\bar Z_j^n)) + R(\gamma_{q_nm_n}^{2,n}(\cdot)||\rho_\sigma(\cdot))\right]\right\},
\end{aligned} \tag{4.13}$$

where q_n is such that $n\tau^n = q_nm_n + r_n$, with $0\le r_n<m_n$ and $q_n, r_n\in\mathbb{N}_0$.

To continue our estimates on the running costs, we must prove the following claim: for each $j\le n\tau^n-1$, $lm_n\le j\le (l+1)m_n-1$ for some $l\in\{0,\dots,q_n\}$, and n large enough,

$$R(\gamma_{lm_n}^{1,n}(\cdot|\bar X_{lm_n}^n,\bar Z_j^n)\|p(\cdot|\bar X_j^n,\bar Z_j^n)) \le R(\gamma_{lm_n}^{1,n}(\cdot|\bar X_{lm_n}^n,\bar Z_j^n)\|p(\cdot|\bar X_{lm_n}^n,\bar Z_j^n)) + \sigma. \tag{4.14}$$

We first note that part (c) of Hypothesis H.1 implies that for any $x,y\in\Delta$, there exists $\delta>0$ such that for $\|x-y\|<\delta$

$$\frac{\tilde p^y(\xi,\zeta)}{\tilde p^x(\xi,\zeta)} = 1+\frac{\tilde p^y(\xi,\zeta)-\tilde p^x(\xi,\zeta)}{\tilde p^x(\xi,\zeta)} \le 1+\frac{\tilde p^y(\xi,\zeta)-\tilde p^x(\xi,\zeta)}{a} \le e^\sigma. \tag{4.15}$$

Then taking n large enough so that $m_n\|b\|_\infty/n<\delta$, we have $\|\bar X_{lm_n+i}^n - \bar X_{lm_n}^n\|\le \frac{i}{n}\|b\|_\infty<\delta$ for $0\le i<m_n$, and hence

$$\begin{aligned}
&\gamma_{lm_n}^{1,n}(B_1|\bar X_{lm_n}^n,\bar Z_j^n) \\
&= \int_{B_1} e^{\langle\alpha,b(\bar X_{lm_n}^n,\zeta)\rangle+\Psi_\sigma(\bar X_{lm_n}^n;\alpha,\zeta)-\Psi_\sigma(\bar X_{lm_n}^n;\alpha,\xi)-\Lambda(\bar X_{lm_n}^n,\alpha)} p(d\zeta|\bar X_{lm_n}^n,\bar Z_j^n) \\
&\le e^\sigma\int_{B_1} e^{\langle\alpha,b(\bar X_{lm_n}^n,\zeta)\rangle+\Psi_\sigma(\bar X_{lm_n}^n;\alpha,\zeta)-\Psi_\sigma(\bar X_{lm_n}^n;\alpha,\xi)-\Lambda(\bar X_{lm_n}^n,\alpha)} \tilde p^{\bar X_j^n}(\bar Z_j^n,\zeta)\vartheta(d\zeta)
\end{aligned}$$

for any $B_1 \in \mathcal{B}(\mathcal{S})$. From the above we get that $\gamma^{1,n}_{lm_n}(\cdot|\bar{X}^n_{lm_n}, \bar{Z}^n_j)$ is absolutely continuous with respect to $p(\cdot|\bar{X}^n_j, \bar{Z}^n_j)$ and that

$$\frac{d\gamma^{1,n}_{lm_n}(\cdot|\bar{X}^n_{lm_n}, \bar{Z}^n_j)}{dp(\cdot|\bar{X}^n_j, \bar{Z}^n_j)} \leq e^{\sigma} \cdot \frac{d\gamma^{1,n}_{lm_n}(\cdot|\bar{X}^n_{lm_n}, \bar{Z}^n_j)}{dp(\cdot|\bar{X}^n_{lm_n}, \bar{Z}^n_j)},$$

which implies (4.14).

Now fix $\bar{\xi} \in \mathcal{S}$ and normalize Ψ_σ in such a way that $\Psi_\sigma(x;\alpha,\bar{\xi}) = 0$. Then, observing that

$$\frac{a}{A} e^{\Psi_\sigma(x;\alpha,\xi_1)} \leq e^{\Psi_\sigma(x;\alpha,\xi_2)} \leq \frac{A}{a} e^{\Psi_\sigma(x;\alpha,\xi_1)}$$

for all $\xi_1, \xi_2 \in \mathcal{S}$, and taking $\xi_1 = \bar{\xi}$, we get that

$$\frac{a}{A} \leq e^{\Psi_\sigma(x;,\alpha,\xi)} \leq \frac{A}{a} \quad \forall \xi \in \mathcal{S},\ x,\ \alpha \in \mathbb{R}^d.$$

Then, from (4.7),

$$\gamma^{1,n}_{lm_n}(d\zeta|x,\xi) = e^{\langle \alpha, b(x,\zeta)\rangle + \Psi_\sigma(x;\alpha,\zeta) - \Psi_\sigma(x;\alpha,\xi) - \Lambda(x,\alpha)} \tilde{p}^x(\xi,\zeta)\vartheta(d\zeta),$$

with $\alpha = \alpha(x, \bar{\beta}^n_{lm_n})$, $x \in \Delta$, and $\bar{\beta}^n_{lm_n}$ to satisfy (4.6), and hence

$$\gamma^{1,n}_{lm_n}(d\zeta|x,\xi) \geq \frac{a^3}{A^2} e^{\langle \alpha, b(x,\zeta)\rangle - \Lambda(x,\alpha)} \vartheta(d\zeta) \geq \frac{a^3}{A^2} e^{-2||b||_\infty \max\limits_{x\in\Delta,\beta\in\Theta}\{\|\alpha(x,\beta)\|\}} \vartheta(d\zeta),$$

where $\Theta \doteq \cup_{t\in[0,1]}\{\beta \in \mathbb{R}^d : ||\beta - \dot{\psi}^*(t)|| \leq K\}$. Using the fact that $(x,\beta) \mapsto \alpha(x,\beta)$ is continuous (part (f) of Lemma B.3), we get that for $x \in \Delta$, $\bar{\beta}^n_{\ell m_n}$ belongs to Θ and, moreover, that $\max_{x\in\Delta,\beta\in\Theta}\{\|\alpha(x,\beta)\|\}$ is bounded. Denoting the jth iteration of the kernel $\gamma^{1,n}_{lm_n}$ by $\gamma^{1,n,j}_{lm_n}$, we observe that [13, Theorem 16.0.2]

$$\|\gamma^{1,n,j}_{lm_n}(\cdot|x,\zeta) - \mu^n_{lm_n}(\cdot)\| \leq \left(1 - \frac{a^3}{A^2} e^{-2\max_{x,\zeta}|\langle \alpha, b(x,\zeta)\rangle|}\right)^j. \tag{4.16}$$

We complete the estimate on the running cost for our admissible control sequence in the inequalities that follow, using (4.14), standard properties of conditional expectation, and (4.7). We have that (4.13) is less than or equal to

$$\begin{aligned}
&\bar{E}_{x,\xi}\left\{\sum_{l=0}^{q_n-1} k_n \left[\frac{1}{m_n}\sum_{j=lm_n}^{(l+1)m_n-1} R(\gamma^{1,n}_{lm_n}(\cdot|\bar{X}^n_{lm_n}, \bar{Z}^n_j)||p(\cdot|\bar{X}^n_{lm_n}, \bar{Z}^n_j)) + R(\gamma^{2,n}_{lm_n}(\cdot)||\rho_\sigma(\cdot))\right]\right.\\
&\quad \left. + \frac{1}{n}\sum_{j=q_nm_n}^{n\tau^n-1}\left[R(\gamma^{1,n}_{q_nm_n}(\cdot|\bar{X}_{q_nm_n}, \bar{Z}^n_{q_nm_n})||p(\cdot|\bar{X}^n_{q_nm_n}, \bar{Z}^n_j)) + R(\gamma^{2,n}_{q_nm_n}(\cdot)||\rho_\sigma(\cdot))\right]\right\}\\
&\quad + \sigma\\
&\leq \sum_{r=1}^{1/k_n} k_n \bar{E}_{x,\xi}\left\{\sum_{l=0}^{r-1}\left[\frac{1}{m_n}\sum_{j=lm_n}^{(l+1)m_n-1} R(\gamma^{1,n}_{lm_n}(\cdot|\bar{X}^n_{lm_n}, \bar{Z}^n_j)||p(\cdot|\bar{X}^n_{lm_n}, \bar{Z}^n_j))\right.\right.\\
&\qquad \left.\left. + R(\gamma^{2,n}_{lm_n}(\cdot)||\rho_\sigma(\cdot))\right] 1_{[(r-1)m_n < n\tau^n \leq rm_n]}\right\} + \sigma
\end{aligned} \tag{4.17}$$

$$
\begin{aligned}
&= \sum_{r=1}^{1/k_n} k_n \sum_{l=0}^{r-1} \bar{E}_{x,\xi} \Bigg\{ 1_{[(r-1)m_n < n\tau^n \le r m_n]} \\
&\qquad \cdot \bar{E}_{x,\xi} \Bigg\{ \frac{1}{m_n} \sum_{j=lm_n}^{(l+1)m_n-1} R(\gamma^{1,n}_{lm_n}(\cdot|\bar{X}^n_{lm_n}, \bar{Z}^n_j)||p(\cdot|\bar{X}^n_{lm_n}, \bar{Z}^n_j)) \\
&\qquad\qquad + \; R(\gamma^{2,n}_{lm_n}(\cdot)||\rho_\sigma(\cdot))|\bar{Z}^n_0, \ldots, \bar{Z}^n_{lm_n}, \bar{X}^n_0, \ldots, \bar{X}^n_{rm_n} \Bigg\} \Bigg\} \\
&\quad + \sigma \\
&= \sum_{r=1}^{1/k_n} k_n \sum_{l=0}^{r-1} \bar{E}_{x,\xi} \Bigg\{ 1_{[(r-1)m_n < n\tau^n \le r m_n]} \\
&\qquad \cdot \bar{E}_{x,\xi} \Bigg\{ \frac{1}{m_n} \sum_{j=lm_n}^{(l+1)m_n-1} \Bigg[\Big\langle \alpha_{lm_n}, \int_{\mathcal{S}} b(\bar{X}^n_{lm_n}, \zeta) \gamma^{1,n}_{lm_n}(d\zeta|\bar{X}^n_{lm_n}, \bar{Z}^n_j) \Big\rangle \\
&\qquad\qquad + \int_{\mathcal{S}} \Psi_\sigma(\bar{X}^n_{lm_n}; \alpha_{lm_n}, \zeta) \gamma^{1,n}_{lm_n}(d\zeta|\bar{X}^n_{lm_n}, \bar{Z}^n_j) \\
&\qquad\qquad - \Psi_\sigma(\bar{X}^n_{lm_n}; \alpha_{lm_n}, \bar{Z}^n_j) \\
&\qquad\qquad - \Lambda(\bar{X}^n_{lm_n}, \alpha_{lm_n}) \Bigg] \\
&\qquad\qquad + \; R(\gamma^{2,n}_{lm_n}(\cdot)||\rho_\sigma(\cdot))|\bar{Z}^n_0, \ldots, \bar{Z}^n_{lm_n}, \bar{X}^n_0, \ldots, \bar{X}^n_{lm_n} \Bigg\} \Bigg\} \\
&\quad + \sigma.
\end{aligned}
$$

Now, adding and subtracting $\int_{\mathcal{S}} R(\gamma^{1,n}_{lm_n}(\cdot|\bar{X}^n_{lm_n}, \xi)||p(\cdot|\bar{X}^n_{lm_n}, \xi))\mu^n_{lm_n}(d\xi)$ inside the expectation and collecting terms, we get that the above expression is less than or equal to

$$
\begin{aligned}
&\sum_{r=1}^{1/k_n} k_n \sum_{l=0}^{r-1} \bar{E}_{x,\xi} \Bigg\{ \Bigg[\frac{1}{m_n} \sum_{j=lm_n}^{(l+1)m_n-1} \Big\langle \alpha_{lm_n}, \int_{\mathcal{S}} b(\bar{X}^n_{lm_n}, \zeta) \gamma^{1,n,j+1}_{lm_n}(d\zeta|\bar{X}^n_{lm_n}, \bar{Z}^n_{lm_n}) \\
&\qquad\qquad - \int_{\mathcal{S}} b(\bar{X}^n_{lm_n}, \zeta) \mu^n_{lm_n}(d\zeta) \Big\rangle \\
&\qquad + \int_{\mathcal{S}} R(\gamma^{1,n}_{lm_n}(\cdot|\bar{X}^n_{lm_n}, \xi)||p(\cdot|\bar{X}^n_{lm_n}, \xi)) \mu_{lm_n}(d\xi) \\
&\qquad + \; R(\gamma^{2,n}_{lm_n}(\cdot)||\rho_\sigma(\cdot)) \Bigg] 1_{[(r-1)m_n < n\tau^n \le r m_n]} \Bigg\} + \frac{4}{n} \ln \frac{A}{a} + \sigma \qquad (4.18) \\
&\le \frac{4}{n} \ln \frac{A}{a} + \sigma + \sum_{r=1}^{1/k_n} k_n \sum_{l=0}^{r-1} \bar{E}_{x,\xi} \Big\{ 1_{[(r-1)m_n < n\tau^n \le r m_n]} \\
&\quad \cdot \Big[\frac{||b||_\infty A^2}{m_n a^3} \max_{x \in \Delta, \beta \in \Theta} \{\|\alpha(x,\beta)\|\} e^{2||b||_\infty \max_{x \in \Delta, \beta \in \Theta} \{\|\alpha(x,\beta)\|\}}
\end{aligned}
$$

$$
\begin{aligned}
&+ \int_{\mathcal{S}} R(\gamma^{1,n}_{lm_n}(\cdot|\bar{X}^n_{lm_n},\xi)||p(\cdot|\bar{X}^n_{lm_n},\xi))\mu^n_{lm_n}(d\xi) + R(\gamma^{2,n}_{lm_n}(\cdot)||\rho_\sigma(\cdot))\Big]\Big\} \\
&\le \frac{4}{n}\ln\frac{A}{a} + \sigma + \frac{||b||_\infty A^2}{na^3}\max_{x\in\Delta,\beta\in\Theta}\{||\alpha(x,\beta)||\}e^{2||b||_\infty \max_{x\in\Delta,\beta\in\Theta}\{\|\alpha(x,\beta)\|\}} \\
&\quad + \sum_{r=1}^{1/k_n} k_n \sum_{l=0}^{r-1} \bar{E}_{x,\xi}\Big\{1_{[(r-1)m_n<n\tau^n\le rm_n]}\Big[L_\sigma(\bar{X}^n_{lm_n},\bar{\beta}^n_{lm_n})\Big]\Big\} \\
&\le \sum_{r=1}^{1/k_n} k_n \sum_{l=0}^{r-1} \bar{E}_{x,\xi}\Big\{1_{[(r-1)m_n<n\tau^n\le rm_n]}L_\sigma(\psi^*(lm_n/n),\dot{\psi}^*(lm_n/n))\Big\} \\
&\quad + \frac{4}{n}\ln\frac{A}{a} + 2\sigma + \frac{||b||_\infty A^2}{na^3}\max_{x\in\Delta,\beta\in\Theta}\{\|\alpha(x,\beta)\|\}e^{2||b||_\infty \max_{x\in\Delta,\beta\in\Theta}\{\|\alpha(x,\beta)\|\}} \\
&\le \bar{E}_{x,\xi}\left\{\sum_{l=0}^{[\frac{\tau^n}{k_n}]} k_n L_\sigma(\psi^*(lm_n/n),\dot{\psi}^*(lm_n/n))\right\} + 3\sigma \ \text{ for } n \text{ large enough,} \\
&\le k_n \sum_{l=0}^{[1/k_n]} L_\sigma(\psi^*(lm_n/n),\dot{\psi}^*(lm_n/n)) + 3\sigma.
\end{aligned}
$$

In the first, second, and third inequalities we have used (4.16), (4.8), and (4.6), respectively. We conclude that the admissible control sequence that we constructed has a running cost which is nearly optimal, as we had claimed.

We can now return to the proof of (4.5). Using (A.3), (4.18), and Lemma B.3(e) (with $\varepsilon = \sigma$), we get that

$$
\begin{aligned}
&\limsup_{n\to\infty} W^n_\sigma(x,\xi) \\
&\le \limsup_{n\to\infty} \bar{E}_{x,\xi}\left\{\frac{1}{n}\sum_{j=1}^{n-1}[R(\nu^{1,n}_j(\cdot)||p(\cdot|\bar{X}^n_j,\bar{Z}^n_j)) + R(\nu^{2,n}_j(\cdot)||\rho(\cdot))] + h(\bar{Y}^n)\right\} \\
&\le \int_0^1 L_\sigma(\psi^*(t),\dot{\psi}^*(t))dt + 3\sigma + \limsup_{n\to\infty}\bar{E}_{x,\xi}\{h(\bar{Y}^n)\} \\
&\le \int_0^1 L_\sigma(\psi(t),\dot{\psi}(t))dt + 4\sigma + \limsup_{n\to\infty}\bar{E}_{x,\xi}\{h(\bar{Y}^n)\}.
\end{aligned}
$$

Thus, the proof of (4.5) will be complete once we prove that

$$
\limsup_{n\to\infty}\bar{E}_{x,\xi}\{h(\bar{Y}^n)\} \le h(\psi) + \tilde{\theta}(\sigma),
$$

with $\tilde{\theta}(\sigma)\to 0$ when $\sigma\to 0$. This in turn will be implied by

$$
\lim_{\sigma\to 0}\limsup_{n\to\infty}\bar{P}_{x,\xi}\left\{\sup_{t\in[0,1]}||\bar{Y}^n(t) - \psi^*(t)|| \ge \sigma\right\} = 0, \tag{4.19}
$$

because of the Lipschitz property of h and part (e) of Lemma B.3.

To show (4.19), it is convenient to define a sequence of control measures associated with the controls $\nu^{1,n}_j$ and $\nu^{2,n}_j$ given in (4.10) and (4.11). For $B_1, B_2 \in \mathcal{B}(S)$, $B \in$

$\mathcal{B}(\mathbb{R}^d)$, define

$$\tilde{\nu}_l^{1,n}(B_1 \times B_2) \doteq \frac{1}{m_n} \sum_{j=l_n}^{(l+1)m_n-1} \delta_{\bar{Z}_j^n}(B_1) \times \nu_j^{1,n}(B_2|\bar{X}_0^n, \ldots, \bar{X}_j^n, \bar{Z}_j^n),$$

$$\tilde{\nu}_l^{2,n}(B) \doteq \frac{1}{m_n} \sum_{j=lm_n}^{(l+1)m_n-1} \nu_j^{2,n}(B|\bar{X}_0^n, \ldots, \bar{X}_j^n),$$

$$\tilde{\nu}^{1,n}(\cdot|t) \doteq \begin{cases} \tilde{\nu}_l^{1,n}(\cdot) & \text{if} \quad t \in [lk_n, (l+1)k_n) \quad \text{for } l = 0, \ldots, 1/k_n - 2, \\ \tilde{\nu}_{(\frac{1}{k_n}-1)}^{1,n}(\cdot) & \text{if} \quad t \in [1-k_n, 1], \end{cases}$$

and

$$\tilde{\nu}^{2,n}(\cdot|t) \doteq \begin{cases} \tilde{\nu}_l^{2,n}(\cdot) & \text{if} \quad t \in [lk_n, (l+1)k_n) \quad \text{for } l = 0, \ldots, 1/k_n - 2, \\ \tilde{\nu}_{(\frac{1}{k_n}-1)}^{2,n}(\cdot) & \text{if} \quad t \in [1-k_n, 1]. \end{cases}$$

With B_1, B_2, and B as before, and with $C \in \mathcal{B}([0,1])$, now define the random measures $\nu^{1,n}$ and $\nu^{2,n}$ on $\mathcal{S} \times \mathcal{S} \times [0,1]$ and $\mathbb{R}^d \times [0,1]$, respectively, by

$$\nu^{1,n}(B_1 \times B_2 \times C) \doteq \int_C \nu^{1,n}(B_1 \times B_2|t)dt \quad \text{and} \quad \nu^{2,n}(B \times C) \doteq \int_C \nu^{2,n}(B|t)dt.$$

Finally, let ν_{prod}^n be the random measure on $\mathcal{S} \times \mathcal{S} \times \mathbb{R}^d \times [0,1]$ defined as

$$\nu_{prod}^n(B_1 \times B_2 \times B \times C) = \int_C \nu^{1,n}(B_1 \times B_2|t)\nu^{2,n}(B|t)dt. \tag{4.20}$$

Let us also define

$$S^{1,n}(t) \doteq x + \int_{\mathcal{S}\times\mathcal{S}\times[0,t]} b(S^{1,n}(s), y)\nu^{1,n}(d\zeta \times dy \times ds)$$

and

$$S^{2,n}(t) \doteq \int_{\mathbb{R}^d\times[0,t]} y\nu^{2,n}(dy \times ds).$$

Since L_σ is continuous, ψ^* is continuous and $\dot{\psi}^*$ has only a finite number of discontinuities,

$$\sup_n \bar{E}_{x,\xi}\left\{\frac{1}{n}\sum_{j=0}^{n-1} R(\nu_j^{1,n}(\cdot) \times \nu_j^{2,n}(\cdot)||p(\cdot|\bar{X}_j^n, \bar{Z}_j^n) \times \rho_\sigma(\cdot))\right\} < \infty.$$

Theorem 5.3.5 in [4], the fact that $\mathcal{S}$ and $[0,1]$ are compact, and arguments analogous to the proof of Theorem C.1(e) then imply that given any subsequence of $\{(\nu^{1,n}, \nu^{2,n}, \bar{X}^n, \bar{U}^n, \tau^n, S^{1,n}, S^{2,n}), n \in \mathbb{N}\}$ there exists a subsubsequence such that $(\nu^{1,n}, \nu^{2,n}, \bar{X}^n, \bar{U}^n, \tau^n, S^{1,n}, S^{2,n})$ converges in distribution to $(\nu^1, \nu^2, \bar{X}, \bar{U}, \tau, \bar{X}, \bar{U})$ when $n \to \infty$. We define $\bar{Y}(t) \doteq \lim_{n\to\infty}(\bar{X}^n(t) + \bar{U}^n(t))$.

From the definition of $\bar{\beta}_j^{1,n}$ and Lemma B.3(d), for each $l \in \{0, \ldots, [\frac{\tau^n}{k_n}]\}$,

$$\begin{aligned} ||\bar{\beta}_{lm_n}^n - \dot{\psi}^*(lm_n/n)|| &= ||\bar{\beta}_{lm_n}^{1,n} + \bar{\beta}_{lm_n}^{2,n} - \dot{\psi}^*(lm_n/n)|| \\ &= \left\| \int_{\mathcal{S}} b(\bar{X}_{lm_n}^n, \xi)\mu_{lm_n}^n(d\xi) + \int_{\mathbb{R}^d} y\tilde{\nu}_{lm_n}^{2,n}(dy) - \dot{\psi}^*(lm_n/n) \right\| \\ &\leq K||\bar{X}_{lm_n}^n - \psi^*(lm_n/n)||. \end{aligned}$$

Then for $t \in [0,\tau]$

$$
\begin{aligned}
||\bar{Y}(t) - \psi^*(t)|| = \lim_{n\to\infty} \Bigg\| & \int_{\mathcal{S}\times\mathcal{S}\times[0,t]} b(S^{1,n}(s),y)\nu^{1,n}(d\zeta\times dy\times ds) \\
& - \int_{\mathcal{S}\times\mathcal{S}\times[0,t]} b(\tilde{X}^n(s),y)\nu^{1,n}(d\zeta\times dy\times ds) \\
& + \int_{\mathcal{S}\times\mathcal{S}\times[0,t]} b(\tilde{X}^n(s),y)\nu^{1,n}(d\zeta\times dy\times ds) \\
& + \sum_{l=0}^{[\frac{t}{k_n}]} k_n \int_{\mathbb{R}^d} y\tilde{\nu}^{2,n}_{lm_n}(dy) - \sum_{l=0}^{[\frac{t}{k_n}]} k_n \int_{\mathcal{S}} b(\tilde{X}^n_{lm_n},\xi)\mu^n_{lm_n}(d\xi) \\
& + \sum_{l=0}^{[\frac{t}{k_n}]} k_n \int_{\mathcal{S}} b(\tilde{X}^n_{lm_n},\xi)\mu^n_{lm_n}(d\xi) - \int_0^t \dot{\psi}^*(s)ds \Bigg\| \\
\le \limsup_{n\to\infty} & \int_0^1 (K||S^{1,n}(s) - \tilde{X}^n(s)|| \wedge 2\|b\|_\infty)ds \\
& + \limsup_{n\to\infty} \sum_{l=0}^{[\frac{t}{k_n}]} k_n \Bigg\| \frac{1}{m_n} \sum_{j=l}^{(l+1)m_n} \int_{\mathcal{S}} b(\bar{X}_{lm_n},\zeta)\gamma^{1,n,j}_{lm_n}(d\zeta|\bar{X}^n_{lm_n},\bar{Z}_{lm_n}) \\
& \qquad\qquad - \int_{\mathcal{S}} b(\bar{X}^n_{lm_n},\xi)\mu^n_{lm_n}(d\xi) \Bigg\| \\
& + \limsup_{n\to\infty} K\int_0^t ||\bar{X}^n(s) - \psi^*(s)||ds \\
\le \limsup_{n\to\infty} & \sum_{l=0}^{[\frac{t}{k_n}]} k_n||b||_\infty \frac{A^2}{na^3} e^{2||b||_\infty \max_{x\in\Delta,\beta\in\Theta}\{\|\alpha(x,\beta)\|\}} \\
& + K\int_0^t ||\bar{X}(s) - \psi^*(s)||ds \\
\le K & \sup_{s\in[0,\tau]} ||\bar{X}(s) - \bar{Y}(s)|| + K\int_0^t ||\bar{Y}(s) - \psi^*(s)||ds.
\end{aligned}
$$

Gronwall's inequality with $\bar{K} \doteq Ke^K$ gives

$$
\sup_{s\in[0,\tau]} ||\bar{Y}(s) - \psi^*(s)|| \le \bar{K} \sup_{s\in[0,\tau]} ||\bar{X}(s) - \bar{Y}(s)|| = \bar{K} \sup_{s\in[0,\tau]} ||\bar{U}(s)||,
$$

which together with Lemma C.2 implies that

$$
\lim_{\sigma\to 0} \bar{P}_{x,\xi}\left\{ \sup_{s\in[0,\tau]} ||\bar{Y}(s) - \psi^*(s)|| \ge \frac{\sigma}{2} \right\} \le \lim_{\sigma\to 0} \bar{P}_{x,\xi}\left\{ \sup_{s\in[0,\tau]} ||\bar{U}(s)|| \ge \frac{\sigma}{2\bar{K}} \right\} = 0.
$$

Finally, writing $\tau = \tau_\sigma$ and following the same arguments given in [4, pp. 205–206], it is proved that $\lim_{\sigma\to 0} \bar{P}_{x,\xi}\{\tau_\sigma < 1\} = 0$. Since $\bar{Y}^n \to \bar{Y}$ w.p.1 uniformly on $[0,1]$, we obtain (4.19), which completes the proof of the lower bound. □

5. The case of noncompact $\mathcal{S}$. Without the assumption of compactness of $\mathcal{S}$, a strong positive recurrence hypothesis on $p(d\zeta|x,\xi)$ is required to guarantee tightness of the measures appearing in the proof of Theorem 2.1. This hypothesis is analogous to Condition 8.2.2 in [4].

Hypothesis H.2. There exists a measurable function $U : \mathcal{S} \to [0,\infty)$ with the following properties:

(a) $\inf_{\zeta\in\mathcal{S}} \left\{U(\zeta) - \log \int_{\mathcal{S}} e^{U(\zeta)} p(d\zeta|x,\xi)\right\} > -\infty$.

(b) For each $M < \infty$ and compact set $\Delta \subset \mathbb{R}^d$, the set

$$Z(M,\Delta) = \left\{(\xi,y) \in \mathcal{S}\times\Delta \,:\, c(\xi,y) := U(\xi) - \log \int_{\mathcal{S}} e^{U(\zeta)} p(d\zeta|y,\xi) \le M\right\}$$

is a compact subset of $\mathcal{S}\times\mathbb{R}^d$.

(c) U is bounded above on every compact subset of $\mathcal{S}$.

Under Hypothesis H.2, part (a) of Theorem C.1 remains valid for $\mathcal{S}$ noncompact, which implies Theorem 2.1 as well. Because the proof requires only small changes in the proofs of Lemma 8.2.4 and Proposition 8.2.5 in [4], we omit the details.

Appendix A. A representation formula. In this appendix we state the variational representation formula required in the proof of the lower bound. It can be derived easily by following the same steps as those given in [4, section 4.4].

Let $p\times\rho_\sigma$ be the stochastic kernel on $S\times\mathbb{R}^d$ given $\xi\in\mathcal{S}$, $x\in\mathbb{R}^d$ defined by $(p\times\rho_\sigma)(d\zeta\times dy|x,\xi) \doteq p(d\zeta|x,\xi)\times\rho_\sigma(dy)$. We consider admissible control sequences consisting of stochastic kernels $\nu_j^n(d\zeta\times dy|\bar X_0^n,\dots,\bar X_j^n,\bar U_0^n,\dots,\bar U_j^n,\bar Z_j^n)$ on $\mathcal{S}\times\mathbb{R}^d$ given $(\mathbb{R}^d)^{j+1}\times(\mathbb{R}^d)^{j+1}\times\mathcal{S}$. For each admissible control sequence $\{\nu_j^n, j=0,\dots,n-1\}$, the controlled system is defined by setting $\bar X_0^n \doteq x$, $\bar U_0^n \doteq 0$ and for $j=0,\dots,n-1$ through

$$\bar X_{j+1}^n \doteq \bar X_j^n + \frac{1}{n}\, b(\bar X_j^n, \bar Z_{j+1}^n),\ \bar U_{j+1}^n \doteq \bar U_j^n + \frac{1}{n}\,\bar G_j^n, \text{ and } \bar Y_j^n = \bar X_j^n + \bar U_j^n, \tag{A.1}$$

where the conditional distribution of $(\bar Z_{j+1}^n, \bar G_j^n)$ is given by

$$\begin{aligned}\bar P_{x,\xi}\left\{(\bar Z_{j+1}^n,\bar G_j^n)\in(d\zeta\times dy)|\bar X_0^n,\dots,\bar X_j^n,\bar U_0^n,\dots,\bar U_j^n,\bar Z_j^n\right\}\\ = \nu_j^n(d\zeta\times dy|\bar X_0^n,\dots,\bar X_j^n,\bar U_0^n,\dots,\bar U_j^n,\bar Z_j^n).\end{aligned} \tag{A.2}$$

We define the processes $\bar X^n \doteq \{\bar X^n(t), t\in[0,1]\}$, $\bar U^n = \{\bar U^n(t), t\in[0,1]\}$, and $\bar Y^n \doteq \{\bar Y^n(t), t\in[0,1]\}$ as the linear interpolations of $\{\bar X_j^n\}$, $\{\bar U_j^n\}$, and $\{\bar Y_j^n\}$, respectively.

THEOREM A.1. *Let $W_\sigma^n(x,\xi) \doteq -1/n \log E_{x,\xi}\{\exp[-nh(Y_\sigma^n)]\}$, where Y_σ^n is defined by* (4.1), *$E_{x,\xi}$ denotes expectation conditioned on $X_0^n = x$ and $Z_0^n = \xi$, and h is a bounded measurable function mapping $\mathcal{C}([0,1]:\mathbb{R}^d)\mapsto\mathbb{R}$. Then for all $n\in\mathbb{N}$, $x\in\mathbb{R}^d$, $\xi\in\mathcal{S}$, and $\sigma>0$, we have the representation*

$$W_\sigma^n(x,\xi) = \inf_{\{\nu_j^n\}} \bar E_{x,\xi}\left\{\frac{1}{n}\sum_{j=0}^{n-1} R(\nu_j^n(\cdot)||(p\times\rho_\sigma)(\cdot|\bar X_j^n,\bar Z_j^n)) + h(\bar Y^n)\right\}. \tag{A.3}$$

Appendix B. Properties of the functions Λ, L, and L_σ. In this appendix we establish properties of the functions $\Lambda(x,\alpha)$ and $L(x,\beta)$ defined in (2.3) and (2.2), respectively, and of the function L_σ defined in (4.2).

LEMMA B.1. *Under Hypothesis* H.1, *the function* $\Lambda(x,\alpha)$ *defined in* (2.3) *satisfies the following properties. For each* $x \in \mathbb{R}^d$, $\Lambda(x,\alpha)$ *is a finite strictly convex function of* $\alpha \in \mathbb{R}^d$ *which is differentiable for all* α. *In addition,* $\Lambda(x,\alpha)$ *is a continuous function of* $(x,\alpha) \in \mathbb{R}^d \times \mathbb{R}^d$.

These properties follow from Lemmas 3.1 and 3.4 in [7] given the relation between the function Λ and the solution to the eigenvalue problem given in (2.1).

Lemma 2.1 in [3] gives a list of properties that are satisfied by the function L, the Legendre–Fenchel transform of Λ. These include convexity and lowersemicontinuity in β, positivity, and uniqueness. Part (a) of the following lemma is also among those properties, and we state it here for use in the proof of part (b), which provides an important variational representation for the function L.

LEMMA B.2. *Under Hypothesis* H.1, *the function* $L(x,\beta)$ *defined in* (2.2) *satisfies the following properties.*

(a) *If* $L(x,\beta)$ *is finite in a neighborhood of* β', *then* $\nabla L(x,\beta')$ *exists and* $L(x,\beta') = \langle \alpha, \beta' \rangle - \Lambda(x,\alpha)$ *if and only if* $\alpha = \nabla L(x,\beta')$.

(b) *For each* x *and* β *in* $\mathbb{R}^d$,

$$L(x,\beta) = \inf \left\{ \int_{\mathcal{S}} R(\gamma(\cdot|x,\xi) \| p(\cdot|x,\xi)) \mu(d\xi) : \gamma\mu = \mu, \int_{\mathcal{S}} b(x,\xi) d\mu = \beta \right\}.$$

If L *is finite, then the infimum is attained uniquely.*

Proof. (a) Since $\Lambda(x,\cdot)$ is strictly convex on $\mathbb{R}^d$, $L(x,\cdot)$ is differentiable on $\mathrm{int}(\mathrm{dom} L(x,\cdot))$. See Theorem D.2.8 in [4]. The last part follows from standard results.

(b) First we consider the case when $\beta \in \mathrm{ri}(\mathrm{dom} L(x,\cdot))$. For $\alpha \in \mathbb{R}^d$ let γ_α be the stochastic kernel defined by

$$\frac{d\gamma_\alpha(\cdot|x,\xi)}{dp(\cdot|x,\xi)}(\zeta) = \frac{e^{\langle \alpha, b(x,\zeta)\rangle + \psi(\zeta)}}{\int_{\mathcal{S}} e^{\langle \alpha, b(x,\zeta)\rangle + \psi(\zeta)} p(d\zeta|x,\xi)}.$$

In terms of the function Λ defined in (2.1) we can write

$$\frac{d\gamma_\alpha(\cdot|x,\xi)}{dp(\cdot|x,\xi)}(\zeta) = e^{-\Lambda(x,\alpha) - \psi(\xi) + \langle \alpha, b(x,\zeta)\rangle + \psi(\zeta)}.$$

Let μ^α be the unique invariant measure of γ_α. (Proposition 4.1 in [7] guarantees that such a measure exists.) Part (a) of the present lemma and the fact that $\Lambda(x,\cdot)$ is strictly convex and differentiable imply that there exists a unique $\alpha = \alpha(x,\beta)$ such that

$$L(x,\beta) = \langle \alpha(x,\beta), \beta \rangle - \Lambda(x, \alpha(x,\beta)) \tag{B.1}$$

with $\alpha(x,\beta) \in \partial L(\beta)$ if and only if $\beta = \nabla \Lambda(x, \alpha(x,\beta))$ (see Corollary 26.3.1 in [14]). Then, Proposition 4.1 in [7] gives

$$E^{\gamma_\alpha}_{\mu^\alpha} b(x,\xi) = \int_{\mathcal{S}} b(x,\xi) \mu^\alpha(d\xi) = \beta. \tag{B.2}$$

Now let γ be any kernel (with corresponding invariant measure μ^γ) satisfying

$$\int_{\mathcal{S}} b(x,\xi) \mu^\gamma(d\xi) = \beta \quad \text{and} \quad \int_{\mathcal{S}} R(\gamma(\cdot|\xi) \| p(\cdot|x,\xi)) \mu^\gamma(\xi) < \infty.$$

Then $\gamma(\cdot|x,\xi) \ll p(\cdot|x,\xi)$ for μ^γ-almost all ξ. Since $\frac{d\gamma_\alpha}{dp}$ is strictly positive, $\gamma(\cdot|x,\xi) \ll \gamma_\alpha(\cdot|x,\xi)$ for almost all ξ (with respect to μ^γ) and

$$\begin{aligned}
&\int_{\mathcal{S}} R(\gamma(\cdot|x,\xi)\|p(\cdot|x,\xi))\mu^\gamma(d\xi)\\
&= \int_{\mathcal{S}}\int_{\mathcal{S}} \log\frac{d\gamma(\cdot|x,\xi)}{dp(\cdot|x,\xi)}(\zeta)\gamma(d\zeta|x,\xi)\mu^\gamma(d\xi)\\
&= \int_{\mathcal{S}}\left[\int_{\mathcal{S}} \log\frac{d\gamma(\cdot|x,\xi)}{d\gamma_\alpha(\cdot|x,\xi)}(\zeta)\gamma(d\zeta|x,\xi) + \int_{\mathcal{S}} \log\frac{d\gamma_\alpha(\cdot|x,\xi)}{dp(\cdot|x,\xi)}(\zeta)\gamma(d\zeta|x,\xi)\right]\mu^\gamma(d\xi)\\
&= \int_{\mathcal{S}} R(\gamma(\cdot|x,\xi)\|\gamma_\alpha(\cdot|x,\xi))\mu^\gamma(d\xi)\\
&\quad + \int_{\mathcal{S}}\int_{\mathcal{S}} \log\Big[\exp\big[\langle\alpha, b(x,\zeta)\rangle + \psi(\zeta) - \psi(\xi) - \Lambda(x,\alpha)\big]\Big]\gamma(d\zeta|x,\xi)\mu^\gamma(d\xi)\\
&= \int_{\mathcal{S}} R(\gamma(\cdot|x,\xi)\|\gamma_\alpha(\cdot|x,\xi))\mu^\gamma(d\xi) - \Lambda(x,\alpha)\\
&\quad - \int_{\mathcal{S}}\psi(\xi)\mu^\gamma(d\xi) + \int_{\mathcal{S}}\int_{\mathcal{S}}\left[\langle\alpha, b(x,\zeta)\rangle + \psi(\zeta)\right]\gamma(d\zeta|x,\xi)\mu^\gamma(d\xi)\\
&= \int_{\mathcal{S}} R(\gamma(\cdot|x,\xi)\|\gamma_\alpha(\cdot|x,\xi))\mu^\gamma(d\xi) - \Lambda(x,\alpha) + \int_{\mathcal{S}}\langle\alpha, b(x,\xi)\rangle\mu^\gamma(d\xi)\\
&= \int_{\mathcal{S}} R(\gamma(\cdot|x,\xi)\|\gamma_\alpha(\cdot|x,\xi))\mu^\gamma(d\xi) + L(x,\beta) \geq L(x,\beta).
\end{aligned}$$

Equality is obtained if and only if $\gamma \equiv \gamma_\alpha$. If β does not belong to $\mathrm{ri}(\mathrm{dom}L(x,\cdot))$, analogous arguments to those given in Appendix C.5 in [4] can be adapted. □

The next result establishes properties of the function L_σ defined in (4.2) that are needed in the proof of the lower bound.

LEMMA B.3. *Given $\sigma > 0$, the function $L_\sigma(x,\beta)$ satisfies the following properties:*

(a) $L_\sigma(x,\beta) = \inf_{z\in\mathbb{R}^d}\{L(x,\beta - z) + \frac{||z||^2}{2\sigma^2}\}$ *and* $L_\sigma(x,\beta) \leq L(x,\beta)$.

(b) $L_\sigma(x,\beta)$ *is a finite, nonnegative, continuous function of* $(x,\beta) \in \mathbb{R}^d \times \mathbb{R}^d$. *Moreover,* $L_\sigma(x,\cdot)$ *is differentiable on* $\mathbb{R}^d$.

(c)

$$\begin{aligned}
L_\sigma(x,\beta) = \inf\Big\{&\int_{\mathcal{S}} R(\gamma(\cdot|x,\xi)\times\upsilon(\cdot)||p(\cdot|x,\xi)\times\rho_\sigma(\cdot))\mu(d\xi) :\\
&\mu\gamma = \mu, \int_{\mathcal{S}} b(x,\xi)\mu(d\xi) + \int_{\mathbb{R}^d} y\upsilon(dy) = \beta\Big\}.
\end{aligned}$$

Further, for each $x,\beta \in \mathbb{R}^d$ there exist a stochastic kernel γ^* and a measure υ^* such that the infimum on the right-hand side is achieved. For Borel sets B_1 of $\mathcal{S}$ and B_2 of $\mathbb{R}^d$, these are given by

$$\gamma^*(B_1|x,\xi) \doteq \int_{B_1} e^{\langle\alpha,b(x,\zeta)\rangle - \Lambda(x,\alpha) - \Psi_\sigma(x;\alpha,\xi) + \Psi_\sigma(x;\alpha,\zeta)} p(d\zeta|x,\xi)$$

and

$$\upsilon^*(B_2) \doteq \int_{B_2} e^{\langle\alpha,y\rangle - \frac{\sigma^2||\alpha||^2}{2}}\rho_\sigma(dy),$$

with $\alpha = \alpha(x,\beta) \in \mathrm{argmax}\{\langle\alpha,\beta\rangle - \Lambda_\sigma(x,\alpha) : \alpha\in\mathbb{R}^d\}$.

(d) *Given any compact set* $\Delta \subset \mathbb{R}^d$ *and* $\varepsilon \in (0,1)$, *there exists* $\eta \in (0,1)$ *such that, whenever* $x, y \in \Delta$, $\beta \in \mathbb{R}^d$, *and* $||x-y|| \le \eta$, *there exists* $\bar{\beta} \in \mathbb{R}^d$ *such that*

$$L_\sigma(y,\bar{\beta}) - L_\sigma(x,\beta) \le \varepsilon \quad \text{and} \quad ||\bar{\beta}-\beta|| \le K||x-y||,$$

where K *is the Lipschitz constant of* b.

(e) *Given* $\psi \in \mathcal{C}([0,1] : \mathbb{R}^d)$ *satisfying* $I_x(\psi) < \infty$ *and* $\varepsilon > 0$, *there exists* $\psi^* \in \mathcal{C}([0,1] : \mathbb{R}^d)$, *with* $\dot{\psi}^*$ *piecewise constant with only finitely many jumps in the interval* $(0,1)$, *such that* $||\psi - \psi^*||_\infty < \varepsilon$ *and*

$$\int_0^1 L_\sigma(\psi^*(t), \dot{\psi}^*(t))dt \le \int_0^1 L_\sigma(\psi(t), \dot{\psi}(t))dt + \varepsilon \le I_x(\psi) + \varepsilon.$$

(f) *The function* $(x,\beta) \to \alpha(x,\beta) \in \operatorname{argmax}\{\langle \alpha, \beta\rangle - \Lambda_\sigma(x,\alpha) : \alpha \in \mathbb{R}^d\}$ *is continuous.*

Proof. (a) The first statement follows from Corollary D.4.2 in [4], while for the second part we take $z = 0$.

(b) From Theorem 26.4 in [14] and Lemma 3.4(iv) in [7], we have that $\operatorname{int}(\operatorname{Dom} L_\sigma(x,\cdot)) = \operatorname{Range}(\nabla\Lambda_\sigma(x,\cdot)) = \mathbb{R}^d$. So $L_\sigma(x,\beta) < \infty$ for all $(x,\beta) \in \mathbb{R}^d \times \mathbb{R}^d$. The nonnegativity of $L_\sigma(x,\beta)$ follows from the nonnegativity of $L(x,\beta)$. The continuity follows from Lemma C.8.1 in [4] and the continuity of $\Lambda(x,\alpha)$ in both variables. Finally, the differentiability follows from the strict convexity of Λ and Theorem D.2.8 in [4].

(c) Let $x, \beta \in \mathbb{R}^d$. From part (b) there exists $\alpha \in \mathbb{R}^d$, with $\alpha = \alpha(x,\beta)$, such that $\beta = \nabla_\alpha \Lambda_\sigma(x,\alpha)$ and $L_\sigma(x,\beta) = \langle \alpha, \beta\rangle - \Lambda_\sigma(x,\alpha)$. Let

$$\gamma_\alpha(d\zeta \times dy|x,\xi) = \exp\Big\{\langle \alpha, b(x,\zeta)\rangle + \Psi(x;\alpha,\zeta) - \Psi(x;\alpha,\xi) - \Lambda(x,\alpha) - \frac{\sigma^2}{2}||\alpha||^2\Big\} p(d\zeta|x,\xi)\rho_\sigma(dy).$$

From Proposition 4.1 in [7],

$$\beta = \int_{\mathcal{S}} b(x,\xi)\mu(d\xi) + \int_{\mathcal{S}} y e^{\langle \alpha, y\rangle - \frac{\sigma^2}{2}||\alpha||^2}\rho_\sigma(dy),$$

where μ is the unique invariant measure of the first marginal of γ_α given by

$$\gamma_{\alpha,1}(d\zeta|x,\xi;\alpha) = \exp\{\langle \alpha, b(x,\zeta)\rangle + \Psi(x;\alpha,\zeta) - \Psi(x;\alpha,\xi) - \Lambda(x,\alpha)\} p(d\zeta|x,\xi).$$

The rest of the proof follows the same arguments given in the proof of Lemma B.2(b) after (B.2).

(d) Let $\Delta \subset \mathbb{R}^d$ compact, $x, y \in \Delta$, $\beta \in \mathbb{R}^d$, and $\varepsilon \in (0,1)$. We know from part (c) that there exist γ^* and υ^* such that the infimum in part (c) is attained for (x,β). Define $\bar{\beta} \doteq \int_{\mathcal{S}} b(y,\xi)\mu^{\gamma^*}(d\xi) + \int_{\mathbb{R}^d} y\nu^*(dy)$. Then, from the representation formula given in part (c),

$$||\beta - \bar{\beta}|| \le \left|\int_{\mathcal{S}} (b(x,\xi) - b(y,\xi))\mu^{\gamma^*}(d\xi)\right| \le K\|x-y\|.$$

Now, for any Borel set B of $\mathcal{S}$, we can use part (c) of Hypothesis H.1 to write

$$\gamma^*(B|x,\xi) = \int_B \exp\Big\{\langle \alpha, b(x,\zeta)\rangle + \Psi_\sigma(x;\alpha,\zeta) - \Psi_\sigma(x;\alpha,\xi) - \Lambda(x,\alpha)\Big\}\tilde{p}^x(\xi,\zeta)\vartheta(d\zeta).$$

From the bound that we have on $\tilde{p}^x(\cdot,\cdot)$, it follows that $\gamma^*(\cdot|x,\xi)$ is absolutely continuous with respect to $p(\cdot|y,\xi)$; from the uniform continuity of $\tilde{p}^x(\xi,\zeta)$, there exists $\eta > 0$ such that $\|x-y\| < \eta$ implies that $\tilde{p}^x(\xi,\zeta) \le \tilde{p}^y(\xi,\zeta)e^{\varepsilon}$ (this is as in (4.15)). Then, from the variational equivalence given in part (c), $L_\sigma(y,\bar{\beta}) < \infty$ and

$$\begin{aligned} L_\sigma(y,\bar{\beta}) &\le \int_{\mathcal{S}} R(\gamma^*(\cdot|x,\xi)\times\nu^*(\cdot)||p(\cdot|y,\xi)\times\rho_\sigma(\cdot))\mu^{\gamma^*}(d\xi) \\ &\le \Big\langle \alpha, \int_{\mathcal{S}} b(x,\xi)\mu^{\gamma^*}(d\xi)\Big\rangle - \Lambda(x,\alpha) + \Big\langle \alpha, \int_{\mathbb{R}^d} y\nu^*(dy)\Big\rangle - \frac{\sigma^2}{2}||\alpha||^2 + \varepsilon \\ &= \langle \alpha,\beta\rangle - \Lambda_\sigma(x,\alpha) + \varepsilon = L_\sigma(x,\beta) + \varepsilon. \end{aligned}$$

(e) The proof of this part is based on Lemmas 6.5.3 and 6.5.5 of [4], which in our case also hold due to the structural properties given in parts (a) and (b).

(f) Given x, $\beta \in \mathbb{R}^d$, part (b) and the differentiability of $\alpha \to \Lambda_\sigma(x,\alpha)$ imply that there exists a unique $\alpha(x,\beta)$ such that $L_\sigma(x,\beta) = \langle \alpha(x,\beta),\beta\rangle - \Lambda_\sigma(x,\alpha(x,\beta))$, $\beta = \nabla_\alpha\Lambda_\sigma(x,\alpha(x,\beta))$, and $\alpha(x,\beta) = \nabla_\beta L_\sigma(x,\beta)$. We observe that $\beta \to L_\sigma(x,\beta)$ is continuously differentiable thanks to [14, Corollary 25.5.1]. Moreover, $x \to \nabla_\beta L_\sigma(x,\beta)$ is continuous [14, Theorem 25.7] and, in fact, $(x,\beta) \to \nabla_\beta L_\sigma(x,\beta)$ is continuous by the same theorem. Therefore, $(x,\beta) \to \alpha(x,\beta)$ is continuous in both variables. This completes the proof of the lemma. □

Appendix C. Proofs of some limit results. This appendix is dedicated to the proofs of some limit results needed in the proof of Theorem 2.1.

THEOREM C.1. *Let $\mathcal{S}$ be compact. For any $x \in \mathbb{R}^d$, $\xi \in \mathcal{S}$ and each $n \in \mathbb{N}$, consider any admissible control sequence such that*

$$\sup_{n\in\mathbb{N}} \bar{E}_{x,\xi}\left\{\frac{1}{n}\sum_{j=0}^{n-1} R(\nu_j^n(\cdot)\|p(\cdot|\bar{X}_j^n,\bar{Z}_j^n))\right\} < \infty,$$

where $\nu_j^n(\cdot) = \nu_j^n(\cdot|\bar{X}_0^n,\dots,\bar{X}_j^n,\bar{Z}_j^n)$. In terms of these sequences we define the piecewise linear interpolation $\{\bar{X}^n\}$, the piecewise constant interpolation $\{\tilde{X}^n\}$, the sequence of admissible control measures $\{\nu^n\}$ and its marginals $\{\hat{\nu}_2^n\otimes\lambda, \hat{\nu}_1^n\otimes\lambda\}$, and the measures $\{\gamma^n\}$ as in section 3. Also, for each $n \in \mathbb{N}$ we define the process $S^n = \{S^n(t), t\in[0,1]\}$ by

$$S^n(t) \doteq x + \int_{\mathcal{S}\times\mathcal{S}\times[0,t]} b(S^n(s),y)\nu^n(d\zeta\times dy\times ds).$$

The following conclusions hold.

(a) *Given any subsequence of $\{(\nu^n, \hat{\nu}_2^n\otimes\lambda, \hat{\nu}_1^n\otimes\lambda, \gamma^n, \bar{X}^n, \tilde{X}^n, S^n), n\in\mathbb{N}\}$, there exist a subsubsequence, a stochastic kernel ν on $\mathcal{S}\times\mathcal{S}\times[0,1]$ (given $\bar{\Omega}$) with marginals μ_1, μ_2, a stochastic kernel γ on $\mathcal{S}\times\mathcal{S}\times[0,1]$, and random variables $\bar{X}$ and S mapping $\bar{\Omega}$ into $\mathcal{C}([0,1]:\mathbb{R}^d)$ such that the subsubsequence converges in distribution to $(\nu,\mu_1,\mu_2,\gamma,\bar{X},\bar{X},S)$.*

(b) *The stochastic kernel ν has the decomposition*

$$\nu(B_1\times B_2\times C) = \int_C \nu(B_1\times B_2|t)dt = \int_C\int_{B_1} \hat{\nu}_1(d\zeta|t)\hat{\nu}_2(B_2|\zeta,t)dt$$

for some stochastic kernels $\nu(\cdot|t)$ on $\mathcal{S}\times\mathcal{S}$ given $[0,1]$, $\hat{\nu}_1(\cdot|t)$ on $\mathcal{S}$ given $[0,1]$, and $\hat{\nu}_2(\cdot|\zeta,t)$ on $\mathcal{S}$ given $\mathcal{S}\times[0,1]$.

(c) *We have the equality $\hat{\nu}_1\otimes\lambda = \hat{\nu}_2\otimes\lambda$.*

(d) *$\hat{\nu}_1(d\zeta|t)$ is an invariant measure of $\hat{\nu}_2(dy|\zeta,t)$ for each $t\in[0,1]$.*

(e) *W.p.*1 *for every* $t \in [0,1]$

$$\bar{X}(t) = x + \int_{\mathcal{S}\times\mathcal{S}\times[0,t]} b(\bar{X}(s), y)\nu(d\zeta \times dy \times ds) \tag{C.1}$$
$$= x + \int_0^t \int_{\mathcal{S}} b(\bar{X}(s), \zeta)\hat{\nu}_1(d\zeta|s)ds.$$

(f) *The stochastic kernel* γ *has the decomposition*

$$\gamma(B_1 \times B_2 \times C) = \int_C \int_{B_1\times B_2} \hat{\nu}_1(d\zeta|t) \otimes p(dy|\bar{X}(t), \zeta)dt.$$

Proof. (a) Given the compactness of $\mathcal{S}$, we immediately get tightness of ν^n, of γ^n, and of all the marginals. Tightness of $\{\bar{X}^n\}$ and of $\{\tilde{X}^n\}$ on $\mathcal{C}([0,1] : \mathbb{R}^d)$ follows from the bound

$$w_{\bar{X}^n}(\delta) \doteq \sup_{\{s,t\in[0,1]:|s-t|\le\delta\}} \|\bar{X}^n(t) - \bar{X}^n(s)\| \le 2\|b\|_\infty\delta, \tag{C.2}$$

and tightness of $\{S^n\}$ can be verified similarly. Since the function mapping ν^n into $(\nu^n, \hat{\nu}_1^n \otimes \lambda, \hat{\nu}_2^n \otimes \lambda)$ is continuous, there exist measures ν and γ over $\mathcal{S}\times\mathcal{S}\times[0,1]$, measures μ_1 and μ_2 over $\mathcal{S}\times[0,1]$, and random variables $\bar{X}$ and S on $\mathcal{C}([0,1] : \mathbb{R}^d)$ such that $(\nu^n, \hat{\nu}_1^n \otimes \lambda, \hat{\nu}_2^n \otimes \lambda, \gamma^n, \bar{X}^n, \tilde{X}^n, S^n) \xrightarrow{\mathcal{D}} (\nu, \mu_1, \mu_2, \gamma, \bar{X}, \bar{X}, S)$ [4, Theorem A.3.6]. Moreover, w.p.1 μ_1 and μ_2 equal the marginals of ν over (ζ, t) and over (y, t), respectively. For the developments below, we note that by the Skorohod representation theorem we can assume that convergence takes place w.p.1 on some probability space, which we also denote by $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{P})$.

(b) We let μ_3 denote the marginal of ν over t. Using the fact that the marginal of ν^n over t is Lebesgue measure λ, we have that w.p.1 for any bounded continuous function g mapping $[0,1]$ into $\mathbb{R}$,

$$\begin{aligned}\int_0^1 g(t)\mu_3(dt) &= \int_{\mathcal{S}\times\mathcal{S}\times[0,1]} g(t)\nu(d\zeta \times dy \times dt)\\ &= \lim_{n\to\infty} \int_{\mathcal{S}\times\mathcal{S}\times[0,1]} g(t)\nu^n(d\zeta \times dy \times dt)\\ &= \lim_{n\to\infty} \int_0^1 g(t)dt\\ &= \int_0^1 g(t)dt.\end{aligned}$$

Since the class of bounded and continuous functions is a measure determining class [1, Theorem 1.3], this implies that w.p.1 $\mu_3(\cdot)$ equals $\lambda(\cdot)$. By Theorem A.5.6 in [4], there exists a stochastic kernel $\nu(d\zeta \times dy|t)$ on $\mathcal{S}\times\mathcal{S}$ given $[0,1]$ such that w.p.1

$$\nu(B_1 \times B_2 \times C) = \int_C \nu(B_1 \times B_2|t)\, dt.$$

Once more Theorem A.5.6 in [4] gives the existence of stochastic kernels $\hat{\nu}_2(dy|\zeta, t)$ on $\mathcal{S}$ given $\mathcal{S}\times[0,1]$ and $\hat{\nu}_1(d\zeta|t)$ on $\mathcal{S}$ given $[0,1]$ (the second and first marginals of $\nu(d\zeta \times dy|t)$, respectively) such that

$$\nu(B_1 \times B_2 \times C) = \int_C \int_{B_1} \hat{\nu}_1(d\zeta|t)\hat{\nu}_2(B_2|\zeta, t)dt.$$

This gives the decomposition of $\nu(d\zeta \times dy \times dt)$ given in part (b).

(c) Consider a function f of the form $f(y,t) = g(y)h(t)$, with $g \in \mathcal{C}(\mathcal{S} : \mathbb{R}^d)$ and $h \in \mathcal{C}([0,1] : \mathbb{R}^d)$. Since

$$\bar{E}_{x,\xi}\left\{g(\bar{Z}^n_{j+1}) - \int_{\mathcal{S}} g(y)\nu^n_j(dy|\bar{X}_0,\dots,\bar{X}^n_j,\bar{Z}^n_j)\right\} = 0,$$

we have that

$$\left\{g(\bar{Z}^n_{j+1}) - \int_{\mathcal{S}} g(y)\nu^n_j(dy|\bar{X}^n_0,\dots,\bar{X}^n_j,\bar{Z}^n_j)\right\}$$

is a martingale difference sequence. Moreover,

$$\begin{aligned}\int_{\mathcal{S}\times[0,1]} f(\zeta,t)(\hat{\nu}^n_1\otimes\lambda)(d\zeta\times dt) &= \sum_{l=0}^{\frac{1}{k_n}-1}\int_{lk_n}^{(l+1)k_n} h(t)dt\cdot\int_{\mathcal{S}} g(\zeta)(\tilde{\nu}^n_l)_1(d\zeta)\\ &= \sum_{l=0}^{\frac{1}{k_n}-1}\int_{lk_n}^{(l+1)k_n} h(t)dt\cdot\frac{1}{m_n}\sum_{j=lm_n}^{(l+1)m_n-1} g(\bar{Z}^n_j)\end{aligned}$$

and similarly

$$\begin{aligned}&\int_{\mathcal{S}\times[0,1]} f(y,t)(\hat{\nu}^n_2\otimes\lambda)(dy\times dt)\\ &= \sum_{l=0}^{\frac{1}{k_n}-1}\int_{lk_n}^{(l+1)k_n} h(t)dt\cdot\left[\frac{1}{m_n}\sum_{j=lm_n}^{(l+1)m_n-1}\int_{\mathcal{S}} g(y)\nu^n_j(dy)\right].\end{aligned}$$

Noting that

$$\left\|\sum_{l=0}^{\frac{1}{k_n}-1}\frac{1}{m_n}\left[\int_{lk_n}^{(l+1)k_n} h(t)dt\right]\left[g(\bar{Z}^n_{lm_n}) - g(\bar{Z}^n_{(l+1)m_n})\right]\right\| \le \frac{2\|g\|\|h\|}{m_n},$$

we have that if $m_n \ge \frac{4\|g\|\|h\|}{\varepsilon}$, then

$$\begin{aligned}&\bar{P}_{x,\xi}\left\{\left|\int_{\mathcal{S}\times[0,1]} f d(\hat{\nu}^n_1\otimes\lambda) - \int_{\mathcal{S}\times[0,1]} f d(\hat{\nu}^n_2\otimes\lambda)\right| \ge \varepsilon\right\}\\ &\quad\le \bar{P}_{x,\xi}\left\{\left|\sum_{l=0}^{\frac{1}{k_n}-1}\frac{1}{m_n}\left[\int_{lk_n}^{(l+1)k_n} h(t)dt\right]\left[\sum_{j=lm_n}^{(l+1)m_n-1}\left(g(\bar{Z}^n_{j+1}) - \int g(y)\nu^n_j(dy)\right)\right]\right| \ge \frac{\varepsilon}{2}\right\}\\ &\quad\le \bar{P}_{x,\xi}\left\{\left|\frac{1}{n}\sum_{j=0}^{n-1}\left[g(\bar{Z}^n_{j+1}) - \int_{\mathcal{S}} g(y)\nu^n_j(dy)\right]\right| \ge \frac{\varepsilon}{2\|h\|}\right\}\\ &\le \frac{4\|h\|^2}{\varepsilon^2}\bar{E}_{x,\xi}\left\{\frac{1}{n^2}\left(\sum_{j=0}^{n-1}\left[g(\bar{Z}^n_{j+1}) - \int_{\mathcal{S}} g(y)\nu^n_j(dy)\right]\right)^2\right\}\\ &\le \frac{4\|h\|^2}{\varepsilon^2}\bar{E}_{x,\xi}\left\{\frac{1}{n^2}\sum_{j=0}^{n-1}\left(g(\bar{Z}^n_{j+1}) - \int_{\mathcal{S}} g(y)\nu^n_j(dy)\right)^2\right\} \le \frac{16\|h\|^2\|g\|^2}{n\varepsilon^2}. \qquad \text{(C.3)}\end{aligned}$$

This implies that $\int_{\mathcal{S}\times[0,1]} f d(\hat{\nu}_1^n \otimes \lambda) - \int_{\mathcal{S}\times[0,1]} f d(\hat{\nu}_2^n \otimes \lambda)$ converges to zero in probability and hence in distribution [1, Theorem 4.3]. Given the convergence w.p.1 of $\hat{\nu}_1^n \otimes \lambda$ and $\hat{\nu}_2^n \otimes \lambda$ to $\hat{\nu}_1 \otimes \lambda$ and $\hat{\nu}_2 \otimes \lambda$, respectively, we get w.p.1

$$\lim_{n\to\infty} \int_{\mathcal{S}\times[0,1]} f(y,t)(\hat{\nu}_r^n \otimes \lambda)(dy \times dt) = \int_{\mathcal{S}\times[0,1]} f(y,t)(\hat{\nu}_r \otimes \lambda)(dy \times dt)$$

for $r = 1, 2$. Therefore w.p.1

$$\int_{\mathcal{S}\times[0,1]} f(y,t)(\hat{\nu}_1 \otimes \lambda)(dy \times dt) = \int_{\mathcal{S}\times[0,1]} f(y,t)(\hat{\nu}_2 \otimes \lambda)(dy \times dt). \tag{C.4}$$

Theorem A.3.14 in [4] implies that we can extend the equality in (C.4) from f of the form $f(y,t) = g(y)h(t)$ to all $f : \mathcal{S} \times [0,1] \mapsto \mathbb{R}$ that are bounded and continuous. Since the class of bounded continuous functions is measure-determining, we have that $\hat{\nu}_1 \otimes \lambda = \hat{\nu}_2 \otimes \lambda$, as we wanted to show.

(d) We now show that for each $t \in [0,1]$ and Borel subset B_1 of $\mathcal{S}$, we have

$$\hat{\nu}_1(B_1|t) = \int_{\mathcal{S}} \hat{\nu}_2(B_1|\zeta, t)\hat{\nu}_1(d\zeta|t). \tag{C.5}$$

Let $\mathcal{U}_b(\mathcal{S})$ denote the space of bounded, uniformly continuous functions mapping $\mathcal{S}$ into $\mathbb{R}$. Since $\mathcal{S}$ is Polish, there exists an equivalent metric m under which $\mathcal{U}_b(\mathcal{S}, m)$ is separable with respect to the uniform metric. Let $\mathcal{E}$ be a countable dense subset of $\mathcal{U}_b(\mathcal{S}, m)$, and let g be any function in $\mathcal{E}$. For each $s \in [0,1]$ let $E_i \subset [0,1]$ be a sequence of sets which *shrinks nicely* to s (see [2, p. 353]), and define

$$f(t,y) \doteq g(y) \cdot \frac{1}{\lambda(E_i)} I_{E_i}(t).$$

Since $\hat{\nu}_1 \otimes \lambda = \hat{\nu}_2 \otimes \lambda$,

$$\frac{1}{\lambda(E_i)} \int_{E_i} \int_{\mathcal{S}} g(y)\hat{\nu}_1(dy|t)\lambda(dt) = \frac{1}{\lambda(E_i)} \int_{E_i} \int_{\mathcal{S}} \int_{\mathcal{S}} g(y)\hat{\nu}_2(dy|\zeta, t)\hat{\nu}_1(d\zeta|t)\lambda(dt).$$

Define $h_1(t) \doteq \int_{\mathcal{S}} g(y)\hat{\nu}_1(dy|t)$ and $h_2(t) \doteq \int_{\mathcal{S}} \int_{\mathcal{S}} g(y)\hat{\nu}_2(dy|\zeta, t)\hat{\nu}_1(d\zeta|t)$. Then we have

$$|h_1(s) - h_2(s)| \le \frac{1}{\lambda(E_i)} \int_{E_i} |h_1(t) - h_1(s)|\, dt + \frac{1}{\lambda(E_i)} \int_{E_i} |h_2(t) - h_2(s)|\, dt,$$

which tends to 0 as $n \to \infty$ for almost all $s \in [0,1]$ (see Theorem C.13 in [2]). This implies that $h_1(s) = h_2(s)$ a.s., so that there exists a set $B_g \in \mathcal{B}([0,1])$ with $\lambda(B_g) = 0$ and such that $\int_{\mathcal{S}} g(y)\hat{\nu}_1(dy|t) = \int_{\mathcal{S}} \int_{\mathcal{S}} g(y)\hat{\nu}_2(dy|\zeta, t)\hat{\nu}_1(d\zeta|t)$ for all $t \notin B_g$. Now define $B \doteq \cup_{g\in\mathcal{E}} B_g$. Then $\lambda(B) = 0$ and for all $t \notin B$ and $g \in \mathcal{E}$ the same equality holds. The equality can then be extended to $g \in \mathcal{U}_b(\mathcal{S}, m)$, which implies that $\hat{\nu}_1(dy|t) = \hat{\nu}_2(dy|t)$ for all $t \notin B$. Finally, redefining $\hat{\nu}_1$ and $\hat{\nu}_2$ in an obvious way for $t \in B$, we get (C.5).

(e) Let $\{\bar{Y}^n, n \in \mathbb{N}\}$ and $\{\tilde{Y}^n, n \in \mathbb{N}\}$ be the sequences of piecewise linear and piecewise constant interpolations, respectively, of the process $\{X_j^n, j = 0, \ldots, n\}$ but when observed only at the endpoints of the intervals of size k_n. That is, they are the interpolations of a process $\{\bar{Y}_j^n, j = 0, \ldots, n\}$ defined through $\bar{Y}_l^n \doteq \bar{X}_{lm_n}^n$ for $l = 0, \ldots, \frac{1}{k_n}$. The intuition behind this idea is described clearly below (2.3). We

will relate $\bar{Y}^n$ to $\bar{X}^n$, $\tilde{Y}^n$ to $\tilde{X}^n$, and both $\bar{Y}^n$ and $\tilde{Y}^n$ to S^n in a way that forces all five sequences to have the same limit [1, Theorem 4.1]. By showing that S and $\bar{X}$ as defined in (C.1) are the same w.p.1, the characterization of the limit process $\bar{X}$ will follow.

By definition, the process $\{\bar{Y}_l^n\}$ follows the evolution

$$\bar{Y}_{l+1}^n = \bar{Y}_l^n + \frac{1}{n} \sum_{j=lm_n}^{(l+1)m_n-1} b(\bar{X}_j^n, \bar{Z}_{j+1}^n).$$

Moreover, for every $l = 0, \ldots, \frac{1}{k_n} - 1$ and with $i = 1, \ldots, m_n - 1$, we have

$$\|\bar{Y}_l^n - \bar{X}_{lm_n+i}^n\| = \left\| \frac{1}{n} \sum_{j=0}^{i-1} b(\bar{X}_{lm_n+j}^n, \bar{Z}_{lm_n+j+1}^n) \right\| \le \frac{i\|b\|_\infty}{n}. \tag{C.6}$$

Hence

$$\begin{aligned} \sup_{t\in[0,1]} \|\bar{Y}^n(t) - \bar{X}^n(t)\| &\le w_{\bar{Y}^n}(k_n) + w_{\bar{X}^n}(1/n) \\ &\quad + \max_{l\in\{0,\ldots,\frac{1}{k_n}-1\}} \max_{i\in\{1,\ldots,m_n-1\}} \|\bar{Y}_l^n - \bar{X}_{lm_n+i}^n\| \\ &\le \left(3k_n + \frac{2}{n}\right) \|b\|_\infty, \end{aligned}$$

where we have used (C.6), (C.2), and the bound

$$w_{\bar{Y}^n}(k_n) \doteq \sup_{\{s,t\in[0,1]:|s-t|\le k_n\}} \|\bar{Y}^n(t) - \bar{Y}^n(s)\| \le 2k_n\|b\|_\infty. \tag{C.7}$$

This implies that in $\mathcal{C}([0,1] : \mathbb{R}^d)$ under the uniform metric, $d(\bar{X}^n, \bar{Y}^n)$ converges to 0 in probability. Similarly, in $\mathcal{D}([0,1] : \mathbb{R}^d)$ under the Skorohod metric, $d(\tilde{X}^n, \tilde{Y}^n)$ converges to 0 in probability.

Next we prove that

$$\lim_{n\to\infty} \bar{P}_{x,\xi} \left\{ \sup_{t\in[0,1]} \left\| S^n(t) - \bar{Y}^n(t) \right\| \ge \varepsilon \right\} = 0, \tag{C.8}$$

which immediately implies the limit $\lim_{n\to\infty} \bar{P}_{x,\xi}\{\sup_{t\in[0,1]} \|S^n(t) - \tilde{Y}^n(t)\| \ge \varepsilon\} = 0$. For any $t \in [0,1]$ with $lk_n \le t < (l+1)k_n$, we have

$$\begin{aligned} &\|\bar{Y}^n(t) - S^n(t)\| \\ &\le w_{\bar{Y}^n}(k_n) + w_{S^n}(k_n) + Kk_n\|b\|_\infty \\ &\quad + \left\| \sum_{j=0}^{l} \frac{1}{n} \sum_{i=jm_n}^{(j+1)m_n} b(\bar{Y}_j^n, \bar{Z}_{i+1}^n) - \int_{\mathcal{S}\times\mathcal{S}\times[0,lk_n]} b(\tilde{Y}^n(s), y)\nu^n(d\zeta \times dy \times ds) \right\| \\ &\quad + \left\| \int_{\mathcal{S}\times\mathcal{S}\times[0,lk_n]} \left[b(\tilde{Y}^n(s), y) - b(\bar{Y}^n(s), y) \right] \nu^n(d\zeta \times dy \times ds) \right\| \\ &\quad + \left\| \int_{\mathcal{S}\times\mathcal{S}\times[0,lk_n]} \left[b(\bar{Y}^n(s), y) - b(S^n(s), y) \right] \nu^n(d\zeta \times dy \times ds) \right\| \\ &\le 3(1+K)k_n\|b\|_\infty + A(t) + \int_0^t K\|\bar{Y}^n(s) - S^n(s)\| ds, \end{aligned}$$

where we have defined

$$A(t) \doteq \left\| \sum_{j=0}^{l} \frac{1}{n} \sum_{i=jm_n}^{(j+1)m_n} b(\bar{Y}_j^n, \bar{Z}_{i+1}^n) - \int_{\mathcal{S}\times\mathcal{S}\times[0,lk_n]} b(\tilde{Y}^n(s), y)\nu^n(d\zeta \times dy \times ds) \right\|.$$

(Note that dependence on t comes through l.) By Gronwall's inequality we can then write

$$\begin{aligned}
&\|\bar{Y}^n(t) - S^n(t)\| \\
&\leq 3(1+K)k_n\|b\|_\infty + A(t) + \int_0^t K\left[3(1+K)k_n\|b\|_\infty + A(s)\right] e^{K(t-s)} ds. \\
&\leq e^K\left[3(1+K)k_n\|b\|_\infty + A(lk_n)\right] + \sum_{j=0}^{l-1} A(jk_n) K e^{Kt} \int_{jk_n}^{(j+1)k_n} e^{-Ks} ds \\
&\leq e^K\left[3(1+K)k_n\|b\|_\infty + (1+K)\max_{j\in\{0,\ldots,1/k_n\}} A(jk_n)\right], \qquad \text{(C.9)}
\end{aligned}$$

where in the last step we have used the inequality $e^{jk_n} - e^{(j+1)k_n} \leq k_n K e^{-Kjk_n}$, valid because of the mean value theorem. Given (C.9), all that remains to show is that $\{\max_{l\in\{0,\ldots,1/k_n\}} A(lk_n), n \in \mathbb{N}\}$ converges to zero in probability as $n \to \infty$.

Now for all $l = 0, \ldots, 1/k_n$ we have that

$$\int_{\mathcal{S}\times\mathcal{S}\times[0,lk_n]} b(\tilde{Y}^n(s), y)\nu^n(d\zeta \times dy \times ds) = \sum_{j=0}^{l} k_n \frac{1}{m_n} \sum_{i=jm_n}^{(j+1)m_n} \int_{\mathcal{S}} b(\bar{Y}_j^n, y)\nu_i^n(dy|\bar{X}_i^n, \bar{Z}_i^n).$$

Moreover, the sequence

$$\left\{ b(\bar{Y}_j^n, \bar{Z}_{i+1}^n) - \int_{\mathcal{S}} b(\bar{Y}_j^n, y)\nu_i^n(dy|\bar{X}_i^n, \bar{Z}_i^n), j = 0, \ldots, \frac{1}{k_n}, i = jm_n, \ldots, (j+1)m_n - 1 \right\}$$

forms a martingale difference sequence with respect to the sequence of sigma fields generated by $\{(\bar{X}_r^n, \bar{Z}_r^n), r = 0, \ldots, i\}$ for $i = 0, \ldots, n-1$. Therefore, the submartingale inequality [5, Lemma 2.2.3] applied to the submartingale

$$\left\{ \left\| b(\bar{Y}_j^n, \bar{Z}_{i+1}^n) - \int_{\mathcal{S}} b(\bar{Y}_j^n, y)\nu_i^n(dy|\bar{X}_i^n, \bar{Z}_i^n) \right\|^2 \right\}$$

implies that for any $\varepsilon > 0$

$$\begin{aligned}
&\bar{P}_{x,\xi}\left\{ \max_{l\in\{0,\ldots,1/k_n\}} A(lk_n) \geq \varepsilon \right\} \\
&\leq \frac{1}{\varepsilon^2} \bar{E}_{x,\xi}\{A(1)^2\} \\
&\leq \frac{1}{\varepsilon^2} \sum_{j=0}^{l} \frac{1}{n^2} \sum_{i=jm_n}^{(j+1)m_n} \bar{E}_{x,\xi} \left\| b(\bar{Y}_j^n, \bar{Z}_{i+1}^n) - \int_{\mathcal{S}} b(\bar{Y}_j^n, y)\nu_i^n(dy|\bar{X}_i^n, \bar{Z}_i^n) \right\|^2 \\
&\leq \frac{4\|b\|_\infty^2}{\varepsilon^2 n},
\end{aligned}$$

which gives (C.8).

Having shown that all five sequences must converge to the same limit, it remains to show that S and $\bar{X}$ as defined in (C.1) are the same w.p.1. We will show that for each fixed t, $S(t) = \bar{X}(t)$ w.p.1. Equality for all $t \in [0,1]$ w.p.1 follows by considering the rationals and then extending by continuity.

Fix $t \in [0.1]$. We have

$$\|S^n(t) - \bar{X}(t)\| \leq \int_0^t K \left\|S^n(s) - \bar{X}(s)\right\| ds$$
$$+ \left\| \int_{\mathcal{S}\times\mathcal{S}\times[0,t]} b(\bar{X}(s), y)\nu^n(d\zeta \times dy \times ds) - \int_{\mathcal{S}\times\mathcal{S}\times[0,t]} b(\bar{X}(s), y)\nu(d\zeta \times dy \times ds) \right\|.$$

Using Gronwall's inequality if follows that

$$\text{(C.10)} \qquad \left\|S^n(t) - \bar{X}(t)\right\|$$
$$\leq \left\| \int_{\mathcal{S}\times\mathcal{S}\times[0,t]} b(\bar{X}(s), y)\nu^n(d\zeta \times dy \times ds) - \int_{\mathcal{S}\times\mathcal{S}\times[0,t]} b(\bar{X}(s), y)\nu(d\zeta \times dy \times ds) \right\|$$
$$+ K \int_0^t \left\| \int_{\mathcal{S}\times\mathcal{S}\times[0,r]} b(\bar{X}(s), y)\nu^n(d\zeta \times dy \times ds) - \int_{\mathcal{S}\times\mathcal{S}\times[0,r]} b(\bar{X}(s), y)\nu(d\zeta \times dy \times ds) \right\| e^{K(t-r)} dr.$$

Weak convergence of ν^n to ν implies that w.p.1

$$\lim_{n\to\infty} \int_{\mathcal{S}\times\mathcal{S}\times[0,t]} b(\bar{X}(s), y)\nu^n(d\zeta \times dy \times ds) = \int_{\mathcal{S}\times\mathcal{S}\times[0,t]} b(\bar{X}(s), y)\nu(d\zeta \times dy \times ds).$$

This statement is true since we can identify the points of discontinuity of the bounded function $b(\bar{X}(s), y)1_{[0,t]}(s)$ to be $\mathcal{S} \times \mathcal{S} \times \{t\}$, which form a set of measure zero under the limit ν. Hence it follows from (C.10) and the dominated convergence theorem that for any given $\varepsilon > 0$

$$\lim_{n\to\infty} \bar{P}_{x,\xi} \left\{ \|S^n(t) - \bar{X}(t)\| \geq \varepsilon \right\} \leq \lim_{n\to\infty} \frac{1}{\varepsilon} \bar{E}_{x,\xi} \left\{ \left\|S^n(t) - \bar{X}(t)\right\| \right\} = 0.$$

Uniqueness of limits implies $S(t) = \bar{X}(t)$ w.p.1, as we wanted to show.

(f) Let $f : \mathcal{S} \times \mathcal{S} \times [0,1]$ be bounded and continuous. Then for each $n \in \mathbb{N}$ the function g_n mapping (ζ, t) to $\int_{\mathcal{S}} f(\zeta, y, t)p(dy|\bar{Y}^n(t), \zeta)$ is also bounded and continuous. Define the bounded and continuous function g mapping (ζ, t) into $\int_{\mathcal{S}} f(\zeta, y, t)$ $p(dy|\bar{X}(t), \zeta)$. Weak continuity of p and a.s. convergence of $\bar{Y}^n$ to $\bar{X}$ imply that for all $t \in [0,1]$ and $\zeta \in \mathcal{S}$ we have $\lim_{n\to\infty} g_n(\zeta, t) = g(\zeta, t)$ a.s. We argue that convergence is, in fact, uniform in $\mathcal{S} \times [0,1]$.

Indeed, let $\varepsilon > 0$ be given. Fix $N \in \mathbb{N}$ satisfying $k_N < \varepsilon/(3\|b\|_\infty)$ so that (see (C.7)) for all $n > N$ and r, s with $|r - s| \leq k_N$, we have $\|\bar{Y}^n(r) - \bar{Y}^n(s)\| < \varepsilon/3$ and $\|\bar{X}(r) - \bar{X}(s)\| < \varepsilon/3$. Further, N can be chosen so that for all $n \geq N$ and $i = 0, \ldots, \frac{1}{k_N} - 1$ we have $\|\bar{Y}^n(ik_N) - \bar{X}(ik_N)\| < \varepsilon/3$. Hence we have that for all $t \in [0,1]$ and any $n \geq N$, $\|\bar{Y}^n(t) - \bar{X}(t)\| < \varepsilon$. By Hypothesis H.1 we can then write

$$g_n(\zeta, t) = \int_{\mathcal{S}} f(\zeta, y, t)p(dy|\bar{Y}^n(t), \zeta) = \int_{\mathcal{S}} f(\zeta, y, t)\tilde{p}^{\bar{Y}^n(t)}(\zeta, y)\vartheta(dy).$$

By the dominated convergence theorem, this last quantity converges (a.s. uniformly in ζ and t) to $\int_{\mathcal{S}} f(\zeta, y, t)\tilde{p}^{\bar{X}(t)}(\zeta, y)\vartheta(dy) = g(\zeta, t)$.

It now follows from the weak convergence of $\hat{\nu}_1^n \otimes \lambda$ to $\hat{\nu}_1 \otimes \lambda$ that

$$\lim_{n\to\infty} \int_0^1 \int_{\mathcal{S}} g_n(\zeta,t)\hat{\nu}_1^n(d\zeta|t)dt = \int_0^1 \int_{\mathcal{S}} g(\zeta,t)\hat{\nu}_1(d\zeta|t)dt.$$

This in turn implies that

$$\lim_{n\to\infty} \int_{\mathcal{S}\times\mathcal{S}\times[0,1]} f(\zeta,y,t)\gamma^n(d\zeta\times dy\times dt) = \int_{\mathcal{S}\times\mathcal{S}\times[0,1]} f(\zeta,y,t)\gamma(d\zeta\times dy\times dt),$$

which completes the proof. □

The last lemma in this section establishes an estimate needed in the proof of the lower bound.

LEMMA C.2. *For any $\delta > 0$,*

$$\lim_{\sigma\to 0} \bar{P}_{x,\xi}\left\{\sup_{t\in[0,1]} ||\bar{U}(t)|| \geq \delta\right\} = 0.$$

Proof. As was discussed in the proof of the lower bound, $S^{2,n} \to \bar{U}$ in distribution in $\mathcal{C}([0,1] : \mathbb{R}^d)$, so that by the Skorohod representation theorem and Fatou's lemma,

$$\bar{E}_{x,\xi}\left\{\left(\sup_{t\in[0,1]} ||\bar{U}(t)||\right)^2\right\} \leq \liminf_{n\to\infty} \bar{E}_{x,\xi}\left\{\left(\sup_{t\in[0,1]} ||S^n(t)||\right)^2\right\}.$$

Hence we have that for any $\delta > 0$,

$$\begin{aligned}
\bar{P}_{x,\xi}\left\{\sup_{t\in[0,1]} ||\bar{U}(t)|| \geq \delta\right\} &\leq \frac{1}{\delta^2}\bar{E}_{x,\xi}\left\{\left(\sup_{t\in[0,1]} ||\bar{U}(t)||\right)^2\right\} \\
&\leq \frac{1}{\delta^2}\liminf_{n\to\infty} \bar{E}_{x,\xi}\left\{\frac{1}{n}\sum_{j=0}^{n-1}\left\|\int_{\mathbb{R}^d} y\nu_j^{2,n}(dy)\right\|^2\right\} \\
&= \frac{1}{\delta^2}\liminf_{n\to\infty} \bar{E}_{x,\xi}\left\{\frac{1}{n}\sum_{j=0}^{n\tau^n-1}||\beta_j^{2,n}||^2\right\} \\
&\leq \frac{2\sigma^2}{\delta^2}\liminf_{n\to\infty} \bar{E}_{x,\xi}\left\{\frac{1}{n}\sum_{j=0}^{n\tau^n-1} R(\gamma_j^{2,n}(\cdot)||\rho_\sigma(\cdot))\right\} \\
&\leq \frac{2\sigma^2}{\delta^2}\liminf_{n\to\infty}\left\{k_n\sum_{l=0}^{[1/k_n]} L_\sigma(\psi^*(lm_n/n),\dot{\psi}^*(lm_n/n)) + 3\sigma\right\} \\
&\leq \frac{2\sigma^2}{\delta^2}[I_x(\psi) + 4\sigma].
\end{aligned}$$

The fourth inequality follows from $\frac{1}{2\sigma^2}||\beta_j^{2,n}||^2 = \hat{L}_\sigma(\beta_j^{2,n}) \leq R(\gamma_j^{2,n}(\cdot)||\rho_\sigma(\cdot))$, where $\hat{L}_\sigma$ is the Legendre–Fenchel of the moment generating function of ρ_σ. The fifth line follows from (4.13) and (4.18), while in line six we have used part (e) of Lemma B.3 with $\varepsilon = \sigma$. Letting $\sigma \to 0$ completes the proof. □

REFERENCES

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

[2] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed., Springer-Verlag, New York, 1998.

[3] P. Dupuis, *Large deviations analysis of some recursive algorithms with state dependent noise*, Ann. Probab., 16 (1988), pp. 1509–1536.

[4] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, John Wiley, New York, 1996.

[5] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.

[6] T. E. Harris, *Theory of Branching Processes*, Springer-Verlag, Berlin, 1963.

[7] I. Iscoe, P. Ney, and E. Nummelin, *Large deviations of uniformly recurrent Markov additive processes*, Adv. in Appl. Math., 6 (1985), pp. 373–412.

[8] H. J. Kushner and F. J. Vázquez-Abad, *Stochastic approximation methods for systems over an infinite horizon*, SIAM J. Control Optim., 34 (1996), pp. 712–756.

[9] H. J. Kushner and J. Yang, *Analysis of adaptive step-size SA algorithms for parameter tracking*, IEEE Trans. Automat. Control, 40 (1995), pp. 1403–1410.

[10] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York, 1997.

[11] L. Ljung, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, 22 (1977), pp. 551–575.

[12] L. Ljung and T. Soderstrom, *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA, 1983.

[13] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.

[14] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 839–870

CONTINUITY AND EXPONENTIAL STABILITY OF MIXED CONSTRAINED MODEL PREDICTIVE CONTROL*

JINHOON CHOI† AND WOOK HYUN KWON‡

Abstract. For the infinite horizon cost function mixed constrained model predictive control, the largest possible stabilizable region for stable plants is the entire state space for both state and output feedback cases. However, for marginal or unstable cases, the largest possible stabilizable region is the constrained m-step stabilizable set for the state feedback case, and it is the region where the estimated state is in the constrained m-step stabilizable set throughout the trajectory for the output feedback case. Only attractivity over the largest stabilizable region is established for the state feedback case with stable or marginal plants and the output feedback case with stable plants [A. Zheng and M. Morari, *IEEE Trans. Automat. Control*, 40 (1995), pp. 1818–1823]. In this paper we show, for both state and output feedback cases, that the closed loop system with the mixed constrained model predictive controller possesses the exponential stability property, much stronger than the attractivity, on the largest possible stabilizable region. Here the exponential stability on the largest possible stabilizable region means that we can find the exponentially converging envelope for any initial condition in the region. Clearly this is much stronger than local exponential stability, for which the region for the envelope is not known and can be arbitrarily small. Moreover, the continuity properties of the mixed constrained model predictive control are also established.

Key words. model predictive control, continuity, exponential stability

AMS subject classifications. 93CD15, 93D05, 90C31

DOI. 10.1137/S0363012901385204

1. Introduction.

Model predictive control, also called receding horizon control, is a technique for implementing the finite horizon open loop optimal control in the infinite horizon feedback settings. Such implementation has been tried since the early 1960s when the open loop optimal control problem was vigorously studied [12], [32]. However, the first complete formulation of stabilizing finite horizon cost function model predictive control and its comprehensive stability result were obtained by Kwon and his coworkers [19], [20] in the late 1970s. They showed that the stability of unconstrained model predictive control is assured if the terminal state is constrained to be zero in the associated open loop optimal control problem. This result is particularly important since, as mentioned in [2], the closed loop stability with the primitive model predictive controller is not guaranteed in general.

Constraints are always present in any practical control problems. For instance, the input cannot assume its value outside a bounded region due to the physical limitations of manipulated variables. Moreover, it is desirable that the states of the plant lie within a designated area in the state space because of safety, environmental regulation, and so on. As a way to handle these ever existing constraints, constrained model predictive control strategies have been proposed and successfully applied to chemical processes [10], [26]. Driven by these successes in practice, the constrained model predictive control problem has been the core research topic within the chem-

*Received by the editors February 20, 2001; accepted for publication (in revised form) November 26, 2002; published electronically June 18, 2003. A preliminary version of this paper appeared in the Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, 1998.
http://www.siam.org/journals/sicon/42-3/38520.html

†Department of Chemical and Biomolecular Engineering, Sogang University, 1 Sinsoo-Dong, Mapo-Koo, Seoul 121-742, Korea (jchoi@ccs.sogang.ac.kr).

‡School of Electrical Engineering, Seoul National University, Silim-Dong, Kwanak-Koo, Seoul, Korea (whkwon@cisl.snu.ac.kr).

ical engineering control community since the late 1970s. Irrespective of this interest in the constrained model predictive control problem within the chemical engineering control community, the asymptotic stability of a hard constrained nonlinear model predictive control law was first established by Keerthi and Gilbert [18] in an effort to extend the terminal state constraint idea by Kwon and his coworkers to the hard constrained nonlinear system case. In the early 1990s, Rawlings and Muske [25] proposed an alternate formulation of hard constrained model predictive control for linear systems employing the infinite horizon cost function instead of terminal state condition and established its attractivity over the set of all initial conditions for which the model predictive control law is defined. Their result was then extended by Zheng and Morari [38] to the mixed constrained stable and marginal linear system cases where, as proposed in [37], hard input constraints and soft state constraints are employed. For output feedback cases, the perturbed system stability theory [16] was employed in [24], [27] to show the local asymptotic stability of the closed loop system with a hard constrained output feedback model predictive controller. On the other hand, Zheng and Morari [38] established that the closed loop system with a mixed constrained output feedback model predictive controller is globally attractive for the stable plant case.

From the pioneering work of maximal output admissible set by Gilbert and Tan [13] and the well-known result by Kalman that the global optimal solution of the unconstrained linear quadratic regulation problem is linear, it is easy to conclude that the mixed constrained model predictive control is linear in a neighborhood of the origin, where the constraints are not active throughout the trajectory and the penalty term for softening is zero, and thus is exponentially stable on the neighborhood. The explicit statement of this result was pointed out first by Sznaier and Damborg [31] and played a key role in solving the hard constrained infinite horizon linear quadratic regulation problem [4], [28] as well as the mixed constrained one [8], [9]. However, this local exponential stability result is not very interesting since the important stability result in the constrained case is the one in the nonlinear region where the constraints play some roles. To this end, in the infinite horizon cost function and mixed constraints cases, no stability over the largest stabilizable region is proven for the state feedback case with unstable plants and the output feedback case with marginal[1] or unstable plants, and attractivity over the largest stabilizable region is only established for the state feedback case with stable or marginal plants and the output feedback case with stable plants. Moreover, exponential stability over the set of all initial conditions for which the model predictive control law is defined cannot be concluded from local exponential stability and attractivity over the set. In this paper we will show, for both state and output feedback cases, that the closed loop system with the mixed constrained model predictive controller possesses the exponential stability property, much stronger than the attractivity, on the set of all initial conditions for which the mixed constrained model predictive control law is defined throughout the trajectory. Moreover, the continuity properties of the mixed constrained model predictive control are also established.

This paper deals with mixed constrained model predictive control, and the control input to the process is limited by the hard constraints such as saturation. Then the stability of mixed constrained model predictive control corresponds to the stabilization of a hard input constrained linear system by mixed constrained model predictive

[1] A linear system is marginal if it has some poles on the unit circle and all the others inside the unit circle.

control. Hence, the general results on the stabilization of a hard input constrained linear system will provide some possible guidelines for study of the stability of the specific constrained model predictive control. Contrary to the unconstrained linear system cases, the global asymptotic stabilization of a hard input constrained linear system is not always possible. Indeed, as shown in [30], [35], a hard input constrained linear discrete time system can be globally asymptotically stabilizable iff all its poles are located inside or on the unit circle. Clearly, a hard input constrained stable linear system is globally exponentially stable without any control and, thus, is globally exponentially stabilizable. It is shown in [5] that the global asymptotic stabilization of a hard input constrained linear system is possible through linear feedback if the linear system is marginally stable; i.e., Jordan blocks associated with eigenvalues on the unit circle are diagonal and all the other eigenvalues are in the open unit disk. However, Yang [36] showed that the global asymptotic stabilization of hard input constrained marginal systems is, in general, impossible through linear feedback. Instead, Yang [35] was able to construct a nonlinear (nonpredictive) globally asymptotically stabilizing control law for such systems. It is shown in [9] that the mixed constrained infinite horizon linear quadratic optimal control that is nonlinear also achieves global asymptotic stability. On the other hand, Lin and Saberi [22] and Lin [21] showed that there exists a linear controller that exponentially stabilizes a hard input constrained marginal system on any bounded subset of the state space. Nevertheless, it is shown in [6] that the global exponential stabilization of a hard input constrained marginal system is not possible by any control law. For unstable systems, it was established in [5] that any input constrained unstable systems can be exponentially stabilized by a linear periodic variable structure controller on any compact subset of the constrained asymptotically stabilizable set,[2] i.e., the set of all initial states whose unstable subspace part can be driven asymptotically to zero with constrained inputs. Moreover, it is shown in [9] that the mixed constrained infinite horizon linear quadratic optimal control that is nonlinear achieves asymptotic stability on the constrained asymptotically stabilizable set. However, Choi [7] showed that the asymptotic stabilization of a hard input constrained unstable system on the constrained asymptotically stabilizable set is not possible in general by linear feedback. Moreover, it is shown in [6] that the exponential stabilization of a hard input constrained unstable system on the constrained asymptotically stabilizable set is not possible by any control law.

In this paper, we first adopt the point-to-set map theory to explore the continuity properties of the optimal solution to the quadratic program associated with the mixed constrained state feedback model predictive control law, which are useful for the rest of the paper. It was brought to the authors' attention by a reviewer that the continuity properties of the hard constrained state feedback model predictive control law were reported using multiparametric programming [1] and a geometric approach [29]. We then establish the exponential stability properties of the mixed constrained model predictive control law under the assumption that the linear plant in the absence of the input saturation is stabilizable and detectable. For the state feedback case, the exponential stability is established by proving that the optimal cost function is an appropriate Lyapunov function for the closed loop system. Indeed, an input constrained stable system is shown to be globally exponentially stabilized in the sense of Lyapunov by the mixed constrained state feedback model predictive control law. Moreover, an input constrained marginal or unstable system is proven to be exponentially stable

[2]Let $\Omega_\infty = \{u : u^{min} \le u_i \le u^{max},\ i = 0, 1, \ldots\}$. Then the maximal Ω_∞-invariant set w.r.t. $\mathbf{R}^n$, proposed in [14], coincides with the constrained asymptotically stabilizable set.

on the constrained m-step stabilizable set that converges to the constrained asymptotically stabilizable set as the control horizon m increases. Hence, a marginal system is semiglobally stabilized by the mixed constrained model predictive control and an unstable system is exponentially stabilized by the mixed constrained model predictive control law on any compact subset of the constrained asymptotically stabilizable set. We also demonstrate that the sum of the optimal objective associated with state feedback model predictive control and a positive definite quadratic term of the observer error forms a Lyapunov function from which one can conclude the exponential stability of the mixed constrained output feedback model predictive control. Even if output is only available for feedback, a stable plant is globally exponentially stabilized. For the marginal and unstable plant cases, the exponential stability is established on the set of all initial conditions for which the mixed constrained model predictive control law is defined throughout the trajectory.

This paper is structured as follows. In the next section, the preliminaries on the point-to-set map theory are summarized. In section 3, the formulation of the mixed constrained model predictive control law is presented. The continuity properties of the quadratic program associated with the mixed constrained model predictive control law are established in section 4. The exponential stability of mixed constrained state feedback model predictive control is established in section 5, whereas the output feedback case is presented in section 6. Finally, the conclusions of the paper are drawn in section 7.

2. Preliminaries. Later in this paper, the point-to-set map theory is employed to establish the continuity properties of the quadratic program associated with the mixed constrained model predictive control law. Hence, in this section, some preliminaries on the point-to-set map theory are summarized as exposed in [17].

DEFINITION 2.1. *A point-to-set map Ω from a set X into a set Y is a map which associates a subset of Y with each point of X.*

DEFINITION 2.2. *Ω is open at a point $\bar{x}$ in X if $\{x_k\}_{k=0}^{\infty} \subset X$, $x_k \to \bar{x}$, and $\bar{y} \in \Omega(\bar{x})$ imply the existence of an integer m and a sequence $\{y_k\}_{k=0}^{\infty} \subset Y$ such that $y_k \in \Omega(x_k)$ for $k \geq m$ and $y_k \to \bar{y}$.*

DEFINITION 2.3. *Ω is closed at a point $\bar{x}$ in X if $\{x_k\}_{k=0}^{\infty} \subset X$, $x_k \to \bar{x}$, $y_k \in \Omega(x_k)$, and $y_k \to \bar{y}$ imply that $\bar{y} \in \Omega(\bar{x})$.*

DEFINITION 2.4. *Ω is continuous at a point $\bar{x}$ in X if it is both open and closed at $\bar{x}$.*

DEFINITION 2.5. *Ω is uniformly compact near $\bar{x}$ if there is a neighborhood N of $\bar{x}$ such that the closure of the set $\cup_{x \in N} \Omega(x)$ is compact.*

Consider the following infimal value function associated with an optimization problem:

$$v(x) = \inf_{y \in \Omega(x)} f(x, y). \tag{1}$$

Within the point-to-set map framework, we have the following theorem for the continuity of a infimal value function.

THEOREM 2.1 (see [17]). *If Ω is continuous at $\bar{x}$ and uniformly compact near $\bar{x}$, and if f is continuous on $\bar{x} \times \Omega(\bar{x})$, then v is continuous at $\bar{x}$.*

Let M be the solution set of (1), defined as

$$M(x) := \{y \in \Omega(x) : v(x) \geq f(x, y)\}.$$

Then we have the following theorem for the continuity of the optimal solution to (1).

THEOREM 2.2 (see [17]). *Suppose Ω is continuous at $\bar{x}$, f is continuous on $\bar{x} \times \Omega(\bar{x})$, M is nonempty and uniformly compact near $\bar{x}$, and M is single-valued. Then M is continuous at $\bar{x}$.*

Before we proceed to the next section, we set up some notations that are used throughout the paper. Given a vector x, $|x|$ and $|x|_\infty$ denote the 2 and ∞ norms of x, respectively. Given a matrix A, $\|A\|$ denotes the induced 2 norm of A. For a given set Ω, intΩ denotes the interior of Ω. Given a p-dimensional vector w, w_l denotes the lth element of w.

3. Formulation. Consider the input constrained linear discrete time system:

$$x_{k+1} = Ax_k + B\sigma(u_k), \tag{2}$$

$$y_k = Cx_k,$$

where $x_k \in \mathbf{R}^n$, $u_k \in \mathbf{R}^q$, $y_k \in \mathbf{R}^l$, and

$$\sigma_i(u_k) := \begin{cases} u_i^{min} & \text{if } u_{ki} < u_i^{min}, \\ u_i & \text{if } u_i^{min} \le u_{ki} \le u_i^{max}, \\ u_i^{max} & \text{if } u_{ki} > u_i^{max}. \end{cases}$$

Throughout the paper, system (2) is assumed to be stabilizable and detectable in the absence of the input saturation.

Assumption 3.1. (A, B) and (A, C) are stabilizable and detectable pairs, respectively.

It is assumed that the control and the state of the system are desired to satisfy the following inequalities:

$$u^{min} \le u_k \le u^{max}, \qquad k = 0, 1, \ldots,$$

$$Gx_k \le g, \qquad k = 0, 1, \ldots,$$

where $G \in \mathbf{R}^{p \times n}$ and $g \in \mathbf{R}^p$.

Associated with system (2), consider the following quadratic program (QP):

$$J_m(x_k) = \min_{U_k^m, \epsilon_k} \sum_{i=1}^{\infty} x_{k+i|k}^T R x_{k+i|k} + \sum_{i=0}^{m-1} u_{k+i|k}^T S u_{k+i|k} + \epsilon_k^T Q \epsilon_k \tag{3}$$

subject to

$$x_{k+i+1|k} = Ax_{k+i|k} + Bu_{k+i|k}, \quad x_{k|k} = x_k, \tag{4}$$

$$u^{min} \le u_{k+i|k} \le u^{max}, \qquad i = 0, 1, \ldots, m-1,$$

$$u_{k+i|k} = 0, \qquad i = m, m+1, \ldots,$$

$$G\hat{x}_{k+i|k} \le g + \epsilon_k, \qquad i = 1, \ldots,$$

$$\epsilon_k \ge 0,$$

where $R > 0$ and $S > 0$ are symmetric matrices, $Q > 0$ is a diagonal matrix, $\epsilon_k \in \mathbf{R}^p$, m is finite, and $U_k^m := [u_{k|k}^T \; u_{k+1|k}^T \; \cdots \; u_{k+m-1|k}^T]^T$.

Then mixed constrained model predictive control with horizon m ($\mathrm{MPC_m}$) is a nonlinear static state feedback control law whose control output at the kth sampling time, u_k, is the first q elements $u_{k|k}^{m*}$ of the optimal solution U_k^{m*} to the quadratic optimization problem $J_m(x_k)$. Since we consider the regulation problem throughout the paper, the following conditions are assumed to hold for the well-posedness of the problem.

Assumption 3.2.

$$0 \in \mathrm{int}\{[v_0 \cdots v_{m-1}]^T : v_i \in \mathbf{R}^q, \;\; u^{min} \le v_i \le u^{max}, \;\; i = 0, 1, \ldots, m-1\},$$

$$0 \in \mathrm{int}\left\{z : z \in \mathbf{R}^n, \;\; Gz \le g\right\}.$$

Note that the first condition implies $u^{min} < 0 < u^{max}$ and the second condition dictates $g > 0$.

The incremental input constraints

$$\Delta u^{min} \le \Delta u_k \le \Delta u^{max}, \qquad k = 0, 1, \ldots,$$

where $\Delta u_k := u_k - u_{k-1}$, are also considered in some formulations of model predictive control. In this paper, the exponential stability of the mixed constrained model predictive control law is established by showing that the optimal cost function is indeed a Lyapunov function. However, if the incremental input constraints are considered, the optimal cost function depends not only on x_k but also on u_{k-1} and, thus, it is not a Lyapunov function for the exponential stability of the mixed constrained model predictive control law. Even if the state is augmented with u_{k-1}, the resulting state constraints are not in the form considered above and, although we consider the difference form of constraints as well, they may not satisfy Assumption 3.2. Hence, the results in this paper are not applicable in the presence of the incremental input constraints.

4. Continuity properties. In this section, we establish the continuity properties of J_m[3] and the associated optimal solution employing the point-to-set map theory presented in the preliminaries section.

4.1. Stable plants. Let

$$\Phi_m(x_k, U_k^m, \epsilon_k) := \sum_{i=1}^{\infty} x_{k+i|k}^T R x_{k+i|k} + \sum_{i=0}^{m-1} u_{k+i|k}^T S u_{k+i|k} + \epsilon_k^T Q \epsilon_k,$$

where

$$x_{k+i|k} = \begin{cases} A^i x_k + \sum_{j=0}^{i-1} A^{i-j-1} B u_{k+j|k}, & i = 1, \ldots, m-1, \\ A^i x_k + A^{i-m} \sum_{j=0}^{m-1} A^{m-j-1} B u_{k+j|k}, & i = m, m+1, \ldots. \end{cases}$$

[3]The Lyapunov theorem for exponential stability doesn't require the continuity of the Lyapunov function. But the continuity of J_m follows trivially in the middle of establishing the continuity of the optimal solution and is included.

Clearly, $\Phi_m(x_k, U_k^m, \epsilon_k)$ is continuous w.r.t. (x_k, U_k^m, ϵ_k) iff $\sum_{i=1}^{\infty} x_{k+i|k}^T R x_{k+i|k}$ is continuous w.r.t. (x_k, U_k^m). To see the continuity of $\sum_{i=1}^{\infty} x_{k+i|k}^T R x_{k+i|k}$, first observe that

$$\sum_{i=1}^{\infty} x_{k+i|k}^T R x_{k+i|k} = \sum_{i=1}^{m-1} x_{k+i|k}^T R x_{k+i|k} + x_{k+m|k}^T \left[\sum_{i=0}^{\infty} (A^T)^i R A^i \right] x_{k+m|k}. \tag{5}$$

Then the first term on right-hand side (RHS) of (5) is clearly continuous w.r.t. (x_k, U_k^m). Moreover, since A is stable, $\sum_{i=0}^{\infty} (A^T)^i R A^i =: T$ is simply a unique positive definite solution of the following Lyapunov equation [3, p. 215]:

$$T = A^T T A + R.$$

Hence, the second term on the RHS of (5) is also continuous w.r.t. (x_k, U_k^m), and, in turn, left-hand side (LHS) of (5) is continuous w.r.t. (x_k, U_k^m).

We now define a point-to-set map. For this, we need the following fact.

Fact 4.1 (see [3, p. 213]). If every eigenvalue of A has magnitude strictly less than one, there exist $\rho \in [0, 1)$ and $D \geq 1$ such that

$$\|A^i\| \leq D\rho^i.$$

Define

$$\Omega_m(x_k) := \left\{ \begin{array}{l} (U_k^m, \epsilon_k) \in \mathbf{R}^{p+mq} : u^{min} \leq u_{k+i|k} \leq u^{max}, \qquad i = 0, 1, \ldots, m-1, \\ \\ GA^i x_k + G \sum_{j=0}^{i-1} A^{i-j-1} B u_{k+j|k} \leq g + \epsilon_k, \qquad i = 1, \ldots, m-1, \\ \\ GA^i x_k + GA^{i-m} \sum_{j=0}^{m-1} A^{m-j-1} B u_{k+j|k} \leq g + \epsilon_k, \qquad i = m, m+1, \ldots, \\ g + \epsilon_k \leq (\max\{|g|_\infty, K_1|x_k| + K_2\} + 1)\mathbf{1}, \\ \epsilon_k \geq 0 \end{array} \right\},$$

where

$$\mathbf{1} := [1\ 1\ \cdots\ 1]^T,$$

$$K_1 := \|G\|D \geq \|G\| \max_{i \geq 1} D\rho^i,$$

$$K_2 := \max \left\{ \begin{array}{c} \max\limits_{1 \leq i \leq m-1} \max\limits_{U_k^m \in \Psi_k^m} \|G\| \sum\limits_{j=0}^{i-1} D\rho^{i-j-1} \|B\| \sqrt{m} |u_{k+j|k}|_\infty \\ \\ \max\limits_{i \geq m} \max\limits_{U_k^m \in \Psi_k^m} \|G\| \sum\limits_{j=0}^{m-1} D\rho^{i-j-1} \|B\| \sqrt{m} |u_{k+j|k}|_\infty \end{array} \right\}$$

$$= \|G\| D \frac{1 - \rho^m}{1 - \rho} \|B\| \sqrt{m} \vartheta,$$

$$\Psi_k^m := \left\{ U_k^m \in \mathbf{R}^{mq} : u^{min} \le u_{k+i|k} \le u^{max}, \quad i = 0, 1, \ldots, m-1 \right\},$$

$$\vartheta := \max \left\{ \begin{array}{l} |u_{min}|_\infty \\ |u_{max}|_\infty \end{array} \right\}.$$

Associated with $\Omega_m(x_k)$, consider the following QP:

$$\tilde{J}_m(x_k) := \min_{(U_k^m, \epsilon_k) \in \Omega_m(x_k)} \Phi_m(x_k, U_k^m, \epsilon_k). \tag{6}$$

QP (6) contains some additional constraints compared to QP (3). Hence, $\Omega_m(x_k)$ is a subset of the feasible region of QP (3) and, thus, $J_m(\cdot) \le \tilde{J}_m(\cdot)$. However, K_1 and K_2 in the aforementioned additional constraints are chosen such that any optimal feasible point for QP (3) satisfies them due to the penalty term of ϵ_k. Hence, $J_m(\cdot) = \tilde{J}_m(\cdot)$.

THEOREM 4.1. *$\tilde{J}_m(\cdot) = J_m(\cdot)$ is continuous on $\mathbf{R}^n$.*

Proof. $\Phi_m(x_k, U_k^m, \epsilon_k)$ is shown to be continuous w.r.t. (x_k, U_k^m, ϵ_k) right before Fact 4.1. Hence, from Theorem 2.1, it suffices to show that $\Omega_m(\cdot)$ is open and closed on $\mathbf{R}^n$, and $\Omega_m(\cdot)$ is uniformly compact near any $\bar{x} \in \mathbf{R}^n$.

We first establish that $\Omega_m(\cdot)$ is open on $\mathbf{R}^n$. To see this, consider $\bar{x} \in \mathbf{R}^n$. Let $\{x_k\}_{k=0}^\infty \subset \mathbf{R}^n$, $x_k \to \bar{x}$, and $(\bar{U}^m, \bar{\epsilon}) \in \Omega_m(\bar{x})$. Let $\{\delta_j\}_{j=0}^\infty$ be a monotonically decreasing sequence of positive integers such that $\delta_0 = 1$ and $\delta_j \to 0$ as $j \to \infty$. Then there exists N_0 such that $|x_k - \bar{x}| < \frac{\delta_0}{3\max\{K_1, \|G\|D\}}$ for all $k \ge N_0$. Moreover, for each $j \ge 1$, there exists $N_j > N_{j-1}$ such that $|x_k - \bar{x}| < \frac{\delta_j}{3\max\{K_1, \|G\|D\}}$ for all $k \ge N_j$. Define

$$f_i(U_k^m) := \begin{cases} G\sum_{j=0}^{i-1} A^{i-j-1} B u_{k+j|k}, & 1 \le i \le m-1, \\ GA^{i-m} \sum_{j=0}^{m-1} A^{m-j-1} B u_{k+j|k}, & i \ge m. \end{cases}$$

For each $1 \le l \le p$, let $\gamma_l \ge 0$ be such that

$$(GA^i \bar{x} + f_i(\bar{U}^m))_l = g_l + \bar{\epsilon}_l - \gamma_l, \qquad i \ge 1.$$

For $N_j \le k < N_{j+1}$, let $U_k^m := \bar{U}^m$ and

$$\epsilon_{kl} := \begin{cases} \bar{\epsilon}_l - \frac{\delta_j}{3} & \text{if } g_l + \bar{\epsilon}_l \ge \max\{|g|_\infty, K_1|\bar{x}| + K_2\} + 1 - \frac{\delta_j}{3}, \\ \bar{\epsilon}_l + \frac{\delta_j}{3} & \text{if } \gamma_l \le \frac{\delta_j}{3}, \\ \bar{\epsilon}_l & \text{otherwise.} \end{cases} \qquad 1 \le l \le p.$$

The first two cases in the definition of ϵ_{kl} are disjoint because $\gamma_l \le \frac{\delta_j}{3}$ implies

$$g_l + \bar{\epsilon}_l \le \max\{|g|_\infty, K_1|\bar{x}| + K_2\} + \gamma_l < \max\{|g|_\infty, K_1|\bar{x}| + K_2\} + 1 - \frac{\delta_j}{3}.$$

It clearly holds that

$$u^{min} \le u_{k+i|k} \le u^{max}, \qquad i = 0, 1, \ldots, m-1.$$

Moreover, all the other constraints are also satisfied for each case considered in the definition of ϵ_{kl} because, for each $N_j \le k < N_{j+1}$, $1 \le l \le p$, the following inequalities hold.

The $g_l + \bar{\epsilon}_l \geq \max\{|g|_\infty, K_1|\bar{x}| + K_2\} + 1 - \frac{\delta_j}{3}$ *case.*

$$(GA^i x_k + f_i(U_k^m))_l \leq (GA^i \bar{x} + f_i(\bar{U}^m))_l + \frac{\delta_i}{3} \leq K_1|\bar{x}| + K_2 + \frac{\delta_j}{3}$$

$$\leq g_l + \bar{\epsilon}_l - 1 + \frac{2\delta_j}{3} \leq g_l + \bar{\epsilon}_l - \frac{\delta_j}{3} = g_l + \epsilon_{kl}, \qquad i \geq 0.$$

$$g_l + \epsilon_{kl} = g_l + \bar{\epsilon}_l - \frac{\delta_j}{3} \leq \max\{|g|_\infty, K_1|\bar{x}| + K_2\} + 1 - \frac{\delta_j}{3} \leq \max\{|g|_\infty, K_1|x_k| + K_2\} + 1.$$

$$\epsilon_{kl} = \bar{\epsilon}_l - \frac{\delta_j}{3} \geq \max\{|g|_\infty, K_1|\bar{x}| + K_2\} + 1 - \frac{2\delta_j}{3} - g_l \geq \max\{|g|_\infty, K_1|x_k| + K_2\} - g_l \geq 0.$$

The $\gamma_l \leq \frac{\delta_j}{3}$ *case.*

$$(GA^i x_k + f_i(U_k^m))_l \leq (GA^i \bar{x} + f_i(\bar{U}^m))_l + \frac{\delta_j}{3} \leq g_l + \bar{\epsilon}_l + \frac{\delta_j}{3} = g_l + \epsilon_{kl}, \quad i \geq 0.$$

$$g_l + \epsilon_{kl} = g_l + \bar{\epsilon}_l + \frac{\delta_j}{3} \leq g_l + \bar{\epsilon}_l + 1 - \frac{2\delta_j}{3} \leq g_l + \bar{\epsilon}_l - \gamma_l + 1 - \frac{\delta_j}{3} = (GA^i \bar{x} + f_i(\bar{U}^m))_l + 1 - \frac{\delta_j}{3}$$

$$\leq \max\{|g|_\infty, K_1|\bar{x}| + K_2\} + 1 - \frac{\delta_j}{3} \leq \max\{|g|_\infty, K_1|x_k| + K_2\} + 1.$$

$$\epsilon_{k,l} = \bar{\epsilon}_l + \frac{\delta_j}{3} \geq 0.$$

The $g_l + \bar{\epsilon}_l < \max\{|g|_\infty, K_1|\bar{x}| + K_2\} + 1 - \frac{\delta_j}{3}$ and $\gamma_l > \frac{\delta_j}{3}$ *case.*

$$(GA^i x_k + f_i(U_k^m))_l \leq (GA^i \bar{x} + f_i(\bar{U}^m))_l + \frac{\delta_j}{3} = g_l + \bar{\epsilon}_l - \gamma_l + \frac{\delta_j}{3} < g_l + \bar{\epsilon}_l = g_l + \epsilon_{kl}, \quad i \geq 0.$$

$$g_l + \epsilon_{kl} = g_l + \bar{\epsilon}_l < \max\{|g|_\infty, K_1|\bar{x}| + K_2\} + 1 - \frac{\delta_j}{3} \leq \max\{|g|_\infty, K_1|x_k| + K_2\} + 1.$$

$$\epsilon_{k,l} = \bar{\epsilon}_l \geq 0.$$

Hence, for $k \geq N_0$, $(U_k^m, \epsilon_k) \in \Omega_m(x_k)$ and $(U_k^m, \epsilon_k) \to (\bar{U}^m, \bar{\epsilon})$.

Next we show that $\Omega_m(\cdot)$ is closed on $\mathbf{R}^n$. For this, let $\{x_k\}_{k=0}^\infty \subset \mathbf{R}^n$, $x_k \to \bar{x}$, $(U_k^m, \epsilon_k) \in \Omega_m(x_k)$, and $(U_k^m, \epsilon_k) \to (\bar{U}^m, \bar{\epsilon})$. Then since all the functionals in the inequalities that define $\Omega_m(\cdot)$ are continuous w.r.t. (x_0, U_k^m, ϵ_k), it holds that $(\bar{U}^m, \bar{\epsilon}) \in \Omega_m(\bar{x})$.

Finally, we show that $\Omega_m(\cdot)$ is uniformly compact near any $\bar{x} \in \mathbf{R}^n$. To see this, consider the closed unit ball $B^n(1)$ around $\bar{x}$. A set in a finite-dimensional space is compact iff it is closed and bounded. Hence, we only prove that $\cup_{x \in B^n(1)} \Omega_m(x)$ is bounded. But, for any $x \in B^n(1)$, $\Omega_m(x)$ is contained in the bounded set $\Psi_k^m \times \{\epsilon_k \in \mathbf{R}^p : \epsilon_k \geq 0, \text{ and } g + \epsilon_k \leq (\max\{|g|_\infty, K_1(|\bar{x}| + 1) + K_2\} + 1)\mathbf{1}\}$. ∎

We now show that U_k^{m*} is continuous on $\mathbf{R}^n$.

THEOREM 4.2. *U_k^{m*} is continuous on $\mathbf{R}^n$.*

Proof. We first show that U_k^{m*} is uniformly compact near any $\bar{x} \in \mathbf{R}^n$. Similar to the uniform compactness proof of $\Omega_m(\cdot)$ in Theorem 4.2, we only prove that $\cup_{x \in B^n(1)} U_k^{m*}$ is bounded. But, for any $x \in B^n(1)$, U_k^{m*} is contained in the bounded set Ψ_k^m.

Since S is positive definite and inputs are upper and lower bounded, it is well known that the optimal solution U_k^{m*} to $J_m = \tilde{J}_m$ for each $x_k \in \mathbf{R}^n$ exists and is unique. This implies U_k^{m*} is nonempty and single-valued on $\mathbf{R}^n$. Since U_k^{m*} is also uniformly compact near any $\bar{x} \in \mathbf{R}^n$, the continuity of U_k^{m*} follows from Theorem 2.2. ∎

4.2. Marginal or unstable plants. If A contains marginal and unstable modes, we partition A as follows:

$$A = V \mathcal{J} V^{-1} = [V_u V_s] \begin{bmatrix} \mathcal{J}_u & 0 \\ 0 & \mathcal{J}_s \end{bmatrix} \begin{bmatrix} \tilde{V}_u \\ \tilde{V}_s \end{bmatrix},$$

where $\mathcal{J}_s$ contains the eigenvalues of A that lie in the open unit disk and $\mathcal{J}_u$ contains the rest. Then for any U_k^m such that the objective of $\mathrm{MPC_m}$ is bounded, the following equation must be satisfied:

$$\tilde{V}_u x_{k+m|k} = 0 \tag{7}$$

since the objective would be unbounded otherwise. Hence, the necessary and sufficient condition for the finiteness of the optimum of QP (3) is the existence of $U_k^m \in \Psi_k^m$ for which (7) is satisfied. Clearly, the addition of (7) as constraints to QP (3) does not alter the optimal solution.

Define the constrained m-step null controllable set as

$$\chi_m := \left\{ x \in \mathbf{R}^n : 0 = A^m x + \sum_{i=0}^{m-1} A^{m-i-1} B u_{k+i|k}, \right.$$
$$\left. u^{min} \le u_{k+i|k} \le u^{max}, \;\; i = 0, 1, \ldots, m-1 \right\}.$$

Assumption 3.2 dictates that χ_m contains an open neighborhood of the origin in the stabilizable subspace of (A, B) for sufficiently large m. Notice that, from Assumption 3.1, the marginal and unstable subspace of (A, B) is contained in the stabilizable subspace of (A, B). Define

$$\mathcal{E} := \{x : \tilde{V}_u A^m x = 0\}.$$

Then $\mathcal{E}$ is independent of m and is indeed the stable subspace of (A, B), because the following equalities hold:

$$\mathcal{E} = \left\{ x : \tilde{V}_u [V_u V_s] \begin{bmatrix} \mathcal{J}_u^m & 0 \\ 0 & \mathcal{J}_s^m \end{bmatrix} \begin{bmatrix} \tilde{V}_u \\ \tilde{V}_s \end{bmatrix} x = 0 \right\}$$

$$= \left\{ x : \tilde{V}_u V_u \mathcal{J}_u^m \tilde{V}_u x + \tilde{V}_u V_s \mathcal{J}_s^m \tilde{V}_s x = 0 \right\}$$

$$= \left\{ x : \mathcal{J}_u^m \tilde{V}_u x = 0 \right\} = \left\{ x : \tilde{V}_u x = 0 \right\}.$$

Notice that the second last equality follows from the facts that $\tilde{V}_u V_u = I$ and $\tilde{V}_u V_s = 0$.

THEOREM 4.3. *There exists* $U_k^m \in \Psi_k^m$ *such that* $\tilde{V}_u x_{k+m|k} = 0$ *iff*

$$x_k \in \pi_m := \{x \in \mathbf{R}^n : x = y + z,\ \ y \in \chi_m,\ z \in \mathcal{E}\},$$

where π_m *is called constrained* m*-step stabilizable set.*[4]

Proof. The sufficiency part is obvious. To show the necessity, suppose $\tilde{V}_u x_{k+m|k} = 0$. Since $U_k^m \in \Psi_k^m$, there exists y in χ_m such that

$$0 = A^m y + \sum_{i=0}^{m-1} A^{m-i-1} B u_{k+i|k}.$$

Let

$$z := x_k - y.$$

Then $z \in \mathcal{E}$ since $\tilde{V}_u z \neq 0$ implies $0 \neq \mathcal{J}_u^m \tilde{V}_u z = \tilde{V}_u A^m z = \tilde{V}_u x_{k+m|k}$. Hence, the theorem follows. □

Notice that π_m contains an open neighborhood of the origin in $\mathbf{R}^n$ for sufficiently large m because $\mathcal{E}$ is the stable subspace of (A, B) and χ_m contains an open neighborhood of the origin in the marginal and unstable subspace of (A, B) for sufficiently large m.

From Fact 4.1 and the stability of $\mathcal{J}_s$, there exist $\nu \in [0, 1)$ and $H \in [1, \infty)$ such that

$$\|\mathcal{J}_s^i\| \leq H\nu^i.$$

Then a compact set of feasible points for QP (3) that contains the optimal solution is

$$\bar{\Omega}_m(x_k) := \left\{ \begin{array}{ll} (U_k^m, \epsilon_k) \in \mathbf{R}^{p+mq} : u^{min} \leq u_{k+i|k} \leq u^{max}, & i = 0, 1, \ldots, m-1, \\ GA^i x_k + G\sum_{j=0}^{i-1} A^{i-j-1} B u_{k+j|k} \leq g + \epsilon_k, & i = 1, \ldots, m-1, \\ GA^i x_k + GA^{i-m}\sum_{j=0}^{m-1} A^{m-j-1} B u_{k+j|k} \leq g + \epsilon_k, & i = m, m+1, \ldots, \\ g + \epsilon_k \leq (\max\{|g|_\infty, \bar{K}_1|x_k| + \bar{K}_2\} + 1)\mathbf{1}, & \\ \epsilon_k \geq 0, & \\ \tilde{V}_u x_{k+m|k} = 0 \end{array} \right\},$$

where

$$\bar{K}_1 := \|G\|\|V_s\|\|\tilde{V}_s\| H \mathcal{A},$$

$$\bar{K}_2 := \|G\|\|V_s\|\|\tilde{V}_s\| H m \mathcal{A}\|B\|\sqrt{q}\vartheta.$$

[4]Let $\Omega_m = \{u : u^{min} \leq u_i \leq u^{max},\ i = 0, 1, \ldots, m-1,\ u_j = 0,\ j = m, m+1, \ldots\}$. Then the maximal Ω_m-invariant set w.r.t. $\mathbf{R}^n$, proposed in [14], coincides with the constrained m-step stabilizable set.

Then, for any $x_k \in \pi_m$ and $(U_k^m, \epsilon_k) \in \bar{\Omega}_m(x_k)$, it holds that

$$\sum_{i=1}^{\infty} x_{k+i|k}^T R x_{k+i|k} = \sum_{i=1}^{m-1} x_{k+i|k}^T R x_{k+i|k} + x_{k+m|k}^T \left[\sum_{i=0}^{\infty} \tilde{V}_s^T \mathcal{J}_s^{iT} V_s^T R V_s \mathcal{J}_s^i \tilde{V}_s \right] x_{k+m|k}.$$

Hence, through the same arguments as the stable plant case, $\Phi_m(\cdot,\cdot,\cdot)$ is continuous w.r.t. (x_k, U_k^m, ϵ_k) on $\{x_k\} \times \bar{\Omega}_m(x_k)$ for each $x_k \in \pi_m$. We now show that $\bar{\Omega}_m(\cdot)$ is open on π_m.

For this, we need the following fact, which is proven similarly to Theorem 6.3.

Fact 4.2. Suppose $x_k, \bar{x} \in \pi_m$, $(\bar{U}^m, \bar{\epsilon}) \in \bar{\Omega}_m(\bar{x})$, and

$$\Xi_m(\bar{U}^m, x_k - \bar{x}) := \min_{V_k^m \in \Upsilon^m(\bar{U}^m)} \sum_{i=1}^{\infty} \xi_{k+i|k}^T R \xi_{k+i|k} + \sum_{i=0}^{m} v_{k+i|k}^T S v_{k+i|k},$$

where

$$\Upsilon^m(\bar{U}^m) := \{V_k^m \in \mathbf{R}^{mq} : u^{min} - \bar{u} \le v_{k+i|k} \le u^{max} - \bar{u}\},$$

$$\xi_{k+i+1|k} = A\xi_{k+i|k} + B v_{k+i|k}, \quad i \ge 0,$$

$$\xi_{k|k} = x_k - \bar{x}.$$

Then the solution V_k^{m*} to $\Xi_m(\bar{U}^m, \cdot)$ is Lipschitz continuous on $\{x : x + \bar{x} \in \pi_m\}$.

THEOREM 4.4. *$J_m(\cdot) = \bar{J}_m(\cdot)$ is continuous on π_m, where*

$$\bar{J}_m(x_k) := \min_{(U_k^m, \epsilon_k) \in \bar{\Omega}_m(x_k)} \Phi_m(x_k, U_k^m, \epsilon_k). \tag{8}$$

Proof. We first show that $\bar{\Omega}_m(\cdot)$ is open on π_m. For this, consider $\bar{x} \in \pi_m$. Let $\{x_k\}_{k=0}^{\infty} \subset \pi_m$, $x_k \to \bar{x}$, $(\bar{U}^m, \bar{\epsilon}) \in \bar{\Omega}_m(\bar{x})$, and $U_k^m = \bar{U}^m + V_k^{m*}$. Let $\{\delta_j\}_{j=0}^{\infty}$ be a monotonically decreasing sequence of positive integers such that $\delta_0 = 1$ and $\delta_j \to 0$ as $j \to \infty$. Define

$$f_i(x_k, U_k^m) := \begin{cases} GA^i x_k + G\sum_{j=0}^{i-1} A^{i-j-1} B u_{k+j|k}, & 1 \le i \le m-1, \\ GA^i x_k + GA^{i-m} \sum_{j=0}^{m-1} A^{m-j-1} B u_{k+j|k}, & i \ge m. \end{cases}$$

Notice that

$$f_i(x_k, \bar{U}^m + V_k^{m*}) - f_i(\bar{x}, \bar{U}^m) = GA^{i-m} \left\{ A^m (x_k - \bar{x}) + \sum_{j=0}^{m-1} A^{m-j-1} B v_{k+j|k}^{m*} \right\}, \quad i \ge m.$$

Hence, from Fact 4.2, $f_i(x_k, \bar{U}^m + V_k^{m*}) - f_i(\bar{x}, \bar{U}^m)$ is Lipschitz continuous w.r.t. $x_k - \bar{x}$. Let κ_i denote the smallest Lipschitz constant for given i. We now claim that $\kappa := \max_i \kappa_i$ is finite. To prove this claim, we first notice that for $i \ge m$,

$$f_i(x_k, \bar{U}^m + V_k^{m*}) - f_i(\bar{x}, \bar{U}^m) = GA^{i-m} \xi_{k+m|k}.$$

Since $\tilde{V}_u \xi_{k+m|k} = 0$, it follows that

$$f_i(x_k, \bar{U}^m + V_k^{m*}) - f_i(\bar{x}, \bar{U}^m) = GV_s \mathcal{J}_s^{i-m} \tilde{V}_s \xi_{k+m|k}.$$

Hence, the claim follows from the stability of $\mathcal{J}_s$.

Now there exists N_0 such that $|x_k - \bar{x}| < \frac{\delta_0}{3\max\{\bar{K}_1,\kappa\}}$ for all $k \geq N_0$. Moreover, for each $j \geq 1$, there exists $N_j > N_{j-1}$ such that $|x_k - \bar{x}| < \frac{\delta_j}{3\max\{\bar{K}_1,\kappa\}}$ for all $k \geq N_j$.

Let γ_l and ϵ_{kl} be defined similarly to the stable plant case. Then, the following inequalities hold.

The $g_l + \bar{\epsilon}_l \geq \max\{|g|_\infty, \bar{K}_1|\bar{x}| + \bar{K}_2\} + 1 - \frac{\delta_j}{3}$ *case.*

$$(f_i(x_k, U_k^m))_l \leq (f_i(\bar{x}, \bar{U}^m))_l + \frac{\delta_i}{3} \leq \bar{K}_1|\bar{x}| + \bar{K}_2 + \frac{\delta_j}{3}$$

$$\leq g_l + \bar{\epsilon}_l - 1 + \frac{2\delta_j}{3} \leq g_l + \bar{\epsilon}_l - \frac{\delta_j}{3} = g_l + \epsilon_{kl}, \qquad i \geq 0.$$

The $\gamma_l \leq \frac{\delta_j}{3}$ *case.*

$$(f_i(x_k, U_k^m))_l \leq (f_i(\bar{x}, \bar{U}^m))_l + \frac{\delta_j}{3} \leq g_l + \bar{\epsilon}_l + \frac{\delta_j}{3} = g_l + \epsilon_{kl}, \qquad i \geq 0.$$

The $g_l + \bar{\epsilon}_l \leq \max\{|g|_\infty, \bar{K}_1|\bar{x}| + \bar{K}_2\} + 1 - \frac{\delta_j}{3}$ and $\gamma_l \geq \frac{\delta_j}{3}$ *case.*

$$(f_i(x_k, U_k^m))_l \leq (f_i(\bar{x}, \bar{U}^m))_l + \frac{\delta_j}{3} \leq g_l + \bar{\epsilon}_l - \gamma_l + \frac{\delta_j}{3} \leq g_l + \bar{\epsilon}_l = g_l + \epsilon_{kl}, \qquad i \geq 0.$$

Moreover, similar to the stable plant case, all the other constraints are satisfied. Hence, for $k \geq N_0$, $(U_k^m, \epsilon_k) \in \bar{\Omega}_m(x_k)$ and $(U_k^m, \epsilon_k) \to (\bar{U}^m, \bar{\epsilon})$.

Similar to the stable plant case, $\bar{\Omega}_m(\cdot)$ is closed and thus continuous on π_m and is uniformly compact near any $\bar{x} \in \pi_m$. Hence, the theorem follows. □

THEOREM 4.5. *U_k^{m*} is uniformly compact on π_m and, thus, is continuous on π_m.*

Proof. Notice that, for any $x_k \in \pi_m$ and $(U_k^m, \epsilon_k) \in \bar{\Omega}_m(x_k)$,

$$\Phi_m = \sum_{i=1}^{m-1} x_{k+i|k}^T R x_{k+i|k} + x_{k+m|k}^T \tilde{V}_s^T \mathcal{R}_s \tilde{V}_s x_{k+m|k} + \sum_{i=0}^{m-1} u_{k+i|k}^T S u_{k+i|k},$$

where $\mathcal{R}_s$ is a unique positive definite solution of the following Lyapunov equation:

$$\mathcal{R}_s = \mathcal{J}_s^T \mathcal{R}_s \mathcal{J}_s + V_s^T R V_s.$$

Hence, the similar arguments to the stable plant case establish that U_k^{m*} is nonempty and single-valued on π_m. Moreover, similar to the stable plant case, it can be shown that U_k^{m*} is uniformly compact on π_m. Hence the theorem follows. □

5. Exponential stability of state feedback model predictive control.

5.1. Stable plants.

THEOREM 5.1. *The closed loop system with* $\mathrm{MPC_m}$ *is globally exponentially stable.*

From the Lyapunov theorem for global exponential stability [34, p. 267], the global exponential stability can be established if we can find a Lyapunov function for which there exist $a, b, c > 0$ such that

$$a|x_k|^2 \leq J_m(x_k) \leq b|x_k|^2,$$

$$\Delta J_m(x_k) \leq -c|x_k|^2.$$

We will show that $J_m(\cdot)$ is such a Lyapunov function for the closed loop system with $\mathrm{MPC_m}$.

We first show that there exists $a > 0$ such that $J_m(x_k) \geq a|x_k|^2$ for all $x_k \in \mathbf{R}^n$. For this, consider the following infinite horizon quadratic optimal control problem:

$$J_{um}(x_k) = \min_{U_k^m} \sum_{i=1}^{\infty} x_{k+i|k}^T R x_{k+i|k} + \sum_{i=0}^{m-1} u_{k+i|k}^T S u_{k+i|k}$$

subject to (4) and

$$u_{k+i|k} = 0, \qquad i = m, m+1, \ldots.$$

It is shown in [25] that

$$J_{um}(x_k) = x_k^T P_{m+1} x_k,$$

where P_{m+1} is a positive definite symmetric matrix which is the solution of the following recursion relation:

$$P_N = R + A^T \left[P_{N-1} - P_{N-1} B (B^T P_{N-1} B + S)^{-1} B^T P_{N-1} \right] A, \quad N > 1,$$

$$P_1 = \sum_{i=1}^{\infty} (A^T)^i R A^i.$$

Hence, it holds that $J_m(x_k) \geq J_{um}(x_k) \geq \lambda_{min}(P_{m+1})|x_k|^2$ for all $x_k \in \mathbf{R}^n$.

We next establish that there exists $b > 0$ such that $J_m(x_k) \leq b|x_k|^2$ for all $x_k \in \mathbf{R}^n$. Suppose $u_{k+i|k} = 0$, $i \geq 0$. Then $x_{k+i|k} = A^i x_k$ and, thus, $(0, \|G\|D|x_k|\mathbf{1})$ is a feasible point for J_m. Thus, it holds that

$$J_m(x_k) \leq \sum_{i=1}^{\infty} x_k^T A^{iT} R A^i x_k + D^2 \|G\|^2 |x_k|^2 \mathbf{1}^T Q \mathbf{1}$$

$$\leq x_k^T A^T T A x_k + D^2 \|G\|^2 |x_k|^2 \mathbf{1}^T Q \mathbf{1} \leq (\lambda_{max}(A^T T A) + D^2 \|G\|^2 \lambda_{max}(Q)\sqrt{q})|x_k|^2.$$

Finally, similar to the result in [38], it can be shown that

$$\Delta J_m(x_k) := J_m(x_k) - J_m(x_{k-1}) \leq -x_k^T R x_k - u_{k-1}^T S u_{k-1} \leq -x_k^T R x_k \leq -\lambda_{max}(R)|x_k|^2.$$

To this end, $J_m(\cdot)$ is a continuous Lyapunov function of the closed loop system with $\mathrm{MPC_m}$, and we have established the global exponential stability of the closed loop system with $\mathrm{MPC_m}$.

Remark 5.1. From Theorem 5.1, $x_k \to 0$ as $k \to 0$. Hence it holds that $u_k \to 0$ as $k \to 0$ since the optimal solution to QP (3) is 0 at $x_k = 0$ and is continuous w.r.t. x_k as established in section 4.

5.2. Marginal or unstable plants.

THEOREM 5.2. *The closed loop system with* $\mathrm{MPC_m}$ *is exponentially stable on* π_m.

To show the exponential stability on π_m of the closed loop system with $\mathrm{MPC_m}$, we will first show that $J_m(\cdot)$ is a Lyapunov function on π_m for the closed loop system

with $\mathrm{MPC_m}$. However, this establishes local exponential stability only. Exploiting the fact that $x_k \in \pi_m$ implies $x_{k+1} \in \pi_m$, we show that an exponentially converging envelope can be found that is valid for all $x_0 \in \pi_m$.

It can be shown similarly to the stable plant case that there exist $a, c > 0$ such that for any $x_k \in \pi_m$,

$$a|x_k|^2 \le J_m(x_k), \qquad \Delta J_m(x_k) \le -c|x_k|^2.$$

However, contrary to the stable plant case, the cost function of J_m with $U_k^m = 0$ is finite only if $x_k \in \mathcal{E}$. Hence, the similar arguments to the stable plant case establish that there exists $b > 0$ such that for any $x_k \in \mathcal{E} \subset \pi_m$,

$$J_m(x_k) \le b|x_k|^2.$$

We now show that such a constant indeed exists for all $x_k \in \pi_m$. Given $x \in \pi_m \setminus \mathcal{E}$, let

$$y := x - P_{\mathcal{E}}x \neq 0,$$

where $P_{\mathcal{E}}x$ denotes the projection of x on $\mathcal{E}$. Then y is perpendicular to $\mathcal{E}$. Let $r_{max} := \max_{x \in \chi_m} |x|$. Then, from Theorem 4.3, it holds that

$$\pi_m \subset \{x \in \mathbf{R}^n : dist(x, \mathcal{E}) \le r_{max}\}.$$

Hence, since $y \neq 0$, there exists $\alpha_x \in [1, \infty)$ such that $\alpha_x x \in \partial\pi_m$. Notice that there exists $(\hat{U}_k^m, \hat{\epsilon}_k) \in \bar{\Omega}_m(\alpha_x x)$ that drives $\alpha_x x$ to a point $z \in \mathcal{E}$ in m steps. Then, from the linearity of the plant, $(\frac{1}{\alpha_x}\hat{U}_k^m, \frac{1}{\alpha_x}\hat{\epsilon}_k) \in \bar{\Omega}_m(x)$ drives x to the point $\frac{1}{\alpha_x} z \in \mathcal{E}$ in m steps. Notice that

$$|\hat{U}_k^m| \le \sqrt{mq}\vartheta.$$

Hence it holds that

$$\left|\frac{1}{\alpha_x}\hat{U}_k^m\right| \le \sqrt{mq}\vartheta\frac{1}{\alpha_x} \le \frac{\sqrt{mq}\vartheta}{r_{min}}|x|,$$

where $r_{min} := \min_{x \in \partial\pi_m} |x|$ is greater than 0 because π_m contains an open neighborhood of the origin in $\mathbf{R}^n$. Hence, for any $x \in \pi_m \setminus \mathcal{E}$, there exists $\tilde{U}_k^m$ such that $\tilde{V}_u x_{k+m|k} = 0$ and

$$|\tilde{U}_k^m| \le \frac{\sqrt{mq}\vartheta}{r_{min}}|x|.$$

Indeed, this property holds for all $x \in \pi_m$ since, for any $x \in \mathcal{E}$, $\tilde{U}_k^m = 0$ results in $\tilde{V}_u x_{k+m|k} = 0$. Notice that

$$\tilde{x}_{k+i|k} = \begin{cases} A^i x_k + \sum_{j=0}^{i-1} A^{i-j-1} B\tilde{u}_{k+j|k}, & i = 1, \ldots, m-1, \\ V_s \mathcal{J}_s^{i-m} \tilde{V}_s A^m x_k + V_s \mathcal{J}_s^{i-m} \tilde{V}_s \sum_{j=0}^{m-1} A^{m-j-1} B\tilde{u}_{k+j|k}, & i = m, m+1, \ldots. \end{cases}$$

Define

$$\tilde{\epsilon}_k := \mathcal{H}\mathcal{A}\left(1 + m\|B\|\frac{\sqrt{mq}\vartheta}{r_{min}}\right)|x_k|\mathbf{1},$$

where

$$\mathcal{H} := \max\{1, \|V_s\|\|\tilde{V}_s\|H\},$$

$$\mathcal{A} := \max_{0\le i\le m} \|A^i\|.$$

Then $(\tilde{U}_k^m, \tilde{\epsilon}_k)$ is a feasible point for $J_m(x_k)$. Hence, for any $x_k \in \pi_m$,

$$J_m(x_k) \le \sum_{i=1}^{m-1} \tilde{x}_{k+i|k}^T R \tilde{x}_{k+i|k} + \sum_{i=0}^{\infty} \tilde{x}_{k+m|k}^T \tilde{V}_s^T \mathcal{J}_s^{iT} V_s^T R V_s \mathcal{J}_s^i \tilde{V}_s \tilde{x}_{k+m|k}$$

$$+ \sum_{i=0}^{m-1} \tilde{u}_{k+i|k}^T S \tilde{u}_{k+i|k} + \tilde{\epsilon}_k^T Q \tilde{\epsilon}_k$$

$$\le \left[\left(m - 1 + \frac{H^2\|V_s\|^2\|\tilde{V}_s\|^2}{1-\nu^2} \right) \mathcal{A}^2 \left(1 + m\|B\| \frac{\sqrt{mq}\vartheta}{r_{min}} \right)^2 \lambda_{max}(R) + \frac{mq\vartheta^2}{r_{min}^2} \lambda_{max}(S) \right.$$

$$\left. + \mathcal{H}^2\mathcal{A}^2 \left(1 + m\|B\| \frac{\sqrt{mq}\vartheta}{r_{min}} \right)^2 p\lambda_{max}(Q) \right] |x_k|^2.$$

To this end, on π_m, $J_m(\cdot)$ is a continuous Lyapunov function of the closed loop system with $\mathrm{MPC_m}$ for which there exist $a, b, c > 0$ such that for all $x_k \in \pi_m$,

$$a|x_k|^2 \le J_m(x_k) \le b|x_k|^2,$$

$$\Delta J_m(x_k) \le -c|x_k|^2.$$

Hence, we have established the exponential stability of the closed loop system with $\mathrm{MPC_m}$. It is now well known that $x_k \in \pi_m$ implies $x_{k+1} \in \pi_m$ because, in the absence of disturbances, $x_{k+1} = x_{k+1|k}^* \in \pi_{m-1} \subset \pi_m$.[5] Hence, following the standard procedure to find the exponentially converging envelope for the trajectory from a, b, c information [34], there exist $\sigma > 0$ and $\rho \in (0, 1)$ such that for all $x_0 \in \pi_m$,

$$|x_k| \le \sigma\rho^k |x_0|.$$

Notice that π_m is the largest possible domain of attraction of the closed loop system because the control law is not defined outside π_m from Theorem 4.3. To this end, we have established Theorem 5.2.

Remark 5.2. Similar to Remark 5.1, it holds that $U_k^{m*} \to 0$ as $k \to \infty$.

We now discuss what kind of constrained stabilization is achieved by the mixed constrained state feedback model predictive controller. In [33], Tsirukis and Morari showed that if all unstable eigenvalues of A are in the closed unit disk, for any $x \in \mathbf{R}^n$ there exists M such that $x \in \pi_m$ for all $m \ge M$. Hence, a marginal system subject to input constraints is semiglobally exponentially stabilized by the mixed constrained state feedback model predictive control law; for any bounded subset W of $\mathbf{R}^n$, there exists $m < \infty$ such that the closed loop with $\mathrm{MPC_m}$ is exponentially stable on W. Moreover, it is shown in [5] that π_m converges to π_∞ as $m \to \infty$. Hence, any unstable system can be exponentially stabilized by the mixed constrained model predictive control law on any compact subset of π_∞.

[5] $\{u_{k+1|k}^*, \ldots, u_{k-m+1|k}^*\}$ forms a possible $m-1$ input sequence for $x_{k+1|k}^* \in \pi_{m-1}$ to be true.

6. Exponential stability of output feedback model predictive control. In many cases, the states of a physical system are not all measurable. Under these circumstances, a state feedback controller is often cascaded with an asymptotic observer to form an output feedback control strategy. In this section, we will investigate the exponential stability properties of the output feedback closed loop system obtained by combining the mixed constrained state feedback model predictive controller and an asymptotic observer.

Suppose the state of the plant is estimated by an asymptotic observer:

$$\hat{x}_{k+1} = A\hat{x}_k + Bu_k + L\{C[A\hat{x}_k + Bu_k] - y_{k+1}\},$$

where $\hat{x}_k$ denotes the state estimated at the sampling time k. Notice that such an observer always exists under Assumption 3.1. Then the dynamics of the error $e_k := \hat{x}_k - x_k$ becomes

$$e_{k+1} = (I + LC)Ae_k.$$

Then mixed constrained output feedback model predictive control with horizon m ($\mathrm{OFMPC_m}$) is a nonlinear output feedback control law whose control output at the kth sampling time, u_k, is the first q elements $u_{k|k}^{m*}$ of the optimal solution U_k^{m*} of the quadratic optimization problem $J_m(\hat{x}_k)$.

For notational convenience, the state evolution constraints (4) in $J_m(\cdot)$ are replaced with

$$\hat{x}_{k+1|k+1} = \hat{x}_{k+1} = A\hat{x}_k + Bu_k + L\{C[A\hat{x}_k + Bu_k] - y_{k+1}\},$$

$$\hat{x}_{k+i+1|k+1} = A\hat{x}_{k+i|k+1} + Bu_{k+i|k+1}, \quad i \geq 1,$$

where $\hat{x}_k = \hat{x}_{k|k}^{m*}$ and $u_k = u_{k|k}^{m*}$. Define

$$\xi_{k+i|k+1} := \hat{x}_{k+i|k+1} - \hat{x}_{k+i|k}^{m*} \qquad \forall i \geq 1.$$

Then

$$\xi_{k+i+1|k+1} = A\xi_{k+i|k+1} + Bv_{k+i|k+1}, \quad i \geq 1, \tag{9}$$

$$\xi_{k+1|k+1} = LCAe_k,$$

where $v_{k+i|k+1} := u_{k+i|k+1} - u_{k+i|k}^{m*}$.

It can be easily shown that $\hat{V}(\hat{x}_k, e_k)$ is a Lyapunov function for the closed loop system of the estimated state and the observer error on $\mathcal{R}$ iff $V(\hat{x}_k, x_k) := \hat{V}(\hat{x}_k, \hat{x}_k - x_k)$ is a Lyapunov function for the closed loop system of the estimated state and the actual state on $\mathcal{P} := \left[\begin{smallmatrix} I & 0 \\ I & -I \end{smallmatrix}\right]\mathcal{R}$. Hence, we will construct a Lyapunov function for the closed loop system of the estimated state and the observer error in what follows.

6.1. Stable plants.

THEOREM 6.1. *If A and $(I + LC)A$ are stable, then the closed loop system with* $\mathrm{OFMPC_m}$ *is globally exponentially stable.*

To establish the global exponential stability of the closed loop system with $\mathrm{OFMPC_m}$, consider the following Lyapunov function candidate:

$$\hat{V}_m(\hat{x}_k, e_k) := J_m(\hat{x}_k) + \sum_{i=0}^{\infty} e_{k+i|k}^T O e_{k+i|k} = J_m(\hat{x}_k) + e_k^T \mathcal{O} e_k,$$

where

$$e_{k+i+1|k} = (I + LC)Ae_{k+i|k}, \qquad e_{k|k} = e_k,$$

O is a positive definite $n \times n$ matrix, and $\mathcal{O}$ is a unique positive definite solution of the following Lyapunov equation [3, p. 215]:

$$\mathcal{O} = A^T(I + LC)^T\mathcal{O}(I + LC)A + O.$$

It can be easily shown that $\hat{V}_m(\cdot, \cdot)$ is continuous and there exist $a, b > 0$ such that

$$a(|\hat{x}_k|^2 + |e_k|^2) \leq \hat{V}_m(\hat{x}_k, e_k) \leq b(|\hat{x}_k|^2 + |e_k|^2)$$

because both $J_m(\hat{x}_k)$ and $e_k^T\mathcal{O}e_k$ have such characteristics.

Now from the Lyapunov stability theorem [34, p. 267], the global exponential stability follows if there exists $c > 0$ such that

$$\Delta\hat{V}_m(\hat{x}_k, e_k) \leq -c(|\hat{x}_k|^2 + |e_k|^2).$$

To show this, we first establish a series of lemmas.

Let F denote the feedback gain of unconstrained model predictive control. Associated with F, consider the maximal output admissible set O_∞ [13]:

$$O_\infty := \left\{ x \in \mathbf{R}^n \,\middle|\, (A - BF)^k x = \bar{A}^k x \in Y, k \geq 0 \right\},$$

where

$$Y := \left\{ x \in \mathbf{R}^n \,\middle|\, Gx \leq g, u^{min} \leq -Fx \leq u^{max} \right\}.$$

From Assumption 3.2, Y contains a neighborhood of the origin. Hence, the maximal output admissible set also contains a neighborhood of the origin. This implies U_k^{m*} is linear w.r.t. $\hat{x}_k$ in the neighborhood of the origin. Thus, U_k^{m*} is Lipschitz continuous in the neighborhood of the origin. Let γ be the radius of a closed ball contained in the maximal output admissible set and κ be the smallest Lipschitz constant for U_k^{m*} on $B(\gamma)$. Then, for all $\hat{x}_k \in B(\gamma)$,

$$|U_k^{m*}| \leq \kappa|\hat{x}_k|.$$

Moreover, since U_k^{m*} in J_m is always contained in a compact set, there exists a constant $\mathcal{K} > 0$ such that for all $\hat{x}_k \in \mathbf{R}^n$,

$$|U_k^{m*}| \leq \mathcal{K}.$$

To this end, it holds that for all $\hat{x}_k \in \mathbf{R}^n$,

$$|U_k^{m*}| \leq K|\hat{x}_k|,$$

where $K := \max\{\kappa, \frac{\mathcal{K}}{\gamma}\}$. To this end, it is trivial to show the following lemma.

LEMMA 6.1.

$$|\hat{x}_{k+1|k}^{m*}| \leq (\|A\| + \|B\|K)\,|\hat{x}_k|.$$

LEMMA 6.2. *Suppose A and $(I + LC)A$ are stable and $v_{k+i|k+1} = 0$ for all $i \geq 1$. Define $\eta_k := \max_{i \geq 1} |G\xi_{k+i|k+1}|_\infty \mathbf{1}$. Then, the following hold:*

(i) $|\xi_{k+i|k+1}| \le D\rho^{i-1}\|LCA\||e_k|$.
(ii) $|\eta_k| \le \sqrt{p}\|G\|D\|LCA\||e_k|$.

Proof. (i) is obvious because $\xi_{k+i|k+1} = A^{i-1}LCAe_k$. Then (ii) follows from the following inequalities:

$$\begin{aligned}|\eta_k| &= \max_{i\ge 1}\sqrt{p}|G\xi_{k+i|k+1}|_\infty \le \max_{i\ge 1}\sqrt{p}\|G\|D\rho^{i-1}\|LCA\||e_k| \\ &= \sqrt{p}\|G\|D\|LCA\||e_k|. \qquad \square\end{aligned}$$

LEMMA 6.3. *Suppose Z is a positive semidefinite $n \times n$ matrix and a, b are n-dimensional vectors. Then given $\delta > 0$,*

$$(a+b)^T Z(a+b) \le (1+\delta)a^T Za + \left(1+\frac{1}{\delta}\right) b^T Zb.$$

Proof.

$$0 \le \left(a - \frac{b}{\delta}\right)^T Z\left(a - \frac{b}{\delta}\right) = a^T Za + \frac{1}{\delta^2}b^T Zb - \frac{2}{\delta}a^T Zb$$

or

$$2a^T Zb \le \delta a^T Za + \frac{1}{\delta}b^T Zb.$$

From this inequality,

$$(a+b)^T Z(a+b) = a^T Za + b^T Zb + 2a^T Zb \le (1+\delta)a^T Za + \left(1+\frac{1}{\delta}\right) b^T Zb. \qquad \square$$

THEOREM 6.2. *There exists $c > 0$ such that*

$$\Delta\hat{V}_m(\hat{x}_k, e_k) \le -c(|\hat{x}_k|^2 + |e_k|^2).$$

Proof. Let

$$u_{k+i|k+1} = u^{m*}_{k+i|k}, \quad i = 1, \dots, m,$$

$$\epsilon_{k+1} = \epsilon^{m*}_k + \eta_k,$$

where $u^{m*}_{k+m|k} = 0$. Then $(U^m_{k+1}, \epsilon_{k+1})$ is a feasible point for $J_m(\hat{x}_{k+1})$. Let

$$\mathcal{U}_k := \sum_{i=1}^{m-1} u^{m*T}_{k+i|k} S u^{m*}_{k+i|k},$$

$$\hat{V}_k := \hat{x}^{m*T}_{k+1|k} R\hat{x}^{m*}_{k+1|k} + u^{m*T}_{k|k} S u^{m*}_{k|k}.$$

Then,

$$J_m(\hat{x}_{k+1}) \le \sum_{i=2}^{\infty} \hat{x}^T_{k+i|k+1} R\hat{x}_{k+i|k+1} + \mathcal{U}_k + \epsilon^T_{k+1} Q\epsilon_{k+1} \tag{10}$$

$$= \sum_{i=2}^{\infty} (\hat{x}^{m*}_{k+i|k} + \xi_{k+i|k+1})^T R(\hat{x}^{m*}_{k+i|k} + \xi_{k+i|k+1}) + \mathcal{U}_k + (\epsilon^{m*T}_k + \eta_k)Q(\epsilon^{m*}_k + \eta_k)$$

$$\leq (1+\delta)\left(\sum_{i=2}^{\infty} \hat{x}^{m*T}_{k+i|k} R \hat{x}^{m*}_{k+i|k} + \mathcal{U}_k + \epsilon^{m*T}_k Q \epsilon^{m*}_k\right)$$

$$+ \left(1 + \frac{1}{\delta}\right)\left(\sum_{i=2}^{\infty} \xi^T_{k+i|k+1} R \xi_{k+i|k+1} + \eta^T_k Q \eta_k\right)$$

$$\leq J_m(\hat{x}_k) - \hat{V}_k + \delta J_m(\hat{x}_k) + \left(1 + \frac{1}{\delta}\right)\left(\sum_{i=2}^{\infty} \lambda_{max}(R)|\xi_{k+i|k+1}|^2 + \lambda_{max}(Q)|\eta_k|^2\right),$$

where δ will be chosen later. As shown in section 5, there exists $\mathcal{F} > 0$ such that $J_m(\hat{x}_k) \leq \mathcal{F}|\hat{x}_k|^2$ for all $\hat{x}_k \in \mathbf{R}^n$. Hence,

$$\Delta J_m(\hat{x}_k) := J_m(\hat{x}_{k+1}) - J_m(\hat{x}_k)$$

$$\leq -\hat{V}_k + \delta J_m(\hat{x}_k) + \left(1 + \frac{1}{\delta}\right)\left(\sum_{i=2}^{\infty} \lambda_{max}(R)|\xi_{k+i|k+1}|^2 + \lambda_{max}(Q)|\eta_k|^2\right)$$

$$\leq -\lambda_{min}(R)|\hat{x}^{m*}_{k+1|k}|^2 + \delta\mathcal{F}|\hat{x}_k|^2 + \left(1 + \frac{1}{\delta}\right)\left(\sum_{i=2}^{\infty} \lambda_{max}(R)|\xi_{k+i|k+1}|^2 + \lambda_{max}(Q)|\eta_k|^2\right)$$

$$\leq -\alpha|\hat{x}_k|^2 + \beta|e_k|^2,$$

where

$$\alpha := \lambda_{min}(R)\left[\|A\| + \|B\|K\right]^2 - \delta\mathcal{F},$$

$$\beta := \left(1 + \frac{1}{\delta}\right)\left[\lambda_{max}(R)D^2\frac{\rho^2}{1-\rho^2}\|LCA\|^2 + \lambda_{max}(Q)p\|G\|^2D^2\|LCA\|^2\right].$$

This implies

$$\Delta\hat{V}_m(\hat{x}_k, e_k) \leq -\alpha|\hat{x}_k|^2 + \beta|e_k|^2 - \lambda_{min}(O)|e_k|^2.$$

Thus, there exists $c > 0$ such that

$$\Delta\hat{V}_m(\hat{x}_k, e_k) \leq -c(|\hat{x}_k|^2 + |e_k|^2)$$

if δ and O are chosen such that $\delta < \frac{\lambda_{min}(R)[\|A\|+\|B\|K]^2}{\mathcal{F}}$ and $\beta < \lambda_{min}(O)$. ∎

To this end, $\hat{V}_m(\cdot,\cdot)$ is a continuous Lyapunov function for the closed loop system with OFMPC$_\mathrm{m}$ for which there exist $a, b, c > 0$ such that

$$a(|\hat{x}_k|^2 + |e_k|^2) \leq \hat{V}_m(\hat{x}_k, e_k) \leq b(|\hat{x}_k|^2 + |e_k|^2),$$

$$\Delta\hat{V}_m(\hat{x}_k, e_k) \leq -c(|\hat{x}_k|^2 + |e_k|^2).$$

Hence, we established Theorem 6.1.

Remark 6.1. Similar to Remark 5.1, it holds that $u_k \to 0$ as $k \to 0$.

6.2. Marginal or unstable plants. For the marginal or unstable plant case, the $\hat{x}_k$ sequence may leave π_m due to the estimation error e_k although $x_k \in \pi_m$. Clearly, J_m is well-posed iff $\hat{x}_k \in \pi_m$. Hence, a trajectory that starts from an initial condition $(\hat{x}_0, e_0)$ is well-posed iff $\hat{x}_k \in \pi_m$ for all k. Suppose $\hat{x}_k \in \pi_m$. Then it holds that $\hat{x}_{k+1} \in \pi_m$ iff $(\hat{x}_k, e_k) \in \Sigma$, where

$$\Sigma := \left\{ (\hat{x}_k, e_k) \in \mathbf{R}^{2n} : \exists V_{k+1}^m \in \Pi_k^m(U_k^{m*}), \right.$$
$$\left. \tilde{V}_u \xi_{k+m+1|k+1} = \tilde{V}_u A^m LCAe_k + \tilde{V}_u \sum_{i=1}^{m} A^{m-i} B v_{k+i|k+1} = 0 \right\},$$

$$V_{k+1}^m := [v_{k+1|k+1} \cdots v_{k+m|k+1}],$$

$$\Pi_k^m(U_k^{m*}) := \{V_{k+1}^m \in \mathbf{R}^{mq} : u^{min} - u_{k+i|k}^{m*} \le v_{k+i|k+1} \le u^{max} - u_{k+i|k}^{m*}, \;\; i = 1, \ldots, m\}.$$

This well-posedness requirement sets up the largest possible stabilizable region with $\mathrm{MPC_m}$. However, the quantification of Σ is not possible in general and is an open question. Notice that $\Sigma \subset \pi_m \times \mathbf{R}^n$. Define

$$\mathcal{X}_m(U_k^{m*}) := \left\{ w \in \mathbf{R}^n : A^m w + \sum_{i=1}^{m} A^{m-i} B v_{k+i|k+1} = 0, \right.$$
$$\left. u^{min} - u_{k+i|k}^{m*} \le v_{k+i|k+1} \le u^{max} - u_{k+i|k}^{m*}, \;\; i = 1, \ldots, m \right\}.$$

Consider the null space of A^m:

$$\mathcal{N}(A^m) = \{\varpi : A^m \varpi = 0\}.$$

Let

$$\mathcal{Q} := \{\varphi : \varphi = w - P_{\mathcal{N}(A^m)} w, \; w \in \mathcal{X}_m(U_k^{m*})\}.$$

Hence given $w \in \mathcal{X}_m(U_k^{m*})$, there exist $\varpi \in \mathcal{N}(A^m)$, $\varphi \in \mathcal{Q}$ such that

$$w = \varpi + \varphi.$$

We now partition A as follows:

$$A = V \mathcal{J} V^{-1} = [V_z V_n] \begin{bmatrix} \mathcal{J}_z & 0 \\ 0 & \mathcal{J}_n \end{bmatrix} \begin{bmatrix} \tilde{V}_z \\ \tilde{V}_n \end{bmatrix},$$

where $\mathcal{J}_z$ contains the zero eigenvalues of A and $\mathcal{J}_n$ contains the rest. Notice that $\mathcal{J}_z$ is nilpotent and $\mathcal{J}_n$ is invertible. In all the practical situations, m is greater than the number of rows of the square matrix $\mathcal{J}_z$. Hence we assume $J_z^m = 0$ in the rest of this paper. Then

$$0 = A^m \varpi = V_n \mathcal{J}_n^m \tilde{V}_n \varpi$$

or equivalently

$$0 = \begin{bmatrix} \tilde{V}_z \\ \tilde{V}_n \end{bmatrix} A^m \varpi = \begin{bmatrix} 0 \\ \mathcal{J}_n^m \tilde{V}_n \varpi \end{bmatrix}.$$

Hence,

$$\mathcal{N}(A^m) = \{\varpi : \tilde{V}_n \varpi = 0\}. \tag{11}$$

Clearly,

$$0 = A^m w + \sum_{i=1}^{m} A^{m-i} B v_{k+i|k+1}$$

is equivalent to

$$\begin{aligned} 0 &= \begin{bmatrix} \tilde{V}_z \\ \tilde{V}_n \end{bmatrix} A^m w + \begin{bmatrix} \tilde{V}_z \\ \tilde{V}_n \end{bmatrix} \sum_{i=1}^{m} A^{m-i} B v_{k+i|k+1} \\ &= \begin{bmatrix} 0 \\ \mathcal{J}_n^m \tilde{V}_n w \end{bmatrix} + \sum_{i=1}^{m} \begin{bmatrix} \mathcal{J}_z^{m-i} \tilde{V}_z B v_{k+i|k+1} \\ \mathcal{J}_n^{m-i} \tilde{V}_n B v_{k+i|k+1} \end{bmatrix}. \end{aligned}$$

Moreover, since $\varphi = w - P_{\mathcal{N}(A^m)} w$ is always in the span of the transposes of rows of $\tilde{V}_n$ from (11),

$$\varphi = \tilde{V}_n^T \varepsilon.$$

Hence,

$$\begin{aligned} \mathcal{Q} = \Bigg\{ & \tilde{V}_n^T \varepsilon : \mathcal{J}_n^m \tilde{V}_n \tilde{V}_n^T \varepsilon + \sum_{i=1}^{m} \mathcal{J}_n^{m-i} \tilde{V}_n B v_{k+i|k+1} = 0, \quad \sum_{i=1}^{m} \mathcal{J}_z^{m-i} \tilde{V}_z B v_{k+i|k+1} = 0, \\ & u^{min} - u^{m*}_{k+i|k} \le v_{k+i|k+1} \le u^{max} - u^{m*}_{k+i|k}, \ \ i = 1, \dots, m \Bigg\} \\ = \{ & \tilde{V}_n^T \varepsilon : \varepsilon \in \mathcal{Q}_n \}, \end{aligned}$$

where

$$\begin{aligned} \mathcal{Q}_n = \Bigg\{ & \varepsilon : \varepsilon = (\tilde{V}_n \tilde{V}_n^T)^{-1} \left(-\mathcal{J}_n^{-1} \tilde{V}_n B v_{k+1|k+1} - \cdots - \mathcal{J}_n^{-m} \tilde{V}_n B v_{k+m|k+1} \right), \\ & \sum_{i=1}^{m} \mathcal{J}_z^{m-i} \tilde{V}_z B v_{k+i|k+1} = 0, \\ & u^{min} - u^{m*}_{k+i|k} \le v_{k+i|k+1} \le u^{max} - u^{m*}_{k+i|k}, \ \ i = 1, \dots, m \Bigg\}. \end{aligned}$$

Let

$$\begin{aligned} \mathcal{L}_k^m (U_k^{m*}) := \Bigg\{ & V_{k+1}^m \in \mathbf{R}^{mq} : \sum_{i=1}^{m} \mathcal{J}_z^{m-i} \tilde{V}_z B v_{k+i|k+1} = 0, \\ & u^{min} - u^{m*}_{k+i|k} \le v_{k+i|k+1} \le u^{max} - u^{m*}_{k+i|k}, \ \ i = 1, \dots, m \Bigg\}. \end{aligned}$$

Since $\mathcal{Q}_n$ is the set of all points that are linear combinations of $-(\tilde{V}_n\tilde{V}_n^T)^{-1}\mathcal{J}_n^{-i-1}\tilde{V}_nB$, $i=0,\dots,m-1$, whose coefficients are in $\mathcal{L}_k^m(U_k^{m*})$, it is a polyhedron. Let $\{w_i\}_{i=1}^{p_m}$ be the set of all vertices of $\mathcal{Q}_n$. For each i, w_i can be represented as

$$w_i = -\sum_{j=0}^{m-1}(\tilde{V}_n\tilde{V}_n^T)^{-1}\mathcal{J}_n^{-i-1}\tilde{V}_nBv_i(j),$$

where $v_i(j)\in\mathcal{L}_k^m(U_k^{m*})$.

THEOREM 6.3. *There exists a positive constant $\mathcal{Z}$ such that for all $y\in\mathcal{X}_m(U_k^{m*})$, there exists $\tilde{V}_{k+1}^m\in\Pi_k^m(U_k^{m*})$ for which*

$$|\tilde{V}_{k+1}^m|\le\mathcal{Z}|y|$$

and

$$A^m y+\sum_{i=1}^{m}A^{i-1}Bv_{k+i|k+1}=0.$$

Proof. Notice that

$$y=\varpi+\tilde{V}_n^T\varepsilon,\qquad \varpi\in\mathcal{N}(A^m),\qquad \varepsilon\in\mathcal{Q}_n,$$

and, thus,

$$|\varepsilon|=|V_n^T\tilde{V}_n^T\varepsilon|\le\|V_n^T\||\tilde{V}_n^T\varepsilon|\le\|V_n^T\||y|.$$

Notice that the last inequality follows from the fact that the angle between ϖ and $\tilde{V}_n^T\varepsilon$ is right. Hence, the above claim follows if there exists a positive constant $\mathcal{Z}_n$ such that for all $\varepsilon\in\mathcal{Q}_n$, there exists $\tilde{V}_{k+1}^m\in\mathcal{L}_k^m(U_k^{m*})$ for which

$$|\tilde{V}_{k+1}^m|\le\mathcal{Z}_n|\varepsilon|$$

and

$$\varepsilon=-\sum_{j=0}^{m-1}(\tilde{V}_n\tilde{V}_n^T)^{-1}\mathcal{J}_n^{-i-1}\tilde{V}_nBv_i(j).$$

Suppose $\varepsilon\in\mathcal{Q}_n$. Then there exists a face,[6] S_j, of the polyhedron $\mathcal{Q}_n$ such that ε is contained in the polyhedral sector defined by the origin and S_j. Let n_1 denote the number of vertices of S_j. Then there exists a unique set of nonnegative real numbers $a_{\varepsilon l}^j$'s for which $\sum_{l=1}^{n_1}a_{\varepsilon l}^j\le 1$ such that

$$\varepsilon=\sum_{l=1}^{n_1}a_{\varepsilon l}^j w_l^j=[w_1^j\cdots w_{n_1}^j]\begin{bmatrix}a_{\varepsilon 1}^j\\ \vdots\\ a_{\varepsilon n_1}^j\end{bmatrix},$$

where $\{w_l^j\}$ is the set of all vertices of S_j. Now define

$$\hat{v}_{k+i|k+1}:=\sum_{l=1}^{n_1}a_{\varepsilon l}^j v_l^j(i-1)=[v_1^j(i-1)\cdots v_{n_1}^j(i-1)][w_1^j\cdots w_{n_1}^j]^{\dagger}\varepsilon,$$

[6] In this paper, a face is a simplice that is obtained by possibly dividing a face into simplices.

where $v_l^j(j) \in \mathcal{L}_k^m(U_k^{m*})$ such that

$$w_l^j = -\sum_{i=0}^{m-1} (\tilde{V}_n \tilde{V}_n^T)^{-1} \mathcal{J}_n^{-i-1} \tilde{V}_n B v_l^j(i),$$

and $M^\dagger$ denotes the pseudo-inverse of M. Notice that

$$[w_1^j \cdots w_{n_1}^j]$$

$$= \left[-\sum_{i=0}^{m-1} (\tilde{V}_n \tilde{V}_n^T)^{-1} \mathcal{J}_n^{-i-1} \tilde{V}_n B v_1^j(i) \;\; \cdots \;\; -\sum_{i=0}^{m-1} (\tilde{V}_n \tilde{V}_n^T)^{-1} \mathcal{J}_n^{-i-1} \tilde{V}_n B v_{n_1}^j(i) \right]$$

$$= -(\tilde{V}_n \tilde{V}_n^T)^{-1} \left[\mathcal{J}_n^{-1} \tilde{V}_n B \cdots \mathcal{J}_n^{-m} \tilde{V}_n B \right] \begin{bmatrix} v_1^j(0) & \cdots & v_{n_1}^j(0) \\ \vdots & \ddots & \vdots \\ v_1^j(m-1) & \cdots & v_{n_1}^j(m-1) \end{bmatrix}.$$

Hence,

$$\tilde{V}_{k+1}^m := \Gamma_j \varepsilon \in \mathcal{L}_k^m(U_k^{m*}),$$

where Γ_j is a linear mapping from $\mathcal{Q}_n$ to $\mathcal{L}_k^m(U_k^{m*})$, defined by

$$\Gamma_j := -\begin{bmatrix} v_1^j(0) & \cdots & v_{n_1}^j(0) \\ \vdots & \ddots & \vdots \\ v_1^j(m-1) & \cdots & v_{n_1}^j(m-1) \end{bmatrix}$$

$$\times \left\{ (\tilde{V}_n \tilde{V}_n^T)^{-1} \left[\mathcal{J}_n^{-1} \tilde{V}_n B \cdots \mathcal{J}_n^{-m} \tilde{V}_n B \right] \begin{bmatrix} v_1^j(0) & \cdots & v_{n_1}^j(0) \\ \vdots & \ddots & \vdots \\ v_1^j(m-1) & \cdots & v_{n_1}^j(m-1) \end{bmatrix} \right\}^\dagger.$$

This implies

$$|\hat{V}_{k+1}^m| \le \mathcal{Z}_n |\varepsilon|,$$

where $\mathcal{Z}_n := \max_j \|\Gamma_j\|$. Moreover,

$$\varepsilon = \sum_{l=1}^{n_1} a_{\varepsilon l}^j w_l^j = \sum_{l=1}^{n_1} -a_{\varepsilon l}^j \sum_{i=0}^{m-1} (\tilde{V}_n \tilde{V}_n^T)^{-1} \mathcal{J}_n^{-i-1} \tilde{V}_n B v_j^l(i)$$

$$= \sum_{i=0}^{m-1} -(\tilde{V}_n \tilde{V}_n^T)^{-1} \mathcal{J}_n^{-i-1} \tilde{V}_n B \left(\sum_{l=1}^{n_1} a_{\varepsilon l}^j v_j^l(i) \right)$$

$$= \sum_{i=0}^{m-1} -(\tilde{V}_n \tilde{V}_n^T)^{-1} \mathcal{J}_n^{-i-1} \tilde{V}_n B \hat{v}_{k+i|k+1}.$$

Hence, the theorem follows. □

Similar to Theorem 4.3, one can show the following theorem.

THEOREM 6.4. *There exists* $V_{k+1}^m \in \Pi_k^m(U_k^{m*})$ *such that* $\tilde{V}_u \xi_{k+m|k} = 0$ *iff*

$$\xi_{k+1|k+1} = LCAe_k \in \mathcal{M}_m(U_k^{m*}),$$

where

$$\mathcal{M}_m(U_k^{m*}) := \{w \in \mathbf{R}^n : w = y + z,\ \ y \in \mathcal{X}_m(U_k^{m*}),\ z \in \mathcal{E}\}.$$

Clearly, $\mathcal{M}_m(U_k^{m*})$ contains a nonempty interior in $\mathbf{R}^n$ for sufficiently large m.

Now we can easily establish the following theorem.

THEOREM 6.5. *Given* e_k *such that* $LCAe_k \in \mathcal{M}_m(U_k^{m*})$, *there exists* $\hat{V}_{k+1}^m \in \Pi_k^m(U_k^{m*})$ *such that*

$$|\hat{V}_{k+1}^m| \le \zeta |e_k|,$$

where $\zeta = \mathcal{Z}\|LCA\|$.

Proof. If $LCAe_k \in \mathcal{E}$, choose $\hat{V}_{k+1}^m = 0 \in \Pi_k^m(U_k^{m*})$. If $LCAe_k \in \mathcal{M}_m(U_k^{m*}) \setminus \mathcal{E}$, let

$$y := LCAe_k - P_{\mathcal{E}} LCAe_k \neq 0.$$

Then y is perpendicular to $\mathcal{E}$ and is contained in $\mathcal{X}_m(U_k^{m*})$. Moreover, it holds that

$$|y| = |LCAe_k - P_{\mathcal{E}} LCAe_k| \le |LCAe_k|.$$

Notice that the inequality follows from the fact that the angle between $LCAe_k$ and $P_{\mathcal{E}} LCAe_k$ is acute. Now choose $\hat{V}_{k+1}^m = \Gamma_j (I - P_{\mathcal{E}}) LCAe_k$. Then it holds that, for all $LCAe_k \in \mathcal{M}_m(U_k^{m*}) \setminus \mathcal{E}$,

$$|\hat{V}_{k+1}^m| \le \mathcal{Z}|y| \le \mathcal{Z}|LCAe_k| \le \mathcal{Z}\|LCA\||e_k|.$$

Hence the theorem follows. □

As mentioned before, a trajectory starting from $(\hat{x}_0, e_0) \in \Sigma$ is well-posed iff $(\hat{x}_k, e_k) \in \Sigma$ for all k. Hence,

$$\mathcal{D} := \{(\hat{x}_0, e_0) \in \Sigma : (\hat{x}_k, e_k) \in \Sigma\ \forall k\}$$

is the largest possible domain of attraction.

THEOREM 6.6. *If* $(I+LC)A$ *is stable, then the closed loop system with* $\mathrm{OFMPC_m}$ *is exponentially stable on* $\mathcal{D}$.

To prove Theorem 6.6, we first show that the following Lyapunov function candidate is indeed a Lyapunov function of the closed loop system on Σ:

$$\hat{V}_m(\hat{x}_k, e_k) := J_m(\hat{x}_k) + \sum_{i=0}^{\infty} e_{k+i|k}^T O e_{k+i|k} = J_m(\hat{x}_k) + e_k^T \mathcal{O} e_k,$$

where O and $\mathcal{O}$ are defined similar to the stable plant case. Then using the fact that $(\hat{x}_k, e_k) \in \mathcal{D}$ implies $(\hat{x}_{k+1}, e_{k+1}) \in \mathcal{D}$, we will show that the exponentially converging envelope is valid for all $(\hat{x}_0, e_0) \in \mathcal{D}$.

As shown in section 4, $J_m(\cdot)$ is continuous on π_m. Hence, $\hat{V}_m(\cdot,\cdot)$ is continuous on $\pi_m \times \mathbf{R}^n$. Now similar to the stable plant case, it is clear that there exists $a > 0$ such that for all $(\hat{x}_k, e_k) \in \pi_m \times \mathbf{R}^n$,

$$a(|\hat{x}_k|^2 + |e_k|^2) \le \hat{V}_m(\hat{x}_k, e_k).$$

Moreover, as shown in section 5, there exists $K > 0$ such that for all $\hat{x}_k \in \pi_m$,

$$|U_k^{m*}| \le K|\hat{x}_k|.$$

Hence, similar to the stable plant case, there exists $b > 0$ such that for all $(\hat{x}_k, e_k) \in \pi_m \times \mathbf{R}^n$,

$$\hat{V}_m(\hat{x}_k, e_k) \le b(|\hat{x}_k|^2 + |e_k|^2)$$

because such constants also exist for $J_m(\cdot)$ and $e_k^T \mathcal{O} e_k$ on π_m and $\mathbf{R}^n$, respectively.

Now the exponential stability of OFMPC$_\mathrm{m}$ follows if there exists $c > 0$ such that for all $(\hat{x}_k, e_k) \in \Sigma$,

$$\Delta\hat{V}_m(\hat{x}_k) \le -c(|\hat{x}_k|^2 + |e_k|^2).$$

To show this, we need the following lemma.

LEMMA 6.4. *Suppose $(I + LC)A$ is stable and $\hat{V}_{k+1}^m = \Gamma_j(I - P_{\mathcal{E}})LCAe_k$. Define*

$$\eta_k := \max_{i \ge 1} |G\xi_{k+i|k+1}|_\infty \mathbf{1},$$

$$\mathcal{V}_k := \sum_{i=1}^{m} \hat{v}_{k+i|k+1}^T S \hat{v}_{k+i|k+1}.$$

Then, it holds that

(i)

$$|\xi_{k+i|k+1}| \le \begin{cases} \mathcal{A}(\|LCA\| + \|B\|\zeta)|e_k| & \text{if } 1 \le i \le m, \\ H\nu^{i-m-1}\|V_s\|\|\tilde{V}_s\|\mathcal{A}(\|LCA\| + \|B\|\zeta)|e_k| & \text{if } i \ge m+1, \end{cases}$$

where $\mathcal{A} = \max_{0 \le i \le m} \|A^i\|$;

(ii) $|\eta_k| \le \sqrt{p}\|G\|H\|V_s\|\|\tilde{V}_s\|\mathcal{A}(\|LCA\| + \|B\|\zeta)|e_k|$;

(iii) $\mathcal{V}_k \le \lambda_{max}(S)\zeta^2|e_k|^2$.

Proof. Notice that the norms of the terms associated with e_k and $\hat{V}_{k+1}^m$ in $\xi_{k+i+1|k+1}$ are bounded by $\mathcal{A}\|LCA\||e_k|$ and $\mathcal{A}\|B\|\zeta|e_k|$ for each $i \in \{1, \ldots, m\}$, whereas they are bounded by $H\nu^{i-m-1}\|V_s\|\|\tilde{V}_s\|\mathcal{A}\|LCA\||e_k|$ and $H\nu^{i-m-1}\|V_s\|\|\tilde{V}_s\|$ $\mathcal{A}\|B\|\zeta|e_k|$ for each $i \ge m+1$. Hence, (i) follows. From (i), it is trivial to show (ii). Finally, (iii) follows from the inequalities

$$\mathcal{V}_k \le \lambda_{max}(S)|\tilde{V}_{k+1}^m|^2 \le \lambda_{max}(S)\zeta^2|e_k|^2. \qquad \square$$

THEOREM 6.7. *There exists $c > 0$ such that for all $(\hat{x}_k, e_k) \in \Sigma$,*

$$\Delta \hat{V}_m(\hat{x}_k, e_k) \leq -c(|\hat{x}_k|^2 + |e_k|^2).$$

Proof. Let

$$u_{k+i|k+1} = u^{m*}_{k+i|k} + \hat{v}_{k+i|k+1}, \qquad i = 1, \ldots, m,$$

$$\epsilon_{k+1} = \epsilon^{m*}_k + \eta_k.$$

Then $(U^m_{k+1}, \epsilon_{k+1})$ is a feasible point for $J_m(\hat{x}_{k+1})$. Suppose $\mathcal{U}_k$ and V_k are defined as before and

$$\mathcal{V}_{k+1} := \sum_{i=1}^{m} \hat{v}^T_{k+i|k+1} S \hat{v}_{k+i|k+1},$$

$$\mathcal{W}_{k+1} := \sum_{i=1}^{m} u^T_{k+i|k+1} S u_{k+i|k+1}.$$

Then,

$$J_m(\hat{x}_{k+1}) \leq \sum_{i=2}^{\infty} \hat{x}^T_{k+i|k+1} R \hat{x}_{k+i|k+1} + \mathcal{W}_{k+1} + \epsilon^T_{k+1} Q \epsilon_{k+1}$$

$$\leq J_m(\hat{x}_k) - \hat{V}_k + \delta J_m(\hat{x}_k) + \left(1 + \frac{1}{\delta}\right) \left(\sum_{i=2}^{\infty} \lambda_{max}(R) |\xi_{k+i|k+1}|^2 + \lambda_{max}(Q) |\eta_k|^2 + \mathcal{V}_{k+1} \right),$$

where δ will be chosen later. Similar to the stable plant case, it is trivial to establish that there exists $\mathcal{F} > 0$ such that $J_m(\hat{x}_k) \leq \mathcal{F}|\hat{x}_k|^2$ for all $\hat{x}_k \in \pi_m$. Hence, it holds that for all $(\hat{x}_k, e_k) \in \Sigma$,

$$\Delta J_m(\hat{x}_k) \leq -V_k + \delta J_m(\hat{x}_k)$$

$$+ \left(1 + \frac{1}{\delta}\right) \left(\sum_{i=2}^{\infty} \lambda_{max}(R) |\xi_{k+i|k+1}|^2 + \lambda_{max}(Q) |\eta_k|^2 + \mathcal{V}_{k+1} \right)$$

$$\leq -\lambda_{min}(R) |\hat{x}^{m*}_{k+1|k}|^2 + \delta \mathcal{F} |\hat{x}_k|^2$$

$$+ \left(1 + \frac{1}{\delta}\right) \left(\sum_{i=2}^{\infty} \lambda_{max}(R) |\xi_{k+i|k+1}|^2 + \lambda_{max}(Q) |\eta_k|^2 + \mathcal{V}_{k+1} \right)$$

$$\leq -\alpha |\hat{x}_k|^2 + \beta |e_k|^2,$$

where

$$\alpha := \lambda_{min}(R) \left[\|A\| + \|B\| K \right] - \delta \mathcal{F},$$

$$\beta := \left(1+\frac{1}{\delta}\right)\lambda_{max}(R)m\mathcal{A}^2(\|LCA\|+\|B\|\zeta)^2$$

$$+\left(1+\frac{1}{\delta}\right)\lambda_{max}(R)H\|V_s\|^2\|\tilde{V}_s\|^2\frac{\nu^2}{1-\nu^2}\mathcal{A}^2(\|LCA\|+\|B\|\zeta)^2$$

$$+\left(1+\frac{1}{\delta}\right)\lambda_{max}(Q)p\|G\|^2H\|V_s\|^2\|\tilde{V}_s\|^2\mathcal{A}^2(\|LCA\|+\|B\|\zeta)^2$$

$$+\left(1+\frac{1}{\delta}\right)\lambda_{max}(S)\zeta^2.$$

This implies for all $(\hat{x}_k, e_k) \in \Sigma$,

$$\Delta\hat{V}_m(\hat{x}_k, e_k) \le -\alpha|\hat{x}_k|^2 + \beta|e_k|^2 - \lambda_{min}(O)|e_k|^2.$$

Thus, there exists $c > 0$ such that for all $(\hat{x}_k, e_k) \in \Sigma$,

$$\Delta\hat{V}_m(\hat{x}_k, e_k) \le -c(|\hat{x}_k|^2 + |e_k|^2)$$

if δ and O are chosen such that $\delta < \frac{\lambda_{min}(R)}{F}$ and $\beta < \lambda_{min}(O)$. □

Since $(\hat{x}_k, e_k) \in \mathcal{D}$ implies $(\hat{x}_{k+1}, e_{k+1}) \in \mathcal{D}$, there exist $\sigma > 0$ and $\rho \in (0,1)$ such that for all $(\hat{x}_0, e_0) \in \mathcal{D}$,

$$|(\hat{x}_k, e_k)| \le \sigma\rho^k|(\hat{x}_0, e_0)|.$$

To this end, we have proven Theorem 6.6.

Remark 6.2. Similar to the comment following Theorem 5.1, it holds that $U_k^{m*} \to 0$ as $k \to \infty$.

The necessary and sufficient condition for constrained stabilizability is $x_k \in \pi_\infty$ for all k. The above claim dictates that $(\hat{x}_0, e_0) \in \mathcal{D}$ implies $x_k \in \pi_\infty$ for all k. However, $x_k \in \pi_m$ for all k is not necessary and there may exist an initial condition $x_0 \in \pi_\infty \setminus \pi_m$ for which there exists e_0 such that (x_0, e_0) is in the domain of attraction of the closed loop system.

7. Comparison with nonconstructive approach. In the previous section, the exponential stability of mixed constrained model predictive control is established by constructing a Lyapunov function of the closed loop system. However, a nonconstructive approach to the asymptotic stability of hard constrained output feedback model predictive control with the state constraint dropping scheme has also been proposed [24], [23] based on the stability theory of perturbed systems [16]. In this section, we comment on this nonconstructive approach to mixed constrained state feedback model predictive control.

When the unperturbed system is not globally stable, the existing stability theory of perturbed systems establishes the local stability of the perturbed system but doesn't give any explicit information on the corresponding domain of attraction. Hence, for the marginal and unstable plant cases, the largest possible domain of attraction cannot be in general identified via the nonconstructive approach. Indeed, such local stability of the closed loop system without explicit quantification is straightforward since, near the origin, the output feedback system is linear.

When a state feedback law is combined with an asymptotic observer, the key requirement for the perturbed system stability theory to be applicable is the Lipschitz

continuity of the state feedback law. The result in [15] is identified as a vehicle to establish the Lipschitz continuity of hard constrained state feedback model predictive control with the state constraint dropping scheme [24]. In the original version of the paper, we have pointed out that the results in [15] are not in general applicable to mixed constrained state feedback model predictive control. The first reason was the requirement of the results in [15] that the gradients of binding constraints must be linearly independent. The example in the appendix was given as the example where this linear independence condition is violated. The second reason was that the result in [15] is concerned with a finite-dimensional QP whereas the QP associated with constrained state feedback model predictive control contains infinite number of constraints. Although all but a finite number of constraints are redundant for a given state [25], [38], the required finite horizon may increase indefinitely with the size of the state. Thus, the applicability of the result in [15] to the global exponential stability of constrained model predictive control for stable plants is not clear. There exists an extension of the result in [15] to infinite-dimensional QPs [11]. However, the result in [11] requires that the linear map in the constraints be surjective and again is not applicable to J_m. From one of the reviews, we found that our first concern has been resolved and the second one has also been pointed out since the submission of this paper. Indeed it was shown in [1] and [29] that the solution of a hard constrained model predictive control with a finite number of constraints is continuous and piecewise affine on a compact set and thus is Lipschitz continuous over the set. Moreover, it is also mentioned in [1] that the same is also true for the mixed constrained case. However, the results in [1] and [29] are only applicable to the problem with a finite number of constraints, and the required global Lipschitz continuity for stable plants hasn't been established yet. We conjecture that the mixed constrained state feedback model predictive control is globally Lipschitz for stable plants. However, this needs to be proven before we discuss the applicability of nonconstructive approach to the stable plant case.

8. Conclusion. In this paper, we established the continuity of mixed constrained model predictive control. Moreover, the stability properties of mixed constrained model predictive control for both state feedback and output feedback cases is addressed as well. The previous efforts in this direction established only the attractivity of the closed loop system for limited class of plants. However, despite the presence of nonlinear elements such as the mixed constrained model predictive controller and the saturation of input, we were able to show that for both state and output feedback cases with any plant, the closed loop system with the mixed constrained model predictive controller possesses the strongest stability property, exponential stability on the largest possible candidate for the domain of attraction, which is the set of all initial conditions for which the mixed constrained model predictive control law is defined throughout the trajectory.

Appendix. Consider the following stable second order plant:

$$x_{k+1} = \begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix} x_k + \begin{bmatrix} 10 \\ 0 \end{bmatrix} \sigma(u_k),$$

where

$$\sigma(u) := \begin{cases} -1 & \text{if } u < -1, \\ u & \text{if } -1 \le u \le 1, \\ 1 & \text{if } u > 1. \end{cases}$$

Suppose the control and the state of the system are desired to satisfy the following inequalities:

$$-1 \le u_k \le 1, \qquad k = 0, 1, \ldots,$$

$$x_k \le \begin{bmatrix} 15 \\ 50 \end{bmatrix}, \qquad k = 0, 1, \ldots.$$

Associated with the plant, consider MPC_2 defined by the following QP:

$$J_2(x_k) = \min_{U_k^m, \epsilon_k} \sum_{i=1}^{\infty} x_{k+i|k}^T x_{k+i|k} + \sum_{i=0}^{1} u_{k+i|k}^T u_{k+i|k} + \epsilon_k^T \epsilon_k$$

subject to

$$x_{k+i+1|k} = \begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix} x_{k+i|k} + \begin{bmatrix} 10 \\ 0 \end{bmatrix} u_{k+i|k}, \qquad x_{k|k} = x_k,$$

$$-1 \le u_{k+i|k} \le 1, \qquad i = 0, 1,$$

$$u_{k+i|k} = 0, \qquad i = 2, \ldots,$$

$$x_{k+i|k} \le \begin{bmatrix} 15 \\ 50 \end{bmatrix} + \epsilon_k, \qquad k = 1, \ldots,$$

$$\epsilon_k \ge 0, \qquad k = 1, \ldots.$$

Suppose

$$x_0 = \begin{bmatrix} 100 \\ 50 \end{bmatrix}.$$

Then the optimal solution of the QP is $u_{0|0}^* = u_{1|0}^* = -1, \epsilon_0 = 0$. Thus the binding constraints are

$$-1 \le u_{0|0}^*,$$

$$-1 \le u_{1|0}^*,$$

$$25 + 10u_{0|0}^* \le 15 + \epsilon_{01},$$

$$50 \le 50 + \epsilon_{02},$$

$$25 + 10u_{1|0}^* \le 15 + \epsilon_{01},$$

$$\epsilon_{01} \ge 0,$$

$$\epsilon_{02} \ge 0.$$

Clearly, the gradients of the binding constraints are not linearly independent.

Acknowledgment. The authors are grateful to the anonymous reviewers for many useful comments in improving the paper.

REFERENCES

[1] A. Bemporad, M. Morari, V. Dua, and E. N. Pistikopoulos, *The explicit linear quadratic regulator for constrained systems*, Automatica J. IFAC, 38 (2002), pp. 3–20.
[2] R. R. Bitmead, M. Gevers, and V. Wertz, *Adaptive Optimal Control*, Prentice–Hall, Englewood Cliffs, NJ, 1990.
[3] F. M. Callier and C. A. Desoer, *Linear System Theory*, Springer-Verlag, New York, 1991.
[4] D. Chmielewski and V. Manousiouthakis, *On constrained infinite-time linear quadratic optimal control*, Systems Control Lett., 29 (1996), pp. 121–129.
[5] J. Choi, *On the stabilization of linear discrete systems subject to input saturation*, Systems Control Lett., 36 (1999), pp. 241–244.
[6] J. Choi, *Constrained exponential stabilizability of marginal and/or unstable linear systems*, Internat. J. Control, 73 (2000), pp. 652–654.
[7] J. Choi, *On the constrained asymptotic stabilizability of unstable linear discrete time systems via linear feedback*, in Proceedings of the 2001 American Control Conference, Vol. 6, IEEE Press, Piscataway, NJ, 2001, pp. 4926–4929.
[8] J. Choi, H. S. Ko, and K. S. Lee, *Constrained linear quadratic optimal control of chemical processes*, Computers and Chemical Engineering, 24 (2000), pp. 823–827.
[9] J. Choi and K. S. Lee, *Mixed constrained infinite horizon linear quadratic optimal control*, Asian J. Control, submitted.
[10] C. R. Cutler and B. L. Ramaker, *Dynamic matrix control - a computer control algorithm*, in Proceedings of the Joint Automated Control Conference, American Automatic Control Council, 1980.
[11] A. L. Dontchev and W. W. Hager, *Lipschitzian stability in nonlinear control and optimization*, SIAM J. Control Optim., 31 (1992), pp. 569–603.
[12] S. E. Dreyfus, *Dynamic Programming and the Calculus of Variations*, Academic Press, New York, 1965.
[13] E. G. Gilbert and K. T. Tan, *Linear systems with state and control constraints: The theory and application of maximal output admissible sets*, IEEE Trans. Automat. Control, 36 (1991), pp. 1008–1020.
[14] P. Gutman and M. Cwikel, *An algorithm to find maximal state constraint sets for discrete-time linear dynamical systems with bounded controls and states*, IEEE Trans. Automat. Control, 32 (1987), pp. 251–254.
[15] W. W. Hager, *Lipschitz continuity for constrained processes*, SIAM J. Control Optim., 17 (1979), pp. 321–338.
[16] A. Halanay, *Quelques questions de la théorie de la stabilité pour les systèmes aux différences finie*, Arch. Ration. Mech. Anal., 12 (1963), pp. 150–154.
[17] W. W. Hogan, *Point-to-set maps in mathematical programming*, SIAM Rev., 15 (1973), pp. 591–603.
[18] S. Keerthi and E. Gilbert, *Optimal infinite-horizon feedback laws for a general class of constrained discrete-time systems: Stability and moving horizon approximations*, J. Optim. Theory Appl., 57 (1988), pp. 265–293.
[19] W. H. Kwon and A. E. Pearson, *A modified quadratic cost problem and feedback stabilization of a linear system*, IEEE Trans. Automat. Control, 22 (1977), pp. 838–842.
[20] W. H. Kwon, A. M. Bruckstein, and T. Kailath, *Stabilizing feedback design via the moving horizon method*, Internat J. Control, 37 (1983), pp. 631–643.
[21] Z. Lin, *Global and Semi-global Control Problems for Linear Systems Subject to Input Saturation and Minimum-Phase Input-Output Linearizable Systems*, Ph.D. dissertation, Washington State University, Pullman, WA, 1994.
[22] Z. Lin and A. Saberi, *Semi-global exponential stabilization of linear discrete-time systems subject to input saturation via linear feedbacks*, Systems Control Lett., 24 (1995), pp. 125–132.
[23] E. S. Meadows and J. Rawlings, *Topics in model predictive control*, in Methods of Model Based Process Control, R. Berber, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1995, pp. 331–347.
[24] K. R. Muske and J. Rawlings, *Linear model predictive control of unstable processes*, J. Process Control, 3 (1993), pp. 85–96.
[25] J. B. Rawlings and K. R. Muske, *The stability of constrained receding horizon control*, IEEE

Trans. Automat. Control, 38 (1993), pp. 1512–1516.
[26] J. Richalet, A. Rault, J. L. Testud, and J. Papon, *Model predictive heuristic control: Application to industrial processes*, Automatica J. IFAC, 14 (1978), pp. 413–428.
[27] P. O. M. Scokaert, J. B. Rawlings, and E. S. Meadows, *Discrete-time stability with perturbations: Application to model predictive control*, Automatica J. IFAC, 33 (1997), pp. 463–470.
[28] P. O. M. Scokaert and J. B. Rawlings, *Constrained linear quadratic regulation*, IEEE Trans. Automat. Control, 43 (1998), pp. 1163–1169.
[29] M. M. Seron, G. C. Goodwin, and J. A. de Don'a, *Geometry of Model Predictive Control for Constrained Linear Systems*, Technical report EE0031, The University of Newcastle, NSW, Australia, 2000.
[30] E. D. Sontag, *An algebraic approach to bounded controllability of linear systems*, Internat. J. Control, 39 (1984), pp. 181–188.
[31] M. Sznaier and M. J. Damborg, *Suboptimal control of linear systems with state and control inequality constraints*, in Proceedings of the 26th IEEE Conference on Decision and Control, IEEE Press, Pisctataway, NJ, 1987, pp. 761–762.
[32] E. Y. Tse and M. Athans, *Adaptive stochastic control for a class of systems*, IEEE Trans. Automat. Control, 17 (1972), pp. 38–52.
[33] A. Tsirukis and M. Morari, *Controller design with actuator constraints*, in Proceedings of the 31st IEEE Conference on Decision and Control, Vol. 3, IEEE Press, Piscataway, NJ, 1992, pp. 2623–2628.
[34] M. Vidyasagar, *Nonlinear Systems Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1993.
[35] Y. Yang, *Global Stabilization of Linear Systems with Bounded Feedback*, Ph.D. dissertation, Rutgers, the State University of New Jersey, New Brunswick, 1993.
[36] Y. Yang, unpublished work, 1993.
[37] E. Zafiriou and H.-W. Chiou, *Output constraint softening for SISO model predictive control*, in Proceedings of the American Control Conference, Vol. 1, IEEE Press, Piscataway, NJ, 1993, pp. 372–376.
[38] A. Zheng and M. Morari, *Stability of model predictive control with mixed constraints*, IEEE Trans. Automat. Control, 40 (1995), pp. 1818–1823.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 871–906

A TWO-LEVEL ENERGY METHOD FOR INDIRECT BOUNDARY OBSERVABILITY AND CONTROLLABILITY OF WEAKLY COUPLED HYPERBOLIC SYSTEMS*

FATIHA ALABAU-BOUSSOUIRA†

Abstract. This work is concerned with the boundary observability of an abstract system of two coupled second order evolution equations, the coupling operator being a compact perturbation of the uncoupled system. We assume that only one of the two components of the unknown is observed. This is indirect observability. We prove that by observing only one component, one can get back a full weakened energy of both components under a compatibility condition linking the operators of each equation and for small coupling. Using the Hilbert uniqueness method, we then establish an indirect exact controllability result. We apply this abstract result to several coupled systems of partial differential equations (wave-wave, coupled elastodynamic systems, Petrowsky-Petrowsky, and wave-Petrowsky systems).

Key words. boundary observability, indirect controllability, hyperbolic systems, abstract linear evolution equations

AMS subject classifications. 34G10, 35B35, 35B37, 35L90, 93D15, 93D20

DOI. 10.1137/S0363012902402608

1. Introduction: Motivations. Let T be a given positive time and Ω be a bounded open set of $\mathbb{R}^N$ with a boundary Γ of class $\mathcal{C}^2$. It is well known that the energy of weak solutions of the wave equation, in Ω,

$$\begin{cases} u_{tt} - \Delta u = 0 \text{ in } \Omega \times (0,T), \\ u(.,0) = u_0(.),\ u_t(.,0) = u_1(.) \text{ in } \Omega, \end{cases} \tag{1.1}$$

when subjected to homogeneous Dirichlet conditions on $\partial\Omega \times (0,T)$, is conserved through time, that is, $E(u(t)) = E(u(0))$ for all $t \geq 0$. Here Δ stands for the Laplacian with respect to the spatial variables and the subscript t stands for the partial derivative with respect to the t-variable. We recall that the energy of the solutions u of the wave equation is defined by

$$E(u(t)) = \frac{1}{2}\int_\Omega (|u_t|^2 + |\nabla u|^2)\, dx.$$

Moreover, these solutions satisfy the following direct inequality (see [25], [26]):

$$\int_0^T \int_\Gamma \left|\frac{\partial u}{\partial \nu}\right|^2 d\gamma\, dt \leq c_2\, E(u(0)), \tag{1.2}$$

where ν stands for the unit outward normal vector to Γ; the positive constant c_2 depends on T and on the geometry of Ω. This result implies, in particular, a hidden regularity result saying that the weak solutions u of the wave equation satisfy, in addition,

$$\frac{\partial u}{\partial \nu} \in L^2(\Gamma \times (0,T)).$$

*Received by the editors February 13, 2002; accepted for publication (in revised form) December 6, 2002; published electronically June 18, 2003.
http://www.siam.org/journals/sicon/42-3/40260.html
†MMAS, Université de Metz et CNRS (UMR 7122), 57045 Metz Cedex 1, France (alabau@poncelet.sciences.univ-metz.fr).

Moreover, if Γ_1 is a part of the boundary which satisfies certain geometric conditions (for instance, it is the exterior boundary of an annulus; see, e.g., [25], [26], [17], and see [6] for more general conditions), then these solutions satisfy the inverse inequality (see [25], [26])

$$\int_0^T \int_{\Gamma_1} \left| \frac{\partial u}{\partial \nu} \right|^2 d\gamma \, dt \geq c_1 \, E(u(0)) \tag{1.3}$$

for $T \geq T_0$, where T_0 is sufficiently large, and for a positive constant c_1 which depends only on T and on the geometry of Ω. This inequality is also called the "observability inequality." In this context, the observed quantity is the L^2-norm of the trace of the solution's normal derivative, and one wants to get back information on the initial state of the solution. In particular, if one observes the same quantities, one wants to make sure that they correspond to the same initial data (unique continuation results). Of course, an inequality such as (1.3) is more precise.

The above results hold for all initial data (u_0, u_1) of finite energy. Moreover, using the Hilbert uniqueness method (HUM), Lions showed that the direct and inverse inequalities lead to the following exact controllability result: for any $T \geq T_0$ and all initial data $(y_0, y_1) \in L^2(\Omega) \times H^{-1}(\Omega)$ there exists a control $v \in L^2(\Gamma_1 \times (0,T))$ such that the solution of

$$\begin{cases} y_{tt} - \Delta y = 0 \text{ in } \Omega \times (0,T), \\ y = v \text{ on } \Sigma_1 = \Gamma_1 \times (0,T), \quad y = 0 \text{ on } \Sigma_0 = (\Gamma - \Gamma_1) \times (0,T), \\ y(.,0) = y_0(.), \quad y_t(.,0) = y_1(.) \text{ in } \Omega \end{cases} \tag{1.4}$$

satisfies, in addition, $y(T,\cdot) = y_t(T,\cdot) = 0$ in Ω; i.e., the control v drives back the system to equilibrium at time T.

In this paper we consider the case of coupled systems. Our goal here is to establish indirect observability estimates for weakly coupled systems. In this case, one observes a single component of the solution on Γ_1, and one wants to get back the energy of both components.

To be more precise, we consider the following weakly coupled system of two wave equations:

$$\begin{cases} u_{1,tt} - \Delta u_1 + \alpha u_2 = 0 \text{ in } \Omega \times (0,T), \\ u_{2,tt} - \Delta u_2 + \alpha u_1 = 0 \text{ in } \Omega \times (0,T), \\ u_1 = 0 \text{ on } \Sigma = \Gamma \times (0,T), \ u_2 = 0 \text{ on } \Sigma, \\ u_i(0) = u_i^0, \ u_{i,t}(0) = u_i^1, \end{cases} \tag{1.5}$$

where α is a coupling parameter. We then wonder if it is possible to get, for sufficiently large time T, the following type of observability inequality:

$$\int_0^T \int_{\Gamma_1} \left| \frac{\partial u_1}{\partial \nu} \right|^2 d\gamma \, dt \geq c \Big(e_1(u_1(0)) + e_2(u_2(0)) \Big), \tag{1.6}$$

where $e_i(u_i(t))$ stands for some energy of the corresponding component of the unknown.

We assume that Ω is a nonempty bounded open set in $\mathbb{R}^N$ having a boundary Γ of class C^2. Moreover, $\{\Gamma_0, \Gamma_1\}$ is a partition of Γ such that $\overline{\Gamma}_0 \cap \overline{\Gamma}_1 = \emptyset$ and x_0 is a

point in $\mathbb{R}^N$ such that $m\cdot\nu\le 0$ on Γ_0 and $m\cdot\nu\ge 0$ on Γ_1, where $m(x)=x-x_0$. We denote by $|\ |$ the L^2-norm on Ω. Then, we prove in this paper the following result.

THEOREM 1.1. *We assume the above hypotheses on* Γ_1 *and* Γ_0. *Then there exists* $\alpha^\star>0$ *such that for all* $0<|\alpha|<\alpha^\star$, *there exists* $T_0=T_0(\alpha)>0$ *such that for all* $T>T_0$ *and all* $U^0=(u_1^0,u_1^1,u_2^0,u_2^1)\in\mathcal{H}=(H_0^1(\Omega)\times L^2(\Omega))^2$ *the solution* (u_1,u_2) *of* (1.5) *satisfies*

$$2\int_0^T\int_{\Gamma_1}\left|\frac{\partial u_1}{\partial\nu}\right|^2 d\gamma\,dt\ge\frac{c_1}{2}\left(|u_1^1|^2+|\nabla u_1^0|^2\right)+\frac{c_3}{2}\left(|u_2^1|^2_{H^{-1}(\Omega)}+|u_2^0|^2\right),\tag{1.7}$$

where the constants c_1, c_3 *are given by*

$$c_1(\alpha,T)=\frac{a_1(T-T_3)}{(1+\alpha T)(1+\alpha T_3)},\qquad c_3(\alpha,T)=\frac{\alpha a_2(T-T_2)(T-T_2^-)}{1+\alpha T},$$

where a_1,a_2 *are constants independent on* α *and* T. *If we denote by* C *a generic positive constant, then* T_0, $|T_2|,|T_2^-|$ *behave as* $C\alpha^{-1}$ *and* T_3 *behaves as* C *as* α *goes to zero.*

Moreover, if the solution of (1.5) *satisfies*

$$\frac{\partial u_1}{\partial\nu}=0\ \text{ on }\Gamma_1\times(0,T),$$

then one has $u_1=u_2=0$ *in* $\Omega\times[0,T]$.

By duality, the above result can be translated into an exact *indirect* controllability result. For this, we consider the system

$$\begin{cases} y_{1,tt}-\Delta y_1+\alpha y_2=0 \quad\text{in}\quad \Omega\times(0,T),\\ y_{2,tt}-\Delta y_2+\alpha y_1=0 \quad\text{in}\quad \Omega\times(0,T),\\ y_1=v \quad\text{on}\quad \Sigma_1=\Gamma_1\times(0,T),\quad y_1=0 \quad\text{on}\quad \Sigma_0=\Gamma_0\times(0,T),\\ y_2=0 \quad\text{on}\quad \Sigma=\Gamma\times(0,T),\\ (y_1,y_{1,t})(0)=(y_1^0,y_1^1),\quad (y_2,y_{2,t})(0)=(y_2^0,y_2^1) \quad\text{on}\quad \Omega. \end{cases}\tag{1.8}$$

We prove the following theorem.

THEOREM 1.2. *Assume the hypotheses of Theorem* 1.1. *Then there exists* $\alpha^\star>0$ *such that for all* $0<|\alpha|<\alpha^\star$, *there exists* $T_0=T_0(\alpha)>0$ *such that for all* $T>T_0$ *and all* $Y^0=(y_1^0,y_1^1,y_2^0,y_2^1)\in L^2(\Omega)\times H^{-1}(\Omega)\times H_0^1(\Omega)\times L^2(\Omega)$, *there exists a control* $v\in L^2([0,T];L^2(\Gamma_1))$ *such that the solution* $Y(t)=(y_1,y_1',y_2,y_2')$ *of* (1.8) *satisfies*

$$y_i(.,T)=\partial_t y_i(.,T)=0\ \text{ in }\Omega\text{ for }i=1,2.$$

For such systems, we remark that several notions of observability (and by duality arguments of controllability) have already been considered. In [25], [26], Lions introduces three notions of observability, namely complete, partial, and simultaneous observability. He proved that such complete and partial observability inequalities can be obtained when one couples a wave equation to a Petrowsky equation, provided that the coupling parameter is sufficiently small. These results have been extended to a larger set of coupling parameters in [18] (case of zero order coupling). In the case of complete observability, one observes both components of the unknown on Γ_1 and one wants to know whether this observation can give back the initial energy of

both components of the solution. More precisely, for system (1.5), this means that one wants an estimate of the form

$$\int_0^T \int_{\Gamma_1} \left|\frac{\partial u_1}{\partial \nu}\right|^2 d\gamma\, dt + \int_0^T \int_{\Gamma_1} \left|\frac{\partial u_2}{\partial \nu}\right|^2 d\gamma\, dt \geq c\left(|u_1^1|^2 + |\nabla u_1^0|^2 + |u_2^1|^2 + |\nabla u_2^0|^2\right).$$

This inequality holds for sufficiently small α (adapting Lions's method given for a wave equation coupled with a Petrowsky one). Note that this inequality has been established in [18] for a set of parameters larger than a single parameter α and with no smallness restrictions on the parameters. In the partial observability case, one observes only a single component on the boundary and one wants to get back only the energy of the first component, while the energy of the second component at the initial time is set equal to zero. More precisely, this means that one wants to establish an inequality for the solutions of system (1.5) of the form

$$\int_0^T \int_{\Gamma_1} \left|\frac{\partial u_1}{\partial \nu}\right|^2 d\gamma\, dt \geq c\left(|u_1^1|^2 + |\nabla u_1^0|^2\right)$$

for initial data $u_2^0 = u_2^1 = 0$ in Ω. Comparing this result to our result, one sees that the inequality given in (1.7) represents a substantial improvement. Another notion of observability, the so-called simultaneous observability, has been introduced in [25], [26]. In this case, one wants to observe simultaneously both components of the system and to get back the total initial energy of the solutions. The purpose of this paper is to consider the case of a fourth observability notion, namely the indirect observability, as was stated before. Since the answer is positive for the system (1.5), then a natural subsequent question of interest is, *Is it possible to obtain a somehow general result for abstract systems of second order evolution equations*? The abstract model we refer to in this paper is

$$\begin{cases} u_1'' + A_1 u_1 + \alpha C u_2 = 0 \quad \text{in } V_1', \\ u_2'' + A_2 u_2 + \alpha C^\star u_1 = 0 \quad \text{in } V_2', \\ (u_1, u_1')(0) = (u_1^0, u_1^1) = U_1^0 \in V_1 \times H, \\ (u_2, u_2')(0) = (u_2^0, u_2^1) = U_2^0 \in V_2 \times H, \end{cases} \tag{1.9}$$

where H, $V_1 \subset H$, and $V_2 \subset H$ are separable Hilbert spaces; A_1, A_2 are coercive self-adjoint unbounded operators in H, whereas the coupling operator C is assumed to be bounded in H; $C^\star$ is the adjoint operator of C; and α is a coupling parameter. The total energy of a solution (u_1, u_2) is defined by

$$E(u_1(t), u_2(t)) = \frac{1}{2}\left(|u_1'(t)|^2 + |u_2'(t)|^2 + |A_1^{1/2} u_1(t)|^2 + |A_2^{1/2} u_2(t)|^2\right) + \alpha(u_1, C u_2),$$

where $|\cdot|$ and $(\cdot)$ denote, respectively, the norm and scalar product in H and $A_i^{1/2}$, $i \in \{1, 2\}$, denotes the usual fractional power of a coercive self-adjoint operator A_i in H (see [28]).

Now the question of interest is, *Is the above full system indirectly observable*? That is, does there exist a time $T_0 > 0$ such that for any T larger than T_0 there exists a positive constant c depending only on T and on Ω such that the inverse inequality for $T \geq T_0$,

$$\int_0^T ||B^\star u_1||_G^2\, dt \geq c\, \widetilde{E}(u_1(0), u_2(0)), \tag{1.10}$$

holds for a certain energy $\widetilde{E}$ and a certain linear observation operator $B^\star$ acting from $D(A_1)$ into a Hilbert space G which involves only the first component u_1 of the solution? Indeed, the energy in (1.10) cannot be the natural one. This is due to the fact that the coupling in (1.9) is compact. One needs couplings which are not compact to hope to obtain the natural energy when observing only one component of the solution.

Of course, we are also interested in the dual exact controllability result, which means that we want to drive back the full coupled physical system to equilibrium by only controlling the first component of the system. Many questions arise then: In which spaces can one hope to obtain such results? If so, can the constants T_0 and c be explicitly given?

Let us also remark that if $\alpha = 0$, then an inequality such as (1.10) does not hold. Hence, the results we are looking for cannot be obtained by a perturbation argument with respect to the case $\alpha = 0$. Let us now tell more about the method for proving such results. Three steps are required. The first step is the estimate given in Lemma 4.1. It gives a hidden property of the system, namely that the second component of the system can be "controlled" in some way by the first one. The coercivity property (2.8) and the restrictive assumption (2.9) are required for this result. Let us also mention that we have already used, in a slightly different way, these properties in [1]. Here, however, it is very important to have an estimate where only the weakened partial energy of the second component is involved, that is, where the hypothesis (2.10) is required. The second step is the obtaining of intermediate estimates, which allows us to compare the quantity to observe to the time integral of the natural partial energy of the first component and to the partial natural energy of the first component and partial weakened energy of the second component at the initial time. Namely, we want to obtain an estimate of the form

$$2\int_0^T ||B^\star u_1||_G^2\,dt \geq \int_0^T e_1(t)\,dt - c\Big[e_1(0) + \tilde{e_2}(0)\Big],$$

where c is an explicit positive constant. Note that the hypotheses (2.12) and (2.13) for the nonhomogeneous uncoupled equation (2.11) are crucial for this second step. Moreover, the fact that the involved constants δ_i and η_i do not depend on time T is essential for the third step. Once the second step is performed, the main difficulty of establishing the final desired estimate clearly appears. How do we balance the lack of information on the second component of the system, since in the above estimate we have a "good" term (which is the time integral of the energy of the first component, together with its energy at the initial time) and a "bad" term (which is the weakened energy of the second component)? This is the third step. The key point to overcome this difficulty is to use a two-level energy balance between the natural and weakened energies. For this, we have to "sacrifice" an ε-portion of the energy of the first component to get a positive contribution of the total weakened energy of both components. The conservation of this total weakened energy is then the crucial point to absorb the "bad" term represented by the weakened energy of the second component at initial time. Note that in the course of this third step, the estimates have to be performed carefully; the sizes of ε and of α have to be chosen with caution as well.

To conclude this introduction, let us recall some results existing in the literature which are related to the present paper. Several papers (see, e.g., [9], [27], [31], [23], [29], [8]) concern the stabilization of hyperbolic-parabolic coupled systems, such as thermoelasticity, thermoplates. For such systems, the main purpose is to determine if the dissipation induced by the heat-type equation is sufficient for stabilizing the

full system obtained by coupling it to an hyperbolic-type equation. Observability results for coupled hyperbolic-parabolic systems have been obtained in [5] (see also references therein). Concerning coupled hyperbolic-hyperbolic systems, several results concerning both stabilization and observability via two control forces have been obtained. Complete and partial observability (respectively, controllability) results for coupled systems either of hyperbolic-hyperbolic type or of hyperbolic-parabolic type can be found in [25], [26]. These results assume that the coupling parameter is sufficiently small. They have been extended in [18] to the cases of arbitrary coupling parameters (assuming bounded coupling operators). For both references, the multiplier method was the main ingredient for obtaining the desired estimates. Complete observability (respectively, controllability) results have also been obtained in [24] for systems of coupled second order hyperbolic equations containing first order terms in both the original and the coupled unknowns. These results are based on Carleman estimates. One can also look at [16], [20], [22] for stabilization results. Stabilization and observability results for hyperbolic-hyperbolic systems via a single control force have been considered more recently. In [14] and [15], wave-wave systems having the same principal part are coupled through velocity terms. The coupling is therefore not compact. Indirect observability and uniform stabilization results are established. In [4] (see also references therein), polynomial decay estimates in the case of indirect internal stabilization case are given. These results have been extended to several cases (wave-wave coupling, Petrowsky-Petrowsky coupling) for the locally distributed indirect stabilization case in [7].

This paper is organized as follows. In section 2, we give the main results of this paper, which are the indirect observability, uniqueness, and indirect exact controllability for the abstract system. We also give two examples of application of these abstract results. In section 3, we establish the well-posedness of the abstract system and the conservation of the total natural and weakened energies of solutions of this abstract system, and we prove that they satisfy the direct inequality. We also establish well-posedness of the inhomogenous abstract system. In section 4, we prove several intermediate estimates, which will be useful for what follows. In section 5, we give the proof of the main results, based on the two-level energy method. Applications of these results to concrete systems such as wave-wave, Petrowsky-Petrowsky, coupled elastodynamic, wave-wave with different speeds of propagation, and wave-Petrowsky are given in section 6.

2. Main results. Let V_i, $i = 1, 2$, and H be separable real Hilbert spaces such that the injections $V_i \subset H$ are dense, compact, and continuous for $i = 1, 2$.

In what follows we identify H with its dual space, so that the injections $V_i \subset H \subset V_i'$ hold and are continuous, dense, and compact. The scalar products on V_i, $i = 1, 2$, and H are, respectively, denoted by $(,)_i$ and $(,)$, whereas the corresponding norms are, respectively, denoted by $|\ |_i$ and $|\ |$. Moreover, we denote by $|\ |_{V_i'}$ the norm on V_i', by $\langle , \rangle_{V_i', V_i}$ the duality product, and by A_i, $i = 1, 2$ the duality mapping from V_i to V_i' defined by

$$\langle A_i w, z\rangle_{V_i', V_i} = (w, z)_i \qquad \forall w, z \in V_i.$$

To avoid too much notation we will still denote by A_i the unbounded operator in H, acting from $D(A_i) = \{u \in V_i, A_i u \in H\}$ into H. Note that these operators are self-adjoint. For $i = 1, 2$, we denote by ν_i the smallest positive constant such that

$$|u|^2 \leq \nu_i^2 \, |u|_i^2 \qquad \forall\, u \in V_i.$$

Hence, we have for $i = 1, 2$

$$|u|^2 \leq \nu_i^2 |A_i^{1/2} u|^2 \qquad \forall\, u \in V_i, \tag{2.1}$$

which is equivalent to

$$|A_i^{-1/2} u|^2 \leq \nu_i^2 |u|^2 \qquad \forall\, u \in H. \tag{2.2}$$

Moreover, let C be a given linear continuous operator on H. We denote by β_2 the smallest positive constant such that

$$|Cu| \leq \beta_2 |u| \qquad \forall\, u \in H. \tag{2.3}$$

Hence, we also have

$$|C^\star u| \leq \beta_2 |u|. \tag{2.4}$$

Set

$$\alpha_0 = (\beta_2 \nu_1 \nu_2)^{-1}. \tag{2.5}$$

Finally, let α be a given nonzero parameter. For the sake of clarity we will assume that α is positive; nevertheless, the results in this paper are valid for negative α as well. We consider the following weakly coupled system:

$$\begin{cases} u_1'' + A_1 u_1 + \alpha C u_2 = 0 \quad \text{in } V_1', \\ u_2'' + A_2 u_2 + \alpha C^\star u_1 = 0 \quad \text{in } V_2', \\ (u_1, u_1')(0) = (u_1^0, u_1^1) = U_1^0 \in V_1 \times H, \\ (u_2, u_2')(0) = (u_2^0, u_2^1) = U_2^0 \in V_2 \times H. \end{cases} \tag{2.6}$$

As will be seen in the next section, this problem is well-posed in the sense of semigroup theory in the energy space $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2$, where $\mathcal{H}_i = V_i \times H$ for $i = 1, 2$. We define an unbounded operator linear operator $\mathcal{A}_\alpha$ on $\mathcal{H}$ by

$$\mathcal{A}_\alpha U = (-v_1, A_1 u_1 + \alpha C u_2, -v_2, A_2 u_2 + \alpha C^\star u_1),$$

$$\begin{aligned} D(\mathcal{A}_\alpha) &= \{U = (u_1, v_1, u_2, v_2) \in V_1 \times V_1 \times V_2 \times V_2,\ A_i u_i \in H, i = 1, 2\} \\ &= D(A_1) \times V_1 \times D(A_2) \times V_2. \end{aligned}$$

We can now reformulate the system (2.6) as the abstract first order equation

$$\begin{cases} U' + \mathcal{A}_\alpha U = 0, \\ U(0) = U^0 \in \mathcal{H}. \end{cases} \tag{2.7}$$

Let $U = (u_1, u_1', u_2, u_2')$ be a solution of (2.6). (See the next section for more details.) Then we define the partial and total (natural) energy of the solution U as

$$e_i(u(t)) = \frac{1}{2}(|u'|^2 + |u|_i^2)$$

and

$$E(U(t)) = e_1(u_1(t)) + e_2(u_2(t)) + \alpha(u_1, C u_2).$$

Moreover, in what follows we will also need to use the partial and total weakened energies defined, respectively, by

$$\tilde{e}_i(u(t)) = \frac{1}{2}(|u'|^2_{V_i'} + |u|^2)$$

and

$$\widetilde{E}(U(t)) = \tilde{e}_1(u_1(t)) + \tilde{e}_2(u_2(t)) + \alpha(u_1, A_1^{-1}Cu_2).$$

We now give the assumptions on A_1, A_2, C and the observability operator $B^\star$ under which our main results are valid.

We first assume the following hypothesis, to which we will refer in what follows as hypothesis (H1):

$$\exists \beta_1 > 0 \text{ such that } \beta_1|u| \le |Cu| \qquad \forall\, u \in H. \tag{2.8}$$

Moreover, we assume the following (strong) compatibility hypothesis on the operators A_1, A_2, and C, to which we will refer in what follows as hypothesis (H2):

$$CV_2 \subset V_1, CD(A_2) \subset D(A_1) \text{ and } A_1Cu = CA_2u \qquad \forall u \in D(A_2)\,, \tag{2.9}$$

$$\exists\, \beta_3 > 0 \text{ such that } |A_1^{1/2}CA_2^{-1/2}u| \le \beta_3|u| \qquad \forall\, u \in H\,. \tag{2.10}$$

Moreover, let G be a given Hilbert space with norm $||\ ||_G$ and scalar product $\langle\ \rangle_G$. The space G will be identified with its dual space in all that follows. Also let $B^\star \in \mathcal{L}(D(A_1), G)$ be the observability operator. We assume that there exist positive constants $\delta_i, i = 1, 2, 3$, $\eta_i, i = 1, 2, 3, 4$, such that for all $T > 0$, all $f \in \mathcal{C}^1([0, T]; H)$, and all $U_1^0 = (u_1^0, u_1^1) \in D(A_1) \times V_1$, the solution u_1 of

$$\begin{cases} u_1'' + A_1u_1 = f, \\ (u_1, u_1')(0) = U_1^0 \end{cases} \tag{2.11}$$

satisfies the next two inequalities (respectively, direct and inverse inequalities), denoted, respectively, by hypothesis (H3),

$$\int_0^T ||B^\star u_1||^2_G\, dt \le \delta_1 \int_0^T e_1(t)\, dt + \delta_2(e_1(T) + e_1(0)) + \delta_3 \int_0^T |f|^2\, dt, \tag{2.12}$$

and by hypothesis (H4),

$$\eta_4 \int_0^T ||B^\star u_1||^2_G\, dt \ge (1 - \eta_1\beta) \int_0^T e_1(t)\, dt - \eta_2(e_1(T) + e_1(0)) - \eta_3\beta^{-1} \int_0^T |f|^2\, dt \quad \forall \beta \in (0, \eta_1^{-1}). \tag{2.13}$$

Before giving our main results, let us make some comments.

Comments. First, the assumption (H4) (i.e., inequality (2.13)) implies that the first equation of (2.6) decoupled from the second one satisfies an observability inequality. This can be seen by setting $f = 0$ in (2.11). Thanks to (2.13), we deduce that e_1, which is conserved through time, satisfies the classical inverse inequality

$$\eta_4 \int_0^T ||B^\star u_1||^2_G\, dt \ge [(1 - \eta_1\beta)T - 2\eta_2]\ e_1(0) \qquad \forall \beta \in (0, \eta_1^{-1}).$$

As will be seen later (in section 6), this property is satisfied for most systems (e.g., wave, Petrowsky, etc.).

Second, the hypotheses (2.9) and (2.10) give a condition on which types of equations can be coupled and for which we are able to prove an indirect observability result for the full system. We will give two abstract conditions, under which these hypotheses are satisfied (see Corollaries 2.2 and 2.3 below).

Third, note that for $u \in H$, $A_2^{-1/2}u \in V_2$. Hence, thanks to the hypothesis (2.9), $CA_2^{-1/2}u \in V_1$, so that the left-hand side of inequality (2.10) makes sense. Moreover, for $v \in V_2$ we have $u = A_2^{-1/2}v \in D(A_2)$. Hence, using hypothesis (2.9) we deduce that

$$A_1 C A_2^{-1/2} v = C A_2^{1/2} v \qquad \forall\, v \in V_2,$$

so that

$$A_1^{1/2} C A_2^{-1/2} v = A_1^{-1/2} C A_2^{1/2} v \qquad \forall\, v \in V_2.$$

Using (2.10) together with this last equality, we obtain

$$|A_1^{-1/2} C A_2^{1/2} u| \le \beta_3 |u| \qquad \forall\, u \in V_2. \tag{2.14}$$

We now state the main result of this paper.

THEOREM 2.1. *Assume that A_1, A_2, and $B^\star$ satisfy the hypotheses* (H1)–(H4). *Then there exists $\alpha^\star > 0$ such that for all $0 < |\alpha| < \alpha^\star$, there exists $T_0 = T_0(\alpha) > 0$ such that for all $T > T_0$ and all $U^0 \in \mathcal{H}$ the solution $U(t) = \exp(-\mathcal{A}_\alpha t)U^0$ of* (2.6) *satisfies*

$$2\eta_4 \int_0^T ||B^\star u_1||_G^2 \, dt \ge c_1(\alpha, T)\mathbf{e_1}(\mathbf{0}) + c_3(\alpha, T)\tilde{\mathbf{e_2}}(\mathbf{0}), \tag{2.15}$$

with

$$c_1(\alpha, T) = \frac{a_1(T - T_3)}{(1 + \alpha T)(1 + \alpha T_3)}, \qquad c_3(\alpha, T) = \frac{\alpha a_2 (T - T_2)(T - T_2^-)}{1 + \alpha T}, \tag{2.16}$$

where T_2, T_2^-, T_3 are explicit constants depending only on α, and the numbers ν_i, $i = 1, 2$, β_i for $i = 1, 2, 3$, and a_1 and a_2 are positive fixed constants.

If we denote by C a generic positive constant independent on α, then $|T_2|$, $|T_2^-|$ behave as $C\alpha^{-1}$ and T_3 behaves as C as α goes to zero.

Hence, indirect observability holds for the full system (2.6).

Remarks. The constants in the above theorem are given as follows. The constants T_2, T_2^- are defined, respectively, as the positive and negative root of the second order polynomial Q_α defined in (5.9). The constant T_3 is defined by (5.10). The time T_0 is defined as $T_0 = \max(T_2, T_3)$. Moreover, the fixed constants a_1 and a_2 are given in (5.5) and (5.6).

Thanks to the well-known hidden regularity property of weak solutions of (2.6), the application J which maps $U^0 \in D(\mathcal{A}_\alpha)$ to $B^\star u_1 \in L^2_{loc}([0, +\infty); G)$ can be extended to all $\mathcal{H}$ (see the proof in Proposition 3.3). We will not recall this extension further but will use it in what follows.

This result, together with the other results of this paper, are valid under larger hypotheses on the observability operator $B^\star$. It may act on both u_1 and u_1' or on

u_1' only. It may also only be defined and continuous in a space smaller than $D(A_1)$, provided that this space is dense in V_1.

We now give, as in [1], two abstract examples for which the assumptions (H1) and (H2) are satisfied.

Example 1. Case $A_1 = A_2$.

COROLLARY 2.2. *Assume the hypotheses* (H3)–(H4). *Assume, moreover, that*

$$A_1 = A_2 \text{ with } D(A_1) = D(A_2) \text{ and } V_2 \subset V_1.$$

Choose $C = \mathrm{Id}$. *Then for all* $0 < |\alpha| < \alpha^\star$, *there exists* $T_0 = T_0(\alpha) > 0$ *such that for all* $T > T_0$ *and all* $U^0 \in \mathcal{H}$ *the solution* $U(t) = \exp(-\mathcal{A}_\alpha t)U^0$ *of* (2.6) *satisfies the estimate* (2.15).

Proof. The proof is immediate since with $C = \mathrm{Id}$, the hypotheses (H1)–(H2) are satisfied with $\beta_1 = \beta_3 = 1$. We then conclude by applying Theorem 2.1. One can remark that one can also choose operators C more general than the identity operator on H (see the next corollary). □

Example 2. Case $A_1 \neq A_2$. We recall the following result partially proved in [1]; the full proof is left to the reader.

COROLLARY 2.3. *Assume the hypotheses* (H3)–(H4). *Assume, moreover, that* $V_2 \subset V_1$ *and that there exists a common orthonormal basis* $\{e_k\}_{k=1}^\infty$ *of eigenfunctions of the operators* A_i *in* H, *for* $i = 1, 2$, *with*

$$A_i e_k = \lambda_{i,k} e_k, \qquad k = 1, \ldots, i = 1, 2.$$

Assume, moreover, that the following hypothesis holds:

$$\text{(H5)} \qquad \begin{cases} \exists r : \mathbb{N}^\star \mapsto \mathbb{N}^\star, \text{ one-to-one, such that} \\ \lambda_{2,k} = \lambda_{1,r(k)} \quad \forall\, k \in \mathbb{N}^\star. \end{cases}$$

Choose the coupling operator C *arbitrarily under the form*

$$Cu = \sum_{k=1}^{\infty} u_k w_k e_{r(k)}, \tag{2.17}$$

where the sequence of real numbers $(w_k)_k$ *satisfies*

$$\exists w_- > 0,\ w_+ > 0 \text{ such that } w_- \leq w_k \leq w_+ \qquad \forall\, k \in \{1, \ldots\}.$$

Then there exists $\alpha^\star > 0$ *such that for all* $0 < |\alpha| < \alpha^\star$, *there exists* $T_0 = T_0(\alpha) > 0$ *such that for all* $T > T_0$ *and all* $U^0 \in \mathcal{H}$ *the solution* $U(t) = \exp(-\mathcal{A}_\alpha t)U^0$ *of* (2.6) *satisfies the estimate* (2.15).

As a corollary of Theorem 2.1, we immediatly deduce the following uniqueness result.

THEOREM 2.4. *Assume either the hypotheses of Theorem* 2.1 *or that of Corollary* 2.2 *(or of Corollary* 2.3*) and define* $\alpha^\star$ *as in Theorem* 2.1. *For all* $0 < |\alpha| < \alpha^\star$, *define* $T_0 = T_0(\alpha)$ *as in Theorem* 2.1; *then for all* $T > T_0$ *and all* $U^0 \in \mathcal{H}$, *we have the following uniqueness (or continuation) result: if the solution* $U(t) = \exp(-\mathcal{A}_\alpha t)U^0$ *of* (2.6) *satisfies, in addition,*

$$B^\star u_1(t) = 0 \text{ a.e. in } t \in [0, T],$$

then $U(t) = 0$ *for all* $t \in [0, T]$.

Proof. The proof follows immediately from the observability estimate (2.15). ∎

We now formulate our exact controllability results for the dual problem. We denote by $\mathcal{H}' = V_1' \times H \times V_2' \times H$ the dual space of $\mathcal{H}$. For $Y^0 = (y_1^1, -y_1^0, y_2^1, -y_2^0) \in \mathcal{H}'$ and $v \in L^2_{loc}([0,+\infty); G)$, we consider the problem

$$\left\{ \begin{array}{l} y_1'' + A_1 y_1 + \alpha C y_2 = Bv, \\ y_2'' + A_2 y_2 + \alpha C^\star y_1 = 0, \\ (y_1, y_1')(0) = (y_1^0, y_1^1), \\ (y_2, y_2')(0) = (y_2^0, y_2^1), \end{array} \right. \tag{2.18}$$

where $B \in \mathcal{L}(G, D(A_1)')$ is the adjoint of $B^\star$. An easy computation allows us to see that the adjoint of $\mathcal{A}_\alpha$ is the unbounded operator $\mathcal{A}_\alpha^\star$ defined by $D(\mathcal{A}_\alpha^\star) = V_1 \times D(A_1) \times V_2 \times D(A_2)$ and

$$\mathcal{A}_\alpha^\star Z = (A_1 w_1 + \alpha C w_2, -z_1, A_2 w_2 + \alpha C^\star w_1, -z_2)$$

for $Z = (z_1, w_1, z_2, w_2) \in D(\mathcal{A}_\alpha^\star)$. We set $Y = (y_1', -y_1, y_2', -y_2)$, and we define an operator $\widetilde{B} \in \mathcal{L}(G, (D(\mathcal{A}_\alpha)')$ by setting $\widetilde{B}v = (Bv, 0, 0, 0)$. Then the coupled system (2.18) can be written under the abstract form

$$\left\{ \begin{array}{l} Y' - \mathcal{A}_\alpha^\star Y = \widetilde{B} v, \\ Y(0) = Y^0 \in \mathcal{H}'. \end{array} \right. \tag{2.19}$$

This is precisely the dual problem of (2.7). The solution of (2.19) is defined as usual by the method of transposition (see, e.g., [25], [26], [19]). That is, fixing $Y^0 \in \mathcal{H}'$ and $v \in L^2_{loc}([0,\infty); G)$ arbitrarily, we define a solution of (2.19) as a continuous function Y defined from $[0,\infty)$ with values in $\mathcal{H}'$, such that the equality

$$\langle Y(T), U(T) \rangle_{\mathcal{H}',\mathcal{H}} = \langle Y^0, U^0 \rangle_{\mathcal{H}',\mathcal{H}} + \int_0^T \langle v(t), B^\star u_1(t) \rangle_G \, dt \tag{2.20}$$

holds for every $U^0 \in \mathcal{H}$ where $U(t) = \exp(-t\mathcal{A}_\alpha) U^0 = (u_1, u_1', u_2, u_2')(t)$. One can remark that we used the fact that $\widetilde{(B)}^\star U = B^\star u_1$. We will establish well-posedness (that is, existence, uniqueness, and continuous dependence with respect to the data Y^0 and v) in the next section by a standard procedure. We are concerned by the exact controllability problem for (2.18) in a controllability space $\mathcal{F}'$ that will be specified later on. That is, we are looking for a time $T > 0$ such that for every initial data $Y^0 \in \mathcal{F}'$ there exists a control $v \in L^2([0,T]; G)$ such that the solution of (2.18) satisfies in addition

$$y_i(T) = y_i'(T) = 0, \qquad i = 1, 2. \tag{2.21}$$

We can now state our exact controllability result.

THEOREM 2.5. *Assume either the hypotheses of Theorem* 2.1 *or that of Corollary* 2.2 *(or of Corollary* 2.3*), and let $\alpha^\star$ be as in Theorem* 2.1. *For $0 < |\alpha| < \alpha^\star$, we define T_0 as in Theorem* 2.1. *Then for all $0 < |\alpha| < \alpha^\star$, all $T > T_0$, and all $Y^0 = (y_1^1, -y_1^0, y_2^1, -y_2^0) \in V_1' \times H \times H \times V_2$, there exists a control $v \in L^2([0,T]; G)$ such that the solution $Y = (y_1', -y_1, y_2', -y_2)$ of* (2.19) *(or equivalently of* (2.18)*) satisfies* (2.21). *Hence, exact indirect controllability in time T holds for initial data $(y_1^0, y_1^1, y_2^0, y_2^1) \in H \times V_1' \times V_2 \times H$ for the coupled system* (2.18).

Remark. The above exact controllability result holds in fact in a larger set $\mathcal{F}'$, but we do not have a characterization of this space in terms of Sobolev spaces. We will see in the course of the proof that it contains the space $V_1' \times H \times H \times V_2$, thanks to the observability inequality (2.15).

3. Abstract coupled model. We denote by $((\cdot))$ the standard scalar product on the product space $\mathcal{H}$, i.e.,

$$((U, \widetilde{U})) = \sum_{i=1}^{2} \Big((u_i, \tilde{u_i})_i + (v_i, \tilde{v_i}) \Big)$$

for $U = (u_1, v_1, u_2, v_2)$ and $\widetilde{U} = (\tilde{u_1}, \tilde{v_1}, \tilde{u_2}, \tilde{v_2})$ in $\mathcal{H}$. The associated norm is denoted by $||\cdot||$. We also consider on $\mathcal{H}$ the bilinear form given by

$$(U, \widetilde{U})_\alpha = ((U, \widetilde{U})) + \alpha(Cu_2, \tilde{u_1}) + \alpha(u_1, C\tilde{u_2})$$

for $U = (u_1, v_1, u_2, v_2)$ and $\widetilde{U} = (\tilde{u_1}, \tilde{v_1}, \tilde{u_2}, \tilde{v_2})$ in $\mathcal{H}$. One can remark that

$$2E(U(t)) = ||U(t)||_\alpha^2. \tag{3.1}$$

Then we have the following.

PROPOSITION 3.1. *Assume the above hypotheses on the spaces V_1, V_2, H and the operators A_1, A_2. Moreover, define $\beta_1, \beta_2, \nu_1, \nu_2$ as, respectively, in* (2.8), (2.3), *and* (2.2) *and α_0 by* (2.5). *Then for all $0 \le |\alpha| < \alpha_0$, there exist constants $c_1(\alpha) > 0$ and $c_2(\alpha) > 0$ such that*

$$c_1(\alpha)||U||^2 \le (U, U)_\alpha \le c_2(\alpha)||U||^2 \qquad \forall U \in \mathcal{H}. \tag{3.2}$$

Hence, for all $0 \le |\alpha| < \alpha_0$, the application

$$U \in \mathcal{H} \mapsto ||U||_\alpha = (U, U)_\alpha^{1/2}$$

defines a norm on $\mathcal{H}$ which is equivalent to the norm $||\ ||$.

Proof. Let $U \in \mathcal{H}$ be given. Then one has, thanks to (2.2)–(2.3),

$$|(U, U)_\alpha - ||U||^2| = 2|\alpha \Re((u_1, Cu_2))| \le 2|\alpha|\beta_2\nu_1\nu_2|u_1|_1|u_2|_2,$$

where $\Re$ denotes the real part of a complex number. Hence, the desired estimates hold with $c_1(\alpha) = (1 - \alpha/\alpha_0)$ and $c_2(\alpha) = (1 + \alpha/\alpha_0)$. □

We recall that we defined an unbounded linear operator $\mathcal{A}_\alpha$ on $\mathcal{H}$ in section 2 by

$$\begin{aligned} \mathcal{A}_\alpha U &= (-v_1, A_1u_1 + \alpha Cu_2, -v_2, A_2u_2 + \alpha C^\star u_1), \\ D(\mathcal{A}_\alpha) &= \{U = (u_1, v_1, u_2, v_2) \in V_1 \times V_1 \times V_2 \times V_2,\ A_iu_i \in H, i = 1, 2\} \\ &= D(A_1) \times V_1 \times D(A_2) \times V_2. \end{aligned}$$

We set $D(\mathcal{A}_i) = D(A_i) \times V_i$ and denote by $\mathcal{A}_i$ the unbounded operator with domain $D(\mathcal{A}_i)$ and defined by $\mathcal{A}_i(u_i, v_i) = (-v_i, A_iu_i)$. We can now reformulate the system (2.6) as the abstract first order equation (2.7) (see section 2). We have the following classical well-posedness result.

PROPOSITION 3.2. *Assume the hypotheses of Proposition* 3.1, *and let α_0 be given as in Proposition* 3.1. *Then $-\mathcal{A}_\alpha$ is a skew-adjoint operator on $\mathcal{H}$ which generates a $\mathcal{C}^0$ unitary group $T_0(t) = \exp(-t\mathcal{A}_\alpha)$, $t \in \mathbb{R}$, on $\mathcal{H}$. As a consequence, for all $0 \le |\alpha| < \alpha_0$*

and for every $U^0 \in \mathcal{H}$, the problem (2.7) *has a unique solution $U \in \mathcal{C}([0,+\infty);\mathcal{H})$. If, in addition, $U^0 \in D({\mathcal{A}_\alpha}^k)$ for $k \in \mathbb{N}^\star$, then the solution is in $\mathcal{C}^{k-j}([0,+\infty);D(\mathcal{A}_\alpha^j))$ for $j = 0,\dots,k$. Moreover, the total natural energy of the solution is conserved, i.e., for $U^0 \in \mathcal{H}$, we have*

$$E(U(t)) = E(U(0)) \qquad \forall\, t \geq 0, \tag{3.3}$$

and the total weakened energy of the solution is also conserved, i.e., for $U^0 \in \mathcal{H}$, we have

$$\widetilde{E}(U(t)) = \widetilde{E}(U(0)) \qquad \forall\, t \geq 0. \tag{3.4}$$

Proof. The operator $\mathcal{A}_\alpha$ satisfies

$$(\mathcal{A}_\alpha U, \widetilde{U})_\alpha = -(U, \mathcal{A}_\alpha \widetilde{U})_\alpha \qquad \forall U, \widetilde{U} \in D(\mathcal{A}_\alpha).$$

Hence, $D(\mathcal{A}_\alpha^\star) = D(\mathcal{A}_\alpha)$ and the adjoint of $\mathcal{A}_\alpha$ in $\mathcal{H}$ is equal to $-\mathcal{A}_\alpha$. Thus, $\mathcal{A}_\alpha$ is skew-adjoint in $\mathcal{H}$. Applying Stone's theorem (see, e.g., [10]), we deduce that $-\mathcal{A}_\alpha$ generates a $\mathcal{C}^0$ unitary group $T_0(t) = \exp(-t\mathcal{A}_\alpha)$, $t \in \mathbb{R}$, on $\mathcal{H}$. The well-posedness and regularity of solutions follow at once from classical semigroup theory. Since $T_0(t)$, $t \in \mathbb{R}$, is unitary on $\mathcal{H}$ and since (3.1) holds, we deduce that the natural energy of all solutions with initial data in $\mathcal{H}$ is conserved. For the conservation of the total weakened energy, we proceed as follows. For $U^0 \in D(\mathcal{A}_\alpha)$ the solution $U = (u_1, u_1', u_2, u_2')$ is such that $u_i(t) \in D(A_i)$ for any $t \geq 0$. Moreover, both equations of (2.6) are satisfied in H. Hence we have

$$\tilde{e_i}'(t) = (u_i'' + A_i u_i, A_i^{-1} u_i')$$

for $i = 1, 2$. Now using the two equations of (2.6), we deduce that

$$\tilde{e_1}'(t) = -\alpha\,(u_1', A_1^{-1} C u_2),$$

$$\tilde{e_2}'(t) = -\alpha\,(u_1, C A_2^{-1} u_2').$$

Now using (2.9), we deduce that $Cu = A_1^{-1} C A_2 u$ for all $u \in D(A_2)$. Remarking that for any $v \in H$, $u = A_2^{-1} v \in D(A_2)$, we get that $C A_2^{-1} v = A_1^{-1} C v$ for all $v \in H$. Using this relation in the two previous equations, we obtain $\widetilde{E}'(t) = 0$. Hence, the total weakened energy of solutions with initial data in $D(\mathcal{A}_\alpha)$ is conserved. By density of $D(\mathcal{A}_\alpha)$ in $\mathcal{H}$, we conclude that the same result holds true for any solution with initial data in $\mathcal{H}$. ∎

Remark. The well-posedness of problem (2.7) holds true for any α, since $\mathcal{A}_\alpha$ is a compact perturbation of the corresponding decoupled operator (obtained by setting $\alpha = 0$). Of course, in this case, $\mathcal{H}$ should be equipped with the scalar product $(U, \widetilde{U}) = ((U, \widetilde{U}))$ and the corresponding norm.

We now prove the following direct inequality.

PROPOSITION 3.3. *Assume the above hypotheses on the spaces V_1, V_2, and H and the operators A_1, A_2, and C. Assume, moreover, that the assumption* (H3) *holds and let α_0 be defined by* (2.5). *Then for all $0 \leq |\alpha| < \alpha_0$ and all $U^0 \in D(\mathcal{A}_\alpha)$ the solution $U(t) = \exp(-t\mathcal{A}_\alpha)U^0 = (u_1, u_1', u_2, u_2')$ of* (2.6) *satisfies*

$$\int_0^T ||B^\star u_1||_G^2\, dt \leq M_T E(U^0), \tag{3.5}$$

where

$$M_T = T\,M + \frac{2\delta_2}{c_1(\alpha)},$$

$c_1(\alpha)$ *being defined as in the proof of Proposition* 3.1*, with*

$$M = \frac{\max(\delta_1, 2\delta_3\alpha^2\beta_2^2\nu_2^2)}{c_1(\alpha)}.$$

Moreover, if we denote by J *the operator defined from* $D(\mathcal{A}_\alpha)$ *equipped with the* $||\,||_\alpha$*-norm into* $L^2_{loc}([0,+\infty);G)$*, which maps* U^0 *to the function* $B^\star u_1$*, where* $U(t) = \exp(-t\mathcal{A}_\alpha)U^0 = (u_1, u_1', u_2, u_2')$*, then one can extend* J *to all* $\mathcal{H}$*.*

Proof. Let $U^0 \in D(\mathcal{A}_\alpha)$ be given and denote by $U(t) = (u_1, u_1', u_2, u_2')$ the corresponding solution of (2.6). Then u_1 satisfies

$$\begin{cases} u_1'' + A_1 u_1 = f, \\ (u_1, u_1')(0) = U_1^0 = (u_1^0, u_1^1) \end{cases}$$

with $f = -\alpha C u_2$. Hence, using the assumption (H3) together with (2.3) and (2.1) we obtain

$$\int_0^T ||B^\star u_1||_G^2\,dt \le \delta_1 \int_0^T e_1(t)\,dt + \delta_2(e_1(T) + e_1(0)) + \delta_3\alpha^2\beta_2^2\nu_2^2 \int_0^T |u_2|_2^2\,dt.$$

Hence we have

$$\int_0^T ||B^\star u_1||_G^2\,dt$$
$$\le \max(\delta_1, 2\delta_3\alpha^2\beta_2^2\nu_2^2) \int_0^T (e_1(t) + e_2(t))\,dt + \delta_2((e_1(T) + e_1(0)) + (e_2(T) + e_2(0))).$$

However, thanks to Proposition 3.1 we have

$$e_1(t) + e_2(t) = \frac{||U(t)||^2}{2} \le \frac{||U(t)||_\alpha^2}{2c_1(\alpha)} = \frac{E(U(t))}{c_1(\alpha)}.$$

Using this last inequality in the previous estimate, together with the conservation of energy given in Proposition 3.2, we obtain the desired estimate. The direct inequality, by a density argument of $D(\mathcal{A}_\alpha)$ in $\mathcal{H}$, allows us to prove a hidden regularity result for the solutions of (2.6) with initial data in $\mathcal{H}$ and to extend the operator J to $\mathcal{H}$. □

PROPOSITION 3.4. *Assume the hypotheses of Proposition* 3.1 *and let* α_0 *be given as in Proposition* 3.1*. Then for all* $0 \le |\alpha| < \alpha_0$ *for every* $Y^0 \in \mathcal{H}'$ *and all* $v \in L^2_{loc}([0,+\infty);G)$*, the problem* (2.19) *has a unique solution* $Y \in \mathcal{C}([0,+\infty);\mathcal{H}')$*. Moreover, for every* $T > 0$ *the application* Ψ *defined from* $\mathcal{H}' \times L^2([0,T)];G)$ *into* $\mathcal{C}([0,T];\mathcal{H}')$*, which maps* (Y^0, v) *to* Y*, is continuous.*

Proof. Let $Y^0 \in \mathcal{H}'$ and $v \in L^2_{loc}([0,+\infty);G)$ be fixed arbitrarily. For $T > 0$, we denote by L_T the linear form defined on $\mathcal{H}$ by

$$L_T(U^0) = \langle Y^0,\, U^0\rangle_{\mathcal{H}',\mathcal{H}} + \int_0^T \langle v(t),\, B^\star u_1\rangle_G dt, \tag{3.6}$$

where $U(t) = \exp(-t\mathcal{A}_\alpha)U^0 = (u_1, u_1', u_2, u_2')$. Thanks to the direct inequality proved in Proposition 3.3, we have the estimate

$$|L_T(U^0)| \le \left(||Y^0||_{\mathcal{H}'} + \sqrt{\frac{M_T}{2}} ||v||_{L^2([0,T];G)} \right) ||U^0||_\alpha \tag{3.7}$$

for all $U^0 \in \mathcal{H}$, where the constant M_T depends continuously on T (see Proposition 3.3). Hence $L_T \in \mathcal{H}'$. On the other hand, we proved in Proposition 3.2 that $-\mathcal{A}_\alpha$ generates a unitary group on $\mathcal{H}$. Hence, we have $U^0 = \exp(T\mathcal{A}_\alpha)U(T)$ so that the linear form on $\mathcal{H}$ which maps $U(T)$ to $L_T(U^0)$ is well defined and continuous on $\mathcal{H}$. Therefore, there exists a unique $Y(T) \in \mathcal{H}'$ such that $L_T(U^0) = \langle Y(T), U(T)\rangle_{\mathcal{H}',\mathcal{H}}$ for all $U^0 \in \mathcal{H}$. Moreover, thanks again to the direct inequality proved in Proposition 3.3, and since M_T is continuous with respect to T, one can check that the application from $[0, +\infty)$ to $\mathcal{H}'$ which maps T to L_T is continuous. Therefore, since $||U(T)||_\alpha = ||U^0||_\alpha$, we deduce that $Y \in \mathcal{C}([0, +\infty); \mathcal{H}')$. Finally, thanks to (3.7) we have for all bounded intervals I in $\mathbb{R}$ and all $T \in I$

$$||Y(T)||_{\mathcal{H}'} \le C\Big(||Y^0||_{\mathcal{H}'} + |v||_{L^2(I;G)} \Big),$$

where C is a constant depending only on I. Hence, the application Ψ defined from $\mathcal{H}' \times L^2_{loc}([0, +\infty); G)$ into $\mathcal{C}([0, +\infty); \mathcal{H}')$, which maps (Y^0, v) to Y, is continuous. □

To prove the main result of this paper we first need to prove several intermediate estimates. This is the purpose of next section.

4. Technical intermediate estimates. In what follows, $U = (u_1, u_1', u_2, u_2')$ will stand for a solution of (2.6) corresponding to the initial data $U^0 \in \mathcal{H}$. Moreover, in the remainder of this section, we will denote by C_i, $i = 1, \ldots,$ positive constants which depend only on $\beta_i, i = 1, 2, 3$, ν_1, ν_2 but not on α.

LEMMA 4.1. *Assume the hypotheses of Theorem* 2.1 *and let* α_0 *be given by* (2.5). *Then for all* $0 < |\alpha| < \alpha_0$ *and all* $U^0 \in \mathcal{H}$ *the solution* $U(t) = \exp(-t\mathcal{A}_\alpha)U^0 = (u_1, u_1', u_1, u_2')$ *of* (2.6) *satisfies*

$$\begin{aligned} &\alpha \int_S^T |u_2|^2 \, dt \\ &\le \alpha C_9 \int_S^T |u_1|^2 \, dt + \frac{C_{10}}{\gamma_1}(e_1(T) + e_1(0)) + C_{10}\gamma_1(\tilde{e_2}(T) + \tilde{e_2}(0)) \qquad \forall\ \gamma_1 > 0. \end{aligned} \tag{4.1}$$

Proof. By density of $D(\mathcal{A}_\alpha)$ in $\mathcal{H}$ and since $||U(t)||_\alpha = ||U^0||_\alpha$ for any $t \ge 0$, it is sufficient to prove the desired inequality for $U^0 \in D(\mathcal{A}_\alpha)$. Hence, let $U^0 \in D(\mathcal{A}_\alpha)$ be given. We evaluate the term

$$\int_S^T (u_1'' + A_1u_1 + \alpha C u_2,\ Cu_2) - (u_2'' + A_2u_2 + \alpha C^\star u_1,\ C^\star u_1)\, dt = 0. \tag{4.2}$$

This gives

$$\int_S^T (u_1'', Cu_2) - (u_1, Cu_2'') + (A_1u_1, Cu_2) - (u_1, CA_2u_2) + \alpha|Cu_2|^2 - \alpha|C^\star u_1|^2 \, dt.$$

Hence, integrating by parts the first two terms and using hypothesis (2.9), saying that $A_1C = CA_2$ on $D(A_2)$, we have

$$\alpha \int_S^T |Cu_2|^2 \, dt = \alpha \int_S^T |C^\star u_1|^2 \, dt + [(u_1, Cu_2') - (u_1', Cu_2)]_0^T. \tag{4.3}$$

Remarking that $u_2' \in V_2$ and using the inequality (2.14), we estimate the right-hand side of this inequality as follows:

$$|(u_1, Cu_2')| = \left|\Big(A_1^{1/2}u_1, (A_1^{-1/2}CA_2^{1/2})(A_2^{-1/2}u_2')\Big)\right| \le \beta_3 |A_1^{1/2}u_1|\,|u_2'|_{V_2'}.$$

Hence, we have

$$|(u_1, Cu_2')| \le \beta_3 \left(\frac{|u_1|_1^2}{2\gamma_1} + \frac{\gamma_1 |u_2'|_{V_2'}^2}{2}\right) \qquad \forall\, \gamma_1 > 0. \tag{4.4}$$

Moreover, thanks to (2.3) we have

$$|(u_1', Cu_2)| \le \beta_2 |u_1'|\,|u_2| \le \beta_2 \frac{|u_1'|^2}{2\gamma_1} + \beta_2 \frac{\gamma_1 |u_2|^2}{2} \qquad \forall\, \gamma_1 > 0.$$

Therefore, we have

$$\left|\Big((u_1, Cu_2') - (u_1', Cu_2)\Big)(t)\right| \le \frac{\max(\beta_3, \beta_2)}{\gamma_1} e_1(t) + \max(\beta_3, \beta_2)\gamma_1 \tilde{e_2}(t) \qquad \forall\, \gamma_1 > 0. \tag{4.5}$$

Using (4.5) together with (2.8) and (2.4) in (4.3), we obtain the desired result. □

LEMMA 4.2. *Assume the hypotheses of Theorem* 2.1 *and set*

$$\alpha_1 = \min(\alpha_0, \nu_2^{-1}), \tag{4.6}$$

where α_0 *is given by* (2.5). *Then for all* $0 < |\alpha| < \alpha_1$ *and all* $U^0 \in \mathcal{H}$ *the solution* $U(t) = \exp(-t\mathcal{A}_\alpha)U^0$ *of* (2.6) *satisfies the following estimates:*

$$e_1(T) + e_1(0) \le C_1(e_1(0) + \tilde{e_2}(0)) + \frac{C_2\alpha}{1-\alpha\nu_2}\int_0^T (|u_1'|^2 + |u_1|^2)\,dt, \tag{4.7}$$

$$\tilde{e_2}(T) + \tilde{e_2}(0) \le \frac{C_3}{1-\alpha\nu_2}(e_1(0) + \tilde{e_2}(0)) + \frac{C_4\alpha}{(1-\alpha\nu_2)^2}\int_0^T (|u_1'|^2 + |u_1|^2)\,dt, \tag{4.8}$$

$$\int_0^T |A_2^{-1/2}u_2'|^2\,dt \le \frac{C_5}{\alpha(1-\alpha\nu_2)}(e_1(0) + \tilde{e_2}(0)) + \frac{C_6}{(1-\alpha\nu_2)^2}\int_0^T (|u_1'|^2 + |u_1|^2)\,dt, \tag{4.9}$$

$$\int_0^T |u_2|^2\,dt \le \frac{C_7}{\alpha(1-\alpha\nu_2)}(e_1(0) + \tilde{e_2}(0)) + \frac{C_8}{(1-\alpha\nu_2)^2}\int_0^T (|u_1'|^2 + |u_1|^2)\,dt. \tag{4.10}$$

Proof. As in the previous result it is sufficient to prove the result for $U^0 \in D(\mathcal{A}_\alpha)$. Hence, let $U^0 \in D(\mathcal{A}_\alpha)$ be given. As a first step, we obtain an estimate of $\int_0^T |A_2^{-1/2}u_2'|^2\,dt$. For this, we evaluate

$$\int_0^T (u_2'' + A_2u_2 + \alpha C^\star u_1, A_2^{-1}u_2)\,dt = 0.$$

This gives

$$\begin{aligned}
\int_0^T |A_2^{-1/2}u_2'|^2\,dt &= \int_0^T |u_2|^2\,dt + \alpha\int_0^T (C^\star u_1, A_2^{-1}u_2)\,dt + [(A_2^{-1/2}u_2', (A_2^{-1/2}u_2)]_0^T \\
&\le \int_0^T |u_2|^2\,dt + \alpha\int_0^T |C^\star u_1|\,|A_2^{-1}u_2|\,dt \\
&\quad + \frac{1}{2\gamma}\Big(|A_2^{-1/2}u_2'|^2(T) + |A_2^{-1/2}u_2'|^2(0)\Big) \\
&\quad + \frac{\gamma}{2}\Big(|A_2^{-1/2}u_2|^2(T) + |A_2^{-1/2}u_2|^2(0)\Big) \qquad \forall\, \gamma > 0.
\end{aligned}$$

Using (2.4) together with (2.2) for $i = 2$, we obtain

$$\int_0^T |A_2^{-1/2}u_2'|^2\,dt \le \left(1+\frac{\alpha\beta_2\nu_2^4}{2}\right)\int_0^T |u_2|^2\,dt + \frac{\alpha\beta_2}{2}\int_0^T |u_1|^2\,dt + \frac{1}{2}\left(\frac{|A_2^{-1/2}u_2'|^2(T)}{\gamma}+\gamma\nu_2^2|u_2|^2(T)\right) + \frac{1}{2}\left(\frac{|A_2^{-1/2}u_2'|^2(0)}{\gamma}+\gamma\nu_2^2|u_2|^2(0)\right) \qquad \forall\,\gamma>0.$$

We now choose $\gamma = \nu_2^{-1}$. This gives

$$\int_0^T |A_2^{-1/2}u_2'|^2\,dt \le \left(1+\frac{\alpha\beta_2\nu_2^4}{2}\right)\int_0^T |u_2|^2\,dt + \frac{\alpha\beta_2}{2}\int_0^T |u_1|^2\,dt + \nu_2(\tilde{e_2}(T)+\tilde{e_2}(0)).$$

Inserting (4.1) into this last inequality, we obtain

$$\int_0^T |A_2^{-1/2}u_2'|^2\,dt \le C_{11}\int_0^T |u_1|^2\,dt + \frac{C_{12}}{\alpha\gamma_1}(e_1(T)+e_1(0)) + \left(\frac{C_{13}\gamma_1}{\alpha}+\nu_2\right)(\tilde{e_2}(T)+\tilde{e_2}(0)) \qquad \forall\,\gamma_1>0. \tag{4.11}$$

As a second step, we estimate $\tilde{e_2}(T)+\tilde{e_2}(0)$. We proceed as follows. Since $U^0 \in D(\mathcal{A}_\alpha)$ we have $u_2 \in D(A_2)$. Moreover, recall that $\tilde{e_2}$ is given by the relation

$$\tilde{e_2} = \frac{1}{2}(|u_2|^2 + |A_2^{-1/2}u_2'|^2).$$

Hence, we have

$$\tilde{e_2}'(t) = (u_2, u_2') + (A_2^{-1/2}u_2'', A_2^{-1/2}u_2').$$

Using the second equation of (2.6), we deduce that

$$\tilde{e_2}'(t) = -\alpha(A_2^{-1/2}C^\star u_1, A_2^{-1/2}u_2'). \tag{4.12}$$

Integrating this last equality between 0 and T, we obtain

$$\tilde{e_2}(T)+\tilde{e_2}(0) = 2\tilde{e_2}(0) - \alpha\int_0^T (A_2^{-1/2}C^\star u_1, A_2^{-1/2}u_2')\,dt \le 2\tilde{e_2}(0) + \frac{\alpha}{2}\int_0^T |A_2^{-1/2}C^\star u_1|^2\,dt + \frac{\alpha}{2}\int_0^T |A_2^{-1/2}u_2'|^2\,dt.$$

We use (4.11) in this last inequality together with (2.4) and (2.2). This gives

$$\Big(1-\frac{\alpha\nu_2}{2}-\frac{C_{13}\gamma_1}{2}\Big)(\tilde{e_2}(T)+\tilde{e_2}(0)) \le 2\tilde{e_2}(0) + \frac{\alpha}{2}\Big(C_{11}+\nu_2^2\beta_2\Big)\int_0^T |u_1|^2\,dt + \frac{C_{12}}{2\gamma_1}(e_1(T)+e_1(0)).$$

Choose

$$\gamma_1 = C_{13}^{-1}$$

in the above inequality. This gives

$$\tilde{e_2}(T)+\tilde{e_2}(0)\leq \frac{4\tilde{e_2}(0)}{1-\alpha\nu_2}+\frac{C_{14}\alpha}{1-\alpha\nu_2}\int_0^T|u_1|^2\,dt + \frac{C_{15}}{1-\alpha\nu_2}(e_1(T)+e_1(0)). \tag{4.13}$$

Using (4.13) in (4.11) and in (4.1) (with $\gamma_1=1$), we can improve both estimates. This gives

$$\int_0^T|A_2^{-1/2}u_2'|^2\,dt\leq \frac{C_{16}}{1-\alpha\nu_2}\int_0^T|u_1|^2\,dt+\frac{C_{17}}{\alpha(1-\alpha\nu_2)}(e_1(T)+e_1(0)) + \frac{C_{18}}{\alpha(1-\alpha\nu_2)}\tilde{e_2}(0) \tag{4.14}$$

and

$$\alpha\int_0^T|u_2|^2\,dt\leq\frac{\alpha C_{19}}{1-\alpha\nu_2}\int_0^T|u_1|^2\,dt+\frac{C_{20}}{1-\alpha\nu_2}(e_1(T)+e_1(0))+\frac{C_{21}}{1-\alpha\nu_2}\tilde{e_2}(0). \tag{4.15}$$

For the third step, we estimate $e_1(T)+e_1(0)$ as follows. Since $U^0\in D(\mathcal{A}_\alpha)$ we have $u_1\in D(A_1)$. Moreover, we recall that e_1 is given by the relation

$$e_1=\frac{1}{2}(|u_1|_1^2+|u_2'|^2).$$

Hence, we have

$$e_1'(t)=(A_1^{1/2}u_1,A_1^{1/2}u_1')+(u_2'',u_2').$$

Using the first equation of (2.6), we deduce that

$$e_1'(t)=-\alpha(Cu_2,u_1'). \tag{4.16}$$

Integrating this last equality between 0 and T, and using (2.3), we obtain

$$\begin{aligned}e_1(T)+e_1(0)&=2e_1(0)-\alpha\int_0^T(Cu_2,u_1')\,dt\\&\leq 2e_1(0)+\frac{\alpha}{2\varepsilon_2}\int_0^T|u_1'|^2\,dt+\frac{\alpha\varepsilon_2\beta_2^2}{2}\int_0^T|u_2|^2\,dt.\end{aligned}$$

We use (4.15) in this last inequality. This gives

$$\begin{aligned}&\left(1-\frac{C_{20}\beta_2^2\varepsilon_2}{2(1-\alpha\nu_2)}(e_1(T)+e_1(0))\right)\\&\quad\leq 2e_1(0)+\frac{\alpha}{2\varepsilon_2}\int_0^T|u_1'|^2\,dt+\frac{C_{19}\varepsilon_2\beta_2^2\alpha}{2(1-\alpha\nu_2)}\int_0^T|u_1|^2\,dt+\frac{C_{21}\varepsilon_2\beta_2^2}{2(1-\alpha\nu_2)}\tilde{e_2}(0).\end{aligned}$$

Choose

$$\varepsilon_2=\frac{1-\alpha\nu_2}{C_{20}\beta_2^2}$$

in the above inequality. This gives the estimate

$$e_1(T) + e_1(0) \leq 4e_1(0) + \frac{C_{22}\alpha}{1-\alpha\nu_2}\int_0^T (|u_1'|^2 + |u_1|^2)\,dt + C_{23}\tilde{e_2}(0).$$

Hence, we have proved the estimate (4.7). We now use the estimate (4.7) successively in (4.13), (4.14), and (4.15). This gives the desired estimates (4.8), (4.9), and (4.10). □

LEMMA 4.3. *Assume the hypotheses of Theorem* 2.1 *and let* α_1 *be defined as in* (4.6). *We set*

$$\alpha_2 = \min(\alpha_1, (2\nu_2)^{-1}). \tag{4.17}$$

Then for all $0 < |\alpha| < \alpha_2$ *and all* $U^0 \in \mathcal{H}$ *the solution* $U(t) = \exp(-t\mathcal{A}_\alpha)U^0$ *of* (2.6) *satisfies for all* $\gamma_2 > 0$

$$\int_0^T e_1(t)\,dt \geq \frac{C_{27}T}{2(1+\alpha T)}(e_1(0) - \tilde{e_2}(0)). \tag{4.18}$$

Proof. Using (4.16), we have

$$\int_0^T e_1(t)\,dt = Te_1(0) - \alpha\int_0^T (T-t)(Cu_2, u_1')\,dt.$$

Hence, using (2.3) in this equality, we obtain for all $\gamma_2 > 0$

$$\int_0^T e_1(t)\,dt \geq Te_1(0) - \frac{\alpha T}{2\gamma_2}\int_0^T |u_1'|^2\,dt - \frac{\gamma_2\beta_2^2\alpha T}{2}\int_0^T |u_2|^2\,dt.$$

Thanks to (4.10) together with (2.1) used in this last inequality, we obtain

$$\begin{aligned}\int_0^T e_1(t)\,dt \geq{}& T\left(1 - \frac{\gamma_2\beta_2^2 C_7}{2(1-\alpha\nu_2)}\right)e_1(0) - T\frac{\gamma_2\beta_2^2 C_7}{2(1-\alpha\nu_2)}\tilde{e_2}(0)\\ &- \frac{\alpha T}{2}\left(\frac{1}{\gamma_2}\int_0^T |u_1'|^2\,dt\frac{\gamma_2\beta_2^2 C_8}{1-\alpha\nu_2}\int_0^T (|u_1'|^2 + \nu_1^2|u_1|_1^2)\,dt\right).\end{aligned}$$

We choose $\gamma_2 = \frac{1-\alpha\nu_2}{\beta_2^2 C_7}$ in the above inequality. This leads to

$$\int_0^T e_1(t)\,dt \geq \frac{T}{2}(e_1(0) - \tilde{e_2}(0)) - \frac{\alpha T C_{28}}{1-\alpha\nu_2}\int_0^T e_1(t)\,dt.$$

Since $\alpha \in (0, \alpha_2)$ and thanks to our choice of α_2, we have $1 \leq (1-\alpha\nu_2)^{-1} \leq 2$. Using these two inequalities in the previous inequality, we obtain the desired inequality with a new constant C_{27}. □

Remark. Note that our choice of α_2 is not optimal, but this choice has the advantage of simplifying the next computations by avoiding keeping track of the dependence of the constants with respect to $(1-\alpha\nu_2)$ in the next estimates. Indeed, from now on we will not care any longer about the restrictions on the size of α. Hence, in what follows α will be assumed to be sufficiently small in order to shorten the computations.

COROLLARY 4.4. *Assume the hypotheses of Theorem* 2.1 *and let* α_2 *be defined as in* (4.17). *Then there exists* $\alpha_3 \in (0, \alpha_2)$ *such that for all* $0 < |\alpha| < \alpha_3$ *and all* $U^0 \in \mathcal{H}$ *the solution* $U(t) = \exp(-t\mathcal{A}_\alpha)U^0$ *of* (2.6) *satisfies*

$$\int_0^T (e_1(t) + \tilde{e_2}(t))\, dt \geq \frac{\delta_2 T}{2}\, (\tilde{e_1}(0) + \tilde{e_2}(0)), \tag{4.19}$$

where $\delta_2 = \min(1, \nu_1^{-2})$. *Moreover, we have*

$$e_1(T) + e_1(0) \leq C_1(e_1(0) + \tilde{e_2}(0)) + C_{29}\alpha \int_0^T e_1(t)\, dt, \tag{4.20}$$

$$\int_0^T |u_2|^2\, dt \leq 2\frac{C_7}{\alpha}(e_1(0) + \tilde{e_2}(0)) + C_{30} \int_0^T e_1(t)\, dt, \tag{4.21}$$

$$\int_0^T \tilde{e_2}(t)\ dt \leq 2\frac{C_{24}}{\alpha}(e_1(0) + \tilde{e_2}(0)) + C_{36} \int_0^T e_1(t)\, dt. \tag{4.22}$$

Proof. We first remark that thanks to (2.1)–(2.2) we have

$$\int_0^T (e_1(t) + \tilde{e_2}(t))\, dt \geq \delta_2 \int_0^T (\tilde{e_1}(t) + \tilde{e_2}(t))\, dt, \tag{4.23}$$

where $\delta_2 = \min(1, \nu_1^{-2})$. On the other hand, we have

$$\widetilde{E}(t) = \tilde{e_1}(t) + \tilde{e_2}(t) + \alpha(A_1^{-1/2}u_1, (A_1^{-1/2}CA_2^{1/2})\, A_2^{-1/2}u_2).$$

Hence, thanks to (2.2), (2.14) we obtain

$$|\widetilde{E}(t) - (\tilde{e_1}(t) + \tilde{e_2}(t))| \leq \alpha\beta_3\nu_1\nu_2 \left(\frac{|u_1|^2 + |u_2|^2}{2}\right) \leq \alpha\delta_1(\tilde{e_1}(t) + \tilde{e_2}(t)),$$

where $\delta_1 = \nu_1\nu_2\beta_3$. Integrating this last estimate between 0 and T, and since the energy $\widetilde{E}$ is conserved through time, we obtain

$$\int_0^T (\tilde{e_1}(t)+\tilde{e_2}(t))\, dt \geq \frac{1}{1+\alpha\delta_1} \int_0^T \widetilde{E}(t)\, dt = \frac{T}{1+\alpha\delta_1}\widetilde{E}(0) \geq T\frac{1-\alpha\delta_1}{1+\alpha\delta_1}(\tilde{e_1}(0)+\tilde{e_2}(0)).$$

Using this last inequality together with (4.23) and choosing $\alpha_3 \in (0, \alpha_2)$ such that for all $\alpha \in (0, \alpha_3)$

$$\frac{1-\alpha\delta_1}{1+\alpha\delta_1} \geq \frac{1}{2},$$

we obtain (4.19). Moreover, recalling the definition of $\tilde{e_2}$, i.e.,

$$\tilde{e_2} = \frac{1}{2}(|u_2|^2 + |A_2^{-1/2}u_2'|^2),$$

we easily obtain from (4.10) and (4.9)

$$\int_0^T \tilde{e_2}(t)\ dt \leq \frac{C_{24}}{\alpha(1-\alpha\nu_2)}(e_1(0) + \tilde{e_2}(0)) + \frac{C_{25}}{(1-\alpha\nu_2)^2} \int_0^T (|u_1'|^2 + |u_1|^2)\, dt. \tag{4.24}$$

The estimates (4.20), (4.21), and (4.22) are easily obtained from the corresponding estimates (4.7), (4.10), and (4.24), using (2.1) since $\alpha \in (0, \alpha_2)$. ∎

5. Proof of the main result. We are now able to prove the main estimate of this paper. We assume from now on that $\alpha \in (0, \alpha^\star)$, where $0 < \alpha^\star < \alpha_3$ will be chosen explicitly sufficiently small in the course of the proof. We first remark that u_1 is the solution of

$$\begin{cases} u_1'' + A_1 u_1 = f, \\ (u_1, u_1')(0) = (u_1^0, u_1^1) \in D(\mathcal{A}_1), \end{cases}$$

where $f = -\alpha C u_2$. Hence, from our assumption (H1), u_1 satisfies the inequality (2.13) with the corresponding f. Using (4.20) and (4.21) in the resulting inequality, we obtain

$$\eta_4 \int_0^T ||B^\star u_1||_G^2 \, dt \geq \left(1 - \eta_1\beta - \frac{C_{31}\alpha}{1-\alpha\nu_2} - \frac{C_{32}\alpha^2}{\beta(1-\alpha\nu_2)^2}\right) \int_0^T e_1(t)\, dt - C_{33}\left(1 + \frac{\alpha}{\beta(1-\alpha\nu_2)}\right)(e_1(0) + \tilde{e_2}(0)) \qquad \forall\, \beta \in (0, \eta_1).$$

We now choose $\beta = \alpha$ in the above inequality, and we remark that

$$\left(1 - \eta_1\beta - \frac{C_{31}\alpha}{1-\alpha\nu_2} - \frac{C_{32}\alpha^2}{\beta(1-\alpha\nu_2)^2}\right) \geq \frac{1}{2}$$

and

$$- C_{33}\left(1 + \frac{\alpha}{\beta(1-\alpha\nu_2)}\right) \geq -\frac{C_{34}}{2}$$

for sufficiently small α. Using this in the above estimate, we obtain

$$2\eta_4 \int_0^T ||B^\star u_1||_G^2 \, dt \geq \int_0^T e_1(t)\, dt - C_{34}\Big[e_1(0) + \tilde{e_2}(0)\Big].$$

Now let ε be an arbitrary fixed number in $(0, \varepsilon^\star)$, where $\varepsilon^\star$ is chosen sufficiently small. (This will be defined in the course of the proof.) Then using the above inequality, we can write

$$2\eta_4 \int_0^T ||B^\star u_1||_G^2 \, dt \geq (1-\varepsilon)\int_0^T e_1(t)\, dt + \varepsilon \int_0^T (e_1(t) + \tilde{e_2}(t))\, dt - \varepsilon \int_0^T \tilde{e_2}(t)\, dt - C_{34}\Big[e_1(0) + \tilde{e_2}(0)\Big]. \tag{5.1}$$

Using then (4.22) together with (4.19) in the above inequality, we deduce that

$$2\eta_4 \int_0^T ||B^\star u_1||_G^2 \, dt \geq (1 - \varepsilon C_{37}) \int_0^T e_1(t)\, dt + \varepsilon \frac{\delta_2}{2} T \tilde{e_1}(0) - C_{38}\frac{\varepsilon+\alpha}{\alpha} e_1(0) + \Big[\varepsilon\frac{\delta_2}{2}\, T - C_{38}\frac{\varepsilon+\alpha}{\alpha}\Big]\tilde{e_2}(0). \tag{5.2}$$

We set $\varepsilon_0 = C_{37}^{-1}$ and assume from now on that

$$0 < \varepsilon^\star < \varepsilon_0, \tag{5.3}$$

and we use (4.18) in (5.2) to give a lower bound for $\int_0^T e_1(t)\,dt$. This gives the following estimate:

$$\begin{aligned} 2\eta_4 \int_0^T ||B^\star u_1||_G^2\,dt \geq & \left[\frac{a_1\,T}{1+\alpha T} - C_{38}\frac{\varepsilon+\alpha}{\alpha}\right] e_1(0) + \varepsilon\frac{\delta_2}{2}\,T\tilde{e_1}(0) \\ & + \left[\left(a_2 - \frac{a_1}{1+\alpha T}\right)T - \frac{C_{38}(\varepsilon+\alpha)}{\alpha}\right]\tilde{e_2}(0), \end{aligned} \tag{5.4}$$

where a_i for $i = 1, 2$ are given by

$$a_1 = \frac{(1 - C_{37}\varepsilon)\,C_{27}}{2} > 0, \tag{5.5}$$

$$a_2 = \frac{\varepsilon\delta_2}{2} > 0. \tag{5.6}$$

We remark that the coefficient of $\tilde{e_1}(0)$ in (5.4) is strictly positive. Hence, the proof of the theorem will be complete if we can prove that the coefficients of $e_1(0)$ and $\tilde{e_2}(0)$ which depend only on T, α, and ε are positive for sufficiently large T and sufficiently small α and $\varepsilon^\star$. We will show now that this is possible and give the order in which the parameters T, α, and $\varepsilon^\star$ have to be made large enough, and small enough, respectively. We set

$$T_1 = T_1(\alpha) = \left(\frac{a_1}{a_2} - 1\right)\alpha^{-1}. \tag{5.7}$$

One can notice that $T_1(\alpha)$ goes to $+\infty$ as either ε or α goes to zero. Then for any $T \geq T_1(\alpha)$ we have

$$0 < a_2 - \frac{a_1}{1+\alpha\,T}. \tag{5.8}$$

Let us now denote by Q_α the second order polynomial with respect to T defined by

$$Q_\alpha(T) = \alpha a_2 T^2 + [a_2 - a_1 - C_{38}(\varepsilon+\alpha)]\,T - \frac{C_{38}(\varepsilon+\alpha)}{\alpha}. \tag{5.9}$$

This polynomial has two real roots. Moreover, one can remark that the coefficient of T in this polynomial is negative for sufficiently small ε independently on α. Hence, one root is negative whereas the other one is positive. We denote the negative root by $T_2^-(\alpha)$ and the positive root by $T_2(\alpha)$. Then the coefficient of $\tilde{e_2}(0)$ in (5.4) is given by

$$\frac{Q_\alpha(T)}{1+\alpha\,T}.$$

Let us further remark that $T_2(\alpha) > T_1(\alpha)$ since Q_α can be rewritten under the form

$$Q_\alpha = \frac{a_1\alpha(T - T_1(\alpha))T}{1+\alpha T_1(\alpha)} - \frac{C_{38}(\varepsilon+\alpha)(1+\alpha T)}{\alpha}.$$

Hence, for $T \geq T_2(\alpha)$ the coefficient of $\tilde{e_2}(0)$ is positive. It is given by

$$\alpha \frac{a_2(T - T_2(\alpha))(T - T_2^-(\alpha))}{1 + \alpha T}.$$

We now turn to the coefficient of $e_1(0)$ in (5.4). This coefficient will be positive for sufficiently large T if $a_1 - C_{38}(\varepsilon + \alpha)$ can be made positive for sufficiently small ε and α. This can be easily realized by assuming that $\alpha^\star < C_{27}(2C_{38})^{-1}$ and by choosing $\varepsilon^\star = \min(\varepsilon_0, \varepsilon_1)$ with

$$\varepsilon_1 = \frac{C_{27} - 2C_{38}\alpha^\star}{2C_{38} + C_{27}C_{37}}.$$

Now for $\alpha \in (0, \alpha^\star)$ and $\varepsilon \in (0, \varepsilon^\star)$ we set

$$T_3(\alpha) = \frac{C_{38}(\varepsilon + \alpha)}{a_1 - C_{38}(\varepsilon + \alpha)} > 0. \tag{5.10}$$

Then remarking that $T_3 = T_3(\alpha)$ is such that

$$\frac{C_{38}(\varepsilon + \alpha)}{\alpha} = \frac{a_1 T_3}{1 + \alpha T_3},$$

the coefficient of $e_1(0)$ in (5.4) is given by

$$\left[\frac{a_1(T - T_3(\alpha))}{(1 + \alpha T)(1 + \alpha T_3(\alpha)}\right]$$

and is positive for $T \geq T_3(\alpha)$. Using the above expressions for the coefficients of $\tilde{e_2}(0)$ and $e_1(0)$ in (5.4), we obtain the following observability inequality with positive coefficients for $T \geq \max(T_2(\alpha), T_3(\alpha))$:

$$\begin{aligned} 2\eta_4 \int_0^T ||B^\star u_1||_G^2 \, dt \geq & \left[\frac{a_1(T - T_3(\alpha))}{(1 + \alpha T)(1 + \alpha T_3(\alpha)}\right] e_1(0) \\ & + \left[\alpha \frac{a_2(T - T_2(\alpha))(T - T_2^-(\alpha))}{1 + \alpha T}\right] \tilde{e_2}(0). \end{aligned} \tag{5.11}$$

□

Proof of Theorem 2.5. We apply the HUM. Let $U^0 \in D(\mathcal{A}_\alpha)$ be given. We denote by $U = (u_1, u_1', u_2, u_2')$ the corresponding solution of (2.7). Thanks to the observability inequality proved in Theorem 2.1, the seminorm defined by

$$||U^0||_{\mathcal{F}} = \left(\int_0^T ||B^\star u_1||_G^2 \, dt\right)^{1/2} \tag{5.12}$$

is a norm on $D(\mathcal{A}_\alpha)$. We denote by $\mathcal{F}$ the completion of $D(\mathcal{A}_\alpha)$ with respect to this norm. Thanks to the direct and inverse inequalities proved, respectively, in Proposition 3.3 and in Theorem 2.1, we have the following continuous and dense imbeddings:

$$D(\mathcal{A}_\alpha) \subset \mathcal{H} \subset \mathcal{F} \subset \widetilde{\mathcal{H}} \subset \widehat{\mathcal{H}}, \tag{5.13}$$

where $\widetilde{\mathcal{H}} = V_1 \times H \times H \times V_2'$ and $\widehat{\mathcal{H}} = H \times V_1' \times H \times V_2'$. Hence, by duality, we have the following continuous imbeddings:

$$\widetilde{\mathcal{H}}' \subset \mathcal{F}' \subset \mathcal{H}'. \tag{5.14}$$

For every $U^0 \in \mathcal{H}$, we associate a linear form on $\mathcal{H}$ defined by

$$\langle \Lambda(U^0), \widetilde{U^0} \rangle = \int_0^T \langle B^\star u_1(t), B^\star \tilde{u_1}(t) \rangle_G dt. \tag{5.15}$$

By definition of the norm on $\mathcal{F}$ we have the estimate

$$|\langle \Lambda(U^0), \widetilde{U^0} \rangle| \leq ||U^0||_{\mathcal{F}} ||\widetilde{U^0}||_{\mathcal{F}} \qquad \forall U^0 \in \mathcal{H},\ \forall \widetilde{U^0} \in \mathcal{H}.$$

Hence, since $\mathcal{H}$ is dense in $\mathcal{F}$ by definition of $\mathcal{F}$, the map $\Lambda(U^0)$ can be extended in a unique way to a continuous map on $\mathcal{F}$ and $\Lambda(U^0) \in \mathcal{F}'$. Moreover, thanks to the above inequality, the linear map Λ that maps $U^0 \in \mathcal{H}$ to $\Lambda U^0 \in \mathcal{F}'$ is continuous when $\mathcal{H}$ is equipped with the norm $||\cdot||_{\mathcal{F}}$. Hence, by density, Λ can be extended in a unique way to a continuous linear map, still denoted by Λ, from $\mathcal{F}$ to $\mathcal{F}'$. Moreover, we have

$$\langle \Lambda(U^0), \widetilde{U^0} \rangle_{\mathcal{F}',\mathcal{F}} = \langle U^0, \widetilde{U^0} \rangle_{\mathcal{F}} \qquad \forall U^0 \in \mathcal{H},\ \forall \widetilde{U^0} \in \mathcal{H}, \tag{5.16}$$

where $\langle , \rangle_{\mathcal{F}}$ denotes the scalar product associated with the norm on $\mathcal{F}$. Therefore Λ is coercive and continuous on $\mathcal{F}$, so that thanks to the Lax–Milgram lemma, Λ is an isomorphism from $\mathcal{F}$ onto $\mathcal{F}'$. Let us now apply the HUM. Let $Y^0 \in \mathcal{F}'$ be fixed arbitrarily. We set $U^0 = \Lambda^{-1} Y^0$. Then $U^0 \in \mathcal{F}$. Since $\mathcal{H}$ is dense in $\mathcal{F}$, there exists a sequence $(U_n^0)_n \subset \mathcal{H}$ which converges to U^0 in $\mathcal{F}$. We set $Y_n^0 = \Lambda U_n^0$ for all $n \in \mathbb{N}$. Then, since Λ is continuous, Y_n^0 converges to Y^0 in $\mathcal{F}'$. Thanks to (5.14), we have that $Y_n^0 \in \mathcal{H}'$ and converges to Y^0 in $\mathcal{H}'$. We set $U_n = \exp(-t\mathcal{A}_\alpha) U_n^0$ and $v_n = -B^\star u_{1,n}$ for all $n \in \mathbb{N}$, where $u_{1,n}$ denotes the first component of U_n. Then by definition of the norm on $\mathcal{F}$ and since $(U_n^0)_n$ converges in $\mathcal{F}$, $(v_n)_n$ is a Cauchy sequence in $L^2([0,T];G)$. Hence, it converges to a function v in $L^2([0,T];G)$. We associate with U_n the solutions Y_n of the dual problem in $\mathcal{H}'$:

$$\begin{cases} Y_n' - \mathcal{A}_\alpha^\star Y_n = \widetilde{B} v_n, \\ Y_n(T) = 0. \end{cases} \tag{5.17}$$

One can remark that this retrograde problem is well-posed, thanks to the change of time variable $t \mapsto T - t$ and since this first order abstract equation is equivalent to a time second order (and therefore reversible in time) equation. Therefore, by definition of the solutions of this dual problem by transposition, and thanks to the definition of Λ and to our choice of v_n, we have

$$\langle Y_n(0), \widetilde{U^0} \rangle_{\mathcal{H}',\mathcal{H}} = \langle \Lambda(U_n^0), \widetilde{U^0} \rangle_{\mathcal{H}',\mathcal{H}} \qquad \forall\ \widetilde{U^0} \in \mathcal{H}.$$

Hence, we have $Y_n(0) = \Lambda(U_n^0) = Y_n^0$. Hence, we solved the exact controllability problem for the initial data Y_n^0, since Y_n satisfies

$$\begin{cases} Y_n' - \mathcal{A}_\alpha^\star Y_n = \widetilde{B} v_n, \\ Y_n(0) = Y_n^0, \quad Y_n(T) = 0. \end{cases}$$

Now for every $T > T_0$ the map Ψ defined in Proposition 3.4 is continuous; hence, $(Y_n)_n$ converges Y in $\mathcal{C}([0,T];\mathcal{H}')$, where $Y = \Psi(Y^0, v)$ is the solution of

$$\begin{cases} Y' - \mathcal{A}_\alpha^\star Y = \widetilde{B} v, \\ Y(0) = Y^0. \end{cases}$$

Therefore we have $(Y_n(T))_n$ in particular, which converges to $Y(T)$ in $\mathcal{H}'$, but since $Y_n(T) = 0$ for all n, we deduce that Y satisfies, in addition, $Y(T) = 0$. Hence, we solved the exact controllability problem for any initial data in $\mathcal{F}'$. Thanks to the first imbedding in (5.14), we conclude the proof. □

Remark. In the course of the above proof, we approximated the initial data $U^0 \in \mathcal{F}$ by a sequence of initial data $(U_n^0)_n \subset \mathcal{H}$. Indeed, using the HUM, it is not necessary to prove well-posedness of the homogeneous problem in $\mathcal{F}$ for solving the exact controllability problem in $\mathcal{F}'$. This is due to the fact that we just need to use the associated control, which is itself well defined for initial data in $\mathcal{F}$ (by extension by continuity). However, we still need to solve the homogeneous equation in a space larger than the energy space. To avoid this step, we prefer to approximate the initial data in $\mathcal{F}$ by a sequence of data in the energy space. We conjecture that well-posedness of (2.7) does not hold in general in $\mathcal{F}$, unless for instance in cases for which one can show that $\mathcal{F}$ does not depend on T. This difficulty does not appear when direct and inverse inequalities are obtained with the usual energy, which is not the case here. Indeed, using extrapolation theory (we refer the reader to [10] for a clear exposition on this subject), one can solve the homogeneous problem in the space $\mathcal{H}_{-1}$. (One can also solve the homogeneous equation in the dual space of the energy space.) Since one can easily show that for $|\alpha| \leq \alpha_0$ $\mathcal{A}_\alpha$ is invertible, $\mathcal{H}_{-1}$ is defined as the completion of $\mathcal{H}$ with respect to the norm $||\ ||_{-1}$, where the norm $||\ ||_{-1}$ is given on $\mathcal{H}$ by

$$||U||_{-1} = ||\mathcal{A}_\alpha^{-1} U||_\alpha.$$

This extrapolated space contains the energy space $\mathcal{H}$ and $\mathcal{F}$ as dense subspaces. Then, one can prove that the semigroup $\exp(-t\mathcal{A}_\alpha)$, $t \geq 0$, can be extended in a unique way to a strongly continuous semigroup $T_{-1}(t)$, $t \geq 0$, on the extrapolated space $\mathcal{H}_{-1}$ for $t \geq 0$. However, since in general the space $\mathcal{F}$ depends on T, one cannot show that $\mathcal{F}$ is invariant under the action of the semigroup $T_{-1}(t)$.

6. Applications.

6.1. The case $A_1 = A_2$. In all that follows, Ω is a nonempty bounded open set in $\mathbb{R}^N$ having a boundary Γ of class C^2, $\{\Gamma_0, \Gamma_1\}$ is a partition of Γ such that $\overline{\Gamma}_0 \cap \overline{\Gamma}_1 = \emptyset$, and x_0 is a point in $\mathbb{R}^N$ such that $m \cdot \nu \leq 0$ on Γ_0 and $m \cdot \nu \geq 0$ on Γ_1, where $m(x) = x - x_0$. We set $\sup_\Omega ||m|| = R$.

Coupled wave equations with same speed of propagation. We consider the following system:

$$\text{(6.1)} \qquad \begin{cases} u_{1,tt} - \Delta u_1 + \alpha u_2 = 0 \quad \text{in} \quad \Omega \times (0,T), \\ u_{2,tt} - \Delta u_2 + \alpha u_1 = 0 \quad \text{in} \quad \Omega \times (0,T), \\ u_1 = u_2 = 0 \quad \text{on} \quad \Sigma = \Gamma \times (0,T), \\ (u_1, u_{1,t})(0) = (u_1^0, u_1^1), \quad (u_2, u_{2,t})(0) = (u_2^0, u_2^1) \quad \text{on} \quad \Omega. \end{cases}$$

We set $H = L^2(\Omega)$ and $V_1 = V_2 = H_0^1(\Omega)$, equipped, respectively, with the L^2 scalar product and the scalar product $(u,z)_i = \int_\Omega \nabla u \cdot \nabla z$ for $i = 1, 2$ and the corresponding norms. We define the duality mappings A_1 and A_2 as in section 2. As we said before, we still denote by $A_1 = A_2 = -\Delta$ the operators A_1, A_2 viewed as unbounded operators in H acting from $D(A_1) = D(A_2) = H^2(\Omega) \cap H_0^1(\Omega)$ into H.

We define the partial natural and weakened energies by

$$e_i(t) = \frac{|u_i'(t)|^2 + |\nabla u_i(t)|^2}{2},$$

$$\tilde{e_i}(t) = \frac{|u_i'(t)|^2_{H^{-1}(\Omega)} + |u_i(t)|^2}{2}.$$

The inequalities (2.1), (2.3), (2.8), (2.9), and (2.10) are satisfied with ν_i equal to the Poincaré's constant and $\beta_1 = \beta_2 = \beta_3 = 1$. Hence, in order to apply Corollary 2.2, we just need to check the assumptions (H3) and (H4). For this, we have to consider the inhomogeneous problem

$$\left\{\begin{array}{l} u_{1,tt} - \Delta u_1 = f \quad \text{in} \quad \Omega \times (0,T), \\ u_1 = 0 \quad \text{on} \quad \Gamma \times (0,T), \\ (u_1, u_{1,t})(0) = (u_1^0, u_1^1) \in D(A_1) \times V_1 \quad \text{on} \quad \Omega, \end{array}\right. \tag{6.2}$$

where $f \in \mathcal{C}^1([0,T]; H)$. We set $G = L^2(\Gamma_1)$. Moreover, we define a continuous linear operator $B^\star$ from $D(A_1)$ to G by

$$B^\star u = \left.\frac{\partial u}{\partial \nu}\right|_{\Gamma_1},$$

where ν stands for the normal derivative to Γ. The direct inequality (2.12) can be easily obtained following [25], [26]. For this, we choose a vector field h with $h \in (\mathcal{C}^1(\overline{\Omega}))^N$ such that $h_{|\Gamma} = \nu$. (See, e.g., [25], [26] for the proof of existence of such a vector field.) Then we evaluate, as usual, the expression

$$\int_0^T \int_\Omega (u_{1,tt} - \Delta u_1) h \,.\, \nabla u_1 \, dx \, dt = \int_0^T \int_\Omega f \, h \,.\, \nabla u_1 \, dx \, dt;$$

we then obtain the usual equality with an additional term due to the fact that f is nonvanishing. For the sake of completeness we give it below:

$$\begin{aligned} \int_0^T \int_\Gamma h \,.\, \nu \left|\frac{\partial u_1}{\partial \nu}\right|^2 d\gamma \, dt = \int_0^T \int_\Omega \Big(\mathrm{div}(h)(|u_1'|^2 - |\nabla u_1|^2) + 2 \frac{\partial h_j}{\partial x_i} \frac{\partial u_1}{\partial x_i} \frac{\partial u_1}{\partial x_j} \Big) \\ + 2 \left[\int_\Omega u_1' h \,.\, \nabla u_1 \right]_0^T - 2 \int_0^T \int_\Omega f \, h \,.\, \nabla u_1 \, dx \, dt, \end{aligned}$$

where we used the convention summation of repeated indices. We then easily obtain (2.12).

We now turn to the proof of (2.13). For this we use the classical multiplier $Mu_1 = m \,.\, \nabla u_1 + \frac{N-1}{2} u_1$ and compute

$$\int_0^T \int_\Omega (u_{1,tt} - \Delta u_1) \, Mu_1 \, dx \, dt = \int_0^T \int_\Omega f \, Mu_1 \, dx \, dt,$$

$$\begin{aligned} \frac{1}{2} \int_0^T \int_\Gamma m \,.\, \nu \left|\frac{\partial u_1}{\partial \nu}\right|^2 d\gamma \, dt = \int_0^T e_1(t) \, dt - \int_0^T (f, Mu_1) \, dt \\ + \left[\int_\Omega u_{1,t} M u_1 \, dx \right]_0^T. \end{aligned} \tag{6.3}$$

Now recalling that the following two inequalities hold for all $t \geq 0$ (see [17, pp. 38–39]),

$$|Mu_1| \leq R|\nabla u_1|, \qquad |(u_{1,t}, Mu_1)| \leq Re_1(t),$$

and using the estimate

$$\int_0^T |(f, Mu_1)|\, dt \leq \int_0^T \frac{R|f|^2}{2\beta}\, dt + \int_0^T \frac{R\beta|\nabla u_1|^2}{2}\, dt \leq \int_0^T \frac{R|f|^2}{2\beta}\, dt + R\beta \int_0^T e_1(t)\, dt$$

for all $\beta > 0$ in the above equality, we obtain (2.13). Hence, assumptions (H3)–(H4) are satisfied, and we can therefore apply Corollary 2.2. This gives Theorem 1.1. By duality, we derive Theorem 1.2 for the exact controllability of system (1.8).

Coupled Kirchhoff–Petrowsky plates. Let $N = 2$ (for physical validity of the model) and assume that Ω is a nonempty bounded open set in $\mathbb{R}^N$ having a boundary Γ of class C^4. We assume as before that $\{\Gamma_0, \Gamma_1\}$ is a partition of Γ such that $\overline{\Gamma}_0 \cap \overline{\Gamma}_1 = \emptyset$ and x_0 is a point in $\mathbb{R}^N$ such that $m \cdot \nu \leq 0$ on Γ_0 and $m \cdot \nu \geq \beta > 0$ on Γ_1, where $m(x) = x - x_0$. We set $\sup_\Omega ||m|| = R$.

We consider the following system:

$$\begin{cases} u_{1,tt} + \Delta^2 u_1 + \alpha u_2 = 0 \quad \text{in} \quad \Omega \times (0,T), \\ u_{2,tt} + \Delta^2 u_2 + \alpha u_1 = 0 \quad \text{in} \quad \Omega \times (0,T), \\ u_1 = u_2 = 0 = \dfrac{\partial u_1}{\partial \nu} = \dfrac{\partial u_2}{\partial \nu} \quad \text{on} \quad \Sigma = \Gamma \times (0,T), \\ (u_1, u_{1,t})(0) = (u_1^0, u_1^1), \quad (u_2, u_{2,t})(0) = (u_2^0, u_2^1) \quad \text{on} \quad \Omega. \end{cases} \tag{6.4}$$

We set $H = L^2(\Omega)$ (equipped with the usual norm and scalar product) and $V_1 = V_2 = H_0^2(\Omega)$ equipped with the scalar product

$$(u, z)_1 = (u, z)_2 = \int_\Omega \Delta u \Delta z \; dx \; dy$$

and the associated norm. We define the duality mappings A_1, A_2 as in section 2. As said before, we still denote by $A_1 = A_2 = -\Delta$ the operators A_1, A_2 viewed as unbounded operators in H acting from $D(A_1) = D(A_2) = H^4(\Omega) \cap H_0^2(\Omega)$ into H.

We define the partial natural and weakened energies by

$$e_i(t) = \frac{|u_i'(t)|^2 + |\Delta u_i(t)|^2}{2},$$

$$\tilde{e_i}(t) = \frac{|u_i'(t)|^2_{H^{-2}(\Omega)} + |u_i(t)|^2}{2}.$$

The inequalities (2.1), (2.3), (2.8), (2.9), and (2.10) are satisfied and $\beta_1 = \beta_2 = \beta_3 = 1$. Hence, in order to apply Corollary 2.2, we just need to check assumptions (H3) and (H4). For this, we have to consider the inhomogeneous problem

$$\begin{cases} u_{1,tt} + \Delta^2 u_1 = f \quad \text{in} \quad \Omega \times (0,T), \\ u_1 = \dfrac{\partial u_1}{\partial \nu} = 0 \quad \text{on} \quad \Gamma \times (0,T), \\ (u_1, u_{1,t})(0) = (u_1^0, u_1^1) \in D(A_1) \times V_1 \quad \text{on} \quad \Omega, \end{cases} \tag{6.5}$$

where $f \in \mathcal{C}^1([0,T];H)$. We set $G = L^2(\Gamma_1)$. Moreover, we define a continuous linear operator $B^\star$ from $D(A_1)$ to G by

$$B^\star u = \Delta u|_{\Gamma_1}.$$

The direct inequality (2.12) can be easily obtained following [25], [26]. For this, we choose as before a vector field h as in the previous example. Then, we evaluate, as usual, the expression

$$\int_0^T \int_\Omega (u_{1,tt} + \Delta^2 u_1) h \,.\, \nabla u_1 \, dx \, dt = \int_0^T \int_\Omega f \, h \,.\, \nabla u_1 \, dx \, dt,$$

and we then obtain the usual equality with an additional term on the right-hand side due to the fact that f is nonvanishing. This additional term is

$$\int_0^T \int_\Omega f \, h \,.\, \nabla u_1 \, dx \, dt.$$

We then easily obtain (2.12).

We now turn to the proof of (2.13). We define μ^2 as the smallest positive constant such that

$$|\nabla u|^2 \le \mu^2 \, |\Delta u|^2 \qquad \forall \, u \in V_1.$$

We then use the multiplier $Mu_1 = m \,.\, \nabla u_1$ and compute

$$\int_0^T \int_\Omega (u_{1,tt} + \Delta^2 u_1) \, Mu_1 \, dx \, dt = \int_0^T \int_\Omega f \, Mu_1 \, dx \, dt.$$

Then proceeding as in [25], [26], one obtains

$$\frac{1}{2} \int_0^T \int_{\Gamma_1} m \,.\, \nu |\Delta u_1|^2 \, d\gamma \, dt = [(u_1', m \,.\, \nabla u_1)]_0^T + 2 \int_0^T e_1(t) \, dt + \int_0^T \int_\Omega f \, m \,.\, \nabla u_1 \, dx \, dt.$$

Using now the estimate

$$|(u_1', m \,.\, \nabla u_1)(t)| \le R\mu e_1(t),$$

together with the definition of R and μ, in the above equality, we obtain for all $\beta > 0$

$$\frac{R}{2} \int_0^T \int_{\Gamma_1} |\Delta u_1|^2 \, d\gamma \, dt \ge (2 - R\beta\mu^2) \in_0^T e_1(t) - R\mu(e_1(0) + e_1(T)) - \frac{1}{2\beta} \int_0^T \int_\Omega |f|^2 \, dx \, dt.$$

Hence, (2.13) holds.

Hence, assumptions (H3)–(H4) are satisfied; we can therefore apply Corollary 2.2. This gives the following results.

THEOREM 6.1. *There then exists $\alpha^\star > 0$ such that for all $0 < |\alpha| < \alpha^\star$, there exists $T_0 = T_0(\alpha) > 0$ such that for all $T > T_0$ and all $U^0 \in \mathcal{H}$ the solution $U(t) = \exp(-\mathcal{A}_\alpha t)U^0 = (u_1, u_1', u_2, u_2')$ of (6.4) satisfies*

$$2 \int_0^T \int_{\Gamma_1} |\Delta u_1|^2 \, d\gamma \, dt \ge \frac{c_1}{2} \left(|u_1^1|^2 + |\Delta u_1^0|^2 \right) + \frac{c_3}{2} \left(|u_2^1|^2_{H^{-2}(\Omega)} + |u_2^0|^2 \right), \tag{6.6}$$

where the constants c_i for $i = 1, 3$ are given in Theorem 2.1.

Moreover, if in addition the solution of (6.4) *satisfies*

$$\Delta u_1 = 0 \text{ on } \Gamma_1 \times (0,T),$$

then one has $u_1 = u_2 = 0$ *in* $\Omega \times [0,T]$.

We now give the dual exact controllability result. For this, we consider the system

$$\text{(6.7)} \quad \begin{cases} y_{1,tt} + \Delta^2 y_1 + \alpha y_2 = 0 \quad \text{in} \quad \Omega \times (0,T), \\ y_{2,tt} + \Delta^2 y_2 + \alpha y_1 = 0 \quad \text{in} \quad \Omega \times (0,T), \\ y_1 = 0 \quad \text{on} \quad \Sigma = \Gamma \times (0,T), \qquad \dfrac{\partial y_1}{\partial \nu} = v \quad \text{on} \quad \Sigma_1 = \Gamma_1 \times (0,T), \\ \qquad y_1 = 0 \quad \text{on} \quad \Sigma_0 = \Gamma_0 \times (0,T), \\ y_2 = \dfrac{\partial y_2}{\partial \nu} = 0 \quad \text{on} \quad \Sigma, \\ (y_1, y_{1,t})(0) = (y_1^0, y_1^1), \quad (y_2, y_{2,t})(0) = (y_2^0, y_2^1) \quad \text{on} \quad \Omega. \end{cases}$$

Since the hypotheses of Theorem 2.5 are satisfied, we have the following result.

THEOREM 6.2. *There exists* $\alpha^\star > 0$ *such that for all* $0 < |\alpha| < \alpha^\star$, *there exists* $T_0 = T_0(\alpha) > 0$ *such that for all* $T > T_0$ *and all* $Y^0 = (y_1^0, y_1^1, y_2^0, y_2^1) \in L^2(\Omega) \times H^{-2}(\Omega) \times H_0^2(\Omega) \times L^2(\Omega)$ *there exists a control* $v \in L^2([),T]; L^2(\Gamma_1))$ *such that the solution* $Y(t) = (y_1, y_1', y_2, y_2')$ *of* (6.7) *satisfies*

$$y_i(.,T) = \partial_t y_i(.,T) = 0 \text{ in } \Omega \text{ for } i = 1,2.$$

Coupled linear elastodynamic systems. In all that follows, we shall use the summation convention for repeated indices. Let (a_{ijkl}) be a tensor such that

$$a_{ijkl} = a_{jikl} = a_{klij} \qquad \forall\, i,j,k,l \in \{1,\dots,N\},$$

satisfying for some $\gamma_0 > 0$

$$a_{ijkl}\varepsilon_{ij}\varepsilon_{kl} \geq \gamma_0 \varepsilon_{ij}\varepsilon_{ij}$$

for every symmetric tensor (ε_{ij}). Given a function u defined from Ω with values in $\mathbb{R}^N$, we set

$$\varepsilon_{ij}(u) = \frac{1}{2}(u_{i,j} + u_{j,i}), \qquad \sigma_{ij}(u) = a_{ijkl}\varepsilon_{kl}(u),$$

where

$$u_{i,j} = \frac{\partial u_i}{\partial x_j}, \qquad u_{j,i} = \frac{\partial u_j}{\partial x_i}.$$

We consider the following system:

$$\text{(6.8)} \quad \begin{cases} u_{i,tt} - \sigma_{ij,j}(u) + \alpha v_i = 0 \quad \text{in} \quad \Omega \times (0,T), \qquad 1 \leq i \leq N, \\ v_{i,tt} - \sigma_{ij,j}(v) + \alpha u_i = 0 \quad \text{in} \quad \Omega \times (0,T), \qquad 1 \leq i \leq N, \\ u_i = v_i = 0 \quad \text{on} \quad \Sigma = \Gamma \times (0,T), \qquad 1 \leq i \leq N, \\ (u_i, u_{i,t})(0) = (u_i^0, u_i^1), (v_i, v_{i,t})(0) = (v_i^0, v_i^1) \quad \text{on} \quad \Omega, \quad 1 \leq i \leq N. \end{cases}$$

We set $H = (L^2(\Omega))^N$ and $V_1 = V_2 = (H_0^1(\Omega))^N$. The space H is equipped with the usual product norm. We recall that thanks to Korn's and Poincaré's inequalities, the bilinear form

$$(u, \tilde{u})_k = \int_\Omega \sigma_{ij}(u)\varepsilon_{ij}(\tilde{u})\, dx$$

is a scalar product on V_k for $k = 1, 2$ and that the associated norm is equivalent to the usual product norm on V_k. We equip V_k with the above bilinear form and the corresponding norm. We define the duality mappings A_1, A_2 as in section 2. Moreover, we still denote by $A_1 = A_2 = -\sigma_{ij,j}$ the operators A_1, A_2 viewed as unbounded operators in H acting from $D(A_1) = D(A_2) = (H^2(\Omega) \cap H_0^1(\Omega))^N$ into H.

We define the partial natural and weakened energies by

$$e_1(t) = \frac{|u'(t)|^2 + |u(t)|_1^2}{2},$$

$$e_2(t) = \frac{|v'(t)|^2 + |v(t)|_2^2}{2},$$

$$\tilde{e_1}(t) = \frac{|u'(t)|_{V_1'}^2 + |u(t)|^2}{2},$$

$$\tilde{e_2}(t) = \frac{|v'(t)|_{V_2'}^2 + |v(t)|^2}{2}.$$

The inequality (2.1) is satisfied thanks to Korn's and Poincaré's inequalities for some constants $\nu_1 = \nu_2 > 0$. The inequalities (2.3), (2.8), (2.9), and (2.10) are satisfied with $\beta_1 = \beta_2 = \beta_3 = 1$. Hence, in order to apply Corollary 2.2, we just need to check the assumptions (H3) and (H4). For this, we have to consider the inhomogeneous problem

$$\text{(6.9)} \qquad \begin{cases} u_{i,tt} - \sigma_{ij,j}(u) = f \quad \text{in} \quad \Omega \times (0,T), \qquad 1 \le i \le N, \\ u_i = 0 \quad \text{on} \quad \Gamma \times (0,T), \qquad 1 \le i \le N, \\ (u_i, u_{i,t})(0) = (u_i^0, u_i^1) \in D(A_1) \times V_1 \quad \text{on} \quad \Omega, \qquad 1 \le i \le N, \end{cases}$$

where $f \in \mathcal{C}^1([0,T]; H)$. We set $G = (L^2(\Gamma_1))^N$. Moreover, we define a continuous linear operator $B^\star$ from $D(A_1)$ to G by

$$(B^\star u)_i = \sigma_{ij}(u)\nu_j|_{\Gamma_1}, \qquad i \in \{1, \dots, N\},$$

where $\nu = (\nu_1, \dots, \nu_N)$ stands for the normal derivative to Γ. We recall the following result proved in [2] (see Lemma 2.1) for any strong solution u of (6.9):

$$\text{(6.10)} \qquad \frac{\gamma_0}{2}\sigma_{ij}(u)\varepsilon_{ij}(u) \le \sum_{i=1}^N |\sigma_{ij}(u)\nu_j|^2 \le \frac{\gamma_1}{\gamma_0}\sigma_{ij}(u)\varepsilon_{ij}(u) \text{ on } \Gamma,$$

where $\gamma_1 = \sum_{i,j,k,l} |a_{ijkl}|^2$. The direct inequality (2.12) can be easily obtained following [2]. For this, we choose a vector field h as before. Then, we evaluate the expression

$$\int_0^T \int_\Omega (u_{i,tt} - \sigma_{ij,j}(u))h \,.\, \nabla u_i \, dx \, dt = \int_0^T \int_\Omega f_i \, h \,.\, \nabla u_i \, dx \, dt.$$

We obtain the usual equality with an additional term due to the fact that f is nonvanishing. This additional term can be easily treated using a Cauchy–Schwarz inequality, whereas the other terms are treated exactly as in [2]. Hence, we obtain the desired inequality (2.12).

We now turn to the proof of (2.13). For this, we use the classical multiplier $Mu_i = m \cdot \nabla u_i + \frac{N-1}{2} u_i$ and compute

$$\int_0^T \int_\Omega (u_{i,tt} - \sigma_{ij,j}(u))\, Mu_i \, dx\, dt = \int_0^T \int_\Omega f_i \, Mu_i \, dx\, dt,$$

Following the computations in [2], we obtain

$$\begin{aligned} \frac{1}{2} \int_0^T \int_\Gamma m \cdot \nu \sigma_{ij}(u) \varepsilon_{ij}(u)\ d\gamma\, dt &= \int_0^T e_1(t)\, dt - \int_0^T (f_i, Mu_i)\, dx\, dt \\ &\quad + \left[\int_\Omega u_{i,t} Mu_i \right]_0^T . \end{aligned} \tag{6.11}$$

Using (6.10) together with Young's inequality in the second term of the right-hand side in the above inequality, we obtain (2.13). (We refer the reader to [2] for estimating the other terms.) Hence, assumptions (H3)–(H4) are satisfied; we can therefore apply Corollary 2.2. This gives the following results.

THEOREM 6.3. *There exists $\alpha^\star > 0$ such that for all $0 < |\alpha| < \alpha^\star$, there exists $T_0 = T_0(\alpha) > 0$ such that for all $T > T_0$ and all $U^0 \in \mathcal{H}$ the solution $U(t) = \exp(-\mathcal{A}_\alpha t) U^0 = (u_1, u_1', u_2, u_2')$ of (6.8) satisfies*

$$\begin{aligned} 2 \int_0^T \int_{\Gamma_1} |\sigma_{ij}(u)\nu_j|^2\, d\gamma\, dt &\geq \frac{c_1}{2} \left(|u^1|^2 + ||u^0||_1^2 \right) \\ &\quad + \frac{c_3}{2} \left(|v^1|_{V_2'}^2 + |v^0|^2 \right), \end{aligned} \tag{6.12}$$

where the constants c_i for $i = 1, 3$ are given in Theorem 2.1.

Moreover, if in addition the solution of (6.8) satisfies

$$\sigma_{ij}(u)\nu_j = 0 \text{ on } \Gamma_1 \times (0,T), \qquad 1 \leq i \leq N,$$

then one has $u = v = 0$ in $\Omega \times [0,T]$.

We now give the dual exact controllability result. For this we consider the system

(6.13)

$$\begin{cases} y_{i,tt} - \sigma_{ij,j}(y) + \alpha z_i = 0 \quad \text{in} \quad \Omega \times (0,T), & 1 \leq i \leq N, \\ z_{i,tt} - \sigma_{ij,j}(z) + \alpha y_i = 0 \quad \text{in} \quad \Omega \times (0,T), & 1 \leq i \leq N, \\ y_i = w_i \quad \text{on} \quad \Sigma_1 = \Gamma_1 \times (0,T),\ y_i = 0 \quad \text{on} \quad \Sigma_0 = \Gamma_0 \times (0,T), & 1 \leq i \leq N, \\ z_i = 0 \quad \text{on} \quad \Sigma = \Gamma \times (0,T), & 1 \leq i \leq N, \\ (y_i, y_{i,t})(0) = (y_i^0, y_i^1), \quad (z_i, z_{i,t})(0) = (z_i^0, z_i^1) \quad \text{on} \quad \Omega, & 1 \leq i \leq N. \end{cases}$$

Since the hypotheses of Theorem 2.5 are satisfied, we have the following result.

THEOREM 6.4. *There exists $\alpha^\star > 0$ such that for all $0 < |\alpha| < \alpha^\star$, there exists $T_0 = T_0(\alpha) > 0$ such that for all $T > T_0$ and all $Y^0 = (y^0, y^1, z^0, z^1) \in (L^2(\Omega))^N \times V_1' \times (H_0^1(\Omega))^N \times (L^2(\Omega))^N$ there exists a control $w = (w_1, \ldots, w_N) \in L^2([0,T]; (L^2(\Gamma_1))^N)$ such that the solution $Y(t) = (y, y', z, z')$ of (6.13) satisfies*

$$y(.,T) = \partial_t y(.,T) = z(.,T) = \partial_t z(.,T) = 0 \text{ in } \Omega.$$

6.2. Case of different operators A_1 and A_2. In order to avoid loss of regularity of solutions, we will assume in all of this subsection that $\Gamma_0 = \emptyset$. Hence, we do not treat the case of mixed boundary conditions. Nevertheless, the results are still valid for a more general situation (see [11], [12], [13]).

Coupled wave equations with different speed of propagation. We consider the following system:

$$\begin{cases} u_{1,tt} - r_1\Delta u_1 + \alpha C u_2 = 0 \quad \text{in} \quad \Omega\times(0,T), \\ u_{2,tt} - r_2\Delta u_2 + \alpha C^\star u_1 = 0 \quad \text{in} \quad \Omega\times(0,T), \\ u_1 = 0,\ u_2 = 0 \quad \text{on} \quad \Sigma = \Gamma\times(0,T), \\ (u_1,u_{1,t})(0) = (u_1^0,u_1^1), \quad (u_2,u_{2,t})(0) = (u_2^0,u_2^1) \quad \text{on} \quad \Omega, \end{cases} \tag{6.14}$$

where $r_i > 0, i = 1,2$. We mainly keep the notation of section 6.1. We set $H = L^2(\Omega)$ and $V_1 = V_2 = H_0^1(\Omega)$ equipped, respectively, with the L^2 scalar product and the scalar product $(u,z)_i = r_i\int_\Omega \nabla u\cdot\nabla z$ and the corresponding norms for $i = 1,2$. We define the duality mappings A_1 and A_2 as in section 2. In order to apply Corollary 2.3, we just need to check that the operators A_1 and A_2 have a common basis of eigenfunctions. In [1], we proved that this holds true only for specific geometries of Ω. More precisely for domains Ω which are N-dimensional intervals $\prod_{i=1}^N(a_i,b_i)$, where $a_i < b_i$, $i = 1,\dots,N$, and $N \le 3$. The operator C is then chosen as in Corollary 2.3.

We still denote by $A_i = -r_i\,\Delta$ for $i = 1,2$ the operators A_i viewed now as unbounded operators in H acting from $D(A_1) = D(A_2) = H^2(\Omega)\cap H_0^1(\Omega)$ in H. Then, Grisvard's results imply in particular that $D(A_1) = D(A_2) = H^2(\Omega)\cap H_0^1(\Omega)$.

We define the partial natural and weakened energies by

$$e_i(t) = \frac{|u_i'(t)|^2 + r_i\,|\nabla u_i(t)|^2}{2},$$

$$\tilde{e_i}(t) = \frac{|u_i'(t)|^2_{H^{-1}(\Omega)} + |u_i(t)|^2}{2}.$$

The inequalities (2.1), (2.3), (2.8), (2.9), and (2.10) are satisfied. We set $G = L^2(\Gamma)$. Moreover, we define a continuous linear operator $B^\star$ from $D(A_1)$ to G by

$$B^\star u = \left.\frac{\partial u}{\partial\nu}\right|_\Gamma.$$

We have the following result.

THEOREM 6.5. *Assume now that Ω is an N-dimensional interval $\prod_{i=1}^N(a_i,b_i)$, where $a_i < b_i$, $i = 1,\dots,N$, with $N \le 3$, and that there exists a positive integer k_0 such that $r_2 = k_0^2 r_1$. Choose, moreover, C as in Corollary 2.3. Then there exists $\alpha^\star > 0$ such that for all $0 < |\alpha| < \alpha^\star$, there exists $T_0 = T_0(\alpha) > 0$ such that for all $T > T_0$ and all $U^0 \in \mathcal{H}$ the solution $U(t) = \exp(-\mathcal{A}_\alpha t)U^0 = (u_1, u_1', u_2, u_2')$ of (6.14) satisfies*

$$2\int_0^T\int_\Gamma \left|\frac{\partial u_1}{\partial\nu}\right|^2 d\gamma\,dt \ge \frac{d_1}{2}\left(|u_1^1|^2 + |\nabla u_1^0|^2\right) + \frac{d_3}{2}\left(|u_2^1|^2_{H^{-1}(\Omega)} + |u_2^0|^2\right), \tag{6.15}$$

where the constants d_i for $i = 1,3$, are of the form $d_i = D_i c_i$ for $i = 1,3$, where the constants c_i are given in Theorem 2.1, and where the constants D_i are independent of α and T.

Moreover, if in addition the solution of (6.14) *satisfies*

$$\frac{\partial u_1}{\partial \nu} = 0 \text{ on } \Gamma \times (0,T),$$

then one has $u_1 = u_2 = 0$ *in* $\Omega \times [0,T]$.

The proof follows the ideas of [1] and will not be detailed here. Of course, we can also derive a dual exact controllability property from the above theorem, which we do not formulate here for the sake of brevity.

Coupled wave-Petrowsky equations. We consider the following system:

$$\text{(6.16)} \qquad \begin{cases} u_{1,tt} - \Delta u_1 + \alpha C u_2 = 0 \quad \text{in} \quad \Omega \times (0,T), \\ u_{2,tt} + \Delta^2 u_2 + \alpha C^\star u_1 = 0 \quad \text{in} \quad \Omega \times (0,T), \\ u_1 = 0,\ u_2 = \Delta u_2 = 0 \quad \text{on} \quad \Sigma = \Gamma \times (0,T), \\ (u_1, u_{1,t})(0) = (u_1^0, u_1^1), \quad (u_2, u_{2,t})(0) = (u_2^0, u_2^1) \quad \text{on} \quad \Omega, \end{cases}$$

where the coupling operator C is chosen in the next theorem. We mainly keep the notation of section 6.1. We set $H = L^2(\Omega)$ and $V_1 = H_0^1(\Omega)$ equipped, respectively, with the L^2 scalar product and the scalar product $(u,z)_1 = \int_\Omega \nabla u \cdot \nabla z$ and the corresponding norms. Moreover, we set $V_2 = H^2(\Omega) \cap H_0^1(\Omega)$ equipped with the scalar product $(u,z)_2 = \int_\Omega \Delta u \cdot \Delta z$ and the associated norm. We define the duality mappings A_1 and A_2 as in section 2. As said before, we still denote by A_i the operators A_i viewed as unbounded operators in H. We already know from the previous wave-wave case with different speeds of propagation that for domains Ω which are N-dimensional intervals, $D(A_1) = H^2(\Omega) \cap H_0^1(\Omega)$. Moreover, we proved in [1], thanks to Grisvard's regularity results, that $D(A_2) = \{u \in H^4(\Omega) \cap H_0^1(\Omega), \Delta u = 0 \text{ on } \Gamma\}$. Hence, all the integration by parts required for the definition of weak solutions of the above coupled wave-Petrowsky system are justified for data in the domain of the operator. We define the partial natural and weakened energies by

$$e_1(t) = \frac{|u_1'(t)|^2 + |\nabla u_1(t)|^2}{2},$$

$$e_2(t) = \frac{|u_2'(t)|^2 + |\Delta u_2(t)|^2}{2},$$

$$\tilde{e_1}(t) = \frac{|u_1'(t)|^2_{H^{-1}(\Omega)} + |u_1(t)|^2}{2},$$

$$\tilde{e_2}(t) = \frac{|u_2'(t)|^2_{V_2'} + |u_2(t)|^2}{2}.$$

The inequalities (2.1), (2.3), (2.8), (2.9), and (2.10) are satisfied. We set $G = L^2(\Gamma)$. Moreover, we define a continuous linear operator $B^\star$ from $D(A_1)$ to G by

$$B^\star u = \left. \frac{\partial u}{\partial \nu} \right|_\Gamma .$$

We have the following result.

THEOREM 6.6. *Let Ω be a N-dimensional interval $\prod_{i=1}^N (a_i, b_i)$, where $a_i < b_i$, $i = 1, \dots, N$, with $N \leq 3$. Assume, moreover, that there exists $d > 0$ for which $b_i - a_i = d$ for all $i = 1, \dots, N$ such that $\frac{\pi}{d\sqrt{N}} \in N^\star$. In addition, assume that for $u = \sum_{k \in (N^\star)^N} u_k e_k$ in H, Cu is defined by*

$$Cu = \sum_{k \in (N^\star)^N} w_k u_k e_{r(k)},$$

where $(w_k)_k$ is an arbitrary sequence of real numbers such that there exist w_-, w_+ with $0 < w_- \leq w_k \leq w_+$ for all k and where r is the one-to-one application defined by $r(k) = (\ell, \dots, \ell)$ for $k \in (N^\star)^N$, where $\ell = \frac{\pi}{d\sqrt{N}} \sum_{i=1}^N k_i^2$. Then, there exists $\alpha^\star > 0$ such that for all $0 < |\alpha| < \alpha^\star$, there exists $T_0 = T_0(\alpha) > 0$ such that for all $T > T_0$ and all $U^0 \in \mathcal{H}$ the solution $U(t) = (u_1, u_1', u_2, u_2')$ of (6.16) satisfies

$$2 \int_0^T \int_\Gamma \left| \frac{\partial u_1}{\partial \nu} \right|^2 d\gamma \, dt \geq \frac{c_1}{2} \left(|u_1^1|^2 + |\Delta u_1^0|^2 \right) + \frac{c_3}{2} \left(|u_2^1|_{V_2'}^2 + |u_2^0|^2 \right), \tag{6.17}$$

where the constants c_i for $i = 1, 3$ are given in Theorem 2.1.

Moreover, if in addition the solution of (6.16) satisfies

$$\frac{\partial u_1}{\partial \nu} = 0 \text{ on } \Gamma \times (0, T),$$

then one has $u_1 = u_2 = 0$ in $\Omega \times [0, T]$.

As before, we do not give the details of the proof. Moreover, a dual exact controllability result can also be stated.

Remarks and open questions. The main restrictive assumption under which the results presented in this paper are valid is the compatibility assumption (2.9). In the case of different operators A_1 and A_2 in the two coupled equations, this assumption requires, in particular, the existence of a common orthonormal basis of eigenfunctions of A_1 and A_2 and that the eigenvalues of A_2 (which in the present paper is the operator acting on the unobserved component) are, in a unique way, eigenvalues of A_1 (see Lemma 2.3). This above-mentioned compatibility assumption is required only in the starting estimate given in Lemma 4.1. The argument in the proof of this lemma is, in some sense, of "algebraic" nature. It would be interesting to know if this estimate still holds true under more general assumptions using more sophisticated tools. If it is not the case, it would be interesting to know if other weaker estimates can still lead to indirect observability and exact controllability results, and if so, in which spaces.

Another important aspect of the abstract model studied in this paper is the coercivity property given by the zero order coupling terms. It would be interesting to know if a coercivity assumption with first order coupling operators can still lead to positive indirect observability estimates. Of course, it would also be interesting to explore others classes of couplings with no coercivity properties, for instance. Finally, in subsection 6.2 we give some applications for which the involved operators do not have the same principal part. Nevertheless, the restrictions on the geometry and on the coupling term are then very severe. We conjecture that positive observability results can still be obtained in more general situations. One can note that the papers treating observability, exact controllability, or stabilization for coupled hyperbolic equations assume in general that both equations have the same principal part (see, e.g., [14], [15]).

Acknowledgments. The author is very grateful to the referees and the associate editor for their valuable comments and suggestions.

REFERENCES

[1] F. Alabau-Boussouira, *Indirect boundary stabilization of weakly coupled hyperbolic systems*, SIAM J. Control Optim., 41 (2002), pp. 511–541.

[2] F. Alabau and V. Komornik, *Boundary observability, controllability, and stabilization of linear elastodynamic systems*, SIAM J. Control Optim., 37 (1998), pp. 521–542.

[3] F. Alabau, *Observabilité frontière indirecte de systèmes faiblement couplés*, C. R. Acad. Sci. Paris Sér. I Math., 333 (2001), pp. 645–650.

[4] F. Alabau, P. Cannarsa, and V. Komornik, *Indirect internal stabilization of weakly coupled systems*, J. Evol. Equ., 2 (2002), pp. 127–150.

[5] P. Albano and D. Tataru, *Carleman estimates and boundary observability for a coupled parabolic-hyperbolic system*, Electron. J. Differential Equations, (2000), paper 22.

[6] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[7] A. Beyrath, *Indirect internal stabilization of coupled systems with locally distributed damping*, C. R. Acad. Sci. Paris Sér I Math., 333 (2001), pp. 451–456.

[8] N. Burq and G. Lebeau, *Mesures de défaut de compacité, application au système de Lamé*, Ann. Sci. École Norm. Sup. (4), 34 (2001), pp. 817–870.

[9] C. M. Dafermos, *On the existence and the asymptotic stability of solutions to the equations of linear thermoelasticity*, Arch. Ration. Mech. Anal., 29 (1968), pp. 241–271.

[10] K.-J. Engel and R. Nagel, *One-Parameter Semigroups for Linear Evolution Equations*, Grad. Texts in Math. 194, Springer-Verlag, New York, 2000.

[11] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monogr. Stud. Math. 24, Pitman, Boston, 1985.

[12] P. Grisvard, *Singularities in Boundary Value Problems*, Rech. Math. Appl. 22, Masson, Paris, Springer-Verlag, Berlin, 1992.

[13] P. Grisvard, *Contrôlabilité exacte des solutions de l'équation des ondes en présence de singularités*, J. Math. Pures Appl., 68 (1989), pp. 215–259.

[14] B. V. Kapitonov, *Uniform stabilization and exact controllability for a class of coupled hyperbolic systems*, Comput. Appl. Math., 15 (1996), pp. 199–212.

[15] B. V. Kapitonov and J. S. Souza, *Observability and uniqueness theorem for a coupled hyperbolic system*, Int. J. Math. Math. Sci., 24 (2000), pp. 423–432.

[16] J. U. Kim and Y. Renardy, *Boundary control of the Timoshenko beam*, SIAM J. Control Optim., 25 (1987), pp. 1417–1429.

[17] V. Komornik, *Exact Controllability and Stabilization. The Multiplier Method*, RAM Res. Appl. Math. 36, Masson, Paris, John Wiley, Chicester, UK, 1994.

[18] V. Komornik and P. Loreti, *Ingham-type theorems for vector-valued functions and observability of coupled linear systems*, SIAM J. Control Optim., 37 (1998), pp. 461–485.

[19] V. Komornik, *Rapid boundary stabilization of linear distributed systems*, SIAM J. Control Optim., 35 (1997), pp. 1591–1613.

[20] V. Komornik and B. Rao, *Boundary stabilization of compactly coupled wave equations*, Asymptot. Anal., 14 (1997), pp. 339–359.

[21] V. Komornik and E. Zuazua, *A direct method for the boundary stabilization of the wave equation*, J. Math. Pures Appl., 69 (1990), pp. 33–54.

[22] J. E. Lagnese and J.-L. Lions, *Modelling Analysis and Control of Thin Plates*, Rech. Math. Appl. 6, Masson, Paris, 1988.

[23] I. Lasiecka, *Uniform decay rates for full von Karman system of dynamic thermoelasticity with free boundary conditions and partial boundary dissipation*, Comm. Partial Differerential Equations, 24 (1999), pp. 1801–1847.

[24] I. Lasiecka and R. Triggiani, *Carleman estimates and exact boundary controllability for a system of coupled, nonconservative second-order hyperbolic equations*, in Partial Differential Equation Methods in Control and Shape Analysis, Lecture Notes in Pure and Appl. Math. 188, Dekker, New York, 1997, pp. 215–243.

[25] J. L. Lions, *Contrôlabilité exacte et stabilisation de systèmes distribués*, Vol. 1, Masson, Paris, 1988.

[26] J. L. Lions, *Contrôlabilité exacte et stabilisation de systèms distribués*, Vol. 2, Masson, Paris, 1988.

[27] K. Liu and Z. Liu, *Exponential stability and analyticity of abstract linear thermoelastic systems*, Z. Angew. Math. Phys., 48 (1997), pp. 885–904.
[28] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.
[29] J. E. Munoz Rivera and R. Racke, *Smoothing properties, decay, and global existence of solutions to nonlinear coupled systems of thermoelastic type*, SIAM J. Math. Anal., 26 (1995), pp. 1547–1563.
[30] D. L. Russell, *A general framework for the study of indirect damping mechanisms in elastic systems*, J. Math. Anal. Appl., 173 (1993), pp. 339–358.
[31] G. Lebeau and E. Zuazua, *Decay rates for the three-dimensional linear system of thermoelasticity*, Arch. Ration. Mech. Anal., 148 (1999), pp. 179–231.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 907–935

HOW TO GET A CONSERVATIVE WELL-POSED LINEAR SYSTEM OUT OF THIN AIR. PART II. CONTROLLABILITY AND STABILITY*

MARIUS TUCSNAK† AND GEORGE WEISS‡

Abstract. Let A_0 be a possibly unbounded positive operator on the Hilbert space H, which is boundedly invertible. Let C_0 be a bounded operator from $\mathcal{D}(A_0^{1/2})$ (with the norm $\|z\|_{1/2}^2 = \langle A_0 z, z\rangle$) to another Hilbert space U. In Part I of this work we have proved that the system of equations

$$\ddot{z}(t) + A_0 z(t) + \frac{1}{2} C_0^* C_0 \dot{z}(t) = C_0^* u(t),$$
$$y(t) = -C_0 \dot{z}(t) + u(t)$$

determines a well-posed linear system Σ with input u and output y, input and output space U, and state space $X = \mathcal{D}(A_0^{1/2}) \times H$. Moreover, Σ is conservative, which means that a certain energy balance equation is satisfied both by the trajectories of Σ and by those of its dual system. In this paper we show that Σ is exactly controllable if and only if it is exactly observable, if and only if it is exponentially stable. Moreover, if we denote by A the generator of the contraction semigroup associated with Σ (which acts on X), then Σ is exponentially stable if and only if one of the entries in the second column of $(i\omega I - A)^{-1}$ is uniformly bounded as a function of $\omega \in \mathbb{R}$. We also show that, under a mild assumption, Σ is approximately controllable if and only if it is approximately observable, if and only if it is strongly stable, if and only if the dual system is strongly stable. We prove many related results and we give examples based on wave and beam equations.

Key words. well-posed linear system, exponential stability, strong stability, conservative system, exact controllability, beam equation, wave equation

AMS subject classifications. 93C25, 93B05, 93B07, 93C20, 35B35

DOI. 10.1137/S0363012901399295

1. Introduction and main results.

This paper is a continuation of our paper [35] in which we have investigated a class of conservative linear systems with a special structure, which occur often in applications. These systems are described by a second order differential equation (in a Hilbert space) and an output equation, and they may have unbounded control and observation operators. The main aim of [35] was to prove the wellposedness, conservativity, and other regularity properties of such systems. Here we investigate conditions under which such systems are exponentially stable or strongly stable. It turns out that these stability properties are equivalent to certain controllability and observability properties as well as to certain estimates.

We recall the construction from the paper [35] in order to be able to state the new results. Let H be a Hilbert space, and let $A_0 : \mathcal{D}(A_0) \to H$ be a self-adjoint, positive, and boundedly invertible operator. We introduce the scale of Hilbert spaces H_α, $\alpha \in \mathbb{R}$, as follows: for every $\alpha \geq 0$, $H_\alpha = \mathcal{D}(A_0^\alpha)$, with the norm $\|z\|_\alpha = \|A_0^\alpha z\|_H$. The space $H_{-\alpha}$ is defined by duality with respect to the pivot space H as follows: $H_{-\alpha} = H_\alpha^*$ for $\alpha > 0$. Equivalently, $H_{-\alpha}$ is the completion of H with respect to the

*Received by the editors December 6, 2001; accepted for publication (in revised form) January 17, 2003; published electronically June 25, 2003. This research was supported by EPSRC (from the UK) under grant GR/R05048/01 and by CNRS (from France) under grant 943706 DRCI.
http://www.siam.org/journals/sicon/42-3/39929.html

†Department of Mathematics, University of Nancy-I, POB 239, Vandoeuvre les Nancy 54506, France (Marius.Tucsnak@iecn.u-nancy.fr).

‡Department of Electrical and Electronic Engineering, Imperial College London, Exhibition Road, London SW7 2BT, UK (G.Weiss@imperial.ac.uk).

norm $\|z\|_{-\alpha} = \left\|A_0^{-\alpha} z\right\|_H$. The operator A_0 can be extended (or restricted) to each H_α such that it becomes a bounded operator

$$A_0 : H_\alpha \to H_{\alpha-1} \qquad \forall\, \alpha \in \mathbb{R}.$$

Let C_0 be a bounded linear operator from $H_{\frac{1}{2}}$ to U, where U is another Hilbert space. We identify U with its dual, so that $U = U^*$. We denote $B_0 = C_0^*$ so that $B_0 \in \mathcal{L}(U, H_{-\frac{1}{2}})$. The class of systems studied in [35] and also here is described by

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} z(t) + A_0 z(t) + \frac{1}{2} B_0 \frac{\mathrm{d}}{\mathrm{d}t} C_0 z(t) = B_0 u(t), \tag{1.1}$$

$$z(0) = z_0, \qquad \dot{z}(0) = w_0, \tag{1.2}$$

$$y(t) = -\frac{\mathrm{d}}{\mathrm{d}t} C_0 z(t) + u(t), \tag{1.3}$$

where $t \in [0, \infty)$ is the time. The equation (1.1) is understood as an equation in $H_{-\frac{1}{2}}$, i.e., all the terms are in $H_{-\frac{1}{2}}$. Most of the linear equations modelling the damped vibrations of elastic structures can be written in the form (1.1), where z stands for the displacement field and the term $B_0 \frac{\mathrm{d}}{\mathrm{d}t} C_0 z(t)$, informally written as $B_0 C_0 \dot{z}(t)$, represents a viscous feedback damping. The signal $u(t)$ is an external input with values in U (often a displacement, a force, or a moment acting on the boundary), and the signal $y(t)$ is the output (measurement) with values in U as well. The state $x(t)$ of this system and its state space X are defined by

$$x(t) = \begin{bmatrix} z(t) \\ \dot{z}(t) \end{bmatrix}, \qquad X = H_{\frac{1}{2}} \times H.$$

We will use some fairly standard notation for certain function spaces: we refer to [35, section 1] for the meaning of $\mathcal{H}^p(0,\infty;W)$, $\mathcal{H}^p_{loc}(0,\infty;W)$ (with $p \in \mathbb{N}$), $C^n(0,\infty;W)$, and $BC^n(0,\infty;W)$ (with $n \in \{0,1,2,\ldots\}$). We write C instead of C^0.

We assume that the reader understands the concepts of a well-posed linear system and of a conservative linear system. These were explained in [35, sections 1, 3, 4] with suitable references to the literature. We will often use results from [35], which we refer to as "Part I." In such cases, we put the prefix I in front of the number of the item quoted. For example, Theorem I.1.4 refers to Theorem 1.4 in Part I, and (I.4.2) refers to formula (4.2) in Part I. The first main result of [35] has been the following (Theorem I.1.1).

THEOREM 1.1. *With the above assumptions, the equations* (1.1)–(1.3) *determine a conservative linear system* Σ *in the following sense:*

There exists a conservative linear system Σ *whose input and output spaces are both* U *and whose state space is* X. *If* $u \in L^2([0,\infty), U)$ *is the input function,* $x_0 = \left[\begin{smallmatrix} z_0 \\ w_0 \end{smallmatrix}\right] \in X$ *is the initial state,* $x = \left[\begin{smallmatrix} z \\ w \end{smallmatrix}\right]$ *is the corresponding state trajectory, and* y *is the corresponding output function, then*

(1)

$$z \in BC(0,\infty; H_{\frac{1}{2}}) \cap BC^1(0,\infty; H) \cap \mathcal{H}^2_{loc}(0,\infty; H_{-\frac{1}{2}}).$$

(2) *The two components of* x *are related by* $w = \dot{z}$.

(3) $C_0 z \in \mathcal{H}^1(0,\infty;U)$ *and the equations* (1.1) *(in* $H_{-\frac{1}{2}}$*) and* (1.3) *(in* U*) hold for almost every* $t \geq 0$ *(hence,* $y \in L^2([0,\infty),U)$*).*

If $\dot{z}$ is a continuous function of t with values in $H_{\frac{1}{2}}$ (see Theorems I.1.2 and I.1.4 for sufficient conditions for this to be true), then (1.1) and (1.3) can be rewritten as

$$\ddot{z}(t) + A_0 z(t) + \frac{1}{2} B_0 C_0 \dot{z}(t) = B_0 u(t), \tag{1.4}$$

$$y(t) = -C_0 \dot{z}(t) + u(t). \tag{1.5}$$

We introduce the space $Z_0 = H_1 + A_0^{-1} B_0 U$, which is a Hilbert space if we define on it a suitable norm; see Theorem I.1.2. We can rewrite the equations (1.4), (1.5) as a first order system as follows:

$$\begin{cases} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= \overline{C}x(t) + u(t), \end{cases} \tag{1.6}$$

where

$$A = \begin{bmatrix} 0 & I \\ -A_0 & -\frac{1}{2} B_0 C_0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ B_0 \end{bmatrix}, \tag{1.7}$$

$$\mathcal{D}(A) = \left\{ \begin{bmatrix} z \\ w \end{bmatrix} \in H_{\frac{1}{2}} \times H_{\frac{1}{2}} \;\middle|\; A_0 z + \frac{1}{2} B_0 C_0 w \in H \right\}, \tag{1.8}$$

$$\overline{C} : Z_0 \times H_{\frac{1}{2}} \to U, \qquad \overline{C} = \begin{bmatrix} 0 & -C_0 \end{bmatrix}. \tag{1.9}$$

We denote by C the restriction of $\overline{C}$ to $\mathcal{D}(A)$. A is the generator of a strongly continuous semigroup of contractions on X, denoted $\mathbb{T} = (\mathbb{T}_t)_{t \geq 0}$. For the concepts of semigroup generator, control operator, observation operator, and transfer function of a well-posed linear system, we refer to Weiss [31, 32] or to section I.3. We denote by $\mathbb{C}_\omega$ the open right half-plane in $\mathbb{C}$ where $\operatorname{Re} s > \omega$. We know from Proposition I.5.3 that for any $s \in \rho(A)$ (in particular, for any $s \in \mathbb{C}_0$) the operator $s^2 I + A_0 + \frac{s}{2} B_0 C_0 \in \mathcal{L}(H_{\frac{1}{2}}, H_{-\frac{1}{2}})$ has a bounded inverse denoted $V(s)$:

$$V(s) = \left(s^2 I + A_0 + \frac{s}{2} B_0 C_0 \right)^{-1} \in \mathcal{L}(H_{-\frac{1}{2}}, H_{\frac{1}{2}}). \tag{1.10}$$

The following proposition is a restatement of a part of Theorem I.1.3.

PROPOSITION 1.2. *With the notation of Theorem* 1.1 *and* (1.7)–(1.10)*, the semigroup generator of* Σ *is* A*, its control operator is* B*, and its observation operator is* C*. The transfer function of* Σ *is given for all* $s \in \mathbb{C}_0$ *by*

$$\mathbf{G}(s) = \overline{C}(sI - A)^{-1} B + I = I - C_0 s V(s) B_0,$$

and we have $\|\mathbf{G}(s)\| \leq 1$ *for all* $s \in \mathbb{C}_0$*.*

Now we have all the necessary ingredients to state the new results of this paper. The following theorems use various controllability, observability, and stability concepts. The precise definition of these concepts is given in section 2.

THEOREM 1.3. *With the above notation, the following assertions are equivalent:*

(1) *The pair* (A, B) *is exactly controllable (in some finite time).*
(2) *The pair* (A, C) *is exactly observable (in some finite time).*
(3) *The semigroup* $\mathbb{T}$ *is exponentially stable.*
(4) *The pair* (A, B) *is optimizable.*
(5) *The pair* (A, C) *is estimatable.*
(6) *We have* $\sup_{s\in\mathbb{C}_0} \|A_0^{\frac{1}{2}} V(s)\|_{\mathcal{L}(H)} < \infty$.
(7) *We have* $\sup_{s\in\mathbb{C}_0} \|sV(s)\|_{\mathcal{L}(H)} < \infty$.
(8) *For a dense subset* E *of* $\mathbb{R}$, *we have* $iE \subset \rho(A)$ *and* $\sup_{\omega\in E} \|A_0^{\frac{1}{2}} V(i\omega)\|_{\mathcal{L}(H)} < \infty$.
(9) *For a dense subset* E *of* $\mathbb{R}$, *we have* $iE \subset \rho(A)$ *and* $\sup_{\omega\in E} \|\omega V(i\omega)\|_{\mathcal{L}(H)} < \infty$.

A more precise statement concerning the equivalence of (1), (2), and (3), which gives some information on the time of exact controllability and observability, and which is valid for any conservative system, is Proposition 3.2. The equivalence of (1)–(5) remains valid for *every* conservative system; see Proposition 3.3.

By a well-known theorem of Prüss and Falun, an operator semigroup $\mathbb{T}$ with generator A is exponentially stable if and only if $(sI - A)^{-1}$ is uniformly bounded on $\mathbb{C}_0$. We refer to section 2 for precise references, further comments, and related results (Propositions 2.4 and 2.5). In the specific case of the semigroup generated by A from (1.7)–(1.8), the resolvent $(sI - A)^{-1}$ can be written as a 2×2 matrix of operators; see Proposition I.5.3 (or formula (4.1) later in this paper). Thus, to verify the exponential stability of $\mathbb{T}$, we would have to verify that the four entries of this 2×2 matrix are all uniformly bounded on $\mathbb{C}_0$. However, conditions (6) and (7) in Theorem 1.3 tell us that, in fact, we have to verify only one of the two entries in the second column of the matrix of $(sI - A)^{-1}$. Conditions (8) and (9) tell us that, in fact, it suffices to check the boundedness of one of these entries on a dense subset of the imaginary axis, and we can still conclude exponential stability.

The version of this theorem corresponding to bounded B and C, i.e., with $C_0 \in \mathcal{L}(H, U)$, is in Liu [22, sections 2–3] but without conditions (4)–(7). Using the boundedness of C_0 (and hence also of B_0), Liu was able to give in [22, Theorem 3.4] also other, Hautus-type conditions which are equivalent to the exponential stability of $\mathbb{T}$. For unbounded C_0, we were only able to obtain a Hautus-type estimate as a necessary condition for exponential stability; see Proposition 4.1.

We mention that semigroups of the type discussed in this paper do not necessarily satisfy the spectrum determined growth condition. For a counterexample (a damped wave equation on a compact manifold) see Lebeau [19].

In the proof of Theorem 1.3 (more precisely, to show that (6)$\Longrightarrow$(3)) we use the following proposition, which is of independent interest. For bounded C_0 this proposition follows easily from [22, Theorem 3.4], but for unbounded C_0 the proof is more delicate (see section 4). Related results for a bounded (possibly not positive) operator in place of $C_0^* C_0$ were given in Liu, Liu, and Rao [23].

PROPOSITION 1.4. *With the above notation, suppose that* C_0 *is bounded from below in the sense that there exists a* $c > 0$ *such that* $\|C_0 z\|_U \geq c\|z\|_H$ *for all* $z \in H_{\frac{1}{2}}$.

Then $\mathbb{T}$ *is exponentially stable.*

A result similar to Theorem 1.3 holds also for strong stability, with an additional assumption on the spectrum $\sigma(A_0)$.

THEOREM 1.5. *With the above notation, assume that* $\sigma(A_0)$ *is countable. (This happens, e.g., if* A_0^{-1} *is compact.) Then the following assertions are equivalent:*

(1) $\mathbb{T}$ *is strongly stable.*
(2) *The pair* (A, C) *is exactly observable in infinite time.*

(3) *The pair* (A, C) *is approximately observable in infinite time.*
(4) $\mathbb{T}$ *is weakly stable (equivalently,* $\mathbb{T}^*$ *is weakly stable).*
(5) $\mathbb{T}^*$ *is strongly stable.*
(6) *The pair* (A, B) *is exactly controllable in infinite time.*
(7) *The pair* (A, B) *is approximately controllable in infinite time.*
(8) *For any* $z \in H_1$*, if* z *is an eigenvector of* A_0*, then* $C_0 z \neq 0$*.*

Note that the statement "A_0^{-1} is compact" does *not* imply that $(sI - A)^{-1}$ is compact (to see this, take $U = H$ and $C_0 = A_0^{1/2}$). This theorem follows from a more general result concerning all conservative systems; see Proposition 3.4 here. In the proof, we also use the famous strong stability theorem of Arendt and Batty [2].

Systems with A and B as above have been studied in Guo and Luo [10, 11], establishing connections between the exponential stability of A and the exact controllability of the undamped system $\ddot{z}(t) + A_0 z(t) = B_0 u(t)$, under the additional hypothesis that the undamped system is well-posed. (Unfortunately, the main result on diagonal systems in [10] (Theorem 4) is incorrectly formulated, and it is also incorrectly quoted in [11].) In [11] the emphasis is on eigenvalues and eigenvectors of A, assuming that the eigenvalues of A_0 satisfy a gap condition and $u(t)$ is scalar.

In section 2 we give the background needed here. Section 3 concerns the stability properties of conservative systems so that the results there refer to a more general context than the main results stated earlier. In section 4 we prove our main results, while section 5 is devoted to two examples: a system involving the beam equation and another one based on the wave equation.

2. Background on controllability, observability, optimizability, estimatability, and stability. In this section we recall some controllability, observability, and stability concepts, quoting the relevant literature. Throughout this section, U, X, and Y are Hilbert spaces and $A : \mathcal{D}(A) \to X$ is the generator of a strongly continuous semigroup $\mathbb{T} = (\mathbb{T}_t)_{t \geq 0}$ on X. The space X_1 is $\mathcal{D}(A)$ with the norm $\|z\|_1 = \|(\beta I - A)z\|$, where $\beta \in \rho(A)$ is fixed, while X_{-1} is the completion of X with respect to the norm $\|z\|_{-1} = \|(\beta I - A)^{-1} z\|$. We assume that the reader understands the concept of an admissible (in particular, infinite-time admissible) control operator for $\mathbb{T}$. This has been presented in section I.2 with suitable references. If $B \in \mathcal{L}(U, X_{-1})$ is admissible, then for every $\tau \geq 0$ we denote by Φ_τ the operator

$$\Phi_\tau u = \int_0^\tau \mathbb{T}_{t-\sigma} B u(\sigma)\, \mathrm{d}\sigma \tag{2.1}$$

as in (I.2.3). We have $\Phi_\tau \in \mathcal{L}(L^2([0,\infty), U), X)$. If B is admissible, then for every $x_0 \in X$ and every $u \in L^2([0,\infty), U)$, the function $x(t) = \mathbb{T}_t x_0 + \Phi_t u$ is called the *state trajectory* corresponding to the initial state x_0 and the input function u. We have $x \in \mathcal{H}^1_{loc}(0,\infty; X)$ and $\dot{x}(t) = Ax(t) + Bu(t)$ (equality in X_{-1}) for almost every $t \geq 0$. If, moreover, B is infinite-time admissible, then we denote, as in (I.2.5),

$$\tilde{\Phi} u = \lim_{\tau \to \infty} \int_0^\tau \mathbb{T}_t B\, u(t)\, \mathrm{d}t, \tag{2.2}$$

and we have $\tilde{\Phi} \in \mathcal{L}(L^2([0,\infty), U), X)$.

Similarly, we assume that the reader understands the concepts of an admissible (in particular, infinite-time admissible) observation operator for $\mathbb{T}$, also presented in section I.2. If $C \in \mathcal{L}(X_1, Y)$ is admissible, then we denote by Ψ the unique continuous

operator from X to $L^2_{loc}([0,\infty),Y)$ such that

$$(\Psi x_0)(t) = C\mathbb{T}_t x_0 \qquad \forall\, x_0 \in \mathcal{D}(A). \tag{2.3}$$

In particular, if C is infinite-time admissible, then $\Psi \in \mathcal{L}(X, L^2([0,\infty),Y))$. Recall that B is an (infinite-time) admissible control operator for $\mathbb{T}$ if and only if B^* is an (infinite-time) admissible observation operator for $\mathbb{T}^*$.

DEFINITION 2.1. *Let A be the generator of a strongly continuous semigroup $\mathbb{T}$ on X, and let $B \in \mathcal{L}(U, X_{-1})$ be an admissible control operator for $\mathbb{T}$.*

The pair (A,B) is exactly controllable *in time $T>0$ if for every $x_0 \in X$, there exists a $u \in L^2([0,T],U)$ such that $\Phi_T u = x_0$.*

(A,B) is exactly controllable *if the above property holds for some $T>0$.*

(A,B) is exactly controllable in infinite time *if B is infinite-time admissible and the operator $\tilde{\Phi}$ from (2.2) is onto.*

(A,B) is approximately controllable *in time $T>0$ if $Ran\Phi_T$ is dense in X.*

(A,B) is approximately controllable in infinite time *if $\cup_{\tau>0}\mathrm{Ran}\,\Phi_\tau$ is dense in X.*

(A,B) is optimizable *if for any $x_0 \in X$, there exists $u \in L^2([0,\infty),U)$ such that the state trajectory corresponding to x_0 and u is in $L^2([0,\infty),X)$.*

Note that the exact (or approximate) controllability in infinite time of (A,B) does *not* imply its exact (or approximate) controllability in time T for some $T>0$. Clearly, exact controllability implies optimizability and also approximate controllability in some finite time. Optimizability is one possible generalization of the concept of stabilizability, as known from finite-dimensional control theory.

Remark 2.1. Let $B \in \mathcal{L}(U, X_{-1})$ be an infinite-time admissible control operator for $\mathbb{T}$. Then (A,B) is approximately controllable in infinite time if and only if the range of $\tilde{\Phi}$ from (2.2) is dense in X. The proof is easy.

Now we introduce the corresponding observability concepts via duality.

DEFINITION 2.2. *Suppose that $C \in \mathcal{L}(X_1,Y)$ is an admissible observation operator for $\mathbb{T}$. (Equivalently, C^* is an admissible control operator for the adjoint semigroup $\mathbb{T}^*$.) We say that (A,C) is* exactly observable *(in time T) (in infinite time) if (A^*,C^*) is exactly controllable (in time T) (in infinite time). Similarly, (A,C) is* approximately observable *(in time T) (in infinite time) if (A^*,C^*) is approximately controllable (in time T) (in infinite time). Finally, the pair (A,C) is called* estimatable *if (A^*,C^*) is optimizable.*

Let Ψ be the operator defined in (2.3), and for every $\tau \geq 0$ put $\Psi_\tau = \mathbf{P}_\tau \Psi$. Then (A,C) is exactly observable in time $T>0$ if and only if Ψ_T is bounded from below. (A,C) is exactly observable in infinite time if and only if C is infinite-time admissible and Ψ is bounded from below. (A,C) is approximately observable in time T (or in infinite time) if and only if $\Psi_T x_0 = 0$ (or $\Psi x_0 = 0$) implies $x_0 = 0$.

Recall that the *growth bound* of a strongly continuous semigroup $\mathbb{T}$ is $\omega_0(\mathbb{T}) = \lim_{t\to\infty} \frac{1}{t}\log\|\mathbb{T}_t\| = \inf_{t>0}\frac{1}{t}\log\|\mathbb{T}_t\|$; see, for example, Pazy [24].

DEFINITION 2.3. *The semigroup $\mathbb{T}$ is* exponentially stable *if its growth bound is negative: $\omega_0(\mathbb{T})<0$. $\mathbb{T}$ is* strongly stable *if*

$$\lim_{t\to\infty}\|\mathbb{T}_t x_0\| = 0 \qquad \forall\, x_0 \in X.$$

Finally, $\mathbb{T}$ is weakly stable *if $\lim_{t\to\infty}\langle \mathbb{T}_t x_0, y_0\rangle = 0$ for all $x_0, y_0 \in X$.*

Let $\mathbb{T}$ be a strongly continuous semigroup on X with generator A. A well-known spectral mapping result of Prüss [25, p. 852] implies that if the function $\|(sI-A)^{-1}\|$ is bounded on $\mathbb{C}_0$, then $\mathbb{T}$ is exponentially stable. A little later and independently, this

result was explicitly stated and proved by Falun [13]. A short proof was given in Weiss [30, section 4]. Here we need a result which is closely related to the one just mentioned, without being an obvious consequence of it. The result is very slightly more general than another result of Falun; see [13, Theorem 3]. Moreover, the proposition below gives an estimate for the growth bound $\omega_0(\mathbb{T})$.

PROPOSITION 2.4. *Let $\mathbb{T}$ be a strongly continuous semigroup on X with generator A. Assume that $\omega_0(\mathbb{T}) \le 0$ and E is a dense subset of $\mathbb{R}$ such that $iE \subset \rho(A)$ and*

$$\|(i\omega I - A)^{-1}\| \le M \qquad \forall\, \omega \in E$$

for some $M > 0$. Then $\mathbb{T}$ is exponentially stable; more precisely, $\omega_0(\mathbb{T}) \le -\frac{1}{M}$.

Proof. By a result in Butzer and Berens [7, p. 31] all numbers $s \in \mathbb{C}$ with $|\operatorname{Re} s| < \frac{1}{M}$ (this is a vertical strip) belong to $\rho(A)$, and we have

$$\|(sI - A)^{-1}\| \le \frac{M}{1 - |\operatorname{Re} s| \cdot M} \quad \text{for} \quad |\operatorname{Re} s| < \frac{1}{M}. \tag{2.4}$$

On the other hand, we know from the Hille–Yosida theorem that $\|(sI - A)^{-1}\|$ is bounded on any half-plane $\mathbb{C}_\gamma$ with $\gamma > 0$. This fact, combined with (2.4), shows that $\|(sI - A)^{-1}\|$ is bounded on any half-plane $\mathbb{C}_\alpha$ with $\alpha > -\frac{1}{M}$. Now by yet another result of Falun [13, Theorem 4] we conclude that $\omega_0(\mathbb{T}) \le -\frac{1}{M}$. (The last step is equivalent to applying the Prüss–Huang result mentioned before the proposition for the semigroups generated by $A + \lambda I$ with $\lambda < \frac{1}{M}$.) ∎

PROPOSITION 2.5. *Suppose that X, $\mathbb{T}$, A, U, and B are as in Definition 2.1. Then the following three statements are equivalent:*

(1) *$\mathbb{T}$ is exponentially stable.*

(2) *(A, B) is optimizable, $\mathbb{C}_0 \subset \rho(A)$, and for some $M > 0$*

$$\left\|(sI - A)^{-1}B\right\|_{\mathcal{L}(U,X)} \le M \qquad \forall\, s \in \mathbb{C}_0.$$

(3) *(A, B) is optimizable, $\omega_0(\mathbb{T}) \le 0$, there exists a dense subset of $\mathbb{R}$, denoted E, such that $iE \subset \rho(A)$, and for some $M > 0$*

$$\left\|(i\omega I - A)^{-1}B\right\|_{\mathcal{L}(U,X)} \le M \qquad \forall\, \omega \in E.$$

Proof. The equivalence of (1) and (2) is exactly Proposition 5.1 in Weiss and Rebarber [33]. It is easy to see that (1) implies (3) with $E = \mathbb{R}$ (by also using (2) and limits as $s \to i\omega$). Now suppose that (3) holds (this implies $\mathbb{C}_0 \subset \rho(A)$). We argue as in the proof of [33, Proposition 5.1], obtaining formula (5.1) from [33] valid for all $s \in \mathbb{C}_0$. Taking limits, we see that this formula holds also with $i\omega$ in place of $s \in \mathbb{C}_0$, where $\omega \in E$. Continuing to reason as in [33], we obtain that $(i\omega I - A)^{-1}$ is uniformly bounded as a function of $\omega \in E$. Since $\omega_0(\mathbb{T}) \le 0$, we can apply Proposition 2.4 to conclude that $\mathbb{T}$ is exponentially stable. ∎

Observability (or dually, controllability) and strong stability concepts are linked to properties of Lyapunov equations, and we state in dual form the following result from section 3 of Hansen and Weiss [12].

PROPOSITION 2.6. *Let A be the generator of the strongly continuous semigroup $\mathbb{T}$ on X, and let $C \in \mathcal{L}(X_1, Y)$ be an admissible observation operator for $\mathbb{T}$. Then the following three statements are equivalent:*

(a) *There exist operators $\Pi \in \mathcal{L}(X)$, $\Pi \ge 0$, which satisfy the following equation:*

$$A^*\Pi z + \Pi A z = -C^*Cz \qquad \forall\, z \in \mathcal{D}(A). \tag{2.5}$$

(b) *C is an infinite-time admissible observation operator for $\mathbb{T}$.*
(c) *There exists an operator $P \in \mathcal{L}(X)$ such that for any $z \in \mathcal{D}(A)$*

$$Pz = \lim_{\tau \to \infty} \int_0^\tau \mathbb{T}_t^* C^* C \mathbb{T}_t z \, \mathrm{d}t. \tag{2.6}$$

Moreover, if C is infinite-time admissible, then the following statements hold:
(d) *P from (2.6) is the smallest nonnegative solution of (2.5).*
(e) *If P is invertible, then $\mathbb{T}$ is strongly stable.*
(f) *If $\mathbb{T}$ is strongly stable, then P is the unique self-adjoint solution of (2.5).*
(g) *If $\mathbb{T}$ is uniformly bounded and $P > 0$, then $\mathbb{T}$ is weakly stable.*

The operator P introduced above is called *the observability Gramian* of (A, C), and (2.5) is called a *Lyapunov equation*. Note that, in terms of the operator Ψ from (2.3), we have $P = \Psi^* \Psi$. The following is well known and easy to prove.

PROPOSITION 2.7. *Suppose that C is an infinite-time admissible observation operator for the semigroup $\mathbb{T}$ generated by A. Then (A, C) is approximately observable in infinite time if and only if $P > 0$ (where P is the observability Gramian of (A, C)).*

The *controllability Gramian* of (A, B) is, by definition, the observability Gramian of (A^*, B^*). Thus, the controllability Gramian of (A, B) is defined by

$$Rx = \lim_{\tau \to \infty} \int_0^\tau \mathbb{T}_t B B^* \mathbb{T}_t^* x \, \mathrm{d}t \qquad \forall\, x \in \mathcal{D}(A^*), \tag{2.7}$$

we have $R = \tilde{\Phi}\tilde{\Phi}^*$, and the Lyapunov equation satisfied by R is

$$RA^* z + ARz = -BB^* z \qquad \forall\, z \in \mathcal{D}(A^*).$$

The dual version of Proposition 2.6 is straightforward.

For more details on Gramians we refer to Hansen and Weiss [12], Jacob and Partington [14], Russell and Weiss [28], and Grabowski [8]. For more details on exact controllability in an operator-theoretic setting we also refer to Avdonin and Ivanov [3], Jacob and Zwart [15], Rebarber and Weiss [27], Tucsnak and Weiss [29], and the references therein. In the PDE setting, the relevant literature is overwhelming, and we mention the books of Lions [20], Lagnese and Lions [17], Bensoussan et al. [6], Komornik [16], and the paper of Bardos, Lebeau, and Rauch [5].

3. Conservative linear systems. Recall from section I.3 that for any well-posed system Σ with input function u, state trajectory x, and output function y,

$$\begin{bmatrix} x(\tau) \\ \mathbf{P}_\tau y \end{bmatrix} = \Sigma_\tau \begin{bmatrix} x(0) \\ \mathbf{P}_\tau u \end{bmatrix}, \tag{3.1}$$

where $\mathbf{P}_\tau$ denotes the truncation of a function to $[0, \tau]$ and

$$\Sigma_\tau = \begin{bmatrix} \mathbb{T}_\tau & \Phi_\tau \\ \Psi_\tau & \mathbb{F}_\tau \end{bmatrix}. \tag{3.2}$$

We denote the input, state, and output spaces of Σ by U, X, and Y, respectively. Then the operators Σ_τ appearing above are bounded from $X \times L^2([0,\tau], U)$ to $X \times L^2([0,\tau], Y)$, which means that for some $c_\tau \geq 0$

$$\|x(\tau)\|^2 + \int_0^\tau \|y(t)\|^2 \, \mathrm{d}t \leq c_\tau^2 \left(\|x(0)\|^2 + \int_0^\tau \|u(t)\|^2 \, \mathrm{d}t \right).$$

As explained in section I.1, the system Σ is *conservative* if the operators Σ_τ are unitary from $X \times L^2([0,\tau],U)$ to $X \times L^2([0,\tau],Y)$. This implies that for any input function $u \in \mathcal{H}^1(0,\infty;U)$ and any initial state $x(0) = x_0 \in X$ with $Ax_0 + Bu(0) \in X$, the function $\|x(t)\|^2$ is in $C^1[0,\infty)$ and

$$\frac{\mathrm{d}}{\mathrm{d}t}\|x(t)\|^2 = \|u(t)\|^2 - \|y(t)\|^2 \qquad \forall\, t \geq 0; \tag{3.3}$$

see Proposition I.4.3. Conversely, if (3.3) holds for both the system Σ and for its dual system Σ^d, then Σ is conservative; see Corollary I.4.4.

PROPOSITION 3.1. *Let Σ be a conservative linear system with input space U, state space X, output space Y, semigroup $\mathbb{T}$, control operator B, observation operator C, and transfer function $\mathbf{G}$. Then the following statements are true:*

(1) *$\mathbb{T}$ is a semigroup of contractions.*
(2) *B is infinite-time admissible.*
(3) *C is infinite-time admissible.*
(4) *$\|\mathbf{G}(s)\| \leq 1$ for all $s \in \mathbb{C}_0$.*

Indeed, the above proposition is an immediate consequence of Proposition I.4.5 (which concerns the larger class of dissipative linear systems). The following proposition is probably well known (especially the equivalence of (2) and (3)), but we are not aware of a reference which states the equivalence of all three conditions.

PROPOSITION 3.2. *With the notation of Proposition* 3.1 *and denoting the generator of $\mathbb{T}$ by A for each $\tau > 0$, the following statements are equivalent:*

(1) *The pair (A,B) is exactly controllable in time τ.*
(2) *The pair (A,C) is exactly observable in time τ.*
(3) *$\|\mathbb{T}_\tau\| < 1$ (in particular, $\mathbb{T}$ is exponentially stable).*

Proof. (3) $\Longrightarrow$ (2) With the notation from (3.2), it is clear that for all $x_0 \in X$

$$\|\mathbb{T}_\tau x_0\|^2 + \|\Psi_\tau x_0\|^2 = \|x_0\|^2 \qquad \forall\, \tau \geq 0. \tag{3.4}$$

If (3) holds, then $\|\mathbb{T}_\tau\|^2 = 1 - \varepsilon^2$ with $\varepsilon > 0$. Now (3.4) implies that $\|\Psi_\tau x_0\|^2 \geq \varepsilon^2\|x_0\|^2$, so that Σ is exactly observable in time τ.

(2) $\Longrightarrow$ (3) If (2) holds, then there exists $\varepsilon > 0$ such that $\|\Psi_\tau x_0\| \geq \varepsilon\|x_0\|$ for all $x_0 \in X$. Now (3.4) implies that $\|\mathbb{T}_\tau x_0\|^2 \leq (1-\varepsilon^2)\|x_0\|^2$; hence $\|\mathbb{T}_\tau\| < 1$.

(3) $\Longleftrightarrow$ (1) (3) is equivalent to the fact that $\|\mathbb{T}_\tau^*\| < 1$. (1) is equivalent to the fact that (A^*,B^*) is exactly observable in time τ. Since the dual system Σ^d is conservative, according to Proposition I.4.2 and the equivalence of (2) and (3) proved earlier, we get that (1) is equivalent to (3). ∎

The equivalence of (1)–(5) in Theorem 1.3 is an immediate consequence of Theorem 1.1, Proposition 1.2, and the following simple result about conservative systems.

PROPOSITION 3.3. *With the notation of Proposition* 3.2*, the following five statements are equivalent:*

(1) *The pair (A,B) is exactly controllable.*
(2) *The pair (A,C) is exactly observable.*
(3) *$\mathbb{T}$ is exponentially stable.*
(4) *The pair (A,B) is optimizable.*
(5) *The pair (A,C) is estimatable.*

Proof. The equivalence of (1)–(3) follows from the previous proposition. It is well known and easy to see that (1) implies (4) and (2) implies (5) (for any well-posed system). Suppose that (5) holds. We know from Proposition 3.1 that C is infinite-time admissible. Now it follows from [33, Proposition 5.5] that (3) holds. The proof of (4)$\Longrightarrow$(3) is similar, by using the dual version of [33, Proposition 5.5]. ∎

Another result linking strong stability, observability, and controllability of conservative systems is the following. It is related to Proposition 6 in [34].

PROPOSITION 3.4. *Let Σ, $\mathbb{T}$, A, B, and C be as in Proposition 3.2. Assume that the intersection $\sigma(A) \cap i\mathbb{R}$ is countable. (This happens, for example, if $(\beta I - A)^{-1}$ is compact for some $\beta \in \rho(A)$.) Then the following seven assertions are equivalent:*

(1) *$\mathbb{T}$ is strongly stable.*

(2) *The pair (A, C) is exactly observable in infinite time.*

(3) *The pair (A, C) is approximately observable in infinite time.*

(4) *$\mathbb{T}$ is weakly stable (equivalently, $\mathbb{T}^*$ is weakly stable).*

(5) *$\mathbb{T}^*$ is strongly stable.*

(6) *The pair (A, B) is exactly controllable in infinite time.*

(7) *The pair (A, B) is approximately controllable in infinite time.*

Proof. (1) $\Longrightarrow$ (2) If $\mathbb{T}$ is strongly stable, then we see from (3.4) that $\lim_{\tau \to \infty} \|\Psi_\tau x_0\| = \|x_0\|$. This implies that Ψ is an isometry from X to $L^2([0,\infty), Y)$.

(2) $\Longrightarrow$ (3) This implication is obvious.

(3) $\Longrightarrow$ (4) The fact that Σ is conservative implies that C is infinite-time admissible. According to Proposition 2.7, (3) means that $P > 0$, where P is the observability Gramian. By the last part of Proposition 2.6, $\mathbb{T}$ is weakly stable.

(4) $\Longrightarrow$ (1) Since $\mathbb{T}$ is weakly stable, A has no eigenvalues on $i\mathbb{R}$. Together with the assumption that $\sigma(A) \cap i\mathbb{R}$ is countable, this means that the conditions of the famous stability theorem of Arendt and Batty [2] are satisfied. According to this theorem, $\mathbb{T}$ is strongly stable.

(4) $\Longleftrightarrow$ (5) $\Longleftrightarrow$ (6) $\Longleftrightarrow$ (7) This is similar to the equivalence of (1)–(4) but with the dual system Σ^d in place of Σ. (Recall that Σ^d is also conservative.) ∎

4. Proof of the main results. In this section we prove Theorems 1.3, 1.4, 1.5 as well as other related results. We use the assumptions and the notation from section 1: The conservative linear system Σ is the one constructed in Theorem 1.1 from the operators $A_0 \in \mathcal{L}(H_1, H)$ and $C_0 \in \mathcal{L}(H_{\frac{1}{2}}, U)$. The spaces H_α with $\alpha \in \mathbb{R}$ are constructed from the fractional powers of A_0. The notation $\|z\|_\alpha$ means the norm of z in H_α; in particular, $\|z\|_0$ is the norm of z in H. We put $B_0 = C_0^*$. The operators A and B are defined in (1.7), (1.8) and $\overline{C}$ is defined in (1.9). C is the restriction of $\overline{C}$ to $\mathcal{D}(A)$. The semigroup of contractions generated by A on $X = H_{\frac{1}{2}} \times H$ is denoted by $\mathbb{T}$, and the transfer function of Σ is denoted by $\mathbf{G}$. Recall also the $\mathcal{L}(H_{-\frac{1}{2}}, H_{\frac{1}{2}})$-valued function $V(s)$ from (1.10) and the space Z_0 defined after (1.5).

PROPOSITION 4.1. *With the above notation, if $\mathbb{T}$ is exponentially stable, then denoting $M = \sup_{\omega \in \mathbb{R}} \|(i\omega I - A)^{-1}\|_{\mathcal{L}(X)}$ we have for every $z \in H_{\frac{1}{2}}$*

$$\|(\omega^2 I - A_0)z\|_{-\frac{1}{2}} + \frac{\omega}{2}\|B_0 C_0 z\|_{-\frac{1}{2}} \geq \frac{1}{M}\|z\|_0 \qquad \forall\, \omega \in [0,\infty).$$

Proof. We know from Proposition I.5.3 that on $H_{\frac{1}{2}} \times H_{-\frac{1}{2}}$ (in particular, on X) we have for all $s \in \rho(A)$ with $s \neq 0$

$$(sI - A)^{-1} = \begin{bmatrix} \frac{1}{s}\,[I - V(s)A_0] & V(s) \\ -V(s)A_0 & sV(s) \end{bmatrix}. \tag{4.1}$$

For $s = 0$ the formula remains valid if we replace the left upper block in the matrix with $\frac{1}{2}A_0^{-1}B_0C_0$; see (I.5.4). Since $\mathbb{T}$ is exponentially stable, any point $i\omega$ (with

$\omega \in \mathbb{R}$) is in $\rho(A)$ and $(i\omega I - A)^{-1}$ is uniformly bounded. Looking at the left lower block of $(i\omega I - A)^{-1}$, we have

$$\sup_{\omega\in\mathbb{R}} \|V(i\omega)A_0^{\frac{1}{2}}\|_{\mathcal{L}(H)} = \sup_{\omega\in\mathbb{R}} \|V(i\omega)A_0\|_{\mathcal{L}(H_{\frac{1}{2}},H)} \le \sup_{\omega\in\mathbb{R}} \|(i\omega I - A)^{-1}\|_{\mathcal{L}(X)} = M.$$

The last estimate means that for any $g \in H$ and any $\omega \in \mathbb{R}$,

$$\|V(i\omega)A_0^{\frac{1}{2}} g\|_0 \le M\|g\|_0 .$$

If we choose $g = A_0^{-\frac{1}{2}}\left(-\omega^2 I + A_0 + \frac{i\omega}{2}B_0C_0\right)z$ with $z \in H_{\frac{1}{2}}$ fixed, then we get that for all $\omega \in \mathbb{R}$

$$M\left\|A_0^{-\frac{1}{2}}\left(-\omega^2 I + A_0 + \frac{i\omega}{2}B_0C_0\right)z\right\|_0 \ge \|z\|_0 .$$

From here, using the triangle inequality, we get the estimate in the proposition. ∎

Proof of Proposition 1.4. Recall that the inner product on X is defined by

$$\left\langle \begin{bmatrix} z_1 \\ w_1 \end{bmatrix}, \begin{bmatrix} z_2 \\ w_2 \end{bmatrix} \right\rangle_X = \langle A_0^{\frac{1}{2}} z_1, A_0^{\frac{1}{2}} z_2\rangle_H + \langle w_1, w_2\rangle_H .$$

In what follows, we drop the subscript H when writing the inner product on H (but we use subscripts for other spaces). We define a new inner product on X by

$$\begin{aligned}\left\langle \begin{bmatrix} z_1 \\ w_1 \end{bmatrix}, \begin{bmatrix} z_2 \\ w_2 \end{bmatrix} \right\rangle_{\text{new}} &= \langle A_0^{\frac{1}{2}} z_1, A_0^{\frac{1}{2}} z_2\rangle + \delta\langle w_1, z_2\rangle + \delta\langle z_1, w_2\rangle + \langle w_1, w_2\rangle \\ &= \left\langle \begin{bmatrix} I & \delta A_0^{-1} \\ \delta I & I \end{bmatrix} \begin{bmatrix} z_1 \\ w_1 \end{bmatrix}, \begin{bmatrix} z_2 \\ w_2 \end{bmatrix} \right\rangle_X ,\end{aligned}$$

where $\delta > 0$ is such that $\delta\|A_0^{-\frac{1}{2}}\| < 1$. (Later we shall impose further restrictions on δ.) The 2×2 matrix J appearing above defines a self-adjoint and positive bounded operator on X. Indeed, $J \ge 0$ follows from the Cauchy–Schwarz inequality:

$$\begin{aligned}\left\langle \begin{bmatrix} I & \delta A_0^{-1} \\ \delta I & I \end{bmatrix} \begin{bmatrix} z \\ w \end{bmatrix}, \begin{bmatrix} z \\ w \end{bmatrix} \right\rangle_X &\ge \|A_0^{\frac{1}{2}} z\|^2 - 2\delta\|z\| \cdot \|w\| + \|w\|^2 \\ &\ge \|A_0^{\frac{1}{2}} z\|^2 - 2\delta\|A_0^{-\frac{1}{2}}\| \cdot \|A_0^{\frac{1}{2}} z\| \cdot \|w\| + \|w\|^2 \\ &\ge \|A_0^{\frac{1}{2}} z\|^2 - 2\|A_0^{\frac{1}{2}} z\| \cdot \|w\| + \|w\|^2 \\ &= \left(\|A_0^{\frac{1}{2}} z\| - \|w\|\right)^2 \ge 0.\end{aligned}$$

It is easy to check that J is boundedly invertible, hence $J > 0$, which shows that our definition of a new inner product is correct, and the new norm on X defined by $\|x\|_{\text{new}} = \sqrt{\langle x, x\rangle_{\text{new}}}$ is equivalent to the original norm. Thus, it will suffice to prove that $\mathbb{T}$ is exponentially stable with respect to the new norm.

We shall estimate $\langle Ax, x\rangle_{\text{new}}$, where $x = \left[\begin{smallmatrix} z \\ w \end{smallmatrix}\right] \in \mathcal{D}(A)$. We have

$$\operatorname{Re}\langle Ax, x\rangle_{\text{new}} = \operatorname{Re}\langle (J - I)Ax, x\rangle_X + \operatorname{Re}\langle Ax, x\rangle_X .$$

We know from (I.5.3) that $\operatorname{Re}\langle Ax,x\rangle_X = -\frac{1}{2}\|C_0w\|_U^2$. Computing the product $(J-I)A$, we get that

$$\operatorname{Re}\langle Ax,x\rangle_{\text{new}} = -\delta\left\|A_0^{\frac{1}{2}}z\right\|^2 - \frac{\delta}{2}\operatorname{Re}\langle C_0w,C_0z\rangle_U + \delta\|w\|^2 - \frac{1}{2}\|C_0w\|_U^2.$$

Now remember that C_0 is bounded from below, so that $\|w\| \le \frac{1}{c}\|C_0w\|_U$. Therefore, the above estimate and the Cauchy–Schwarz inequality imply

$$\operatorname{Re}\langle Ax,x\rangle_{\text{new}} \le -\delta\left\|A_0^{\frac{1}{2}}z\right\|^2 + \frac{\delta}{2}\|C_0w\|_U\cdot\|C_0z\|_U - \left(\frac{1}{2}-\frac{\delta}{c^2}\right)\|C_0w\|_U^2.$$

Let $k>0$ be such that $\|C_0z\|_U \le k\|A_0^{\frac{1}{2}}z\|$ for all $z\in H_{\frac{1}{2}}$. Then

$$\operatorname{Re}\langle Ax,x\rangle_{\text{new}} \le -\delta\left\|A_0^{\frac{1}{2}}z\right\|^2 + \frac{\delta k}{2}\|C_0w\|_U\cdot\|A_0^{\frac{1}{2}}z\| - \left(\frac{1}{2}-\frac{\delta}{c^2}\right)\|C_0w\|_U^2.$$

The right-hand side above is a quadratic form in the two numbers $\|A_0^{\frac{1}{2}}z\|$ and $\|C_0w\|_U$. The matrix of this quadratic form is

$$Q = -\begin{bmatrix} \delta & -\frac{\delta k}{4} \\ -\frac{\delta k}{4} & \frac{1}{2}-\frac{\delta}{c^2}\end{bmatrix}.$$

This Q will be negative definite if

$$\frac{1}{2}-\frac{\delta}{c^2}>0 \quad\text{and}\quad 16\left(\frac{1}{2}-\frac{\delta}{c^2}\right) > \delta k^2.$$

Both of these conditions can be satisfied if we choose δ sufficiently small. Suppose that δ has been correctly chosen so that $Q\le -\gamma I$ for some $\gamma>0$. Then we obtain

$$\begin{aligned}\operatorname{Re}\langle Ax,x\rangle_{\text{new}} &\le -\gamma\left(\left\|A_0^{\frac{1}{2}}z\right\|^2 + \|C_0w\|_U^2\right)\\ &\le -\gamma\left(\left\|A_0^{\frac{1}{2}}z\right\|^2 + c^2\|w\|^2\right) \le -\gamma\min(1,c^2)\|x\|_X^2.\end{aligned}$$

Recall that the two norms on X are equivalent so that $\|x\|_X \ge m\|x\|_{\text{new}}$ for some $m>0$. Denoting $\eta = \gamma\min(1,c^2)m^2$ (so that $\eta>0$), we obtain

$$\operatorname{Re}\langle Ax,x\rangle_{\text{new}} \le -\eta\|x\|_{\text{new}}^2$$

so that $A+\eta I$ is dissipative with respect to the new inner product. Hence, the growth bound of $\mathbb{T}$ (which does not depend on the norm) is $\omega_0(\mathbb{T})\le -\eta$. □

LEMMA 4.2. *If we define $B_b = \left[\begin{smallmatrix}0\\ I\end{smallmatrix}\right] \in \mathcal{L}(H,X)$, then (A,B_b) is optimizable.*

Proof. Consider a new conservative linear system $\tilde{\Sigma}$ obtained from the same operator A_0 on the same Hilbert space H but with a larger input space $\tilde{U}$ and with C_0 replaced by $\tilde{C}_0$, which are defined as follows:

$$\tilde{U} = U\times H, \qquad \tilde{C}_0 = \begin{bmatrix} C_0 \\ I\end{bmatrix}.$$

Thus, following the standard construction from section 1, B_0 will be replaced by $\tilde{B}_0 = \tilde{C}_0^* = [B_0\ I]$. According to (1.7), the semigroup $\tilde{\mathbb{T}}$ of $\tilde{\Sigma}$ is generated by

$$\tilde{A} = \begin{bmatrix} 0 & I \\ -A_0 & -\frac{1}{2}B_0C_0 - \frac{1}{2}I \end{bmatrix},$$

with $\mathcal{D}(\tilde{A}) = \mathcal{D}(A)$ as defined in (1.8). It is clear that $\tilde{C}_0$ is bounded from below, so that $\tilde{\mathbb{T}}$ is exponentially stable according to Proposition 1.4.

Now consider the system Σ_b with input space H, state space X, and output space H described by

$$\begin{cases} \dot{x}(t) = Ax(t) + B_b u(t), \\ y(t) = B_b^* x(t). \end{cases}$$

Clearly Σ_b is well-posed, since B_b is bounded. The static output feedback $u = -\frac{1}{2}y$ applied to this system leads to a closed-loop system whose semigroup generator is $A - \frac{1}{2}B_b^*B_b = \tilde{A}$, which (as we already know) is exponentially stable. In particular, it follows that for any initial state $x_0 \in X$, the functions u and x defined by $u(t) = -\frac{1}{2}B_b^*\tilde{\mathbb{T}}_t x_0$ and $x(t) = \tilde{\mathbb{T}}_t x_0$ are both in L^2. Thus, (A, B_b) is optimizable. □

Proof of Theorem 1.3. According to Theorem 1.1 and Proposition 1.2, Σ is a conservative linear system with semigroup generator A, control operator B, and observation operator C. Now the equivalence of (1)–(5) in Theorem 1.3 follows from Proposition 3.3. It is also easy to see that (3) implies (6), (7), (8), and (9). Indeed, if $\mathbb{T}$ is exponentially stable, then $(sI - A)^{-1}$ exists and is uniformly bounded on $\mathbb{C}_\alpha$ for some $\alpha < 0$; see the proof of Proposition 2.4. Looking at the right column of $(sI - A)^{-1}$ in (4.1), we obtain that (6)–(9) all hold.

We prove the equivalence of (6) and (7). Suppose that (7) is false; i.e., there is a sequence (s_n) in $\mathbb{C}_0$ such that $\|s_nV(s_n)\| \to \infty$. Since $\mathbb{T}$ is uniformly bounded, $\|(sI - A)^{-1}\|$ is bounded on any right half-plane $\mathbb{C}_\gamma$ with $\gamma > 0$. Since $sV(s)$ is one of the entries of $(sI - A)^{-1}$, it follows that for large n the sequence (s_n) must be outside $\mathbb{C}_\gamma$. Since this is true for each $\gamma > 0$, we must have $\operatorname{Re} s_n \to 0$. Since $0 \in \rho(A)$, $\|sV(s)\|$ is bounded on a neighborhood of 0. Thus, without loss of generality we may assume that $|\operatorname{Im} s_n| \geq \varepsilon > 0$.

By the uniform boundedness theorem, there exists a vector $x \in H$ such that

$$\lambda_n = \|s_nV(s_n)x\| \to \infty.$$

Denote $z_n = \frac{1}{\lambda_n}V(s_n)x$; then clearly $z_n \in H_{\frac{1}{2}}$, $\|s_nz_n\| = 1$ (hence $\|z_n\|$ is bounded), and

$$\frac{1}{\lambda_n}\, x = s_n^2 z_n + A_0 z_n + \frac{s_n}{2}B_0C_0z_n \to 0 \ \text{ in } \ H.$$

Taking inner products with z_n, we obtain

$$\frac{1}{\lambda_n}\,\langle x, z_n\rangle = s_n^2\|z_n\|^2 + \left\|A_0^{\frac{1}{2}}z_n\right\|^2 + \frac{s_n}{2}\|C_0z_n\|^2 \to 0. \tag{4.2}$$

Looking here only at the imaginary parts and dividing by $\operatorname{Im} s_n$, we obtain

$$2(\operatorname{Re} s_n)\|z_n\|^2 + \frac{1}{2}\|C_0z_n\|^2 \to 0; \tag{4.3}$$

in particular, $\|C_0 z_n\| \to 0$. Now we look at the real parts of the terms in (4.2):

$$-|s_n|^2\|z_n\|^2 + 2(\operatorname{Re} s_n)^2\|z_n\|^2 + \left\|A_0^{\frac{1}{2}} z_n\right\|^2 + \frac{\operatorname{Re} s_n}{2}\|C_0 z_n\|^2 \to 0. \tag{4.4}$$

Recalling that $\operatorname{Re} s_n \to 0$ and $\|s_n z_n\| = 1$, we conclude that $\lim_{n\to\infty} \|A_0^{\frac{1}{2}} z_n\| = 1$. Thus, $\lim_{n\to\infty} \frac{1}{\lambda_n}\|A_0^{\frac{1}{2}} V(s_n)x\| = 1$ so that $\|A_0^{\frac{1}{2}} V(s_n)\| \to \infty$. We have obtained that assertion (6) is false, so (6) implies (7). The proof of the fact that (7) implies (6) is similar with the following modifications: x and λ_n are now chosen such that $\lambda_n = \|A_0^{\frac{1}{2}} V(s_n)x\| \to \infty$. We take again $z_n = \frac{1}{\lambda_n} V(s_n)x$, and now $\|A_0^{\frac{1}{2}} z_n\| = 1$ (instead of $\|s_n z_n\| = 1$). Now the reasoning up to (4.4) remains the same, and from (4.4) we conclude that $\lim_{n\to\infty} \|s_n z_n\| = 1$, which implies that $\|s_n V(s_n)\| \to \infty$.

To prove the equivalence of (8) and (9), we argue similarly as in the proof of the equivalence of (6) and (7), but now $\operatorname{Re} s_n = 0$. This makes the proof simpler, since now we do not need (4.3) and in (4.4) two terms disappear.

We prove that (6) implies (3). If (6) (and hence also (7)) holds, then we see from (4.1) that $(sI - A)^{-1}B_b$ is uniformly bounded on $\mathbb{C}_0$, where B_b is the operator from Lemma 4.2. Since, by the same lemma, (A, B_b) is optimizable, we can apply Proposition 2.5 (the equivalence of points (1) and (2) in that proposition) to conclude that $\mathbb{T}$ is exponentially stable, i.e., (3) holds.

Finally, we prove that (8) implies (3). If (8) (and hence also (9)) holds, then we see from (4.1) that $(i\omega I - A)^{-1}B_b$ is uniformly bounded for $\omega \in E$, where B_b is the operator from Lemma 4.2. Since (A, B_b) is optimizable, and since $\mathbb{T}$ is uniformly bounded, the conditions in point (3) of Proposition 2.5 are satisfied. According to Proposition 2.5, $\mathbb{T}$ is exponentially stable. □

In order to prove Theorem 1.5, we have to prove several preliminary results. We denote by $\sigma_p(A)$ the set of eigenvalues (the point spectrum) of A.

LEMMA 4.3. *If $\lambda \in \sigma_p(A)$ and $x \in \mathcal{D}(A)$ is a corresponding eigenvector (i.e., $(\lambda I - A)x = 0$ and $x \neq 0$), then x is of the form*

$$x = \begin{bmatrix} z \\ \lambda z \end{bmatrix}, \qquad z \in Z_0, \tag{4.5}$$

where $Z_0 = H_1 + A_0^{-1}B_0 U$ and

$$\left(\lambda^2 I + A_0 + \frac{\lambda}{2} B_0 C_0\right) z = 0. \tag{4.6}$$

If $\lambda \notin \mathbb{R}$, then this implies

$$\|C_0 z\|^2 = 4|\operatorname{Re}\lambda| \cdot \|z\|^2, \qquad \langle A_0 z, z\rangle = |\lambda|^2 \cdot \|z\|^2. \tag{4.7}$$

Proof. The formulas (4.5) and (4.6) are an immediate consequence of $\mathcal{D}(A) \subset Z_0 \times H_{\frac{1}{2}}$ (which follows from (1.8)) and of $(\lambda I - A)x = 0$. If we take the scalar product of the sides of (4.6) with z and use the extension of the scalar product to the duality pairing between $H_{-\frac{1}{2}}$ and $H_{\frac{1}{2}}$, we obtain

$$\left\langle \left(\lambda^2 I + A_0\right) z, z\right\rangle + \frac{\lambda}{2}\|C_0 z\|^2 = 0.$$

Since $\mathbb{T}$ is a contraction semigroup, λ must be in the closed left half-plane. Denoting $\lambda = -\sigma + i\omega$ with $\sigma \geq 0$ and $\omega \in \mathbb{R}$, this means

$$\left\langle \left((\sigma^2 - \omega^2)I + A_0\right) z, z\right\rangle - 2i\sigma\omega\|z\|^2 + \frac{-\sigma + i\omega}{2}\|C_0 z\|^2 = 0. \tag{4.8}$$

Looking at the imaginary part of this, we see that $\omega \neq 0$ implies

$$-2\sigma\|z\|^2 + \frac{1}{2}\|C_0 z\|^2 = 0\,,$$

which is the same as the first equality in (4.7). Now we look at the real part of (4.8), using the expression for $\|C_0 z\|^2$ that we have just found, obtaining (after a short computation) the second equality in (4.7). □

LEMMA 4.4. *Suppose that $\omega \in \mathbb{R}$ is such that $i\omega \in \sigma_p(A)$. Then also $-i\omega \in \sigma_p(A)$ and $\omega^2 \in \sigma_p(A_0)$. In this case, z from (4.5) is an eigenvector of A_0 corresponding to the eigenvalue ω^2 (in particular, $z \in H_\alpha$ for all $\alpha > 0$) and we have $C_0 z = 0$.*

Proof. Suppose that $i\omega \in \sigma_p(A)$, and let $z \in Z_0$ be the first component of a corresponding eigenvector as in (4.5). We know from (I.5.4) that $0 \in \rho(A)$ so that $\omega \neq 0$. According to the first part of (4.7) we have $C_0 z = 0$. Now (4.6) (with $\lambda = i\omega$) shows that z is an eigenvector of A_0 corresponding to the eigenvalue ω^2. It is now easy to see that the vector with components z and $-i\omega z$ is also an eigenvector of A corresponding to the eigenvalue $-i\omega$. □

We denote by $\sigma_a(A)$ the set of those $\lambda \in \sigma(A)$ for which λ is not an eigenvalue of A, but $\lambda I - A$ is not bounded from below. In other words, $\lambda \in \sigma_a(A)$ if $\lambda \notin \sigma_p(A)$ and there exists a sequence (x_n) in $\mathcal{D}(A)$ with

$$\|x_n\|_X = 1 \quad \text{and} \quad \lim_{n\to\infty} \|(\lambda I - A)x_n\|_X = 0\,. \tag{4.9}$$

LEMMA 4.5. *If $\omega \in \mathbb{R}$ is such that $i\omega \in \sigma(A)$, then*

$$i\omega \in \sigma_p(A) \cup \sigma_a(A)\,.$$

Proof. Suppose that $i\omega \in \sigma(A)$. We prove that $i\omega \in \sigma_p(A) \cup \sigma_a(A)$ by showing that the contrary statement leads to a contradiction. Indeed, the contrary statement means that $i\omega I - A$ is bounded from below. In this case, the range of $i\omega I - A$ is not dense in X (because if it were dense, then it were all of X, and hence $i\omega I - A$ would have a bounded inverse). Let N be the orthogonal complement of the range of $i\omega I - A$; then it is easy to see that N is invariant under $\mathbb{T}^*$: $\mathbb{T}_t^* N \subset N$ for all $t \geq 0$. Considering the restriction of $\mathbb{T}^*$ to N, we see that $\mathcal{D}(A^*) \cap N$ must be dense in N, so that, in particular, there exist elements $q \in \mathcal{D}(A^*) \cap N$ with $q \neq 0$. From the definition of N we now see that for such q we have $(-i\omega I - A^*)q = 0$ so that $-i\omega \in \sigma_p(A^*)$. Introduce the isomorphism $J \in \mathcal{L}(X)$ defined by the matrix

$$J = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}.$$

We have $J^{-1} = J$ and $A^* = JAJ$; see the fourth step in the proof of Theorems I.1.1 and I.1.3 (in section I.6). Thus, A and A^* have the same eigenvalues so that $-i\omega \in \sigma_p(A)$. According to Lemma 4.4, we obtain that $i\omega \in \sigma_p(A)$, which contradicts our "contrary statement" at the beginning of this proof. □

LEMMA 4.6. *If $\omega \in \mathbb{R}$ is such that $i\omega \in \sigma_a(A)$, then also $-i\omega \in \sigma_a(A)$ and, moreover, $\omega^2 \in \sigma_p(A_0) \cup \sigma_a(A_0)$.*

Proof. Assume that $i\omega \in \sigma_a(A)$ so that for some sequence (x_n) in $\mathcal{D}(A)$ we have (4.9) (with $\lambda = i\omega$). Denoting $x_n = \left[\begin{smallmatrix} z_n \\ w_n \end{smallmatrix}\right]$ and $(i\omega I - A)x_n = \left[\begin{smallmatrix} \nu_n \\ \varepsilon_n \end{smallmatrix}\right]$ so that $\nu_n \to 0$ (in $H_{\frac{1}{2}}$) and $\varepsilon_n \to 0$ (in H), we have

$$\begin{bmatrix} i\omega I & -I \\ A_0 & i\omega I + \frac{1}{2}B_0 C_0 \end{bmatrix} \begin{bmatrix} z_n \\ w_n \end{bmatrix} = \begin{bmatrix} \nu_n \\ \varepsilon_n \end{bmatrix}.$$

From the first row we have $w_n = i\omega z_n - \nu_n$. Substituting this into the equation representing the second row, we get

$$A_0 z_n - \omega^2 z_n - i\omega\,\nu_n + \frac{i\omega}{2} B_0 C_0 z_n - \frac{1}{2} B_0 C_0 \nu_n = \varepsilon_n\,. \tag{4.10}$$

The two sides of this equation are in H, but some terms are in $H_{-\frac{1}{2}}$. Taking the scalar product with z_n and using the duality pairing between $H_{-\frac{1}{2}}$ and $H_{\frac{1}{2}}$, we get

$$\left\langle \left(A_0 - \omega^2 I\right) z_n, z_n \right\rangle - i\omega \langle \nu_n, z_n\rangle + \frac{i\omega}{2} \|C_0 z_n\|^2 - \frac{1}{2} \langle C_0\nu_n, C_0 z_n\rangle = \langle \varepsilon_n, z_n\rangle\,,$$

which shows that

$$\lim_{n\to\infty} \left[\left\langle \left(A_0 - \omega^2 I\right) z_n, z_n \right\rangle + \frac{i\omega}{2} \|C_0 z_n\|^2 \right] = 0\,.$$

Remember that A is invertible (see (I.5.4)) so that $\omega \neq 0$. Taking the imaginary part of the last limit, we conclude that $\lim_{n\to\infty} C_0 z_n = 0$. Now going back to (4.10), we conclude that

$$\lim_{n\to\infty} \left\| \left(\omega^2 I - A_0\right) z_n \right\|_{-\frac{1}{2}} = 0\,. \tag{4.11}$$

Now recall from (4.9) that

$$\|x_n\|_X^2 = \|z_n\|_{\frac{1}{2}}^2 + \|w_n\|_0^2 = 1\,.$$

Since $w_n = i\omega z_n - \nu_n$ with $\lim_{n\to\infty} \|\nu_n\|_{\frac{1}{2}} = 0$, we have $\lim_{n\to\infty} \left(\|w_n\|_0^2 - \omega^2 \|z_n\|_0^2\right) = 0$ so that $\lim_{n\to\infty} (\|z_n\|_{\frac{1}{2}}^2 + \omega^2 \|z_n\|_0^2) = 1$. Since $\|z_n\|_0 \leq \|A_0^{-\frac{1}{2}}\| \cdot \|z_n\|_{\frac{1}{2}}$ and $\|A_0^{-\frac{1}{2}}\|^2 = \|A_0^{-1}\|$, we obtain $\liminf_{n\to\infty} \left(1 + \omega^2 \|A_0^{-1}\|\right) \cdot \|z_n\|_{\frac{1}{2}}^2 \geq 1$ so that the sequence $(\|z_n\|_{\frac{1}{2}})$ is eventually bounded from below:

$$\|z_n\|_{\frac{1}{2}} \geq m > 0 \quad \text{for} \quad n \geq n_0\,.$$

We have for $n \geq n_0$

$$m \leq \|z_n\|_{\frac{1}{2}} = \|A_0 z_n\|_{-\frac{1}{2}} = \|(\omega^2 I - A_0) z_n - \omega^2 z_n\|_{-\frac{1}{2}}$$

$$\leq \|(\omega^2 I - A_0) z_n\|_{-\frac{1}{2}} + \omega^2 \|z_n\|_{-\frac{1}{2}}\,.$$

Now (4.11) implies that the sequence $(\|z_n\|_{-\frac{1}{2}})$ is eventually bounded from below. This together with (4.11) implies that $\omega^2 I - A_0$ is not bounded from below as an (unbounded) operator on $H_{-\frac{1}{2}}$ (with domain $H_{\frac{1}{2}}$). Since $\omega^2 I - A_0$ as an operator on $H_{-\frac{1}{2}}$ is isomorphic to $\omega^2 I - A_0$ as an operator on H (with domain H_1), we conclude that the latter is also not bounded from below. Thus, $\omega^2 \in \sigma_p(A_0) \cup \sigma_a(A_0)$.

It remains to prove that $-i\omega \in \sigma_a(A)$. Define

$$\xi_n = \begin{bmatrix} z_n \\ -w_n \end{bmatrix} \in H_{\frac{1}{2}} \times H_{\frac{1}{2}}$$

so that $\|\xi_n\|_X = \|x_n\|_X = 1$. We will use the extension of A which maps X into X_{-1}. In particular, this extension maps $H_{\frac{1}{2}} \times H_{\frac{1}{2}} \subset X$ into $H_{\frac{1}{2}} \times H_{-\frac{1}{2}} \subset X_{-1}$, as it is easy to see (see Proposition I.5.2 for the inclusion $H_{\frac{1}{2}} \times H_{-\frac{1}{2}} \subset X_{-1}$). We have

$$(-i\omega I - A)\xi_n = \begin{bmatrix} -i\omega I & -I \\ A_0 & -i\omega I + \frac{1}{2}B_0C_0 \end{bmatrix} \begin{bmatrix} z_n \\ -w_n \end{bmatrix} = \begin{bmatrix} -\nu_n \\ \varphi_n \end{bmatrix},$$

where $\varphi_n \in H_{-\frac{1}{2}}$ is given by a formula similar to (4.10):

$$\varphi_n = (A_0 - \omega^2 I)z_n - i\omega\nu_n - \frac{i\omega}{2}B_0C_0z_n + \frac{1}{2}B_0C_0\nu_n.$$

Remember that $\nu_n \to 0$ (in $H_{\frac{1}{2}}$) and $C_0z_n \to 0$ (in U). Using also (4.11), we conclude from the above formula that $\varphi_n \to 0$ (in $H_{-\frac{1}{2}}$). From here we see that

$$\lim_{n\to\infty}(-i\omega I - A)\xi_n = 0 \quad \text{in} \quad H_{\frac{1}{2}} \times H_{-\frac{1}{2}}.$$

Since the sequence (ξ_n) is bounded from below in $H_{\frac{1}{2}} \times H_{\frac{1}{2}}$, it follows that $(-i\omega I - A)$ is not bounded from below (as an operator from $H_{\frac{1}{2}} \times H_{\frac{1}{2}}$ to $H_{\frac{1}{2}} \times H_{-\frac{1}{2}}$). According to point (1) of Proposition I.5.3, it follows that $-i\omega \in \sigma(A)$. Now we can apply Lemma 4.5 to conclude that $-i\omega \in \sigma_p(A) \cup \sigma_a(A)$. If we would have $-i\omega \in \sigma_p(A)$, then it would follow from Lemma 4.4 that also $i\omega \in \sigma_p(A)$, which would contradict our assumption that $i\omega \in \sigma_a(A)$. Thus, we must have $-i\omega \in \sigma_a(A)$. □

Proof of Theorem 1.5. According to Lemmas 4.4, 4.5, and 4.6, if $\omega \in \mathbb{R}$ is such that $i\omega \in \sigma(A)$, then $\omega^2 \in \sigma(A_0)$. Thus, if $\sigma(A_0)$ is countable, as assumed in the theorem, then also $\sigma(A) \cap i\mathbb{R}$ must be countable, as required in Proposition 3.4. Now the equivalence of points (1)–(7) in the theorem follows from Proposition 3.4.

It remains to prove the equivalence between point (8) and the other points. Suppose that (8) holds. We claim that A has no eigenvalues on the imaginary axis. Indeed, if $\omega \in \mathbb{R}$ were such that $i\omega \in \sigma_p(A)$, then according to Lemma 4.4 $\omega^2 \in \sigma_p(A_0)$ and for a corresponding eigenvector $z \in H_1$ we would have $C_0z = 0$, which would contradict (8). Now we can apply the main theorem of Arendt and Batty [2]: our earlier claim together with the fact that $\sigma(A) \cap i\mathbb{R}$ is countable implies that $\mathbb{T}$ is strongly stable. Thus, point (1) of the theorem holds and with it all the others.

Conversely, suppose that point (8) is false, i.e., there exists an $\omega \in \mathbb{R}$ such that $\omega^2 \in \sigma_p(A_0)$, and for a corresponding eigenvector $z \in H_1$ we have $C_0z = 0$. Let x be defined by (4.5) with $\lambda = i\omega$. Then it is easy to verify that $x \in \mathcal{D}(A)$ and $Ax = i\omega x$, whence $\mathbb{T}_t x = e^{i\omega t}x$. This shows that $\mathbb{T}$ is not strongly stable, so in this case the points (1)–(7) are all false. □

5. Examples. The aim of this section is to apply the general stability results derived earlier to some models based on PDEs. We will consider an Euler–Bernoulli beam with an exponentially stabilizing feedback acting in one point followed by an n-dimensional wave equation with boundary control and a nonlocal feedback term entering a Dirichlet boundary condition.

5.1. An Euler–Bernoulli beam with pointwise control. The physical system that we have in mind consists of two rigidly joined beams with both velocity and angular velocity damping at the joint. The other end of both beams is hinged. The inputs are a force and a torque acting at the joint (in addition to the damping force and torque), and the measurements depend linearly on the velocity and the angular

velocity at the joint as well as on the inputs. The fact that we apply both a force and a torque feedback at the joint prevents the well-known lack of robustness of stability concerning the location of the joint: we obtain exponential stability for any location. For the case when only force feedback or only torque feedback is applied, lack of robustness of exponential stability was demonstrated in Rebarber [26]. (We refer to [26] for earlier references on this subject.)

Our system consists of two homogeneous Euler–Bernoulli beams situated along the intervals $[0,\xi]$ and $[\xi,\pi]$ with the joint at the point $\xi \in (0,\pi)$. Denoting by $[f]_\xi$ the jump of the function f at $x=\xi$, we get the equations

$$\frac{\partial^2 z}{\partial t^2}+\frac{\partial^4 z}{\partial x^4}=0,\quad x\in(0,\pi)\setminus\{\xi\},\quad [z]_\xi=0,\quad \left[\frac{\partial z}{\partial x}\right]_\xi=0, \tag{5.1}$$

$$\left[\frac{\partial^3 z}{\partial x^3}\right]_\xi+\frac{\alpha^2}{2}\frac{\partial z}{\partial t}(\xi,t)=\alpha u_1(t),\quad -\left[\frac{\partial^2 z}{\partial x^2}\right]_\xi+\frac{\beta^2}{2}\frac{\partial^2 z}{\partial x\partial t}(\xi,t)=\beta u_2(t), \tag{5.2}$$

$$\left[\frac{\partial^3 z}{\partial x^3}\right]_\xi-\frac{\alpha^2}{2}\frac{\partial z}{\partial t}(\xi,t)=\alpha y_1(t),\quad -\left[\frac{\partial^2 z}{\partial x^2}\right]_\xi-\frac{\beta^2}{2}\frac{\partial^2 z}{\partial x\partial t}(\xi,t)=\beta y_2(t), \tag{5.3}$$

$$z(x,0)=z_0(x),\quad \frac{\partial z}{\partial t}(x,0)=w_0(x),\quad x\in(0,\pi)\setminus\{\xi\}, \tag{5.4}$$

$$z(0,t)=z(\pi,t)=0,\quad \frac{\partial^2 z}{\partial x^2}(0,t)=\frac{\partial^2 z}{\partial x^2}(\pi,t)=0. \tag{5.5}$$

Here, z stands for the transverse displacement of the beam and $\alpha,\beta>0$ are damping coefficients. The external force is αu_1 and the external torque is βu_2, both acting at ξ. The output signals are y_1 and y_2.

An equivalent formulation of (5.1)–(5.5) can be obtained by considering a single homogeneous Euler–Bernoulli beam situated along the interval $[0,\pi]$ with a force and a torque acting at the point $\xi\in(0,\pi)$. (The equivalence is proved in Proposition 5.2 below.) In this case the equations (5.1)–(5.2) are replaced by

$$\begin{aligned}\frac{\partial^2 z}{\partial t^2}+\frac{\partial^4 z}{\partial x^4}+\frac{\alpha^2}{2}\frac{\partial}{\partial t}(z(\xi,t))\,\delta_\xi-\frac{\beta^2}{2}\frac{\partial}{\partial t}\left(\frac{\partial z}{\partial x}(\xi,t)\right)\frac{\mathrm{d}\delta_\xi}{\mathrm{d}x}\\ =\alpha u_1(t)\delta_\xi-\beta u_2(t)\frac{\mathrm{d}\delta_\xi}{\mathrm{d}x}.\end{aligned} \tag{5.6}$$

Here, δ_ξ is the Dirac mass concentrated at ξ and (5.6) is understood as an equation in $\mathcal{D}'(0,\pi)$ (distributions on $(0,\pi)$). The outputs are equivalently given by

$$y_1(t)=-\alpha\frac{\partial z}{\partial t}(\xi,t)+u_1(t),\quad y_2(t)=-\beta\frac{\partial^2 z}{\partial t\,\partial x}(\xi,t)+u_2(t). \tag{5.7}$$

The well-posedness of the system described by (5.4)–(5.7) can be obtained by using Theorem 1.1 if we introduce the appropriate spaces and operators. We start by defining $H=L^2[0,\pi]$ and the operator $A_0:\mathcal{D}(A_0)\to H$ by

$$\mathcal{D}(A_0)=\left\{\phi\in\mathcal{H}^4(0,\pi)\;\middle|\;\phi(0)=\phi(\pi)=0,\ \frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}(0)=\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}(\pi)=0\right\},$$

$$A_0\phi = \frac{\mathrm{d}^4\phi}{\mathrm{d}x^4} \qquad \forall\ \phi \in \mathcal{D}(A_0).$$

It is well known that A_0 is self-adjoint, positive, and boundedly invertible. As in section 1, we put $H_1 = \mathcal{D}(A_0)$ and we introduce the spaces H_α ($\alpha \in \mathbb{R}$) by considering fractional powers of A_0 and duality. A simple calculation shows that

$$H_{\frac{1}{2}} = \mathcal{D}(A_0^{\frac{1}{2}}) = \mathcal{H}^2(0,\pi) \cap \mathcal{H}_0^1(0,\pi)$$

with the norm

$$\|z\|_{\frac{1}{2}}^2 = \int_0^\pi \left|\frac{\mathrm{d}^2 z}{\mathrm{d}x^2}(x)\right|^2 \mathrm{d}x.$$

PROPOSITION 5.1. *The equations* (5.4)–(5.7) *determine a conservative linear system* Σ *with input and output space* $\mathbb{C}^2$ *and with state space* $X = H_{\frac{1}{2}} \times H$. *For* $z_0 \in H_{\frac{1}{2}}$, $w_0 \in H$, *and* $u_1, u_2 \in L^2[0,\infty)$ *the equations* (5.4)–(5.6) *admit a unique solution*

$$z \in BC\left(0,\infty; H_{\frac{1}{2}}\right) \cap BC^1\left(0,\infty; H\right) \cap \mathcal{H}_{loc}^2\left(0,\infty; H_{-\frac{1}{2}}\right).$$

Moreover, we have $z(\xi,\cdot), \frac{\partial z}{\partial x}(\xi,\cdot) \in \mathcal{H}^1(0,\infty)$.

Proof. We take $U = \mathbb{C}^2$, and $B_0 \in \mathcal{L}(U, H_{-\frac{1}{2}})$ is defined by

$$B_0 \begin{bmatrix} \mathrm{v}_1 \\ \mathrm{v}_2 \end{bmatrix} = \alpha\,\mathrm{v}_1\delta_\xi - \beta\,\mathrm{v}_2\frac{\mathrm{d}\delta_\xi}{\mathrm{d}x} \qquad \forall\ \begin{bmatrix} \mathrm{v}_1 \\ \mathrm{v}_2 \end{bmatrix} \in \mathbb{C}^2.$$

By the definition of B_0, for each $\mathrm{v}_1, \mathrm{v}_2 \in \mathbb{C}$, $B_0\left[\begin{smallmatrix} \mathrm{v}_1 \\ \mathrm{v}_2 \end{smallmatrix}\right]$ is a bounded linear functional acting on $H_{\frac{1}{2}}$ so that indeed B_0 maps $\mathbb{C}^2$ into $H_{-\frac{1}{2}}$. The adjoint of B_0 is

$$C_0\phi = B_0^*\phi = \begin{bmatrix} \alpha\phi(\xi) \\ \beta\frac{\mathrm{d}\phi}{\mathrm{d}x}(\xi) \end{bmatrix} \qquad \forall\ \phi \in H_{\frac{1}{2}}. \tag{5.8}$$

Now it is clear that the problem (5.4)–(5.7) can be written in the form (1.1)–(1.3). Thus, this proposition follows from Theorem 1.1. ∎

Note that since Σ is conservative, for every $t \geq 0$ we have

$$\left\|\begin{bmatrix} z(t) \\ \dot{z}(t) \end{bmatrix}\right\|_X^2 - \left\|\begin{bmatrix} z(0) \\ \dot{z}(0) \end{bmatrix}\right\|_X^2 = \int_0^t \left[|u_1(\sigma)|^2 + |u_2(\sigma)|^2\right] \mathrm{d}\sigma - \int_0^t \left[|y_1(\sigma)|^2 + |y_2(\sigma)|^2\right] \mathrm{d}\sigma.$$

The space Z_0 defined after (1.5) (see also Theorem I.1.2) is now given by

$$Z_0 = \left\{ z \in H_{\frac{1}{2}} \;\middle|\; z|_{(0,\xi)} \in \mathcal{H}^4(0,\xi),\ z|_{(\xi,1)} \in \mathcal{H}^4(\xi,1) \right\}.$$

The systems (5.1)–(5.5) and (5.4)–(5.7) are equivalent in the following sense.

PROPOSITION 5.2. *Consider the functions*

$$z \in C\left(0,\infty; Z_0\right) \cap C^1\left(0,\infty; H_{\frac{1}{2}}\right) \cap C^2\left(0,\infty; H\right), \qquad u, y \in C\left(0,\infty; U\right).$$

Then z, u, *and* y *satisfy* (5.1)–(5.5) *if and only if they satisfy* (5.4)–(5.7).

Proof. We denote by $z^{(\mathrm{iii})}$ (respectively, by $z^{(\mathrm{iv})}$) the third (respectively, the fourth) derivative of z computed in $\mathcal{D}'((0,\pi)\setminus\{\xi\})$. Roughly speaking, these derivatives are calculated almost everywhere, ignoring the possible jumps at $x=\xi$. Hence they are, in general, different from $\frac{\mathrm{d}^3 z}{\mathrm{d}x^3}$ and $\frac{\mathrm{d}^4 z}{\mathrm{d}x^4}$ computed in the sense of distributions, i.e., in $\mathcal{D}'(0,\pi)$. We define $L_0 \in \mathcal{L}\left(Z_0, L^2[0,\pi]\right)$ and $G_0 \in \mathcal{L}(Z_0, U)$ by

$$L_0 z = z^{(\mathrm{iv})} \quad \text{and} \quad G_0 z = \begin{bmatrix} \frac{1}{\alpha}\left[z^{\mathrm{iii}}\right]_\xi & -\frac{1}{\beta}\left[\frac{d^2 z}{dx^2}\right]_\xi \end{bmatrix}^T \tag{5.9}$$

(T stands for transpose). Simple calculations show that

$$L_0 = A_0 - B_0 G_0, \quad G_0 H_1 = \{0\}, \quad G_0 A_0^{-1} B_0 = I\,. \tag{5.10}$$

By Theorem I.1.4 we obtain that the functions z, u, y satisfy (5.4)–(5.7) (which are the same, in this example, as (1.1)–(1.3)) if and only if they satisfy

$$\ddot{z}(t) + L_0 z(t) = 0\,, \quad G_0 z(t) + \frac{1}{2} C_0 \dot{z}(t) = u(t), \quad G_0 z(t) - \frac{1}{2} C_0 \dot{z}(t) = y(t)$$

(which are the same as (5.1)–(5.3) and (5.5)) as well as (5.4). □

The main result of this section is the following.

THEOREM 5.3. *For all $\xi \in (0,\pi)$, (5.4)–(5.7) define an exponentially stable well-posed system with input and output space $U = \mathbb{C}^2$ and state space X.*

Proof. Let $A : \mathcal{D}(A) \to X$ be defined as in (1.7), (1.8). If $\psi = \left[\begin{smallmatrix} z \\ w \end{smallmatrix}\right] \in \mathcal{D}(A)$, then $w \in H_{\frac{1}{2}}$ and z is piecewise in H^4 (on $(0,\xi)$ and on (ξ,π)). Thus, $\mathcal{D}(A)$ is compactly included in X so that A has a compact resolvent, hence $\sigma(A)$ consists of eigenvalues of A. We prove that $\sigma(A) \cap i\mathbb{R} = \emptyset$. Arguing by contradiction, suppose that $i\beta \in \sigma(A)$ with $\beta \in \mathbb{R}$. According to Lemma 4.4, if ψ is an eigenvector of A corresponding to $i\beta$, then $\psi = \left[\begin{smallmatrix} z \\ i\beta z \end{smallmatrix}\right]$, where $(\beta^2 I - A_0)z = 0$. This implies that

$$\frac{\mathrm{d}^4 z}{\mathrm{d}x^4} - \beta^2 z = 0, \quad z(0) = z(\pi) = 0, \quad \frac{\mathrm{d}^2 z}{\mathrm{d}x^2}(0) = \frac{\mathrm{d}^2 z}{\mathrm{d}x^2}(\pi) = 0.$$

It follows that $z(x) = c\sin(nx)$ for some $c \in \mathbb{C}$ and $n \in \mathbb{N}$ (with $n^4 = \beta^2$). Moreover, by Lemma 4.4 we also have $C_0 z = 0$ which, by (5.8), yields that $z(\xi) = 0$, $\frac{\mathrm{d}z}{\mathrm{d}x}(\xi) = 0$. This implies that $z = 0$, which is a contradiction. Thus, $\sigma(A) \cap i\mathbb{R} = \emptyset$.

By the equivalence of (3) and (9) in Theorem 1.3, it suffices to show that

$$\sup_{\omega\in\mathbb{R}} \|\omega V(i\omega)\|_{\mathcal{L}(H)} < \infty. \tag{5.11}$$

Suppose that this condition is false. By the uniform boundedness theorem, there exist a sequence of real numbers (β_n) with $|\beta_n| \to \infty$ and $h \in H$ such that $\lambda_n = \|\beta_n V(i\beta_n)h\| \to \infty$. Denoting $z_n = \frac{1}{\lambda_n} V(i\beta_n)h$ and $g_n = \frac{1}{\lambda_n} h$ (so that $g_n \to 0$ in H), we have

$$\|\beta_n z_n\|_H = 1 \qquad \forall\, n \in \mathbb{N}, \tag{5.12}$$

$$-\beta_n^2 z_n + A_0 z_n + \frac{i\beta_n}{2} B_0 C_0 z_n = g_n \in H\,. \tag{5.13}$$

We show (in four steps) that (5.12) and (5.13) lead to a contradiction.

First step. Taking the inner product in H of the sides of (5.13) with $\beta_n z_n$, taking the imaginary parts, and using (5.12), we obtain that $\beta_n C_0 z_n \to 0$. According to (5.8) this means

$$|\beta_n z_n(\xi)| \to 0, \qquad \left|\beta_n \frac{\mathrm{d}z_n}{\mathrm{d}x}(\xi)\right| \to 0. \tag{5.14}$$

On the other hand, (5.13) implies that $z_n \in Z_0$ so that by (5.10) we have

$$A_0 z_n = L_0 z_n + B_0 G_0 z_n .$$

Substituting this into (5.13) and using the fact that $B_0 u \in H$ iff $u = 0$, we obtain

$$-\beta_n^2 z_n + L_0 z_n = g_n \quad \text{and} \quad G_0 z_n = \frac{-i\beta_n}{2} C_0 z_n .$$

The above relations together with (5.8) and (5.9) imply that

$$-\beta_n^2 z_n + z_n^{(\mathrm{iv})} = g_n \to 0 \ \text{ in } L^2[0,\pi], \tag{5.15}$$

$$z_n^{(\mathrm{iii})}(\xi^+) - z_n^{(\mathrm{iii})}(\xi^-) = -i\beta_n \frac{\alpha^2}{2} z_n(\xi), \tag{5.16}$$

$$\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^+) - \frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^-) = i\beta_n \frac{\beta^2}{2} \frac{\mathrm{d}z_n}{\mathrm{d}x}(\xi). \tag{5.17}$$

Second step. Define $f_n \in H$ by $f_n(x) = x \frac{\mathrm{d}z_n}{\mathrm{d}x}(x)$ for $x \in [0,\xi]$ and $f_n(x) = 0$ for $x > \xi$. We take the inner product of the sides of (5.15) with f_n to get

$$\int_0^\xi \left(-\beta_n^2 z_n(x) + z_n^{(\mathrm{iv})}(x)\right) x \frac{\mathrm{d}\bar{z}_n}{\mathrm{d}x}(x)\mathrm{d}x = \int_0^\xi x\, g_n(x) \frac{\mathrm{d}\bar{z}_n}{\mathrm{d}x}(x)\mathrm{d}x . \tag{5.18}$$

By a straightforward calculation, the terms on the left-hand side become

$$\mathrm{Re}\left\{\int_0^\xi -\beta_n^2 z_n(x) x \frac{\mathrm{d}\bar{z}_n}{\mathrm{d}x}(x)\mathrm{d}x\right\} = -\frac{1}{2}\xi|\beta_n z_n(\xi)|^2 + \frac{1}{2}\int_0^\xi |\beta_n z_n(x)|^2 \mathrm{d}x,$$

$$\mathrm{Re}\left\{\int_0^\xi z_n^{(\mathrm{iv})}(x) x \frac{\mathrm{d}\bar{z}_n}{\mathrm{d}x}(x)\mathrm{d}x\right\} = \mathrm{Re}\left[\left(\xi z_n^{(\mathrm{iii})}(\xi^-) - \frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^-)\right)\frac{\mathrm{d}\bar{z}_n}{\mathrm{d}x}(\xi)\right] - \frac{\xi}{2}\left|\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^-)\right|^2 + \frac{3}{2}\int_0^\xi \left|\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(x)\right|^2 \mathrm{d}x .$$

The two relations above and (5.18) yield

$$-\frac{1}{2}\xi|\beta_n z_n(\xi)|^2 + \frac{1}{2}\int_0^\xi |\beta_n z_n(x)|^2 \mathrm{d}x + \mathrm{Re}\left[\left(\xi z_n^{(\mathrm{iii})}(\xi^-) - \frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^-)\right)\frac{\mathrm{d}\bar{z}_n}{\mathrm{d}x}(\xi)\right]$$

$$-\frac{\xi}{2}\left|\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^-)\right|^2 + \frac{3}{2}\int_0^\xi \left|\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(x)\right|^2 \mathrm{d}x = \mathrm{Re}\int_0^\xi x\, g_n(x) \frac{\mathrm{d}\bar{z}_n}{\mathrm{d}x}(x)\mathrm{d}x .$$

Since g_n converges to zero in $L^2[0,\xi]$, we deduce that the right-hand side of the last formula converges to zero. This implies that

$$-\frac{1}{2}\xi|\beta_n z_n(\xi)|^2+\frac{1}{2}\int_0^\xi|\beta_n z_n(x)|^2\mathrm{d}x+\mathrm{Re}\left[\left(\xi z_n^{(\mathrm{iii})}(\xi^-)-\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^-)\right)\frac{\mathrm{d}\bar{z}_n}{\mathrm{d}x}(\xi)\right]$$

$$-\frac{\xi}{2}\left|\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^-)\right|^2+\frac{3}{2}\int_0^\xi\left|\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(x)\right|^2\mathrm{d}x\to 0. \tag{5.19}$$

On the other hand, from (5.15) and the fact that $\|\beta_n z_n\|_H=1$ we deduce that $(z_n^{(\mathrm{iv})}/\beta_n)$ is a bounded sequence in $L^2[0,\xi]$. This further implies the boundedness of the sequences of complex numbers $(z_n^{(\mathrm{iii})}(\xi^-)/\beta_n)$ and $(\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^-)/\beta_n)$. Then, by (5.14) we obtain that

$$-\frac{1}{2}\xi|\beta_n z_n(\xi)|^2+\mathrm{Re}\left[\left(\xi z_n^{(\mathrm{iii})}(\xi^-)-\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^-)\right)\frac{\mathrm{d}\bar{z}_n}{\mathrm{d}x}(\xi)\right]\to 0. \tag{5.20}$$

Relations (5.19) and (5.20) lead to

$$\frac{1}{2}\int_0^\xi|\beta_n z_n(x)|^2\,\mathrm{d}x+\frac{3}{2}\int_0^\xi\left|\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}\right|^2\mathrm{d}x-\frac{\xi}{2}\left|\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^-)\right|^2\to 0.$$

Similarly, we take the inner product of the sides of (5.15) with $(x-\pi)\frac{\mathrm{d}z_n}{\mathrm{d}x}$ truncated to $[\xi,\pi]$, and then we repeat the above argument. This gives

$$\frac{1}{2}\int_\xi^\pi|\beta_n z_n(x)|^2\,\mathrm{d}x+\frac{3}{2}\int_\xi^\pi\left|\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}\right|^2\mathrm{d}x-\frac{\pi-\xi}{2}\left|\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^+)\right|^2\to 0.$$

If we add the last two formulas and we use (5.13), we obtain

$$\int_0^\pi|\beta_n z_n(x)|^2\,\mathrm{d}x+3\int_0^\pi\left|\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(x)\right|^2\mathrm{d}x-\frac{\xi}{2}\left|\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^-)\right|^2-\frac{\pi-\xi}{2}\left|\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^+)\right|^2\to 0. \tag{5.21}$$

Third step. We show that both $|\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^-)|$ and $|\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^+)|$ converge to zero. We use an idea of Liu; see, for instance, [1]: we take the inner product of the sides of (5.15) with $\frac{1}{\phi_n}e^{-\phi_n(\xi-x)}$ truncated to $[0,\xi]$, where $\phi_n=\sqrt{|\beta_n|}$. Thus, we obtain

$$-\int_0^\xi\phi_n^3 z_n(x)e^{-\phi_n(\xi-x)}\mathrm{d}x+\int_0^\xi\frac{z_n^{(\mathrm{iv})}(x)}{\phi_n}e^{-\phi_n(\xi-x)}\mathrm{d}x=\int_0^\xi g_n(x)\frac{e^{-\phi_n(\xi-x)}}{\phi_n}\mathrm{d}x.$$

By using (5.12), (5.15) again and the fact that $\phi_n e^{-\phi_n(\xi-x)}$ is bounded in $L^1[0,\pi]$, we obtain that

$$-\int_0^\xi\phi_n^3 z_n(x)e^{-\phi_n(\xi-x)}\mathrm{d}x+\int_0^\xi\frac{z_n^{(\mathrm{iv})}(x)}{\phi_n}e^{-\phi_n(\xi-x)}\mathrm{d}x\to 0. \tag{5.22}$$

Integrating by parts the second term above four times, we obtain

$$\int_0^\xi\frac{z_n^{(\mathrm{iv})}(x)}{\phi_n}e^{-\phi_n(\xi-x)}\mathrm{d}x$$

$$= \left(\frac{z_n^{(\mathrm{iii})}(x)}{\phi_n} - \frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(x) + \phi_n \frac{\mathrm{d}z_n}{\mathrm{d}x}(x) - \phi_n^2 z_n \right) e^{-\phi_n(\xi - x)} \Bigg|_0^{\xi} + \int_0^{\xi} \phi_n^3 z_n e^{-\phi_n(\xi - x)} \mathrm{d}x .$$

By using (5.22) we obtain from here that

$$\left(\frac{z_n^{(\mathrm{iii})}(x)}{\phi_n} - \frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(x) + \phi_n \frac{\mathrm{d}z_n}{\mathrm{d}x}(x) - \phi_n^2 z_n \right) e^{-\phi_n(\xi - x)} \Bigg|_0^{\xi} \to 0 . \tag{5.23}$$

Among the boundary terms in (5.23), those at $x = 0$ all converge to zero due to the exponential decay of $e^{-\phi_n \xi}$ and to the boundedness of $(z_n^{(\mathrm{iv})}/\beta_n)$ in $L^2[0, \xi]$; the terms containing $z_n(\xi)$, $\frac{\mathrm{d}z_n}{\mathrm{d}x}(\xi)$ also converge to zero due to (5.14). Thus, (5.23) yields

$$\frac{z_n^{(\mathrm{iii})}}{\phi_n}(\xi^-) - \frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^-) \to 0. \tag{5.24}$$

Similarly, we take the inner product of the sides of (5.15) with $\frac{1}{\phi_n} e^{-\phi_n(x - \xi)}$ in $L^2[\xi, \pi]$. Repeating the above analysis, we obtain

$$\frac{z_n^{(\mathrm{iii})}}{\phi_n}(\xi^+) + \frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^+) \to 0 . \tag{5.25}$$

On the other hand, by using (5.14) and (5.16), we obtain

$$z_n^{(\mathrm{iii})}(\xi^+) - z_n^{(\mathrm{iii})}(\xi^-) \to 0 . \tag{5.26}$$

Then the difference of (5.24) and (5.25) leads, by using (5.26), to

$$\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^+) + \frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^-) \to 0 .$$

Recall that, by (5.14) and (5.17), we have $\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^+) - \frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^-) \to 0$. Therefore,

$$\frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^-) \to 0, \qquad \frac{\mathrm{d}^2 z_n}{\mathrm{d}x^2}(\xi^+) \to 0. \tag{5.27}$$

Fourth step. Relations (5.21) and (5.27) imply that $\|\beta_n z_n\|_H \to 0$, which clearly contradicts (5.12). This contradiction shows that (5.11) holds. □

Theorem 5.3 can be generalized for coupled beams described by a version of (5.6) containing variable coefficients; see Ammari, Liu, and Tucsnak [1].

Theorem 5.3 together with Theorem 1.3 implies the following.

PROPOSITION 5.4. *The system from Theorem* 5.3 *is exactly controllable.*

5.2. The wave equation with Dirichlet-type nonlocal boundary feedback. The physical system that we have in mind consists of a vibrating membrane which is fixed on a part of the boundary, while on the other part of the boundary the vibrations are damped by a feedback control acting on the Dirichlet boundary condition. The input is the displacement field on the controlled part of the boundary, and the measurement depends linearly on the velocity field as well as on the input. A membrane could be modelled in a domain in $\mathbb{R}^2$, but we consider a more general wave equation on an n-dimensional (possibly unbounded) domain Ω. The boundary Γ of $\Omega \subset \mathbb{R}^n$ is assumed to be of class C^2 and to have a decomposition

as $\Gamma = \overline{\Gamma_0 \cup \Gamma_1}$, where Γ_0, Γ_1 are disjoint open parts of Γ with $\Gamma_1 \neq \emptyset$. We also assume that the Poincaré inequality holds for all $f \in \mathcal{H}^1_0(\Omega)$, which is always true if Ω is bounded, but it also holds for some unbounded domains (see section I.7). The operator $G : \mathcal{H}^{-1}(\Omega) \to \mathcal{H}^1_0(\Omega)$ is defined by

$$Gf = \phi \text{ if and only if } \phi \in \mathcal{H}^1_0(\Omega) \text{ and } -\Delta\phi = f.$$

Thus, in a certain sense, $G = -\Delta^{-1}$. We denote by $\gamma_0 : \mathcal{H}^1(\Omega) \to \mathcal{H}^{\frac{1}{2}}(\Gamma)$ the Dirichlet trace operator, which is onto, and by $\gamma_1 : \mathcal{H}^2(\Omega) \to \mathcal{H}^{\frac{1}{2}}(\Gamma)$ the outward normal derivative operator. Using Green's formula,

$$\langle \gamma_1 f, \gamma_0 g\rangle_{L^2} = \langle \Delta f, g\rangle_{L^2} + \langle \nabla f, \nabla g\rangle_{L^2} \tag{5.28}$$

for all $f \in \mathcal{H}^2(\Omega)$ and $g \in \mathcal{H}^1(\Omega)$, γ_1 can be extended so that $\gamma_1 f$ is defined as a distribution in $\mathcal{H}^{-\frac{1}{2}}(\Gamma)$ for every $f \in \mathcal{H}^1(\Omega)$ for which $\Delta f \in L^2(\Omega)$. Here, Δf denotes the Laplacian of f in the sense of distributions on Ω, i.e., in the space $\mathcal{D}'(\Omega)$.

We consider the system described by the equations

$$\ddot{z} = \Delta z \text{ in } \Omega \times (0,\infty), \tag{5.29}$$

$$z = 0 \quad \text{on } \Gamma_0 \times (0,\infty), \tag{5.30}$$

$$z - \frac{1}{2}\gamma_1(G\dot{z}) = u \text{ on } \Gamma_1 \times (0,\infty), \tag{5.31}$$

$$z(x,0) = z_0(x), \quad \dot{z}(x,0) = w_0(x) \text{ for } x \in \Omega. \tag{5.32}$$

The input of this system is the function u in (5.31). The output associated with this system is

$$y = z + \frac{1}{2}\gamma_1(G\dot{z}) \text{ on } \Gamma_1 \times (0,\infty). \tag{5.33}$$

Some comments about the domain Ω follow: It is not really necessary to assume that Γ is of class C^2. What we really need is that G maps $L^2(\Omega)$ onto $\mathcal{H}^2(\Omega) \cap \mathcal{H}^1_0(\Omega)$ and that γ_1 maps $\mathcal{H}^2(\Omega) \cap \mathcal{H}^1_0(\Omega)$ onto $\mathcal{H}^{\frac{1}{2}}(\Gamma)$. These properties hold also for some less regular domains, for example, for convex sets in $\mathbb{R}^2$ (see Grisvard [9, Theorem 3.2.1.2, p. 147]) or for rectangular domains in $\mathbb{R}^n$. We will need the following simple result, a direct consequence of the Riesz representation theorem in $L^2(\Omega)$.

PROPOSITION 5.5. *For every* $\mathrm{v} \in L^2(\Gamma)$ *there exists a unique function* $D\mathrm{v} \in L^2(\Omega)$ *such that*

$$\int_\Omega (D\mathrm{v})(x)\overline{\psi(x)}\,dx = -\int_\Gamma \mathrm{v}\,\overline{\gamma_1(G\psi)}\,d\Gamma \qquad \forall\, \psi \in L^2(\Omega). \tag{5.34}$$

Moreover, the operator D *defined above (called the Dirichlet map) is linear and bounded from* $L^2(\Gamma)$ *into* $L^2(\Omega)$.

Due to the Poincaré inequality, the norm on $\mathcal{H}^1_0(\Omega)$ can be defined as $\|f\|_{\mathcal{H}^1_0} = \|\nabla f\|_{L^2}$. Then it is easy to see (using (5.28)) that the corresponding dual norm on $\mathcal{H}^{-1}(\Omega)$ (with respect to the pivot space $L^2(\Omega)$) can be written as

$$\|g\|_{\mathcal{H}^{-1}} = \sup_{\|f\|_{\mathcal{H}^1_0}=1} \langle g, f\rangle_{\mathcal{H}^{-1},\mathcal{H}^1_0} = \|\nabla(Gg)\|_{L^2}.$$

Define $W = \{f \in L^2(\Omega) \mid \Delta f \in \mathcal{H}^{-1}(\Omega)\}$. Since γ_1 maps $\mathcal{H}^2(\Omega) \cap \mathcal{H}_0^1(\Omega)$ onto $\mathcal{H}^{\frac{1}{2}}(\Gamma)$ (see Lions and Magenes [21, Chapter 2, Theorem 5.4]), the Dirichlet trace operator γ_0 can be extended to an operator $\gamma_0 : W \to \mathcal{H}^{-\frac{1}{2}}(\Gamma)$ by putting

$$\langle \gamma_0 f, \gamma_1(G\psi)\rangle_{\mathcal{H}^{-\frac{1}{2}},\mathcal{H}^{\frac{1}{2}}} = -\langle \Delta f, G\psi\rangle_{\mathcal{H}^{-1},\mathcal{H}_0^1} - \langle f, \psi\rangle_{L^2} \qquad \forall\, \psi \in L^2(\Omega).$$

The operators D and γ_0 defined above are related as follows.

PROPOSITION 5.6. *If* $\mathrm{v} \in L^2(\Gamma)$, *then* $\Delta(D\mathrm{v}) = 0$ *in* $\mathcal{D}'(\Omega)$ *(which implies that* $D\mathrm{v} \in W$*). Moreover, we have that* $\gamma_0(D\mathrm{v}) = \mathrm{v}$.

Proof. The fact that $\Delta(D\mathrm{v}) = 0$ (in the sense of distributions) follows directly from (5.34) by taking $\psi = \Delta\varphi$, where $\varphi \in \mathcal{D}(\Omega) = C_0^\infty(\Omega)$. Now using the definition of the extended γ_0, we have that

$$\langle \gamma_0(D\mathrm{v}), \gamma_1(G\psi)\rangle_{\mathcal{H}^{-\frac{1}{2}},\mathcal{H}^{\frac{1}{2}}} = -\langle D\mathrm{v}, \psi\rangle_{L^2} \qquad \forall\, \psi \in L^2(\Omega). \tag{5.35}$$

Since $\gamma_1(G\psi)$ in (5.35) can be any function in $\mathcal{H}^{\frac{1}{2}}(\Gamma)$ and $\mathcal{H}^{\frac{1}{2}}(\Gamma)$ is dense in $L^2(\Omega)$, equations (5.34) and (5.35) imply that $\gamma_0(D\mathrm{v}) = \mathrm{v}$. ∎

To discuss the well-posedness and conservativity of the system (5.29)–(5.33) (using Theorem 1.1), we have to introduce the appropriate spaces and operators. Denote

$$X = L^2(\Omega) \times \mathcal{H}^{-1}(\Omega), \quad U = L^2(\Gamma_1).$$

We also define the Hilbert space

$$Z_0 = \left\{ f \in W \;\middle|\; \gamma_0 f \in L^2(\Gamma) \text{ and } \gamma_0 f|_{\Gamma_0} = 0 \right\}, \tag{5.36}$$

with the norm $\|f\|_{Z_0}$ given by

$$\|f\|_{Z_0}^2 = \|f\|_{L^2}^2 + \|\Delta f\|_{\mathcal{H}^{-1}}^2 + \|\gamma_0 f\|_{L^2}^2.$$

The precise statement of the well-posedness and conservativity of the system described by (5.29)–(5.33) is given in the following proposition.

PROPOSITION 5.7. *The equations* (5.29)–(5.33) *determine a conservative linear system* Σ *with input and output space* U *and state space* X. *If* $z_0 \in Z_0$, $w_0 \in L^2(\Omega)$, $u \in \mathcal{H}^1(0,\infty;U)$, *and the compatibility condition*

$$z_0(x) - \frac{1}{2}\gamma_1\left(Gw_0\right)(x) = u(x,0) \quad \text{for} \quad x \in \Gamma_1$$

holds, then (5.29)–(5.33) *have a unique solution* z, y *satisfying*

$$z \in BC\left(0,\infty;Z_0\right) \cap BC^1\left(0,\infty;L^2(\Omega)\right) \cap BC^2\left(0,\infty;\mathcal{H}^{-1}(\Omega)\right),$$

$$y \in \mathcal{H}^1(0,\infty;U).$$

Proof. We define the following spaces and operators:

- The space H is defined by $H = \mathcal{H}^{-1}(\Omega)$ endowed with the norm

$$\|f\|_H = \|\nabla(Gf)\|_{L^2(\Omega)}.$$

- The operator $A_0 : \mathcal{D}(A_0) \to H$ is defined by

$$\mathcal{D}(A_0) = \mathcal{H}_0^1(\Omega), \quad A_0\phi = -\Delta\phi \qquad \forall\, \phi \in \mathcal{D}(A_0).$$

It is well known that A_0 is self-adjoint, positive, and boundedly invertible.

- As in section 1, we put $H_1 = \mathcal{D}(A_0)$ and we introduce the spaces H_α ($\alpha \in \mathbb{R}$) by considering powers of A_0 and duality so that $A_0 : H_\alpha \to H_{\alpha-1}$. In order to identify the space $H_{\frac{1}{2}} = \mathcal{D}(A_0^{\frac{1}{2}})$, we recall that $H_{\frac{1}{2}}$ is the completion of $\mathcal{D}(A_0)$ with respect to the norm

$$\|z\|_{\frac{1}{2}} = \left\|A_0^{\frac{1}{2}} z\right\|_H = \sqrt{\langle A_0 z, z\rangle_H}.$$

  Since for any $z \in \mathcal{D}(A_0) = \mathcal{H}_0^1(\Omega)$, we have

$$\langle A_0 z, z\rangle_H = -\left\langle \nabla z, \nabla\left(\Delta^{-1} z\right)\right\rangle_{L^2(\Omega)} = \|z\|^2_{L^2(\Omega)};$$

  the space $H_{\frac{1}{2}}$ is given by $H_{\frac{1}{2}} = L^2(\Omega)$.
- Notice that $H_{\frac{3}{2}} = \mathcal{H}^2(\Omega) \cap \mathcal{H}_0^1(\Omega)$. We define the space

$$\widetilde{\mathcal{H}}^{-2}(\Omega) = \left(\mathcal{H}^2(\Omega) \cap \mathcal{H}_0^1(\Omega)\right)'.$$

  (This dual is computed with respect to the pivot space $L^2(\Omega)$.) It can be checked that G can be extended so that $G : \widetilde{\mathcal{H}}^{-2}(\Omega) \to L^2(\Omega)$ and the norm on $\widetilde{\mathcal{H}}^{-2}(\Omega)$ is

$$\|g\|_{\widetilde{\mathcal{H}}^{-2}} = \sup_{\|\Delta f\|_{L^2(\Omega)}=1} \langle g, f\rangle = \|Gg\|_{L^2}.$$

- By definition, the space $H_{-\frac{1}{2}}$ is the completion of H with respect to the norm $\|z\|_{-\frac{1}{2}} = \left\|A_0^{-\frac{1}{2}} z\right\|_H$. Since $G = A_0^{-1}$ on H, the last equality and the definition of the norm on H imply that

$$\|z\|_{-\frac{1}{2}} = \left\|\nabla\left(A_0^{-\frac{3}{2}} z\right)\right\|_{L^2}.$$

  The above relation and the obvious fact that $\|A_0^{\frac{1}{2}} w\|_{L^2} = \|\nabla w\|_{L^2}$ for any $w \in \mathcal{H}_0^1(\Omega)$ imply that

$$\|z\|_{-\frac{1}{2}} = \|A_0^{-1} z\|_{L^2} = \|Gz\|_{L^2} \qquad \forall\, z \in H.$$

  We have thus shown that $H_{-\frac{1}{2}} = \widetilde{\mathcal{H}}^{-2}(\Omega)$.
- We denote by $P \in \mathcal{L}\left(L^2(\Gamma), L^2(\Gamma_1)\right)$ the operator of truncation to Γ_1. Then $P^* \in \mathcal{L}\left(L^2(\Gamma_1), L^2(\Gamma)\right)$ is the operator defined by extending the functions in $L^2(\Gamma_1)$ by zero outside Γ_1. The operator $B_0 \in \mathcal{L}(U, H_{-\frac{1}{2}})$ is defined by

$$B_0 \mathrm{v} = A_0 D P^* \mathrm{v},$$

  where A_0 is considered as an operator from $H_{\frac{1}{2}}$ to $H_{-\frac{1}{2}}$ and $D : U \to L^2(\Omega)$ is the Dirichlet map defined in Proposition 5.5.
- Let $\phi \in \mathcal{H}_0^1(\Omega) = H_1$ and $\mathrm{v} \in L^2(\Gamma_1)$. Then we have

$$\langle B_0 \mathrm{v}, \phi\rangle_{H_{-\frac{1}{2}}, H_{\frac{1}{2}}} = \langle D P^* \mathrm{v}, A_0 \phi\rangle_H = \langle D P^* \mathrm{v}, \phi\rangle_{H_{\frac{1}{2}}}.$$

  By using the fact that $H_{\frac{1}{2}} = L^2(\Omega)$, the density of $\mathcal{H}_0^1(\Omega)$ in $L^2(\Omega)$, and (5.34), it follows that for every $\phi \in L^2(\Omega)$ and for every $\mathrm{v} \in L^2(\Gamma_1)$

$$\langle B_0 \mathrm{v}, \phi\rangle_{H_{-\frac{1}{2}}, H_{\frac{1}{2}}} = \langle D P^* \mathrm{v}, \phi\rangle_{L^2} = -\int_{\Gamma_1} \mathrm{v}\overline{\gamma_1\left(G\phi\right)}\,\mathrm{d}\Gamma.$$

We conclude that the adjoint of B_0 (using the pivot space $H = \mathcal{H}^{-1}(\Omega)$) is $C_0 \in \mathcal{L}(H_{\frac{1}{2}}, U)$ given by

$$C_0\phi = B_0^*\phi = -P\gamma_1(G\phi) \qquad \forall\, \phi \in H_{\frac{1}{2}} = L^2(\Omega).$$

It is clear that the spaces U, H, H_α, X and the operators A_0, C_0, B_0 fit into the simple general framework of section 1. Thus, by Theorem 1.1, they determine via the equations (1.1)–(1.3) a conservative linear system Σ. The state of this system is

$$\xi(t) = \begin{bmatrix} z(t) \\ \dot{z}(t) \end{bmatrix}.$$

To show that Σ is described by (5.29)–(5.33), first we notice that the space $Z_0 = H_1 + A_0^{-1}B_0U$ defined after (1.5) (see also Theorem I.1.2) is given in our case by $Z_0 = H_1 + DP^*U$. By using Lemma 5.6, it can be checked that this space coincides with that defined by (5.36). We define the operator $G_0 \in \mathcal{L}(Z_0, U)$ by

$$G_0 f = P\gamma_0 f \qquad \forall\, f \in Z_0.$$

Clearly, we have $G_0H_1 = \{0\}$, and by Lemma 5.6 we have in $\mathcal{L}(U)$

$$G_0A_0^{-1}B_0 = P\gamma_0A_0^{-1}(A_0DP^*) = P\gamma_0DP^* = I.$$

Hence, all the assumptions in Theorem I.1.4 are satisfied. We define on Z_0 the operator $L_0 = A_0 - B_0G_0$ as in Theorem I.1.4, and it is easy see that $L_0z = -\Delta z$.

If we write the system of equations (I.1.16) in our specific framework, we obtain the system of equations (5.29)–(5.33). Hence, by Theorem I.1.4 the compatibility condition in our proposition is equivalent to (I.1.9), and the equations (5.29)–(5.33) are equivalent to (I.1.11) and (I.1.12). Now by Theorem I.1.2, the equations have a solution z, y with the claimed smoothness properties. □

This system has also been considered in Lasiecka and Triggiani [18, pp. 669–671] but without considering outputs. They have proved the well-posedness of the mapping from the input function to the state, and they have discussed the exponential stability of the system for suitable Γ_1.

THEOREM 5.8. *If Ω is bounded, then the system defined by* (5.29)–(5.33) *is strongly stable, exactly controllable in infinite time, and exactly observable in infinite time.*

Proof. The fact that the equations (5.29)–(5.33) define a conservative linear system with input space U, state space X, and output space Y has been said in Proposition 5.7. The boundedness of Ω implies that the spectrum of A_0 is countable. Thus, according to assertion (8) of Theorem 1.5, in order to check the properties claimed in the theorem it suffices to prove that for any $\phi \in H_1$ if ϕ is an eigenvector of A_0, then $C_0\phi \neq 0$. Due to the particular form of A_0 and C_0, this means that we have to show that if $\phi \in \mathcal{H}_0^1(\Omega)$ is such that for some $\lambda > 0$

$$-\Delta\phi = \lambda\phi \ \text{ in } \ \mathcal{H}^{-1}(\Omega), \qquad \frac{\partial}{\partial\nu}(G\phi) = 0 \ \text{ on } \ \Gamma_1, \tag{5.37}$$

then $\phi = 0$. By denoting $\psi = G\phi$, we see that (5.37) is equivalent to

$$-\Delta\psi = \lambda\psi \ \text{ in } \ \mathcal{H}_0^1(\Omega), \qquad \frac{\partial\psi}{\partial\nu} = 0 \ \text{ on } \ \Gamma_1. \tag{5.38}$$

By classical elliptic regularity results, (5.38) implies that $\psi \in \mathcal{H}^2(\Omega)$ (see, for instance, Grisvard [9, Theorem 2.2.2.5]). This fact combined with (5.38) implies, by a classical unique continuation argument (see, for instance, Komornik [16, Corollary 6.2]), that $\psi = 0$. Thus, (5.37) implies that $\phi = \Delta\psi = 0$ in $\mathcal{H}_0^1(\Omega)$. □

The exponential stability of the system (5.29)–(5.32) was studied in Bardos et al. [4]. By combining Proposition 5.7 and [4, Theorem 1], we obtain the following.

THEOREM 5.9. *Suppose that Ω is bounded and there exists a time $T_0 > 0$ such that every geometric ray in $\Omega \times (0, T_0)$ intersects $\Gamma_1 \times (0, T_0)$ in a nondiffractive point. Then the equations* (5.29)–(5.33) *define an exponentially stable, conservative system with input space U, state space X, and output space Y. This system is also exactly controllable and exactly observable in time T_0.*

The last sentence of the above theorem follows from the exponential stability (stated in the first part of the theorem) using Theorem 1.3.

Acknowledgments. We have had useful discussions on this research with Olof Staffans, Arjan van der Schaft, and Peng-Fei Yao.

REFERENCES

[1] K. AMMARI, M. TUCSNAK, AND Z. LIU, *Decay rates for a beam with pointwise force and moment feedback*, Math. Control Signals Systems, 15 (2002), pp. 229–255.

[2] W. ARENDT AND C.J.K. BATTY, *Tauberian theorems and stability of one-parameter semigroups*, Trans. Amer. Math. Soc., 306 (1988), pp. 837–841.

[3] S.A. AVDONIN AND S.A. IVANOV, *Families of Exponentials*, Cambridge University Press, Cambdridge, UK, 1995.

[4] C. BARDOS, L. HALPERN, G. LEBEAU, J. RAUCH, AND E. ZUAZUA, *Stabilisation de l'équation des ondes au moyen d'un feedback portant sur la condition aux limites de Dirichlet*, Asymptot. Anal., 4 (1991), pp. 285–291.

[5] C. BARDOS, G. LEBEAU, AND J. RAUCH, *Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary*, SIAM J. Control. Optim., 30 (1992), pp. 1024–1065.

[6] A. BENSOUSSAN, G. DA PRATO, M.C. DELFOUR, AND S.K. MITTER, *Representation and Control of Infinite Dimensional Systems*, Vol. 1, Birkhäuser, Boston, 1992.

[7] P.L. BUTZER AND H. BERENS, *Semi-Groups of Operators and Approximation*, Grundlehren Math. Wiss. 145, Springer-Verlag, New York, 1967.

[8] P. GRABOWSKI, *On the spectral Lyapunov approach to parametric optimization of distributed parameter systems*, IMA J. Math. Control Inform., 7 (1990), pp. 317–338.

[9] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Monogr. Stud. Math. 24, Pitman, Boston, 1985.

[10] B.-Z. GUO AND Y.-H. LUO, *Controllability and stability of a second-order hyperbolic system with collocated sensor/actuator*, Systems Control Lett., 46 (2002), pp. 45–65.

[11] B.-Z. GUO AND Y.-H. LUO, *Riesz basis property of a second-order hyperbolic system with collocated scalar input-output*, IEEE Trans. Automat. Control, 47 (2002), pp. 693–698.

[12] S. HANSEN AND G. WEISS, *New results on the operator Carleson measure criterion*, IMA J. Math. Control Inform., 14 (1997), pp. 3–32.

[13] H. FALUN, *Characteristic conditions for exponential stability of linear dynamical systems in Hilbert spaces*, Ann. Differential Equations, 1 (1985), pp. 43–56.

[14] B. JACOB AND J. PARTINGTON, *The Weiss conjecture on admissibility of observation operators for contraction semigroups*, Integral Equations Operator Theory, 40 (2001), pp. 231–243.

[15] B. JACOB AND H. ZWART, *Exact observability of diagonal systems with a finite dimensional output operator*, Systems Control Lett., 43 (2001), pp. 101–109.

[16] V. KOMORNIK, *Exact Controllability and Stabilization - The Multiplier Method*, John Wiley, Chichester, UK, Masson, Paris, 1994.

[17] J.E. LAGNESE AND J.-L. LIONS, *Modelling, Analysis and Control of Thin Plates*, Masson, Paris, Springer-Verlag, Berlin, 1988.

[18] I. LASIECKA AND R. TRIGGIANI, *Control Theory for Partial Differential Equations: Continuous and Approximation Theories*, Cambridge University Press, Cambridge, UK, 2000.

[19] G. Lebeau, *Equation des ondes amorties*, in Algebraic and Geometric Methods in Mathematical Physics, A. Boutet de Monvel and V. Marchenko, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996, pp. 73–109.

[20] J.-L. Lions, *Controlabilité Exacte Perturbations et Stabilisation de Systèmes Distribués*, Vol. 1, Masson, Paris, 1988.

[21] J.-L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. I, Grundlehren Math. Wiss. 181, Springer-Verlag, Berlin, 1972.

[22] K. Liu, *Locally distributed control and damping for the conservative systems*, SIAM J. Control Optim., 35 (1997), pp. 1574–1590.

[23] K. Liu, Z. Liu, and B. Rao, *Exponential stability of an abstract nondissipative linear system*, SIAM J. Control Optim., 40 (2001), pp. 149–165.

[24] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

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

[26] R. Rebarber, *Exponential stability of coupled beams with dissipative joints: A frequency domain approach*, SIAM J. Control Optim., 33 (1995), pp. 1–28.

[27] R. Rebarber and G. Weiss, *Necessary conditions for exact controllability with a finite-dimensional input space*, Systems Control Lett., 40 (2000), pp. 217–227.

[28] D.L. Russell and G. Weiss, *A general necessary condition for exact observability*, SIAM J. Control Optim., 32 (1994), pp. 1–23.

[29] M. Tucsnak and G. Weiss, *Simultaneous exact controllability and some applications*, SIAM J. Control Optim., 38 (2000), pp. 1408–1427.

[30] G. Weiss, *Weak L^p-stability of a linear semigroup on a Hilbert space implies exponential stability*, J. Differential Equations, 76 (1988), pp. 269–285.

[31] G. Weiss, *Transfer functions of regular linear systems. Part* I: *Characterizations of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.

[32] G. Weiss, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.

[33] G. Weiss and R. Rebarber, *Optimizability and estimatability for infinite-dimensional linear systems*, SIAM J. Control Optim., 39 (2000), pp. 1204–1232.

[34] G. Weiss, O.J. Staffans, and M. Tucsnak, *Well-posed linear systems—a survey with emphasis on conservative systems*, Int. J. Appl. Math. Comput. Sci., 11 (2001), pp. 7–33.

[35] G. Weiss and M. Tucsnak, *How to get a conservative well-posed linear system out of thin air. Part* I. *Well-posedness and energy balance*, ESAIM Control Optim. Calc. Var., 9 (2003), pp. 247–274.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 936–965

A CONVERSE LYAPUNOV THEOREM FOR NONUNIFORM IN TIME GLOBAL ASYMPTOTIC STABILITY AND ITS APPLICATION TO FEEDBACK STABILIZATION*

I. KARAFYLLIS† AND J. TSINIAS†

Abstract. Lyapunov-like characterizations for the concepts of nonuniform in time robust global asymptotic stability and input-to-state stability for time-varying systems are established. The main result of our work enables us to derive (1) necessary and sufficient conditions for feedback stabilization for affine in the control systems and (2) sufficient conditions for the propagation of the input-to-state stability property through integrators.

Key words. nonuniform in time asymptotic stability, input-to-state stability, Lyapunov functions, feedback stabilization

AMS subject classifications. 93D20, 93D30, 37B55, 93D15

DOI. 10.1137/S0363012901392967

1. Introduction.

The notion of nonuniform in time *robust global asymptotic stability* (RGAS) is basically motivated by the problem of feedback stabilization for a class of nonlinear systems that, although fail to be stabilized at a specific equilibrium by continuous static time-invariant feedback, a *time-varying* feedback controller can be constructed in such a way that the equilibrium for the resulting closed-loop time-varying system is asymptotically stable, in general being nonuniform with respect to the initial values of time. The notion of RGAS—without uniformity with respect to time—is also motivated by problems related to feedback stabilization, such as

- stabilization of systems with uncertainties,
- stabilization of systems at a reference trajectory.

In the problems mentioned above, the analysis is reduced to studying asymptotic stability at a specific equilibrium of a time-varying system, whose dynamics are in general unbounded with respect to time. Particularly, in [40, 41] it is shown that for a class of triangular systems whose dynamics contain time-varying unknown parameters, it is possible to find, by applying a backstepping design procedure, a smooth time-varying feedback controller in such a way that the equilibrium of the resulting closed-loop system is RGAS, in general nonuniform with respect to initial values of time. Further progress has been obtained in [12, 13, 14, 15, 16, 17] for a large class of nonlinear systems that in general fail to be uniformly asymptotically stabilized by smooth static time-invariant feedback at a specific equilibrium. It is worthwhile to note that among other things in the works [12, 14], by employing the concept of nonuniform in time RGAS and its Lyapunov characterizations, we derive sufficient conditions for the solvability of the state feedback tracking control problem for a class of nonholonomic systems that includes the nonholonomic case in chained form. The corresponding results generalize those obtained in the literature for the same problem, since they are based on much weaker hypotheses. We finally mention the

*Received by the editors July 23, 2001; accepted for publication (in revised form) January 3, 2003; published electronically June 25, 2003. Part of this work appears in *Nonlinear Control Systems* 2001, Elsevier, New York, 2002, pp. 801–805.

http://www.siam.org/journals/sicon/42-3/39296.html

†Department of Mathematics, National Technical University of Athens, Zografou Campus 15780, Athens, Greece (jtsin@central.ntua.gr).

recent work [16], where various equivalent descriptions of nonuniform in time input-to-state stability are proposed and a generalization of the well-known "small-gain theorem" of Jiang, Teel, and Praly in [11] is established for time-varying composite systems.

The main purpose of the present paper is to establish a Lyapunov characterization for the notion of nonuniform in time RGAS. Lyapunov functions play an important role to synthesis and design in control theory, and several important results have been recently established concerning Lyapunov-like descriptions of *robust uniform global asymptotic stability* (RUGAS) and *input-to-state stability* (ISS) (see [2, 4, 5, 6, 8, 9, 19, 20, 21, 24, 25, 33, 34, 43]), *forward completeness* [1], and *asymptotic controllability* (see, for instance, [23, 30]). Our goal is to establish converse Lyapunov theorems for the concepts of *nonuniform in time RGAS* and *nonuniform in time ISS* and give some applications to feedback stabilization. In [42] a converse Lyapunov theorem is established for the particular case of nonuniform in time exponential robust stability and exp-ISS. In the present paper, by extending the approach employed in [20, 34], we establish a Lyapunov characterization for the general concept of RGAS for time-varying systems:

$$\begin{aligned} &\dot{x} = f(t,x,d) \\ &x \in \Re^n, \quad d \in D, \quad t \geq 0. \end{aligned} \tag{1.1}$$

We assume that $D \subset \Re^m$ is a nonempty compact set and $f : \Re^+ \times \Re^n \times D \to \Re^n$ is mapping with $f(t,0,d) = 0$ for all $(t,d) \in \Re^+ \times D$ that satisfies the following hypotheses:

H1. The function $f(t,x,d)$ is measurable in t for all $(x,d) \in \Re^n \times D$.

H2. The function $f(t,x,d)$ is continuous in d for all $(t,x) \in \Re^+ \times \Re^n$.

H3. The function $f(t,x,d)$ is locally Lipschitz with respect to x, uniformly in $d \in D$, in the sense that for every bounded interval $I \subset \Re^+$ and for every compact subset S of $\Re^n$, there exists a constant $L \geq 0$ such that

$$\begin{aligned} &|f(t,x,d) - f(t,y,d)| \leq L|x-y| \\ &\forall t \in I, \quad (x,y) \in S \times S, \quad d \in D. \end{aligned}$$

It turns out from H3 that there exists a positive C^0 function $L : \Re^+ \times \Re^+ \to \Re^+$ such that for each fixed $s \geq 0$ the mappings $L(\cdot, s)$ and $L(s, \cdot)$ are nondecreasing and the following holds:

$$\begin{aligned} &|f(t,x,d) - f(t,y,d)| \leq L(t, |x| + |y|)|x-y| \\ &\forall (t,x,y,d) \in \Re^+ \times \Re^n \times \Re^n \times D. \end{aligned} \tag{1.2}$$

In section 2, we provide some equivalent characterizations for the concept of RGAS for systems (1.1) (Proposition 2.2), and in section 3, we establish its Lyapunov characterization (Theorem 3.1). Section 4 is devoted to various equivalent descriptions of the nonuniform in time ISS property based on the results obtained for RGAS. The results of section 4 are applicable to the ISS feedback stabilization problem. In section 5.1 we derive a necessary and sufficient Lyapunov-based condition for ISS feedback stabilization for systems of the form

$$\begin{aligned} &\dot{x} = f(t,x,v) + g(t,x)u, \\ &x \in \Re^n, \quad v \in \Re^l, \quad u \in \Re^m, \quad t \geq 0, \end{aligned} \tag{1.3}$$

where the dynamics $f(\cdot)$ and $g(\cdot) = (g_1(\cdot), g_2(\cdot), \dots, g_m(\cdot))$ are both C^0 and locally Lipschitz with respect to (x, v) with $f(\cdot, 0, 0) = 0$. (Throughout this paper, given a map $F : \Re^+ \times \Re^{l_1} \to \Re^{l_2}$, we say that it is locally Lipschitz with respect to $x \in \Re^{l_1}$ if for every bounded interval $I \subset \Re^+$ and for every compact subset S of $\Re^{l_1}$, there exists a constant $L \geq 0$ such that $|F(t, x) - F(t, y)| \leq L|x - y|$ for every $(t, x, y) \in I \times S \times S$.) The main results of section 5.1 (Theorem 5.1 and Proposition 5.2) constitute extensions of the well-known Artstein–Sontag theorem [3, 27, 35] for autonomous systems and guarantee existence of a C^∞ mapping $u = k(t, x)$ in such a way that the resulting system

$$\dot{x} = f(t, x, v) + g(t, x)k(t, x) \tag{1.4}$$

satisfies the nonuniform in time ISS property with v as input. An explicit formula for a time-varying feedback stabilizer is proposed in Proposition 5.2. We also prove that, even for autonomous systems for which uniform in time asymptotic stabilization is not feasible, it is possible to exhibit nonuniform in time asymptotic stabilization by means of a time-varying feedback. In section 5.2 we establish an extension of a well-known result concerning the autonomous case (see [11, 36]) for systems of the following form:

$$\dot{x} = f(t, x, y), \tag{1.5a}$$

$$\dot{y} = g(t, x, y) + h(t, x, y)u, \tag{1.5b}$$

$$x \in \Re^n, \quad y \in \Re, \quad u \in \Re, \quad t \geq 0,$$

where $f(\cdot)$, $g(\cdot)$, $h(\cdot)$ are C^0 and locally Lipschitz with respect to (x, y), with $f(\cdot, 0, 0) = 0$ and $g(\cdot, 0, 0) = 0$. Particularly, we show that, under the presence of the (nonuniform in time) ISS for the subsystem (1.5a) with y as input, there exists a feedback law exhibiting ISS stabilization for (1.5) (Proposition 5.6). This result enables us to examine the partial-state feedback stabilization problem for triangular systems. Particularly, by exploiting a Lyapunov function based approach we re-establish the main result in [40] for a special class of triangular systems whose dynamics are time-dependent.

Notations. Throughout this paper we adopt the following notations:

- By M_D we denote the set of all measurable functions from $\Re^+ := [0, +\infty)$ to D, where D is any given compact subset of $\Re^m$.
- For any $x \in \Re^n$, x^T denotes its transpose and $|x|$ its usual Euclidean norm.
- K^+denotes the class of positive nondecreasing C^∞ functions $\phi : \Re^+ \to (0, +\infty)$, and $\mathbf{E}$ denotes the class of nonnegative C^0 functions $\mu : \Re^+ \to \Re^+$, for which $\int_0^{+\infty} \mu(t)dt < +\infty$ and $\lim_{t \to +\infty} \mu(t) = 0$ hold.
- $\mathbf{L}^\infty_{loc}$ denotes the set of all measurable functions $u : \Re^+ \to \Re^m$ that are essentially bounded on any nonempty compact subset of $\Re^+$, and $\mathbf{L}^\infty$ denotes the set of all measurable functions $u : \Re^+ \to \Re^m$ that are essentially bounded on $\Re^+$.
- By $B[x, r]$, where $x \in \Re^n$ and $r > 0$, we denote the closed sphere in $\Re^n$ of radius r centered at x.
- By $x(t) = x(t, t_0, x_0; d)$ we denote the solution of (1.1) at time t that corresponds to some input $d \in M_D$ initiated from x_0 at time t_0. For convenience, in certain parts of the text we prefer the notation $\phi(\cdot)$ instead of $x(\cdot)$.
- For definitions of classes K, K_∞, KL, see [18, 20].
- By Π we denote the subclass of K_∞ consisting of all functions $r : \Re^+ \to \Re^+$, for which $r(s) = \sum_{i=1}^m a_i s^i$ with $a_i \geq 0$ for $i = 1, \dots, m$, $a_1 > 0$ for some positive integer m.

2. The notion of RGAS. In this section we provide a general concept of global asymptotic stability (GAS) and establish some facts that will be used in proofs of main results in sections 3 and 5.

DEFINITION 2.1. *We say that zero $0 \in \Re^n$ is RGAS for* (1.1) *if for every $t_0 \geq 0$, $d \in M_D$, and $x_0 \in \Re^n$, the corresponding solution $x(\cdot)$ of* (1.1) *exists for all $t \geq t_0$ and satisfies the following properties:*

P1 (stability). *For every $\varepsilon > 0$, $T \geq 0$, it holds that*

$$\sup\{|x(t)| : d \in M_D, t \geq t_0, |x_0| \leq \varepsilon, t_0 \in [0,T]\} < +\infty \quad \text{(Lagrange stability)} \tag{2.1a}$$

and there exists a $\delta := \delta(\varepsilon, T) > 0$ such that

$$|x_0| \leq \delta, \quad t_0 \in [0,T] \Rightarrow |x(t)| \leq \varepsilon \quad \forall t \geq t_0, \quad d \in M_D \quad \text{(Lyapunov stability).} \tag{2.1b}$$

P2 (attractivity). *For every $\varepsilon > 0$, $T \geq 0$, and $R \geq 0$, there exists a $\tau := \tau(\varepsilon, T, R) \geq 0$ such that*

$$|x_0| \leq R, \quad t_0 \in [0,T] \Rightarrow |x(t)| \leq \varepsilon \quad \forall t \geq t_0 + \tau, \quad d \in M_D. \tag{2.1c}$$

As in the case of uniform in time RUGAS (see [20]) we have the following proposition.

PROPOSITION 2.2. *The origin $0 \in \Re^n$ is RGAS for* (1.1) *if and only if there exist a pair of functions a_1, a_2 of class K_∞, a_1 being locally Lipschitz on $(0, +\infty)$, and a function β of class K^+ such that for every $d \in M_D$, $t_0 \geq 0$, and $x_0 \in \Re^n$ the following holds:*

$$a_1(|x(t)|) \leq \exp(-t + t_0)\beta(t_0)a_2(|x_0|) \quad \forall t \geq t_0. \tag{2.2}$$

The proof of Proposition 2.2 requires the following technical result.

LEMMA 2.3. *Let $a : \Re^+ \times \Re^+ \to \Re^+$ be a function with $a(\cdot, 0) = 0$ that satisfies the following properties:*

(1) *For each fixed $t \geq 0$, the mapping $a(t, \cdot)$ is nondecreasing.*
(2) *For each fixed $s \geq 0$, the mapping $a(\cdot, s)$ is nondecreasing.*
(3) $\lim_{s \to 0^+} a(t,s) = 0$ *for all $t \geq 0$.*

Then there exists a pair of functions $\zeta \in K_\infty$ and $\gamma \in K^+$ such that

$$a(t,s) \leq \zeta(\gamma(t)s) \qquad \forall (t,s) \in (\Re^+)^2. \tag{2.3}$$

Proof of Lemma 2.3. Without loss of generality we may assume that a is $C^0(\Re^+ \times \Re^+)$. Indeed, otherwise we may consider the function

$$\hat{a}(t,s) := \begin{cases} \dfrac{1}{s} \displaystyle\int_s^{2s} \int_t^{t+1} a(\tau, \xi) d\tau \, d\xi & \text{for } s > 0, \\ 0 & \text{for } s = 0, \end{cases}$$

which by virtue of the inequality $a(t,s) \leq \hat{a}(t,s) \leq a(t+1, 2s)$ is $C^0(\Re^+ \times \Re^+)$ and satisfies $\hat{a}(\cdot, 0) = 0$. Notice that $\hat{a}$ has the same properties (1)–(3) of our statement with a. By invoking property (3), there exists a C^0 strictly decreasing function $\eta : \Re^+ \to (0, +\infty)$ with $\lim_{t \to +\infty} \eta(t) = 0$ such that

$$s \leq \eta(t) \Rightarrow a(t,s) \leq \frac{1}{t+1}. \tag{2.4a}$$

Let μ be the inverse function of η defined on $(0, \eta(0)]$ being nonnegative, continuous, strictly decreasing with $\lim_{t\to 0^+} \mu(t) = +\infty$. Define

$$\tilde{\mu}(s) := \begin{cases} \mu(s) & \text{if} \quad s \in (0, \eta(0)], \\ 0 & \text{if} \quad s > \eta(0). \end{cases} \tag{2.4b}$$

It turns out that $\tilde{\mu} : (0, +\infty) \to \Re^+$ is nonincreasing, continuous, and nonnegative and satisfies $\lim_{t\to 0^+} \tilde{\mu}(t) = +\infty$. Additionally, define

$$\beta(s) := s + \begin{cases} 0 & \text{if} \quad s = 0, \\ \sup\limits_{0<\tau\leq s} a(\tilde{\mu}(\tau), \tau) & \text{if} \quad s > 0. \end{cases} \tag{2.5}$$

We show that $\beta \in K_\infty$. Indeed, by definition (2.5) it follows that $\beta(0) = 0$ and β is strictly increasing with $\lim_{s\to+\infty} \beta(s) = +\infty$. Continuity of β on $(0, +\infty)$ follows from the fact that both a and $\tilde{\mu}$ are C^0 on $(0, +\infty)$. Furthermore, notice that (2.4a) and (2.4b) imply

$$a(\tilde{\mu}(\tau), \tau) \leq \frac{1}{\tilde{\mu}(\tau) + 1} \leq \frac{1}{\tilde{\mu}(s) + 1} \qquad \forall \tau \in (0, s] \text{ and } s \leq \eta(0). \tag{2.6}$$

Since $\lim_{s\to 0^+} \tilde{\mu}(s) = +\infty$ it follows from (2.6) that $\lim_{s\to 0^+} \beta(s) = 0$, and this establishes continuity of β at zero. Let $\zeta(s) := a(s, s) + \beta(s)$. Obviously, $\zeta(\cdot)$ is of class K_∞. Moreover, when $s \geq t$, by virtue of property (2) it holds that $a(t, s) \leq a(s, s) \leq \zeta(s)$, which implies

$$\sup_{s\geq t>0} \frac{\zeta^{-1}(a(t, s))}{s} \leq 1. \tag{2.7}$$

Also, when $0 < s \leq \eta(t)$, it follows from (2.4b) that $\tilde{\mu}(s) \geq t$; hence, by virtue of property (2) and (2.5), $a(t, s) \leq a(\tilde{\mu}(s), s) \leq \zeta(s)$. The latter implies that

$$\sup_{0<s\leq\eta(t)} \frac{\zeta^{-1}(a(t, s))}{s} \leq 1. \tag{2.8}$$

Using property (1), (2.7), and (2.8) we get

$$\sup_{s>0} \frac{\zeta^{-1}(a(t, s))}{s} \leq 1 + \sup_{\eta(t)\leq s\leq t} \frac{\zeta^{-1}(a(t, s))}{s} \leq 1 + \frac{\zeta^{-1}(a(t, t))}{\eta(t)}. \tag{2.9}$$

Finally let γ be any function of class K^+ which satisfies

$$\gamma(t) \geq \frac{\zeta^{-1}(a(t, t))}{\eta(t)} + 1 \qquad \forall t \geq 0. \tag{2.10}$$

The desired (2.3) is a consequence of (2.9) and (2.10). □

We are in a position to establish Proposition 2.2. Its proof is based on Lemma 2.3 and is inspired by the analysis made in [32].

Proof of Proposition 2.2. ($\Rightarrow$) Suppose that $0 \in \Re^n$ is RGAS for (1.1). Let $\xi, T, s \geq 0$ and define

$$a(T, s) := \sup\{|x(t)| : d \in M_D, t \geq t_0, |x_0| \leq s, 0 \leq t_0 \leq T\}, \tag{2.11a}$$

$$M(\xi, T, s) := \sup\{|x(t_0 + \xi)| : d \in M_D, |x_0| \leq s, 0 \leq t_0 \leq T\}. \tag{2.11b}$$

Obviously, our hypothesis that $0 \in \Re^n$ is RGAS guarantees that both $a(\cdot)$ and $M(\cdot)$ are well defined. Moreover, $a(\cdot)$ satisfies all hypotheses of the Lemma 2.3; namely, for each fixed $s \geq 0$, $a(\cdot, s)$ is nondecreasing, and for each fixed $T \geq 0$, $a(T, \cdot)$ is nondecreasing and satisfies $a(\cdot, 0) = 0$. Furthermore, stability of zero asserts that, for every $T \geq 0$, $\lim_{s \to 0^+} a(T, s) = 0$. It turns out from Lemma 2.3 that there exist functions $\zeta_1 \in K_\infty$ and $\gamma_1 \in K^+$ such that

$$a(T, s) \leq \zeta_1(\gamma_1(T)s) \qquad \forall (T, s) \in (\Re^+)^2. \tag{2.12}$$

The previous inequality in conjunction with (2.11a) and (2.11b) implies

$$M(\xi, T, s) \leq \zeta_1(\gamma_1(T)s) \qquad \forall (\xi, T, s) \in (\Re^+)^3. \tag{2.13}$$

Moreover, attractivity of zero guarantees that for every $\varepsilon > 0$, $T \geq 0$, and $R \geq 0$, there exists a $\tau = \tau(\varepsilon, T, R) \geq 0$ such that

$$M(\xi, T, s) \leq \varepsilon \qquad \forall \xi \geq \tau(\varepsilon, T, R) \text{ and } 0 \leq s \leq R. \tag{2.14}$$

Let

$$g(s) := \sqrt{s} + s^2 \tag{2.15a}$$

and let p be a function of class K^+ with $p(0) = 1$ and

$$\lim_{t \to +\infty} p(t) = +\infty. \tag{2.15b}$$

Define

$$\mu(\xi) := \sup \left\{ \frac{M(\xi, T, s)}{p(T)g(\zeta_1(\gamma_1(T)s))}, T \geq 0, s > 0 \right\}. \tag{2.16}$$

Obviously, by (2.12) and (2.15a), the function $\mu : \Re^+ \to \Re^+$ is well defined and satisfies $\mu(\cdot) \leq 1$. We show that $\lim_{\xi \to +\infty} \mu(\xi) = 0$; equivalently, we establish that for any given $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon) \geq 0$ such that

$$\mu(\xi) \leq \varepsilon \text{ for } \xi \geq \delta(\varepsilon). \tag{2.17}$$

Notice first that for any given $\varepsilon > 0$ there exist constants $a := a(\varepsilon)$ and $b := b(\varepsilon)$ with $0 < a < b$ such that

$$x \notin (a, b) \Rightarrow \frac{x}{\sqrt{x} + x^2} \leq \varepsilon. \tag{2.18}$$

We next recall (2.15b), which asserts that, for the above ε for which (2.18) holds, there exists a $c := c(\varepsilon) \geq 0$ such that $p(T) \geq \frac{1}{\varepsilon}$ for all $T \geq c$. This by virtue of (2.13) and (2.15a) yields

$$\frac{M(\xi, T, s)}{p(T)g(\zeta_1(\gamma_1(T)s))} \leq \varepsilon \qquad \forall \xi \geq 0 \tag{2.19a}$$

$$\text{when either } T \geq c \text{ or } \zeta_1(\gamma_1(T)s) \notin (a, b). \tag{2.19b}$$

Hence, in order to establish (2.17), it remains to consider the case

$$a \leq \zeta_1(\gamma_1(T)s) \leq b \quad \text{and} \quad 0 \leq T \leq c. \tag{2.20}$$

Since, for each fixed $(\xi, s) \in (\Re^+)^2$, the mappings $M(\xi, \cdot, s)$, $M(\xi, T, \cdot)$, $\gamma_1(\cdot)$, and $p(\cdot)$ are nondecreasing, we have that

$$\frac{M(\xi, T, s)}{p(T)g(\zeta_1(\gamma_1(T)s))} \leq \frac{M\left(\xi, c, \frac{\zeta_1^{-1}(b)}{\gamma_1(0)}\right)}{g(a)} \tag{2.21}$$

provided that (2.20) holds. By using (2.14) and (2.21) with

$$\varepsilon := \varepsilon g(a), \quad T := c, \quad R := \frac{\zeta_1^{-1}(b)}{\gamma_1(0)},$$

it follows that

$$M\left(\xi, c, \frac{\zeta_1^{-1}(b)}{\gamma_1(0)}\right) \leq \varepsilon g(a) \quad \text{for } \xi \geq \delta(\varepsilon) := \tau\left(\varepsilon g(a), c, \frac{\zeta_1^{-1}(b)}{\gamma_1(0)}\right). \tag{2.22}$$

By taking into account (2.19), (2.20), (2.21), (2.22), and definition (2.16) of $\mu(\cdot)$, it follows that (2.17) holds with $\delta = \delta(\varepsilon)$ as selected in (2.22). Since $\varepsilon > 0$ was arbitrary we conclude that $\lim_{\xi \to +\infty} \mu(\xi) = 0$. Consequently, there exists a continuous strictly decreasing function $\bar{\mu} : \Re^+ \to (0, +\infty)$ such that $\bar{\mu}(\xi) \geq \mu(\xi)$ for all $\xi \geq 0$ and $\lim_{\xi \to +\infty} \bar{\mu}(\xi) = 0$. Thus, by recalling definition (2.16) we obtain

$$M(\xi, T, s) \leq \bar{\mu}(\xi)\theta(T, s) \qquad \forall (T, s) \in (\Re^+)^2, \quad \forall \xi \geq 0, \tag{2.23}$$

where $\theta(T, s) := p(T)g(\zeta_1(\gamma_1(T)s))$. Clearly, θ satisfies all hypotheses of Lemma 2.3 and therefore there exist $\zeta_2 \in K_\infty$ and $\gamma_2 \in K^+$ such that

$$\theta(T, s) \leq \zeta_2(\gamma_2(T)s) \qquad \forall (T, s) \in (\Re^+)^2. \tag{2.24}$$

Moreover, by recalling Proposition 7 in [32] there exist functions a_1, ρ of class K_∞, a_1, being locally Lipschitz on $(0, +\infty)$, such that the KL function $\mu(t)\zeta_2(s)$ is dominated by $a_1^{-1}(\exp(-t)\rho(s))$. Thus, by taking into account (2.11b), (2.23), and (2.24) we have

$$|x(t)| \leq a_1^{-1}(\exp(-t + t_0)\rho(\gamma_2(t_0)|x_0|)) \quad \forall t \geq t_0 \geq 0,\ x_0 \in \Re^n,\ d \in M_D. \tag{2.25}$$

By Corollary 10 in [32] a pair of functions a_2, $\tilde{\beta}$ of class K_∞ can be found such that

$$\rho(rs) \leq \tilde{\beta}(r)a_2(s) \qquad \forall r, s \geq 0, \tag{2.26}$$

and finally, let β be a function of class K^+ with

$$\tilde{\beta}(\gamma_2(t)) \leq \beta(t), \qquad t \geq 0. \tag{2.27}$$

The desired (2.2) is a consequence of (2.25), (2.26), and (2.27).

($\Leftarrow$) Conversely, assume that (2.2) holds. Existence of $x(\cdot)$ for all $t \geq t_0$ as well as (2.1a) are both immediate consequences of (2.2). Let $\varepsilon > 0$ and $T \geq 0$ be arbitrary constants. By selecting $\delta(\varepsilon, T) := a_2^{-1}(\frac{a_1(\varepsilon)}{\beta(T)})$ the desired (2.1b) is fulfilled; thus property P1 holds (stability). Moreover, for any arbitrary positive constants R, ε, T, we may select $\tau = \tau(\varepsilon, T, R) := -\log(\frac{a_1(\varepsilon)}{\beta(T)a_2(R)})$, and by using (2.2) it follows that (2.1c) is fulfilled, and this establishes property P2 (attractivity). ∎

Remark 2.4.

- The notion of RGAS above is an extension of the well-known Sontag's robust uniform GAS (RUGAS) for autonomous systems, namely, when the solution $x(\cdot)$ satisfies $|x(t)| \le G(|x_0|, t-t_0)$ for certain G of class KL (see, for instance, [18, 20]). To justify this, we may recall Proposition 7 in [32], which asserts that for any $G \in KL$ there exist functions a_1 and a_2 of class K_∞ with $G(s,t) \le a_1^{-1}(\exp(-t)a_2(s))$. It turns out that RUGAS is characterized by the inequality $a_1(|x(t)|) \le \exp(-t+t_0)a_2(|x_0|)$, which obviously is a special case of (2.2).
- It is also straightforward to see that, if (2.2) holds with β being bounded over $\Re^+$, then zero is RUGAS and thus it turns out that for this case RGAS is equivalent to RUGAS.

Finally, we provide the following proposition, which generalizes the well-known fact that for autonomous differential equations equi-attractivity implies stability (see [10]). The result of this proposition will be used in sections 3 and 5.

PROPOSITION 2.5. *The origin $0 \in \Re^n$ is RGAS for* (1.1) *if for every $t_0 \ge 0$, $d \in M_D$, and $x_0 \in \Re^n$, the corresponding solution $x(\cdot)$ of* (1.1) *exists for all $t \ge t_0$ and satisfies property* P2 *(attractivity) of Definition* 2.1 *and* (1.1) *is Lagrange stable; namely, for every $\varepsilon > 0$ and $T \ge 0$,* (2.1a) *holds. It turns out that, if there exist a constant $M \ge 0$, functions $a_2 \in K_\infty$, $\sigma \in KL$, and $\beta \in K^+$ such that the estimate*

(2.28)
$$|x(t)| \le \sigma(a_2(\beta(t_0)|x_0|) + M, t-t_0) \qquad \forall t \ge t_0,\ (t_0,x_0) \in \Re^+ \times \Re^n,\ d(\cdot) \in M_D,$$

holds for the solution $x(\cdot)$ of (1.1)*, then $0 \in \Re^n$ is RGAS for* (1.1).

Proof. It suffices to show that for every $\varepsilon > 0$, $T \ge 0$, there exists a $\delta := \delta(\varepsilon, T) > 0$ such that (2.1b) holds. Let $\varepsilon > 0$, $T \ge 0$ be arbitrary. Define

$$R(\varepsilon, T) := \sup\{|x(t)| : d \in M_D, t \ge t_0, |x_0| \le \varepsilon, t_0 \in [0,T]\}. \tag{2.29}$$

By taking into account (1.2), (2.29), completeness of solutions, and our assumption that zero $0 \in \Re^n$ is an equilibrium for (1.1), it follows by use of Gronwall's inequality that

(2.30)
$$|x(t)| \le \exp\left(\int_{t_0}^{t} L(s, R(\varepsilon,T))ds\right)|x_0| \quad \forall t \ge t_0,\ d(\cdot) \in M_D,\ |x_0| \le \varepsilon,\ t_0 \in [0,T].$$

Moreover, property P2 of Definition 2.1 implies that for every $\varepsilon > 0$, $T \ge 0$, there exists a $\tau := \tau(\varepsilon, T) \ge 0$ such that

$$|x_0| \le \varepsilon,\ t_0 \in [0,T] \Rightarrow |x(t)| \le \varepsilon \quad \forall t \ge t_0 + \tau,\ d \in M_D. \tag{2.31}$$

Define

$$\delta(\varepsilon, T) := \varepsilon \exp\left(-\int_0^{T+\tau(\varepsilon,T)} L(s, R(\varepsilon,T))ds\right) \le \varepsilon \tag{2.32}$$

and notice that estimate (2.30) and definition (2.32) guarantee the following implication:

$$|x_0| \le \delta(\varepsilon,T),\ t_0 \in [0,T] \Rightarrow |x(t)| \le \varepsilon \quad \forall t \in [t_0, t_0 + \tau(\varepsilon,T)],\ d(\cdot) \in M_D. \tag{2.33}$$

The desired implication (2.1b) is an immediate consequence of (2.31) and (2.33).

Finally, notice that when estimate (2.28) holds, then property P2 holds and (1.1) is Lagrange stable; hence zero is RGAS. ∎

3. A converse Lyapunov theorem for RGAS. We next establish a Lyapunov characterization of the notion of RGAS, which constitutes generalization of the main result in [20] for the RUGAS case. Its proof is inspired from the analysis employed in [6, 20, 34].

THEOREM 3.1. *For the system* (1.1) *suppose that* H1, H2, H3 *are fulfilled and further* $f \in C^0(\Re^+ \times \Re^n \times D; \Re^n)$. *Then the following statements are equivalent:*

(i) *Zero* $0 \in \Re^n$ *is RGAS.*

(ii) *There exist a* C^∞ *function* $V : \Re^+ \times \Re^n \to \Re^+$, *functions* $\bar{a}_1$, $\bar{a}_2$ *of class* K_∞, $\bar{\beta}$ *of class* K^+ *such that for all* $(t, x, d) \in \Re^+ \times \Re^n \times D$ *it holds that*

$$\bar{a}_1(|x|) \leq V(t,x) \leq \bar{a}_2(\bar{\beta}(t)|x|), \tag{3.1a}$$

$$\dot{V}(t,x,d)\big|_{(1.1)} := \frac{\partial V}{\partial t}(t,x) + \frac{\partial V}{\partial x}(t,x) f(t,x,d) \leq -V(t,x). \tag{3.1b}$$

(iii) *There exist a* C^1 *function* $V : \Re^+ \times \Re^n \to \Re^+$, *functions* $\bar{a}_1$, $\bar{a}_2$ *of class* K_∞, $\bar{\beta}$ *of class* K^+, μ *of class* $\mathbf{E}$ *(see notations for the definition of class* $\mathbf{E}$*), and a* C^0 *positive definite function* $\rho : \Re^+ \to \Re^+$ *such that for all* $(t, x, d) \in \Re^+ \times \Re^n \times D$ *it holds that*

$$\bar{a}_1(|x|) \leq V(t,x) \leq \bar{a}_2(\bar{\beta}(t)|x|), \tag{3.2a}$$

$$\dot{V}(t,x,d)\big|_{(1.1)} := \frac{\partial V}{\partial t}(t,x) + \frac{\partial V}{\partial x}(t,x) f(t,x,d) \leq -\rho(V(t,x)) + \mu(t). \tag{3.2b}$$

For the proof of Theorem 3.1 we need a pair of technical lemmas. The first constitutes an extension of [20, Lemma 4.4] and was inspired by the main result in [22].

LEMMA 3.2. *Let* $y_d : \Re^+ \to \Re^+$ *be a family of absolutely continuous functions parameterized by* $d \in A$ *that satisfies the following differential inequality for almost all* $t \geq t_0$*:*

$$\dot{y}_d(t) \leq -\rho(y_d(t)) + \mu(t), \tag{3.3}$$

where $\rho : \Re^+ \to \Re^+$ *is a* C^0 *positive definite function and* μ *is of class* $\mathbf{E}$. *Then there exists a* KL *function* $\sigma : (\Re^+)^2 \to \Re^+$ *such that for all* $y_d(t_0) = y_0 \geq 0$ *and* $d \in A$ *it holds that*

$$y_d(t) \leq \sigma\left(y_0 + \int_{\Re^+} \mu(t)dt, t - t_0\right) \qquad \forall t \geq t_0. \tag{3.4}$$

Proof. Without loss of generality we may assume that $\int_{\Re^+} \mu(t)dt > 0$ (otherwise $\mu(t) = 0$ for all $t \geq 0$ and this is exactly the case of [20, Lemma 4.4]). First, notice that (3.3) yields

$$y_d(t) \leq y_0 + M \qquad \forall t \geq t_0,\ d \in A, \tag{3.5}$$

$$M := \int_{\Re^+} \mu(t)dt, \tag{3.6}$$

and this shows that $y_d(t)$ is bounded. Let $R \geq 0$ and $0 < \varepsilon \leq R + M$. Since $\lim_{t \to +\infty} \mu(t) = 0$ for any constants $r, \varepsilon > 0$ there exists a time $\tau := \tau(\varepsilon, r) \geq 0$ such that

$$t \geq \tau \Rightarrow \mu(t) \leq \min\left\{\frac{1}{2}\rho(s); \frac{\varepsilon}{2} \leq s \leq r\right\}. \tag{3.7}$$

We now show that the region

$$L_{\varepsilon,R} := \big\{(t,y) \in \Re^+ \times \Re^+ : y \le \varepsilon, t \ge \tau(\varepsilon, R+M)\big\} \tag{3.8}$$

is positively invariant. To see this, notice that, when $R + M \ge y_d(t) \ge \frac{\varepsilon}{2}$ and $t \ge \tau(\varepsilon, R+M)$ for some $d \in A$, then by (3.3) and (3.7) we have

$$\dot{y}_d(t) \le -\rho(y_d(t)) + \mu(t) \le -\frac{1}{2}\rho(y_d(t)) < 0 \tag{3.9}$$

and this establishes positive invariance of $L_{\varepsilon,R}$. We next establish that, if we define

$$T(\varepsilon, r) := \tau(\varepsilon, r) + \frac{2r}{\min_{\varepsilon \le s \le r} \rho(s)}, \tag{3.10}$$

then the following is fulfilled:

$$\text{For every } t \ge t_0 + T(\varepsilon, R+M), \quad d \in A \text{ and } y_0 \le R \Rightarrow (t, y_d(t)) \in L_{\varepsilon,R}. \tag{3.11}$$

Indeed, otherwise, by positive invariance of $L_{\varepsilon,R}$ there would exist $d \in A$ and $y_0 \le R$ such that

$$(t, y_d(t)) \notin L_{\varepsilon,R} \qquad \forall t \in [t_0 + \tau(\varepsilon, R+M), t_0 + T(\varepsilon, R+M)],$$

and since $t \ge \tau(\varepsilon, R, M)$, we would have

$$y_d(t) > \varepsilon \qquad \forall t \in [t_0 + \tau, t_0 + T]. \tag{3.12}$$

On the other hand, by (3.3), (3.5), (3.7) and (3.12), it follows that

$$\dot{y}_d(t) \le -\frac{1}{2} \min_{\varepsilon \le s \le R+M} \rho(s) \qquad \forall t \in [t_0 + \tau, t_0 + T]. \tag{3.13}$$

It turns out from (3.12) and (3.13) that

$$\varepsilon < y_d(t) \le R + M - \frac{1}{2}(t - t_0 - \tau) \min_{\varepsilon \le s \le R+M} \rho(s) \qquad \forall t \in [t_0 + \tau, t_0 + T]. \tag{3.14}$$

Using (3.14) and taking into account definition (3.10) of $T(\cdot)$ we get $\varepsilon < y_d(t_0 + T) \le 0$, which is a contradiction. This establishes (3.11).

Positive invariance of $L_{\varepsilon,R}$ and property (3.11) guarantee that the following attractivity property holds:

$$\begin{aligned} &\text{For all } (\varepsilon, R, t_0, d) \in (0, +\infty) \times \Re^+ \times \Re^+ \times A \text{ and} \\ &t \ge t_0 + T(\varepsilon, R+M), y_0 \le R \Rightarrow 0 \le y_d(t) \le \varepsilon. \end{aligned} \tag{3.15}$$

In order to establish inequality (3.4), we exploit (3.15) and apply an approach similar to that used in Proposition 2.2. We proceed as follows. Define

$$g(s) := \sqrt{s} + s^2, \tag{3.16a}$$

$$v(t) := \sup\left\{\frac{y_d(t_0 + \xi)}{g(y_0 + M)}; d \in A, y_0 \ge 0, t_0 \ge 0, \xi \ge t\right\}, \tag{3.16b}$$

where $M > 0$ is defined by (3.6). Since $M > 0$, the denominator in (3.16b) is strictly positive and (3.5), (3.16a) imply that $v(t) \le 1$ for all $t \ge 0$. We show that

$\lim_{t\to+\infty} v(t) = 0$. Let $\varepsilon > 0$ and let $a := a(\varepsilon)$, $b := b(\varepsilon)$ be a pair of constants with $0 < a < b$ and being defined in such a way that $x \notin [a, b] \Rightarrow x/(\sqrt{x} + x^2) < \varepsilon$. Then by (3.5) it follows that

$$\begin{aligned} &\frac{y_d(t_0+\xi)}{g(y_0+M)} < \varepsilon \quad \forall \xi \geq 0,\ d \in A,\ \text{and } t_0 \geq 0, \\ &\text{provided that either } y_0 + M < a \text{ or } y_0 + M > b. \end{aligned} \tag{3.17a}$$

It remains to consider the case $a \leq y_0 + M \leq b$. By (3.15) we get

$$\frac{y_d(t_0+\xi)}{g(y_0+M)} \leq \frac{y_d(t_0+\xi)}{g(a)} \leq \varepsilon \quad \forall \xi \geq T(\varepsilon g(a), b),\ \forall (t_0, d) \in \Re^+ \times A. \tag{3.17b}$$

It turns out from (3.16b), (3.17a), and (3.17b) that

$$v(t) \leq \varepsilon \qquad \forall t \geq T(\varepsilon g(a), b). \tag{3.18}$$

Since $\varepsilon > 0$ was arbitrary, (3.18) asserts that $\lim_{t\to+\infty} v(t) = 0$. Finally, let $\bar{v} : \Re^+ \to (0, +\infty)$ be a C^0, strictly decreasing function, with $v(t) \leq \bar{v}(t)$ for all $t \geq 0$ and in such a way that $\lim_{t\to+\infty} \bar{v}(t) = 0$. Then, obviously (3.4) is fulfilled with $\sigma(s, t) := g(s)\bar{v}(t)$. □

The second technical lemma provides a Lyapunov characterization of RGAS for (1.1) when its dynamics $f(\cdot)$ satisfy hypotheses H1, H2, and H3.

LEMMA 3.3. *Consider system* (1.1) *where its dynamics satisfy hypotheses* H1, H2 *and* H3 *and assume that* $0 \in \Re^n$ *is RGAS for* (1.1)*; particularly, there exists a pair of functions* a_1, a_2 *of class* K_∞, a_1 *being locally Lipschitz on* $(0, +\infty)$, *and a function* β *of class* K^+ *in such a way that* (2.2) *is satisfied. Then there exists a* C^0 *function* $U : \Re^+ \times \Re^n \to \Re^+$, *which is locally Lipschitz on* $\Re^+ \times (\Re^n \setminus \{0\})$, *such that for all* $(t, x) \in \Re^+ \times \Re^n$, $h \geq 0$, *and* $d(\cdot) \in M_D$ *it holds that*

$$a_1(|x|) \leq U(t, x) \leq \beta(t) a_2(|x|), \tag{3.19a}$$

$$\begin{aligned} &U(t+h, \phi(t+h, t, x; d)) \leq \exp\left(-\frac{h}{2}\right) U(t, x) \\ &\forall (t, x) \in \Re^+ \times \Re^n,\ h \geq 0,\ d(\cdot) \in M_D, \end{aligned} \tag{3.19b}$$

where for convenience we adopt the notation $\phi(\cdot, t, x; d)$ *to denote the solution of* (1.1) *that corresponds to the input* $d(\cdot) \in M_D$, *with* $\phi(t, t, x; d) = x$.

Proof. For the proof we need the following elementary properties for the solution of (1.1), which are immediate consequences of (1.2) and (2.2):

$$|\phi(t, t_0, x; d) - \phi(t, t_0, y; d)| \leq \exp\left(\int_{t_0}^{t} \tilde{L}(s, |x| + |y|) ds\right) |x - y|, \tag{3.20}$$

$$|\phi(t, t_0, x; d) - x| \leq \left(\exp\left(\int_{t_0}^{t} \tilde{L}(s, |x|) ds\right) - 1\right) |x|, \tag{3.21}$$

$$|\phi(t, t_0, x; d)| \geq \exp\left(-\int_{t_0}^{t} \tilde{L}(s, |x|) ds\right) |x|, \tag{3.22}$$

$$\begin{aligned} &|\phi(t, t_1, x; d) - \phi(t, t_2, x; d)| \leq \exp\left(\int_{\min(t_1,t_2)}^{t} \tilde{L}(s, |x|) ds\right) \tilde{L}(\max(t_1, t_2), |x|) |x| |t_1 - t_2| \\ &\forall t \geq t_0 \text{ and } (t_0, x, y; d) \in \Re^+ \times \Re^n \times \Re^n \times M_D, \end{aligned} \tag{3.23}$$

where

$$\tilde{L}(t,s) := L(t, 2a_1^{-1}(\beta(t)a_2(s))).$$

We define

$$U(t,x) := \sup\left\{a_1(|\phi(\tau,t,x;d)|)\exp\left(\frac{1}{2}(\tau-t)\right) : \tau \geq t, d \in M_D\right\}. \tag{3.24}$$

The desired properties (3.19a) and (3.19b) are then immediate consequences of (2.2) and definition (3.24). Inequality (3.19a) asserts that $U : \Re^+ \times \Re^n \to \Re^+$ is continuous at $x = 0$ with $U(t,0) = 0$ for all $t \geq 0$. We next establish that $U(\cdot)$ is locally Lipschitz on $\Re^+ \times (\Re^n \setminus \{0\})$. By (2.2) and (3.24) it follows that for any $T > 0$ the following holds:

(3.25)
$$\begin{aligned} U(t,x) = \max\Bigg\{&\sup\left\{a_1(|\phi(\tau,t,x;d)|)\exp\left(\frac{1}{2}(\tau-t)\right) : t \leq \tau \leq t+T, d \in M_D\right\}, \\ &\sup\left\{a_1(|\phi(\tau,t,x;d)|)\exp\left(\frac{1}{2}(\tau-t)\right) : \tau \geq t+T, d \in M_D\right\}\Bigg\} \\ \leq \max\Bigg\{&\sup\left\{a_1(|\phi(\tau,t,x;d)|)\exp\left(\frac{1}{2}(\tau-t)\right) : t \leq \tau \leq t+T, d \in M_D\right\}, \\ &\beta(t)a_2(|x|)\exp\left(-\frac{1}{2}T\right)\Bigg\}. \end{aligned}$$

Let $T_i : \Re^+ \times (0,+\infty) \to (0,+\infty)$, $i = 1,2$, be a pair of positive, C^0 functions, defined as

$$T_1(t,s) := 2\log\left(\frac{2\beta(t)a_2(s)}{a_1(s)}\right), \qquad T_2(t,s) := 2\log\left(\frac{2\beta(t)a_2(2s)}{a_1\left(\frac{s}{2}\right)}\right). \tag{3.26}$$

Notice that for every $s > 0$ each $T_i(\cdot,s)$ is nondecreasing and the following holds:

$$T_2(t,|x|) \geq \sup\left\{T_1(t,|y|) : y \in B\left[x, \frac{1}{2}|x|\right]\right\}, \qquad x \neq 0. \tag{3.27}$$

It turns out from (3.25) and (3.26) that

$$\begin{aligned} U(t,x) \leq \max\Bigg\{&\sup\left\{a_1(|\phi(\tau,t,x;d)|)\exp\left(\frac{1}{2}(\tau-t)\right) : t \leq \tau \leq t+\xi, d \in M_D\right\}, \\ &\frac{1}{2}a_1(|x|)\Bigg\} \qquad \text{for } \xi \geq T_1(t,|x|),\ x \neq 0, \end{aligned} \tag{3.28}$$

which by virtue of (3.19a) gives

$$\begin{aligned} U(t,x) = \sup&\left\{a_1(|\phi(\tau,t,x;d)|)\exp\left(\frac{1}{2}(\tau-t)\right) : t \leq \tau \leq t+\xi, d \in M_D\right\} \\ &\text{for } \xi \geq T_1(t,|x|),\ x \neq 0. \end{aligned} \tag{3.29}$$

It follows by taking into account (2.2), (3.22), (3.27) and (3.29) that

(3.30)
$$\begin{aligned}|U(t,y)-U(t,x)| &= \left|\sup\left\{a_1(|\phi(\tau,t,y;d)|)\exp\left(\frac{1}{2}(\tau-t)\right): t\le\tau\le t+T_2(t,|x|), d\in M_D\right\}\right.\\ &\qquad \left.-\sup\left\{a_1(|\phi(\tau,t,x;d)|)\exp\left(\frac{1}{2}(\tau-t)\right): t\le\tau\le t+T_2(t,|x|), d\in M_D\right\}\right|\\ &\le \sup\left\{\exp\left(\frac{1}{2}(\tau-t)\right)|a_1(|\phi(\tau,t,y;d)|)-a_1(|\phi(\tau,t,x;d)|)|: t\le\tau\le t+T_2(t,|x|), d\in M_D\right\}\\ &\le M_I\sup\left\{\exp\left(\frac{1}{2}(\tau-t)\right)|\phi(\tau,t,y;d)-\phi(\tau,t,x;d)|: t\le\tau\le t+T_2(t,|x|), d\in M_D\right\}\\ &\qquad \forall y\in B\left[x,\frac{1}{2}|x|\right],\ x\neq 0,\end{aligned}$$

where M_I is any Lipschitz constant for $a_1(\cdot)$ on the interval

$$I := \left[\frac{1}{2}\exp\left\{-\int_t^{t+T_2(t,|x|)}\tilde{L}\left(s,\frac{3}{2}|x|\right)ds\right\}|x|, a_1^{-1}\left(\beta(t)a_2\left(\frac{3}{2}|x|\right)\right)\right],$$

namely, $|a_1(s_1)-a_1(s_2)|\le M_I|s_1-s_2|$ for every $s_1,s_2\in I$. From (3.20) and (3.30) we deduce

$$|U(t,y)-U(t,x)|\le G_1(t,|x|)|y-x|, \qquad \forall y\in B\left[x,\frac{1}{2}|x|\right],\ x\neq 0, \tag{3.31a}$$

$$G_1(t,|x|) := M_I\exp\left(\frac{1}{2}T_2(t,|x|)+\int_t^{t+T_2(t,|x|)}\tilde{L}\left(s,\frac{5}{2}|x|\right)ds\right). \tag{3.31b}$$

This establishes that, for each $t\ge 0$, $U(t,\cdot)$ is locally Lipschitz on $\Re^n\setminus\{0\}$.

Likewise, we may establish that for each fixed nonzero $x\in\Re^n$, the map $U(\cdot,x)$ is locally Lipschitz on $\Re^+$. Indeed, consider a compact interval $I\subset\Re^+$ and let $t_1,t_2\in I$. Then, according to (3.29), for any $\varepsilon>0$, there exists a $d_\varepsilon\in M_D$ and time τ with $t_2\le\tau\le t_2+T_1(t_2,|x|)$ such that

$$U(t_2,x)-\varepsilon\le a_1(|\phi(\tau,t_2,x;d_\varepsilon)|)\exp\left(\frac{1}{2}(\tau-t_2)\right)\le U(t_2,x). \tag{3.32}$$

We distinguish three cases. The first is

$$t_1<t_2\le\tau. \tag{3.33}$$

It then follows by virtue of definition (3.24) that

$$a_1(|\phi(\tau,t_1,x;d_\varepsilon)|)\exp\left(\frac{1}{2}(\tau-t_1)\right)\le U(t_1,x); \tag{3.34}$$

hence, by (3.32) and (3.34) we get

(3.35)
$$U(t_2,x)-U(t_1,x)\le\exp\left(\frac{1}{2}(\tau-t_2)\right)|a_1(|\phi(\tau,t_2,x;d_\varepsilon)|)-a_1(|\phi(\tau,t_1,x;d_\varepsilon)|)|+\varepsilon.$$

Using (3.22) and (3.23) and exploiting the fact that $a_1(\cdot)$ is locally Lipschitz on $(0,+\infty)$, we deduce from (3.35) that for any compact $\Delta \subset \Re^n \setminus \{0\}$ a constant $L_1 > 0$ (being independent of ε and τ) can be found such that

$$\begin{aligned} U(t_2,x) - U(t_1,x) &\le L_1|t_2 - t_1| + \varepsilon \\ \forall t_2 > t_1, t_1,\ t_2 \in I,\ x \in \Delta. \end{aligned} \tag{3.36}$$

The second case is

$$t_2 \le t_1 \le \tau. \tag{3.37}$$

We may recall again (3.32), (3.34) and estimate

$$\begin{aligned} U(t_2,x) - U(t_1,x) &\le a_1(|\phi(\tau,t_2,x;d_\varepsilon)|)\exp\left(\frac{1}{2}(\tau - t_2)\right) \\ &\quad - a_1(|\phi(\tau,t_1,x;d_\varepsilon)|)\exp\left(\frac{1}{2}(\tau - t_1)\right) + \varepsilon \\ &= \exp\left(\frac{1}{2}(\tau - t_2)\right)\left(a_1(|\phi(\tau,t_2,x;d_\varepsilon)|) - a_1(|\phi(\tau,t_1,x;d_\varepsilon)|)\right) \\ &\quad + a_1(|\phi(\tau,t_1,x;d_\varepsilon)|)\left(\exp\left(\frac{1}{2}(\tau - t_2)\right) - \exp\left(\frac{1}{2}(\tau - t_1)\right)\right) + \varepsilon, \end{aligned} \tag{3.38}$$

and, as previously, it follows by (3.23) and (3.38) that there exists a constant $L_2 > 0$ (being independent of ε and τ) such that

$$\begin{aligned} U(t_2,x) - U(t_1,x) &\le L_2|t_2 - t_1| + \varepsilon \\ \forall x \in \Delta, \quad t_1, t_2 \in I,& \text{ provided that (3.37) holds.} \end{aligned} \tag{3.39}$$

Finally, consider the case

$$t_2 \le \tau < t_1 \tag{3.40}$$

for certain τ and d_ε such that (3.32) holds. We now invoke the left-hand-side inequality of (3.19a):

$$a_1(|x|) \le U(t_1,x). \tag{3.41}$$

It follows by virtue of (3.32), (3.40), and (3.41) that

$$\begin{aligned} U(t_2,x) - U(t_1,x) &\le a_1(|\phi(\tau,t_2,x;d_\varepsilon)|)\exp\left(\frac{1}{2}(\tau - t_2)\right) - a_1(|x|) \\ &\le \exp\left(\frac{1}{2}(\tau - t_2)\right)\left(a_1(|\phi(\tau,t_2,x;d_\varepsilon)|) - a_1(|x|)\right) \\ &\quad + a_1(|x|)\left(\exp\left(\frac{1}{2}(t_1 - t_2)\right) - 1\right). \end{aligned} \tag{3.42}$$

Using (3.21) and (3.22) and the fact that $a_1(\cdot)$ is locally Lipschitz on $(0,+\infty)$, we deduce from (3.42) that for any compact $\Delta \subset \Re^n \setminus \{0\}$ a constant $L_3 > 0$ (being independent of ε and τ) can be found such that

$$\begin{aligned} U(t_2,x) - U(t_1,x) &\le L_3|t_2 - t_1| + \varepsilon \\ \forall x \in \Delta, \quad t_1, t_2 \in I,& \text{ provided that (3.40) holds.} \end{aligned} \tag{3.43}$$

From (3.37), (3.39), and (3.43) it follows that $U(t_2,x)-U(t_1,x)\leq \max(L_1,L_2,L_3)|t_2-t_1|+\varepsilon$ for all $t_1,t_2\in I$, $\varepsilon>0$, and $x\in\Delta$. Similarly, we handle the case $U(t_1,x)-U(t_2,x)$ and conclude that for any compact sets $I\subset\Re^+$ and $\Delta\subset\Re^n\setminus\{0\}$, there is a constant $C>0$ such that

$$|U(t_2,x)-U(t_1,x)|\leq C|t_2-t_1|+\varepsilon \quad \forall t_1,t_2\in I,\ x\in\Delta. \tag{3.44}$$

Since $\varepsilon>0$ is arbitrary, inequalities (3.31) and (3.44) establish that $U(\cdot)$ is locally Lipschitz. The proof is complete. □

We are now in a position to prove Theorem 3.1.

Proof of Theorem 3.1. (i) ⇒ (ii) For convenience we still adopt here the notation $\phi(t,t_0,x_0;d)$ to denote the solution of (1.1) that corresponds to the input $d\in M_D$, initiated from $x_0\in\Re^n$ at time $t_0\geq 0$. Suppose first that $0\in\Re^n$ is RGAS and establish existence of $V(\cdot)$ satisfying (3.1). Since $0\in\Re^n$ is RGAS for (1.1), it follows by Lemma 3.3 that there exists a C^0 function $U:\Re^+\times\Re^n\to\Re^+$, which is locally Lipschitz on $\Re^n\times(\Re^n\setminus\{0\})$; a pair of functions a_1, a_2 of class K_∞, a_1 being locally Lipschitz on $(0,+\infty)$; and a function β of class K^+, such that for all $(t,x)\in\Re^+\times\Re^n$, $h\geq 0$, and $d(\cdot)\in M_D$ both (3.19a), (3.19b) hold. The proof is divided into two parts. In Part I we construct a function $W:\Re^+\times\Re^n\to\Re^+$ of class $C^0(\Re^+\times\Re^n)\cap C^\infty(\Re^+\times(\Re^n\setminus\{0\}))$, which satisfies

$$\frac{1}{2}a_1(|x|)\leq W(t,x)\leq\frac{3}{2}\beta(t)a_2(|x|)\qquad\forall(t,x)\in\Re^+\times\Re^n, \tag{3.45a}$$

$$\frac{\partial W}{\partial t}(t,x)+\frac{\partial W}{\partial x}(t,x)f(t,x,d)\leq-\frac{1}{4}W(t,x)$$
$$\forall(t,x,d)\in\Re^+\times(\Re^n\setminus\{0\})\times D, \tag{3.45b}$$

where a_1, $a_2\in K_\infty$ and $\beta\in K^+$ are the functions defined in (3.19a), (3.19b).

In Part II, by exploiting (3.45), we build the desired Lyapunov function $V:\Re^+\times\Re^n\to\Re^+$ that satisfies (3.1) for appropriate functions $\bar a_1$, $\bar a_2$ of class K_∞ and $\bar\beta$ of class K^+.

Part I. We proceed to the construction of an "almost smooth" W satisfying (3.45a), (3.45b). If the dynamics $f(\cdot)$ were Lipschitzian in both t and x, then the smoothing approach of [20] applied to the time-extended system $\dot x=f(t,x,d)$, $\dot t=1$, would lead to the existence of a function W satisfying both (3.45a) and (3.45b). However, we have assumed that $f(\cdot)$ is continuous in t, so we need to make a modification of the approach in [20]. We proceed as follows. Let $\psi_1:\Re^n\to\Re^+$, $\psi_2:\Re\to\Re^+$ be a pair of C^∞ functions with $\psi_1(\xi)=0$ and $\psi_2(\tau)=0$ when $|\xi|\geq 1$ and $\tau\notin(0,1)$, respectively, in such a way that

$$\int_{\Re^n}\psi_1(\xi)d\xi=\int_{\Re}\psi_2(\tau)d\tau=1.$$

Let S be a compact subset of $\Re^+\times(\Re^n\setminus\{0\})$. We consider the following family of functions:

$$W_\sigma(t,x):=\int_{\Re}\int_{\Re^n}U(t+\sigma\tau,x+\sigma\xi)\psi_1(\xi)\psi_2(\tau)d\xi d\tau,\qquad\sigma>0, \tag{3.46}$$

where $U(\cdot)$ is the function provided by Lemma 3.3. Let

$$r:=\min_{(t,x)\in S}|x|>0,$$
$$\tilde S:=\left\{(t+c\tau,x+c\xi)\in\Re^+\times\Re^n:(t,x)\in S,c\in\left[0,\frac{1}{2}r\right],\xi\in B[0,1],\tau\in[0,1]\right\}. \tag{3.47}$$

Obviously, $S \subseteq \tilde{S} \subseteq \Re^+ \times (\Re^n \setminus \{0\})$, $\tilde{S}$ is compact, and let C be a Lipschitz constant for U on $\tilde{S}$. It follows by virtue of (3.46) and (3.47) that for $\sigma < \frac{1}{2}r$, W_σ is well defined and C^∞ on S and satisfies

$$|W_\sigma(t,x) - U(t,x)| \le C\sigma \quad \forall (t,x) \in S, \quad \sigma < \frac{1}{2}r. \tag{3.48}$$

We also obtain the following by recalling (3.21) and (3.47):

$$(t+h+\sigma\tau, \phi(t+h,t,x;d) + \sigma\xi) \in \tilde{S}, \quad (t+h+\sigma\tau, \phi(t+h+\sigma\tau, t+\sigma\tau, x+\sigma\xi; d)) \in \tilde{S} \tag{3.49a}$$

$$\forall (t,x) \in S,\ d \in M_D,\ (\tau,\xi) \in [0,1] \times B[0,1], \quad \sigma \le \frac{1}{4}r,\ h > 0, \text{ sufficiently small.} \tag{3.49b}$$

Then by using (3.19b), (3.46), (3.49a) and (3.49b) we get

$$\begin{aligned} W_\sigma(t+h, \phi(t+h,t,x;d)) - W_\sigma(t,x) &\le \left(\exp\left(-\frac{h}{2}\right) - 1\right) W_\sigma(t,x) \\ &+ \int_{\Re}\int_{\Re^n} (U(t+h+\sigma\tau, \phi(t+h,t,x;d)+\sigma\xi) \\ &\qquad - U(t+h+\sigma\tau, \phi(t+h+\sigma\tau, t+\sigma\tau, x+\sigma\xi; d)))\psi_1(\xi)\psi_2(\tau)d\xi d\tau \\ &\le \left(\exp\left(-\frac{h}{2}\right) - 1\right) W_\sigma(t,x) \\ &+ C\int_{\Re}\int_{\Re^n} |\phi(t+h,t,x;d)+\sigma\xi - \phi(t+h+\sigma\tau, t+\sigma\tau, x+\sigma\xi; d)|\psi_1(\xi)\psi_2(\tau)d\xi d\tau \\ &\qquad \forall (t,x) \in S,\ d \in M_D,\ h > 0 \text{ sufficiently small.} \end{aligned} \tag{3.50}$$

Since f is C^0 and therefore uniformly continuous on compact sets, there exists a function $\delta_1 : \Re^+ \to \Re^+$ of class K such that

$$\sup\{|f(t+\sigma\tau, x, d) - f(t,x,d)| : (t,x) \in \tilde{S}, \tau \in [0,1], d \in D\} \le \delta_1(\sigma). \tag{3.51}$$

Using (1.2) and (3.51) and applying Gronwall's inequality, a function $\delta_2 : \Re^+ \to \Re^+$ of class K can be found such that

$$\begin{aligned} &|\phi(t+h,t,x;d) + \sigma\xi - \phi(t+h+\sigma\tau, t+\sigma\tau, x+\sigma\xi; d)| \le \delta_2(\sigma)h \\ &\forall (t,x) \in S,\ d \in M_D,\ (\tau,\xi) \in [0,1] \times B[0,1],\ h > 0 \text{ sufficiently small.} \end{aligned} \tag{3.52}$$

Specifically, in order to establish (3.52), define $p(s) := |\phi(t+s,t,x;d) + \sigma\xi - \phi(t+s+\sigma\tau, t+\sigma\tau, x+\sigma\xi; d)|$ and let L be a Lipschitz constant for f on $\tilde{S} \times D$, namely, $|f(t,x,d) - f(t,y,d)| \le L|x-y|$ for all $(t,x,d) \in \tilde{S} \times D$ and $(t,y,d) \in \tilde{S} \times D$. We then obtain by (3.51)

$$\begin{aligned} p(h) &\le \int_0^h |f(t+s, \phi(t+s,t,x;d), d) - f(t+\sigma\tau+s, \phi(t+\sigma\tau+s, t+\sigma\tau, x+\sigma\xi; d), d)| ds \\ &\le \delta_1(\sigma)h + \int_0^h |f(t+\sigma\tau+s, \phi(t+s,t,x;d), d) \\ &\qquad - f(t+\sigma\tau+s, \phi(t+\sigma\tau+s, t+\sigma\tau, x+\sigma\xi; d), d)|\, ds \\ &\le \delta_1(\sigma)h + L\int_0^h |\phi(t+s,t,x;d) - \phi(t+\sigma\tau+s, t+\sigma\tau, x+\sigma\xi; d)|\, ds \\ &\le \delta_1(\sigma)h + L\int_0^h p(s) ds + \sigma L h. \end{aligned}$$

The desired (3.52) is then a straightforward consequence of the previous inequality and Gronwall's lemma.

From (3.50) and (3.52) it follows that

$$\lim_{h\to 0^+} \frac{W_\sigma(t+h,\phi(t+h,t,x;d))-W_\sigma(t,x)}{h} = \frac{\partial W_\sigma}{\partial t}(t,x)+\frac{\partial W_\sigma}{\partial x}(t,x)f(t,x,d) \leq -\frac{1}{2}W_\sigma(t,x)+C\delta_2(\sigma). \tag{3.53}$$

By (3.48) and (3.53) we conclude that for any compact $S \subseteq \Re^+ \times (\Re^n \setminus \{0\})$ and $\varepsilon > 0$, there exists a constant $\sigma_0 > 0$ such that for every $\sigma < \sigma_0$ the function W_σ is well defined and C^∞ on S and satisfies for all $(t,x,d) \in S \times D$

$$|W_\sigma(t,x)-U(t,x)| \leq \varepsilon, \tag{3.54a}$$

$$\frac{\partial W_\sigma}{\partial t}(t,x)+\frac{\partial W_\sigma}{\partial x}(t,x)f(t,x,d) \leq -\frac{1}{2}W_\sigma(t,x)+\varepsilon. \tag{3.54b}$$

We may use (3.19a), (3.54a) and (3.54b) and apply partition of unity, as in the proof of [20, Theorem B.1], to build a function $W : \Re^+ \times \Re^n \to \Re^+$ of class $C^0(\Re^+ \times \Re^n) \cap C^\infty(\Re^+ \times (\Re^n \setminus \{0\}))$ that satisfies both (3.45a) and (3.45b).

Part II. We finally proceed to the construction of an everywhere C^∞ function V satisfying (3.1a), (3.1b). This part of proof is based on [34, Lemma 17], which in conjunction with (3.45a) and (3.45b) guarantees the existence of a function $\eta : \Re^+ \to \Re^+$ of class K_∞ with $\eta(s) \leq \frac{d\eta}{ds}(s)s$, such that the map

$$V(t,x) := (\eta(W(t,x)))^4 \tag{3.55}$$

is everywhere C^∞ and satisfies (3.1b). Furthermore, by Lemma 2.3 there exist functions $\tilde{a}_2 \in K_\infty$, $\bar{\beta} \in K^+$ such that

$$\frac{3}{2}\beta(t)a_2(s) \leq \tilde{a}_2(\bar{\beta}(t)s) \qquad \forall t,s \geq 0. \tag{3.56a}$$

Define

$$\bar{a}_1(s) := \left(\eta\left(\frac{1}{2}a_1(s)\right)\right)^4, \qquad \bar{a}_2(s) := (\eta(\tilde{a}_2(s)))^4. \tag{3.56b}$$

By using (3.56a), (3.56b) and invoking (3.45a), (3.45b), it follows that the function V as defined by (3.55) satisfies the desired inequalities (3.1a), (3.1b).

(ii) $\Rightarrow$ (iii) The implication is obvious since (3.1a), (3.1b) implies (3.2a), (3.2b) with $\rho(s) = s$, $\mu(t) \equiv 0 \in \mathbf{E}$, and some $\bar{a}_1 \in K_\infty$, $\bar{a}_2 \in K_\infty$, and $\bar{\beta} \in K^+$.

(iii) $\Rightarrow$ (i) We finally establish the converse part of our theorem, namely, that $0 \in \Re^n$ is RGAS with respect to (1.1) when both (3.2a) and (3.2b) are fulfilled. Define $A := M_D$ and let us again denote the solution of (1.1) by $\phi(t,t_0,x_0;d)$. Then using (3.2a), (3.2b) and applying the result of Lemma 3.2 with $y_d(t) := V(t,\phi(t,t_0,x_0;d))$, it follows that (3.3) holds; thus there exists a KL function σ and a constant $M = \int_0^{+\infty} \mu(t)dt \geq 0$ such that

$$|\phi(t,t_0,x_0;d)| \leq \bar{a}_1^{-1}(\sigma(\bar{a}_2(\bar{\beta}(t_0)|x_0|)+M,t-t_0)) \quad \forall t \geq t_0,\ d(\cdot) \in M_D$$

The latter estimate in conjunction with the result of Proposition 2.5 implies that $0 \in \Re^n$ is RGAS with respect to (1.1). The proof is complete. □

4. The nonuniform in time ISS property for time-varying systems. The results of the previous section enable us to characterize the nonuniform in time notion of ISS in terms of Lyapunov functions. We first introduce the notion of (nonuniform in time) ISS, as an extension of the notion of uniform in time ISS as presented in [36, 37]. In [16] we establish further equivalent descriptions of nonuniform in time ISS that constitute extensions of Sontag's ISS.

DEFINITION 4.1. *Consider the system*

$$\begin{aligned} \dot{x} &= f(t,x,u), \\ x \in \Re^n, &\quad u \in \Re^m, \quad t \geq 0, \end{aligned} \tag{4.1}$$

where $f(t,x,u)$ is measurable in $t \geq 0$ for all $(x,u) \in \Re^n \times \Re^m$ and is locally Lipschitz with respect to (x,u) with $f(\cdot,0,0) = 0$; denote $x(t) = x(t,t_0,x_0;u)$ its solution at time t that corresponds to some input $u \in \mathbf{L}^\infty_{loc}$, initiated from x_0 at time t_0. Let $\gamma(t,s) : (\Re^+)^2 \to \Re^+$ be a C^0 function, which is locally Lipschitz in s, such that for each fixed $t \geq 0$ the map $\gamma(t,\cdot)$ is positive definite. We say that (4.1) *satisfies the weak (nonuniform in time) input-to-state stability property (wISS) from the input u with gain $\gamma(\cdot)$ if each solution of* (4.1) *exists for all $t \geq t_0$ and satisfies properties* P1 *and* P2 *of Definition* 2.1 *provided that*

$$|u(t)| \leq \gamma(t,|x(t)|) \quad \text{a.e. for } t \geq t_0. \tag{4.2}$$

We say that (4.1) *satisfies the (nonuniform in time) ISS from the input u with gain $\gamma(\cdot)$ if it is wISS from the input u with gain $\gamma(\cdot)$ and in addition for each fixed $t \geq 0$ the map $\gamma(t,\cdot)$ is of class K_∞.*

As in the autonomous case (see [29, 37]) we can easily establish the following elementary fact.

Fact 4.2. System (4.1) satisfies the nonuniform in time wISS property from the input u with gain $\gamma(\cdot)$ if and only if $0 \in \Re^n$ is RGAS for the system

$$\begin{aligned} \dot{x} &= f(t,x,\gamma(t,|x|)d), \\ x \in \Re^n, &\quad d \in B[0,1] \subset \Re^m, \quad t \geq 0. \end{aligned} \tag{4.3}$$

The following theorem summarizes some useful equivalent descriptions of nonuniform in time wISS. Its proof is a direct consequence of Proposition 2.2, Lemma 3.2, Theorem 3.1, and Fact 4.2.

PROPOSITION 4.3. *Consider the system* (4.1) *whose dynamics satisfy the regularity assumptions of Definition* 4.1*, and let $\gamma(t,s) : (\Re^+)^2 \to \Re^+$ be a C^0 function, which is locally Lipschitz in s, such that for each fixed $t \geq 0$ the map $\gamma(t,\cdot)$ is positive definite. Then the following statements are equivalent:*

(i) *System* (4.1) *satisfies the nonuniform in time wISS property from the input with gain $\gamma(\cdot)$.*

(ii) *There exists a pair of functions a_1, a_2 of class K_∞, a_1 being locally Lipschitz on $(0,+\infty)$, and a function β of class K^+ such that the following property holds:*

$$\begin{aligned} &|u(t)| \leq \gamma(t,|x(t)|) \text{ a.e. for } t \geq t_0 \Rightarrow a_1(|x(t)|) \leq \exp(-t+t_0)\beta(t_0)a_2(|x_0|), \\ &\qquad\qquad \forall t \geq t_0,\ x_0 \in \Re^n. \end{aligned} \tag{4.4}$$

(iii) *There exists a C^0 function $U : \Re^+ \times \Re^n \to \Re^+$, which is locally Lipschitz on $\Re^+ \times (\Re^n \setminus \{0\})$ and satisfies*

$$a_1(|x|) \leq U(t,x) \leq \beta(t)a_2(|x|) \qquad \forall (t,x) \in \Re^+ \times \Re^n, \tag{4.5a}$$

$$|u(t)| \leq \gamma(t,|x(t)|) \text{ a.e. for } t \geq t_0 \Rightarrow U(t,x(t)) \leq \exp\left(-\frac{1}{2}(t-t_0)\right) U(t_0,x_0) \quad \forall (t_0,x_0) \in \Re^+ \times \Re^n \text{ and } t \geq t_0 \tag{4.5b}$$

with the some a_1, a_2, and β as defined in (4.4).
If in addition $f(\cdot) \in C^0(\Re^+ \times \Re^n; \Re^n)$, then the following are equivalent to the previous statements:

(iv) *There exist a C^∞ function $V : \Re^+ \times \Re^n \to \Re^+$ and functions $\bar{a}_1$, $\bar{a}_2 \in K_\infty$, $\bar{\beta} \in K^+$ such that the following hold for all $(t,x,u) \in \Re^+ \times \Re^n \times \Re^m$:*

$$\bar{a}_1(|x|) \leq V(t,x) \leq \bar{a}_2(\bar{\beta}(t)|x|), \tag{4.6a}$$

$$|u| \leq \gamma(t,|x|) \Rightarrow \dot{V}(t,x,u)|_{(4.1)} \leq -V(t,x). \tag{4.6b}$$

(v) *There exist a C^∞ function $V : \Re^+ \times \Re^n \to \Re^+$ and functions $\bar{a}_1$, $\bar{a}_2 \in K_\infty$, $\bar{\beta} \in K^+$, $\mu \in \mathbf{E}$ and a C^0 positive definite function $\rho : \Re^+ \to \Re^+$ such that the following hold for all $(t,x,u) \in \Re^+ \times \Re^n \times \Re^m$:*

$$\bar{a}_1(|x|) \leq V(t,x) \leq \bar{a}_2(\bar{\beta}(t)|x|), \tag{4.7a}$$

$$|u| \leq \gamma(t,|x|) \Rightarrow \dot{V}(t,x,u)|_{(4.1)} \leq -\rho(V(t,x)) + \mu(t). \tag{4.7b}$$

5. Applications to feedback stabilization. In this section we apply the converse Lyapunov Theorem 3.1 in order to derive necessary and sufficient conditions for ISS-feedback stabilization for affine in the control time-varying systems. For the general case (1.3) we extend the Artstein–Sontag theorem by introducing the concept of time-varying control Lyapunov function (Theorem 5.1). Among other things we establish that, even for a class of autonomous systems, it is possible to achieve nonuniform in time ISS stabilization by smooth time-varying feedback, although an everywhere smooth time-independent feedback exhibiting uniform in time stabilization does not exist (Corollary 5.4).

For the special case (1.5) an extension of a well-known result concerning autonomous systems (see [11, 36]) is established (Proposition 5.6). This result generalizes [40, Lemma 2.3] since is based on weaker hypotheses. Its Lyapunov function based establishment extremely simplifies the analysis made in [40].

5.1. A necessary and sufficient condition for ISS-feedback stabilization. The following theorem is an extension of the Artstein–Sontag theorem (see, for instance, [3, 27, 35]). We consider here the time-varying case (1.3) and in what follows assume that the dynamics f, g are C^0 and locally Lipschitz with respect to $(x,v) \in \Re^n \times \Re^l$, with $f(\cdot,0,0) = 0$.

THEOREM 5.1. *Consider the system* (1.3) *and let $\gamma(t,s) : (\Re^+)^2 \to \Re^+$ be a function, which is C^0, locally Lipschitz in s, and in such a way that for each $t \geq 0$ the mapping $\gamma(t,\cdot)$ is positive definite. Then the following statements are equivalent:*

(i) *There exists a C^∞ function $k : \Re^+ \times \Re^n \to \Re^m$ with $k(t,0) = 0$ for all $t \geq 0$, in such a way that the resulting system* (1.3) *with $u = k(t,x)$, namely, system* (1.4)*, satisfies the nonuniform in time wISS property with gain $\gamma(\cdot)$ from the input v. It turns out that* (1.4) *satisfies the nonuniform in time ISS property with gain $\gamma(\cdot)$ from the input v, provided that for each fixed $t \geq 0$ the map $\gamma(t,\cdot)$ is of class K_∞.*

(ii) *There exists a* C^0 *function* $k : \Re^+ \times \Re^n \to \Re^m$ *with* $k(t,0) = 0$ *for all* $t \geq 0$, *being locally Lipschitz in* x, *in such a way that the resulting system* (1.4) *satisfies the same property as in statement* (i).

(iii) *System* (1.3) *admits a "control Lyapunov function," namely, there exists a* C^1 *function* $V : \Re^+ \times \Re^n \to \Re^+$, *functions* $a_1, a_2 \in K_\infty$, $\beta \in K^+$, $\mu \in \mathbf{E}$, *and a* C^0 *positive definite function* $\rho : \Re^+ \to \Re^+$, *such that*

$$a_1(|x|) \leq V(t,x) \leq a_2(\beta(t)|x|), \tag{5.1a}$$

$$\begin{aligned} &\frac{\partial V}{\partial x}(t,x)g(t,x) = 0, \ |v| \leq \gamma(t,|x|) \\ &\qquad \Rightarrow \frac{\partial V}{\partial t}(t,x) + \frac{\partial V}{\partial x}(t,x)f(t,x,v) \leq -\rho(V(t,x)) + \mu(t). \end{aligned} \tag{5.1b}$$

Proof. The implication (i) $\Rightarrow$ (ii) is obvious. We establish implication (ii) $\Rightarrow$ (iii). Suppose that there exists a map $k(\cdot)$, as in statement (ii) of the theorem, such that system (1.4) satisfies the wISS property with gain $\gamma(\cdot)$. By recalling (iv) of Proposition 4.3, there exists a C^∞ function $V : \Re^+ \times \Re^n \to \Re^+$ in such a way that (5.1a) holds and

$$|v| \leq \gamma(t,x) \Rightarrow \frac{\partial V}{\partial t}(t,x) + \frac{\partial V}{\partial x}(t,x)(f(t,x,v) + g(t,x)k(t,x)) \leq -V(t,x). \tag{5.2}$$

The latter implies (5.1b) with $\mu(t) \equiv 0 \in \mathbf{E}$ and $\rho(s) = s$. We next establish (iii) $\Rightarrow$ (i). Consider the functions a_1, a_2, β, V, and μ as defined in (5.1a), (5.1b) and without any loss of generality assume

$$\mu(t) > 0 \qquad \forall t \geq 0. \tag{5.3}$$

Notice, by virtue of (5.1a), that

$$\frac{\partial V}{\partial t}(t,0) = 0, \qquad \frac{\partial V}{\partial x}(t,0) = 0. \tag{5.4}$$

Condition (5.1b) in conjunction with (5.3) and (5.4) enables us to build by standard partition of unity arguments a C^∞ map $k : \Re^+ \times \Re^n \to \Re^m$ with $k(\cdot,0) = 0$ such that

$$\begin{aligned} &\frac{\partial V}{\partial t}(t,x) + \frac{\partial V}{\partial x}(t,x)f(t,x,v) + \frac{\partial V}{\partial x}(t,x)g(t,x)k(t,x) \leq -\rho(V(t,x)) + \mu(t) \\ &\qquad \forall (t,x) \in \Re^+ \times \Re^n, \quad |v| \leq \gamma(t,|x|). \end{aligned} \tag{5.5}$$

Define $v = \gamma(t,|x|)d$, where $d(\cdot) \in A := M_{B[0,1]}$. Then using (5.5) it follows that for the solution $x(t,t_0,x_0;d)$ of the system $\dot{x} = f(t,x,\gamma(t,|x|)d) + g(t,x)k(t,x)$ it holds that $\dot{y}_d(t) \leq -\rho(y_d(t)) + \mu(t)$ for all $t \geq t_0$, where $y_d(t) := V(t,x(t,t_0,x_0;d))$. It turns out from Lemma 3.2 that there exists a KL function $\sigma : (\Re^+)^2 \to \Re^+$ such that

$$V(t,x(t,t_0,x_0;d)) \leq \sigma(M + V(t_0,x_0), t - t_0) \qquad \forall t \geq t_0, \quad d(\cdot) \in M_{B[0,1]},$$

where $M := \int_{\Re^+} \mu(t)dt$, and thus by virtue of (5.1a)

$$|x(t,t_0,x_0;d)| \leq a_1^{-1}(\sigma(M + a_2(\beta(t_0)|x_0|), t - t_0)) \qquad \forall t \geq t_0, \quad d(\cdot) \in M_{B[0,1]}, \tag{5.6}$$

for any initial (t_0,x_0). Inequality (5.6) in conjunction with Proposition 2.5 implies that $0 \in \Re^n$ is RGAS with respect to (1.4). The desired wISS property for system (1.4) is a consequence of Fact 4.2. ∎

The next proposition establishes the existence of an *explicit* formula of a feedback law exhibiting ISS stabilization for system (1.3).

PROPOSITION 5.2. *Consider the system* (1.3) *and suppose that statement* (iii) *of Theorem* 5.1 *is fulfilled for some positive function* $\mu \in \mathbf{E}$*, certain* $V : \Re^+ \times \Re^n \to \Re^+$ *of class* $C^2(\Re^+ \times \Re^n)$*, and some positive definite, locally Lipschitz function* $\rho : \Re^+ \to \Re^+$*. Let* $\theta : \Re \to \Re^+$ *be any* C^∞ *nondecreasing map with*

$$\theta(s) \begin{cases} = 0, & s \le 0, \\ < 1, & s < 1, \\ = 1, & s \ge 1, \end{cases} \tag{5.7}$$

and let

$$\zeta(t,x) := \left| \frac{\partial V}{\partial t}(t,x) + \max_{|v| \le \gamma(t,|x|)} \frac{\partial V}{\partial x}(t,x) f(t,x,v) + \rho(V(t,x)) \right|, \tag{5.8a}$$

$$W(t,x) := \frac{\partial V}{\partial t}(t,x) + \max_{|v| \le \gamma(t,|x|)} \frac{\partial V}{\partial x}(t,x) f(t,x,v) + \frac{1}{2}\rho(V(t,x)) - \mu(t). \tag{5.8b}$$

Then the feedback law

$$k(t,x) := -\left\{ \frac{\left(\frac{\partial V}{\partial x}(t,x) g(t,x)\right)^T}{1 - \theta\left(\frac{W(t,x)}{\mu(t)}\right) + \left|\frac{\partial V}{\partial x}(t,x) g(t,x)\right|^2} \right\} \zeta(t,x), \tag{5.9}$$

which is everywhere continuous and locally Lipschitz with respect to x *and satisfies* $k(\cdot,0) = 0$*, exhibits wISS stabilization for* (1.4) *with gain* $\gamma(\cdot)$ *from the input* v*.*

Proof. From (5.1b) and definition (5.8b) of $W(\cdot)$ it follows that

$$\frac{\partial V}{\partial x}(t,x) g(t,x) = 0 \Rightarrow W(t,x) \le 0, \tag{5.10a}$$

$$W(t,x) \le \mu(t) \Rightarrow \frac{\partial V}{\partial t}(t,x) + \max_{|v| \le \gamma(t,|x|)} \frac{\partial V}{\partial x}(t,x) f(t,x,v) \le -\frac{1}{2}\rho(V(t,x)) + 2\mu(t). \tag{5.10b}$$

Notice that k is well defined for all (t,x), since the denominator in (5.9) is strictly positive for all $(t,x) \in \Re^+ \times \Re^n$, and is of class $C^0(\Re^+ \times \Re^n)$. Indeed, $\theta\left(\frac{W(t,x)}{\mu(t)}\right) \le 1$ for all $(t,x) \in \Re^+ \times \Re^n$, and suppose that $\theta\left(\frac{W(t,x)}{\mu(t)}\right) = 1$ for certain $(t,x) \in \Re^+ \times \Re^n$. It then follows from (5.7) that $W(t,x) \ge \mu(t)$, and thus by virtue of (5.8a) $\frac{\partial V}{\partial x}(t,x) g(t,x)$ is nonzero. Furthermore, according to regularity assumptions made for $V(\cdot)$, $f(\cdot)$, $\gamma(\cdot)$, $g(\cdot)$, and $\rho(\cdot)$, the map $k(t,x)$ as defined by (5.9) is C^0 on $\Re^+ \times \Re^n$ and locally Lipschitz with respect to $x \in \Re^n$, with $k(t,0) = 0$ for all $t \ge 0$. We next estimate the derivative $\dot{V}(\cdot)$ of $V(\cdot)$ along the trajectories of the solutions of the closed-loop system (1.4). We find

$$\begin{aligned} \dot{V}(t,x) &:= \frac{\partial V}{\partial t}(t,x) + \max_{|v| \le \gamma(t,|x|)} \frac{\partial V}{\partial x}(t,x) f(t,x,v) + \frac{\partial V}{\partial x}(t,x) g(t,x) k(t,x) \\ &\le -\frac{1}{2}\rho(V(t,x)) + 2\mu(t). \end{aligned} \tag{5.11}$$

Indeed, for those t, x for which $W(t,x) \le \mu(t)$, we have by taking into account (5.9) and (5.10b) that

$$\frac{\partial V}{\partial t}(t,x) + \max_{|v| \le \gamma(t,|x|)} \frac{\partial V}{\partial x}(t,x) f(t,x,v) \le -\frac{1}{2}\rho(V(t,x)) + 2\mu(t),$$

$$\frac{\partial V}{\partial x}(t,x) g(t,x) k(t,x) \le 0,$$

which implies (5.11). On the other hand, for those t, x for which $W(t,x) \geq \mu(t)$, it follows from (5.7), (5.9), and (5.10a) that

$$\frac{\partial V}{\partial x}(t,x)g(t,x) \neq 0,$$
$$\frac{\partial V}{\partial x}(t,x)g(t,x)k(t,x) = -\zeta(t,x),$$

and thus by taking into account definition (5.8a) of $\zeta(\cdot)$ it follows that

$$\dot{V}(t,x) \leq -\rho(V(t,x)) \leq -\frac{1}{2}\rho(V(t,x)) + 2\mu(t).$$

This establishes (5.11). We complete the proof by applying Lemma 3.2 as exactly done in the proof of Theorem 5.1. □

We next specialize the result of Theorem 5.1 to the following case of time-varying systems:

$$\begin{aligned} \dot{x} &= f(t,x) + g(t,x)u, \\ x \in \Re^n, \quad u \in \Re^m, \quad t \geq 0, \end{aligned} \tag{5.12}$$

where the mappings f, g are C^0 and locally Lipschitz with respect to x with $f(t,0) = 0$ for all $t \geq 0$.

COROLLARY 5.3. *The following statements are equivalent:*

(i) *There exist a C^1 function $V : \Re^+ \times \Re^n \to \Re^+$, functions $a_1, a_2 \in K_\infty$, $\beta \in K^+$, $\mu \in \mathbf{E}$, and a C^0 positive definite map $\rho : \Re^+ \to \Re^+$, such that for all $(t,x) \in \Re^+ \times \Re^n$*

$$a_1(|x|) \leq V(t,x) \leq a_2(\beta(t)|x|), \tag{5.13a}$$
$$\frac{\partial V}{\partial x}(t,x)g(t,x) = 0 \Rightarrow \frac{\partial V}{\partial t}(t,x) + \frac{\partial V}{\partial x}(t,x)f(t,x) \leq -\rho(V(t,x)) + \mu(t). \tag{5.13b}$$

(ii) *There exists a C^∞ function $k : \Re^+ \times \Re^n \to \Re^m$ with $k(t,0) = 0$ for all $t \geq 0$, such that $0 \in \Re^n$ is GAS for the system*

$$\dot{x} = f(t,x) + g(t,x)k(t,x).$$

(iii) *For every C^0 function $\gamma(t,s) : (\Re^+)^2 \to \Re^+$, being locally Lipschitz in s and such that, for each $t \geq 0$, $\gamma(t,\cdot)$ is positive definite, there exists a C^∞ function $\tilde{k} : \Re^+ \times \Re^n \to \Re^m$ with $\tilde{k}(t,0) = 0$ for all $t \geq 0$, in such a way that the system*

$$\dot{x} = f(t,x) + g(t,x)\left(\tilde{k}(t,x) + v\right)$$

satisfies the wISS property with gain $\gamma(\cdot)$ from the input $v \in \Re^m$.

Proof. Equivalence between (i) and (ii) is an immediate consequence of Theorem 5.1. In order to establish (i) ⇔ (iii) consider the system

$$\dot{x} = \tilde{f}(t,x,v) + g(t,x)u, \tag{5.14}$$

where $\tilde{f}(t,x,v) := f(t,x) + g(t,x)v$, which has the form (1.3). The equivalence between (i) and (iii) follows directly from Theorem 5.1 and the obvious consequence of (5.14):

$$\frac{\partial V}{\partial x}(t,x)g(t,x) = 0, |v| \leq \gamma(t,|x|) \Rightarrow \frac{\partial V}{\partial t}(t,x) + \frac{\partial V}{\partial x}(t,x)\tilde{f}(t,x,v) \leq -\rho(V(t,x)) + \mu(t).$$

The rest part of proof is straightforward and is left to the reader. □

COROLLARY 5.4. *Consider the system*

$$\begin{aligned} &\dot{x} = f(x) + g(x)u, \\ &x \in \Re^n, \quad u \in \Re, \end{aligned} \tag{5.15}$$

where f and g are locally Lipschitz with $f(0) = 0$, and suppose that (5.15) is globally uniformly asymptotically stabilized at the origin by means of a C^0 static feedback $u = k(x)$ with $k(0) = 0$. Then for every C^0 function $\gamma(t, s) : (\Re^+)^2 \to \Re^+$, being locally Lipschitz in s and such that, for each $t \geq 0$, $\gamma(t, \cdot)$ is positive definite, there exists a C^∞ time-varying feedback law $u = k(t, x)$ with $k(\cdot, 0) = 0$ such that the system

$$\dot{x} = f(x) + g(x)(k(t, x) + u)$$

satisfies the wISS property with gain $\gamma(\cdot)$ from the input $u \in \Re$.

Proof. Using Kurzweil's converse Lyapunov theorem in [19] we may find a C^1 radially unbounded, positive definite function $V : \Re^n \to \Re^+$ that satisfies $\frac{\partial V}{\partial x}(x)(f(x) + g(x)k_0(x)) < 0$ for $x \neq 0$. It then follows that

$$\frac{\partial V}{\partial x}(x)g(x) = 0 \Rightarrow \frac{\partial V}{\partial x}(x)f(x) \leq -\rho(V(x)) + \mu(t) \tag{5.16}$$

for a certain C^0 positive definite function $\rho : \Re^+ \to \Re^+$ and for arbitrary $\mu \in \mathbf{E}$. The rest of the proof is straightforward consequence of (5.16) and Corollary 5.3 (implication (i) $\Rightarrow$ (iii)). □

Example 5.5. Consider the affine in the control system

$$\begin{aligned} &\dot{x} = x + y^3, \\ &\dot{y} = u, \\ &(x, y) \in \Re^2, \quad u \in \Re. \end{aligned} \tag{5.17}$$

It is known that there is no C^1 static feedback exhibiting uniform in time asymptotic stabilization at the origin for (5.17). However, a C^0 static feedback law exhibiting global uniform in time asymptotic stability exists, and several approaches can be used to obtain such a feedback. Alternatively, we may apply Corollary 5.3 to establish existence of a locally Lipschitz time-varying feedback $k(t, x, y)$ that guarantees nonuniform in time ISS for any given gain function $\gamma(\cdot)$ for the resulting system:

$$\begin{aligned} &\dot{x} = x + y^3, \\ &\dot{y} = k(t, x, y) + u, \\ &(x, y) \in \Re^2, \quad u \in \Re, \end{aligned} \tag{5.18}$$

with u as input. We may also use Proposition 5.2 to determine an explicit formula for a stabilizing feedback. Indeed, let $f(t, x, y) = (x + y^3, 0)$, $g(t, x, y) = (0, 1)$ and define

$$V(t, x, y) := 2\exp(2t)x^2 + (y + \exp(t)x)^2. \tag{5.19}$$

A simple calculation shows that

$$\frac{\partial V}{\partial y}(t, x, y) = 0 \Leftrightarrow y = -\exp(t)x. \tag{5.20}$$

For those (t,x,y) for which (5.20) holds we have $V(t,x,y) = 2\exp(2t)x^2$ and thus

$$\begin{aligned}\frac{\partial V}{\partial t}(t,x,y) + \frac{\partial V}{\partial x}(t,x,y)(x+y^3) &= 8\exp(2t)x^2 - 4\exp(5t)x^4 \\ &= 4V(t,x,y) - \exp(t)V^2(t,x,y) \\ &\le -\frac{1}{2}V^2(t,x,y) + 4\exp(-t).\end{aligned}$$

Therefore, both (5.13a) and (5.13b) are satisfied with $\rho(s) = \frac{1}{2}s^2$, $\mu(t) = 4\exp(-t)$, $a_1(s) := \frac{1}{2}s^2$, $a_2(s) := 4s^2$, and $\beta(t) := \exp(t)$, and thus, according to Corollary 5.3, for any gain function $\gamma(t,|(x,y)|)$, there exists a C^∞ time-varying feedback $k(t,x,y)$ with $k(\cdot,0,0) = 0$ such that the ISS property with gain $\gamma(\cdot)$ is fulfilled for (5.18). Finally, we may invoke Proposition 5.2 to find an explicit formula for a locally Lipschitz time-varying feedback. Indeed, by (5.7), (5.8), and (5.9) we find

$$k(t,x,y) = \frac{-(y+\exp(t)x)|16\exp(-t)D(t,x,y) + V^4(t,x,y)|}{2 - 2\theta(D(t,x,y)-1) + 8(y+\exp(t)x)^2},$$

where $V(\cdot)$ is defined by (5.19) and

$$\begin{aligned}D(t,x,y) := {} & 3\exp(3t)x^2 + \exp(2t)xy + \frac{3}{2}\exp(3t)xy^3 + \frac{1}{4}\exp(2t)y^4 \\ & + \frac{1}{2}\exp(t)|y+\exp(t)x|\gamma(t,|(x,y)|) + \frac{1}{16}\exp(t)V^4(t,x,y). \qquad \square\end{aligned}$$

5.2. Propagating the ISS property through integrators. In this section we apply Proposition 4.3 in order to derive sufficient conditions for ISS feedback stabilization for the particular class of systems (1.5), where $f(\cdot)$, $g(\cdot)$, $h(\cdot)$ are C^0 and locally Lipschitz with respect to (x,y) with $f(\cdot,0,0) = 0$ and $g(\cdot,0,0) = 0$. In addition to the regularity assumptions made for f,g,h, we further assume that there exists an everywhere strictly positive C^0 function $h_0 : \Re^+ \to (0,+\infty)$, such that

$$h(t,x,y) \ge h_0(t) \qquad \forall (t,x,y) \in \Re^+ \times \Re^n \times \Re. \tag{5.21}$$

As in the time-invariant case (see, for instance, [11, 36]), we impose ISS for the subsystem (1.5a); particularly, we make the following assumptions:

(A1) There exists a C^∞ function $k : \Re^+ \times \Re^n \to \Re$, with $k(\cdot,0) = 0$, such that the system

$$\dot{x} = f(t,x,k(t,x)+y) \tag{5.22}$$

satisfies the nonuniform in time ISS property from the input y. Specifically, assume that there exist functions a_1, a_2 of class K_∞, with a_1 being a locally Lipschitz function; a function β of class K^+; and a C^0 function $\gamma(t,s) : (\Re^+)^2 \to \Re^+$, which is locally Lipschitz in s and for each fixed $t \ge 0$ the map $\gamma(t,\cdot)$ is of class K_∞, in such a way that the following holds:

(5.23)
$$|y(t)| \le \gamma(t,|x(t)|) \text{ a.e. for } t \ge t_0 \Rightarrow a_1(|x(t)|) \le \exp(-t+t_0)\beta(t_0)a_2(|x_0|),$$

where $x(t) := x(t,t_0,x_0;y)$ denotes the trajectory of (5.22) with input y.

(A2) For the function $k(\cdot)$ above we make the following additional hypothesis. There exists a function $E : (\Re^+)^2 \to \Re^+$, with $E(\cdot, 0) = 0$, being nondecreasing in s for each fixed $t \geq 0$ in such a way that

$$|k(t,x)| \leq E(t,|x|) \qquad \forall (t,x) \in \Re^+ \times \Re^n, \tag{5.24a}$$

$$\lim_{t \to +\infty} E\left(t, a_1^{-1}\left(c \exp\left(-\frac{1}{2}t\right)\right)\right) = 0 \qquad \forall c \geq 0. \tag{5.24b}$$

(A3) There exist constants $R > 0$, $m \geq 1$ and a C^0 function $M : \Re^+ \to (0, +\infty)$ such that

$$a_2(s) \leq R s^{2m} \quad \text{for } s \text{ near zero}, \tag{5.25a}$$

$$M(t) s^m \leq \gamma(t,s) \quad \forall t \geq 0, \quad s \text{ near zero}. \tag{5.25b}$$

The following proposition generalizes a well-known result concerning ISS-feedback stabilization for autonomous systems under the presence of uniform in time ISS (see, for instance, [36]). It also constitutes an extension of the main result in [40] under the presence of "exponential," nonuniform in time ISS.

PROPOSITION 5.6. *Under* (A1), (A2), *and* (A3), *for any gain function* $\bar{\gamma}(t,s) : (\Re^+)^2 \to \Re^+$ *with the same properties as* γ *there exists an everywhere* C^∞ *function* $\bar{k} : \Re^+ \times \Re \to \Re$, *with* $\bar{k}(t,0) = 0$ *for all* $t \geq 0$, *such that system* (1.5) *with* $u := \bar{k}(t, y - k(t,x)) + u$ *satisfies the nonuniform in time ISS property with gain* $\bar{\gamma}$ *from the input* u.

Proof. The proof is based on the Lyapunov characterization of wISS (Proposition 4.3). The corresponding analysis is similar to that employed in [38, 39] and extremely simplifies the approach in [40], where ISS stabilization is exhibited under stricter assumptions. We proceed as follows. Our hypothesis (A1) guarantees, according to Proposition 4.3(iii), the existence of a C^0 function $U : \Re^+ \times \Re^n \to \Re^+$, which is locally Lipschitz on $\Re^+ \times (\Re^n \setminus \{0\})$, such that

$$a_1(|x|) \leq U(t,x) \leq \beta(t) a_2(|x|) \qquad \forall (t,x) \in \Re^+ \times \Re^n, \tag{5.26a}$$

$$\begin{aligned} &|y(t) - k(t, x(t))| \leq \gamma(t, |x(t)|) \text{ a.e. for } t \geq t_0 \\ &\Rightarrow U(t, x(t)) \leq \exp\left(-\frac{1}{2}(t - t_0)\right) U(t_0, x_0) \qquad \forall (t_0, x_0) \in \Re^+ \times \Re^n \text{ and } t \geq t_0, \end{aligned} \tag{5.26b}$$

where $(x(t), y(t))$ denotes the trajectory of the closed-loop system (1.5) with $u := \bar{k}(t, y - k(t,x)) + u$. Let us denote by $\gamma^{-1}(t,s)$ the inverse function of $\gamma(t,s)$ with respect to s; i.e., $\gamma^{-1}(\cdot)$ satisfies

$$\gamma\left(t, \gamma^{-1}(t,s)\right) = \gamma^{-1}(t, \gamma(t,s)) = s \qquad \forall (t,s) \in \Re^+ \times \Re^+.$$

Clearly, $\gamma^{-1}(t,s)$ is C^0 and for each fixed $t \geq 0$ the mapping $\gamma^{-1}(t,\cdot)$ is of class K_∞ as well. By Lemma 2.3, a pair of functions $a \in K_\infty \cap C^\infty((0,+\infty))$ and $\kappa \in K^+ \cap C^\infty(\Re^+)$ can be found in such a way that

$$\beta(t) a_2\left(\gamma^{-1}(t,s)\right) \leq a(\kappa(t)s) \qquad \forall (t,s) \in \Re^+ \times \Re^+. \tag{5.27}$$

We define

$$W(t,s) := a(\kappa(t)s). \tag{5.28}$$

Notice that, according to (A3), the function $W(\cdot)$ can be constructed in such a way that, in addition to (5.27), the following holds:

$$W(t,s) = \bar{M}(t)s^2, \quad t \geq 0,\ s \text{ near zero}, \tag{5.29}$$

for a certain function $\bar{M}(\cdot)$ of class $K^+ \cap C^\infty(\Re^+)$. Therefore without loss of generality we may assume that $W(\cdot)$ as defined by (5.28) is of class $C^\infty(\Re^+ \times \Re; \Re^+)$. It follows by (5.26a) and (5.27) that

$$W(t,|y-k(t,x)|) \leq U(t,x) \Rightarrow |y-k(t,x)| \leq \gamma(t,|x|), \tag{5.30a}$$

$$U(t,x) \leq W(t,|y-k(t,x)|) \Rightarrow |x| \leq a_1^{-1}(W(t,|y-k(t,x)|)). \tag{5.30b}$$

Next define

$$\begin{aligned} S_1 &:= \{(t,x,y) \in \Re^+ \times \Re^n \times \Re : W(t,|y-k(t,x)|) \leq U(t,x)\}, \\ S_2 &:= (\Re^+ \times \Re^n \times \Re) \setminus S_1, \end{aligned} \tag{5.31}$$

$$\Phi(t,x,y) := \begin{cases} U(t,x), & (t,x,y) \in S_1, \\ W(t,|y-k(t,x)|), & (t,x,y) \in S_2. \end{cases} \tag{5.32}$$

From (5.27), (5.28), (5.31), and definition (5.32) of Φ, it follows that Φ is C^0 and satisfies

$$\begin{aligned} \bar{a}_1(|(x,y-k(t,x))|) \leq \Phi(t,x,y) \leq \bar{\beta}(t)\bar{a}_2(|(x,y-k(t,x))|) \\ \forall (t,x,y) \in \Re^+ \times \Re^n \times \Re \end{aligned} \tag{5.33}$$

for certain $\bar{a}_1$, $\bar{a}_2 \in K_\infty$ and $\bar{\beta} \in K^+$. By taking into account (5.21) and (5.29) and applying standard partition of unity arguments, it follows that for every gain $\bar{\gamma}$ with the same properties as γ, a $C^\infty(\Re^+ \times \Re^n)$ function $\bar{k}(t,z)$ can be determined in such a way that $\bar{k}(t,0) = 0$ for all $t \geq 0$, and furthermore, for every $(t,z) \in \Re^+ \times (\Re \setminus \{0\})$, the following holds:

$$\begin{aligned} &\frac{\partial W}{\partial t}(t,|z|) + \frac{\partial W}{\partial s}(t,|z|)\mathrm{sgn}(z)\Big(g(t,x,k(t,x)+z) + h(t,x,k(t,x)+z)\big(\bar{k}(t,z)+u\big) \\ &\qquad - \frac{\partial k}{\partial t}(t,x) - \frac{\partial k}{\partial x}(t,x) f(t,x,k(t,x)+z)\Big) \\ &\leq -\frac{1}{2}W(t,|z|) + \exp(-t) \qquad \forall |u| \leq \bar{\gamma}(t,|(x,k(t,x)+z)|), |x| \leq a_1^{-1}(W(t,|z|)). \end{aligned} \tag{5.34}$$

We are now in a position to establish the ISS property for the resulting system

$$\begin{aligned} \dot{x} &= f(t,x,y), \\ \dot{y} &= g(t,x,y) + h(t,x,y)\bar{k}(t,y-k(t,x)) + h(t,x,y)u. \end{aligned} \tag{5.35}$$

Particularly, we show that, if $(x(t),y(t))$ denotes the trajectory of (5.35) initiated from (x_0,y_0) at time t_0 with input $v \in \mathbf{L}^\infty_{loc}$, then the following holds:

$$\begin{aligned} &|u(t)| \leq \bar{\gamma}(t,|(x(t),y(t))|) \text{ a.e. for } t \geq t_0 \\ &\qquad \Rightarrow \Phi(t,x(t),y(t)) \leq \exp\left(-\frac{1}{2}(t-t_0)\right)(\Phi(t_0,x_0,y_0)+2). \end{aligned} \tag{5.36}$$

Indeed, by taking into account (5.26b), (5.30a), (5.31), and (5.32) it follows that

$$\begin{aligned}\Phi(t,x(t),y(t)) = U(t,x(t)) &\le \exp\left(-\frac{1}{2}(t-t_0)\right) U(t_0,x(t_0)) \\ &\le \exp\left(-\frac{1}{2}(t-t_0)\right) \Phi(t_0,x(t_0),y(t_0)) \\ &\text{for the case } (t,x(t),y(t)) \in S_1,\ t \ge t_0,\end{aligned} \tag{5.37a}$$

whereas, by virtue of (5.30b), (5.31), (5.32), and (5.34) we obtain

$$\begin{aligned}\Phi(t,x(t),y(t)) &= W(t,|y(t)-k(t,x(t))|) \\ &\le \exp\left(-\frac{1}{2}(t-t_0)\right)\left(W(t_0,|y(t_0)-k(t_0,x(t_0))|)\right. \\ &\qquad \left. + \int_{t_0}^{t} \exp\left(-\frac{1}{2}(\tau+t_0)\right) d\tau\right) \\ &\le \exp\left(-\frac{1}{2}(t-t_0)\right)\left(\Phi(t_0,x(t_0),y(t_0)) + \int_{t_0}^{t} \exp\left(-\frac{1}{2}(\tau+t_0)\right) d\tau\right) \\ &\quad \text{for the case } (t,x(t),y(t)) \in S_2,\ t \ge t_0.\end{aligned} \tag{5.37b}$$

Combining both cases (5.37a), (5.37b) above and exploiting continuity of Φ, we get (5.36). It turns out by taking into account (5.24a), (5.26a), (5.28), (5.32), and (5.33) that

$$|u(t)| \le \bar{\gamma}(t,|(x(t),y(t))|) \text{ a.e. for } t \ge t_0 \Rightarrow a_1(|x(t)|) \le D(t,t_0,|(x_0,y_0)|), \tag{5.38}$$

$$|u(t)| \le \bar{\gamma}(t,|(x(t),y(t))|) \text{ a.e. for } t \ge t_0 \Rightarrow a(|y-k(t,x)|) \le D(t,t_0,|(x_0,y_0)|), \tag{5.39}$$

where $D(t,t_0,s) := \exp(-\frac{1}{2}(t-t_0))(1+\bar{\beta}(t_0)\bar{a}_2(s+E(t_0,s)))$. It follows from (5.24a), (5.38), and (5.39) that

$$\begin{aligned}|u(t)| \le \bar{\gamma}(t,|(x(t),y(t))|) &\text{ a.e. for } t \ge t_0 \\ &\Rightarrow |y(t)| \le E(t,a_1^{-1}(D(t,t_0,|(x_0,y_0)|))) + a^{-1}(D(t,t_0,|(x_0,y_0)|)),\end{aligned}$$

which by virtue of (5.24b), (5.38), and (5.39) guarantee the ISS property for (5.35) with gain $\bar{\gamma}$ from the input u. □

Conditions (A1), (A2), and (A3) do not in general guarantee that the feedback stabilizer $\bar{k}(\cdot)$ satisfies the same property (A2) imposed for the original feedback $k(\cdot)$. This is a drawback for the achievement of ISS partial-state feedback stabilization for higher dimensional triangular time-varying systems by applying backstepping design. Therefore, some additional conditions should be imposed for the original subsystem (1.5a) and the map $k(\cdot)$ in order to propagate (A2) to the new feedback $\bar{k}(\cdot)$, like those imposed in [40]. For instance, in [40] it was assumed that (1.5a) satisfies an exponential type of ISS from the input y and the dynamics have polynomial structure with respect to (t,x). Further generalizations of Proposition 5.6, as well as conditions weaker than those imposed in [40], which enable us to construct a smooth feedback with the same properties as $k(\cdot)$, are presented in [15]. We limited ourselves instead, to the case examined in [40], by re-establishing ISS stabilization for (1.5) by means of a smooth feedback $\bar{k}(\cdot)$ for which (A2) holds. We next show that the main result in [40] is a straightforward consequence of Proposition 5.6.

PROPOSITION 5.7. *Consider the system* (1.5) *with* $h(t,x,y) \equiv 1$, *and in addition to the regularity properties for* f, g, k, γ *imposed in Proposition* 5.6, *we assume that*

there exists a function r of class Π (see "Notations" for the definition of class Π) and constants a, $K > 0$ such that

$$|f(t,x,y)| + |g(t,x,y)| \le (1+t)^a r(|(x,y)|), \tag{5.40a}$$

$$|k(t,x)| + \left|\frac{\partial k}{\partial t}(t,x)\right| \le (1+t)^a r(|x|), \tag{5.40b}$$

$$\left|\frac{\partial k}{\partial x}(t,x)\right| \le (1+t)^a (1+r(|x|)), \tag{5.40c}$$

$$\frac{1}{K(1+t)^a} s \le \gamma(t,s) \le (1+t)^a r(s). \tag{5.40d}$$

Moreover, assume that subsystem (1.5a) *satisfies assumption* (A1) *with $a_1(s) = a_2(s) = s^2$ and $\beta(t) = M(1+t)^a$ for some constant $M > 0$. Then for any $\Gamma(\cdot) \in \Pi$ there exist a function $\bar r \in \Pi$, constants $\bar a \ge a$ and $\bar M \ge M$, and a feedback $\bar k(\cdot)$ as in statement of Proposition* 5.6 *such that property* (A1) *holds for* (1.5) *with $u := u + \bar k(t, y - k(t,x))$, $\gamma := \Gamma$, and some $a_1(\cdot)$, $a_2(\cdot)$, and $\bar\beta(t) = \bar M(1+t)^{\bar a}$, as well as inequalities* (5.40a), (5.40b), (5.40c), (5.40d), *are fulfilled with $\bar k(t, y - k(t,x))$, Γ, (x,y), $\bar a$, and $\bar r$ instead of $k(t,x)$, γ, x, a, and r, respectively.*

Proof (outline). It can be easily verified that all hypotheses (A1), (A2), and (A3) of Proposition 5.6 are fulfilled for (1.5). Particularly, (A2) holds as a consequence of (5.40b) and the fact that $r \in \Pi$. In order to establish our statement we proceed exactly as in the proof of Proposition 5.6. In our case we may use

$$W(t,s) = C(1+t)^{3a} s^2 \tag{5.41}$$

for some constant $C > 0$ (the constant a is defined in (5.40)) and we can find a polynomial $R \in \Pi$ of the form $R(s) = R_0(s + s^l)$ for $R_0 > 0$ and l being an odd positive integer and a constant $\theta \ge a$ such that (5.34) is fulfilled with

$$\bar k(t,y) := -(1+t)^\theta R(y) \tag{5.42}$$

and with $W(\cdot)$ as given by (5.41). The rest of the proof is the same as that given in proof of Proposition 5.6. Finally, it is immediate to see that, according to definition (5.42), the feedback $\bar k(\cdot)$ satisfies the same properties as those imposed for $k(\cdot)$; hence, it turns out that (A2) holds for the map $\bar k(\cdot)$. □

We may use the result of Proposition 5.7 and apply the induction procedure in order to re-establish Theorem 2.4 in [40], concerning partial-state feedback stabilization for a class of triangular systems.

Corollary 5.8. *Consider the system*

$$\dot x = f(t,x,y), \tag{5.43a}$$

$$\begin{gathered} \dot y_i = g_i(t,x,y_1,\dots,y_i) + y_{i+1}, \qquad i = 1,\dots,m, \\ x \in \Re^n, \quad y = (y_1,\dots,y_m)^T \in \Re^m, \quad t \ge 0, \quad u = y_{m+1} \in \Re, \end{gathered} \tag{5.43b}$$

where f, g_i are C^0 everywhere and locally Lipschitz with respect to (x,y) with $f(t,0,0) = 0$, $g_i(t,0,\dots,0) = 0$ for $i = 1,\dots,m$ and for all $t \ge 0$. Suppose that there exists a function r of class Π and a constant $a > 0$ such that

$$|f(t,x,y)| \le (1+t)^a r(|(x,y)|), \tag{5.44a}$$

$$|g_i(t,x,y_1,\dots,y_i)| \le (1+t)^a r(|(x,y_1,\dots,y_i)|). \tag{5.44b}$$

Moreover, assume that subsystem (5.43a) *satisfies assumption* (A1) *with* $k \equiv 0$, $a_1(s) = a_2(s) = s^2$, *and* $\beta(t) = M(1+t)^a$ *for some constant* $M > 0$ *and gain* $\gamma(t,s)$, *which is* C^0 *on* $\Re^+ \times \Re^+$ *and locally Lipschitz with respect to* $s \geq 0$ *and satisfies* $\gamma(t,\cdot) \in K_\infty$ *for all* $t \geq 0$, *in such a way that the following holds for some constant* $K > 0$*:*

$$\frac{1}{K(1+t)^a} s \leq \gamma(t,s) \leq (1+t)^a r(s). \tag{5.44c}$$

Then for any $\Gamma(\cdot) \in \Pi$ *there exists a* C^∞ *feedback law* $u = \bar{k}(t, y_1, \ldots, y_m)$ *such that system* (5.43) *with* $u := \bar{k}(t,y) + u$ *satisfies the ISS property with gain* Γ *from the input* u.

6. Conclusions. We have provided equivalent characterizations for the concept of robust global asymptotic stability (RGAS) for time-varying systems. Lyapunov characterizations for this concept as well as for the concept of nonuniform in time input-to-state stability (ISS) are given. Moreover, we have provided necessary and sufficient conditions for nonuniform in time ISS stabilization of affine in the control systems by means of a smooth time-varying feedback. An explicit formula for the time-varying feedback stabilizer is also presented. The problem of partial-state nonuniform in time ISS-feedback stabilization for triangular systems is considered.

REFERENCES

[1] D. Angeli and E.D. Sontag, *Forward completeness, unbounded observability and their Lyapunov characterizations*, Systems Control Lett., 38 (1999), pp. 209–217.

[2] D. Angeli, E.D. Sontag, and Y. Wang, *A characterization of integral input-to-state stability*, IEEE Trans. Automat. Control, 45 (2000), pp. 1082–1096.

[3] Z. Artstein, *Stabilization with relaxed controls*, Nonlinear Anal., 7 (1983), pp. 1163–1173.

[4] A. Bacciotti and L. Rosier, *Lyapunov stability and Lagrange stability: Inverse theorems for discontinuous systems*, Math. Control Signals Systems, 11 (1998), pp. 101–125.

[5] A. Bacciotti and L. Rosier, *On the converse of first Lyapunov theorem: The regularity issue*, Systems Control Lett., 41 (2000), pp. 265–270.

[6] A. Bacciotti and L. Rosier, *Liapunov Functions and Stability in Control Theory*, Lecture Notes in Control and Inform. Sci. 267, Springer–Verlag, London, 2001.

[7] J.M. Coron and L. Praly, *Adding an integrator for the stabilization problem*, Systems Control Lett., 17 (1991), pp. 89–104.

[8] W.P. Dayawansa and C.F. Martin, *A converse Lyapunov theorem for a class of dynamical systems which undergo switching*, IEEE Trans. Automat. Control, 44 (1999), pp. 751–760.

[9] L. Grune, *Input-to-state dynamical stability and its Lyapunov function characterization*, to appear.

[10] W. Hahn, *Stability of Motion*, Springer–Verlag, Berlin, 1967.

[11] Z.P. Jiang, A. Teel, and L. Praly, *Small-gain theorem for ISS systems and applications*, Math. Control Signals Systems, 7 (1994), pp. 95–120.

[12] I. Karafyllis and J. Tsinias, *Global stabilization and asymptotic tracking for a class of nonlinear systems by means of time-varying feedback*, Int. J. Robust Nonlinear Control, to appear.

[13] I. Karafyllis and J. Tsinias, *Converse Lyapunov theorems for non-uniform in time global asymptotic stability and stabilization by means of time-varying feedback*, in Nonlinear Control Systems 2001, Elsevier, New York, 2002, pp. 801–805.

[14] I. Karafyllis and J. Tsinias, *Non-uniform in time stabilization for linear systems and tracking control for nonholonomic systems in chained form*, Internat. J. Control, submitted.

[15] I. Karafyllis, *Non-uniform stabilization of control systems*, IMA J. Math. Control Inform., 19 (2002), pp. 419–444.

[16] I. Karafyllis and J. Tsinias, *Characterizations of the non-uniform in time ISS property and applications*, in Proceedings of the Fifteenth International Symposium on Mathematical Theory of Networks and Systems, Notre Dame, IN, 2002.

[17] I. Karafyllis, *Necessary and sufficient conditions for the existence of stabilizing feedback for control systems*, IMA J. Math. Control Inform., 20 (2003), pp. 37–64.
[18] H.K. Khalil, *Nonlinear Systems*, 2nd ed., Prentice–Hall, Englewood Cliffs, NJ, 1996.
[19] J. Kurzweil, *On the inversion of Lyapunov's second theorem on stability of motion*, Amer. Math. Soc. Transl. Ser. 2, 24 (1956), pp. 19–77.
[20] Y. Lin, E.D. Sontag, and Y. Wang, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.
[21] J.L. Mancilla-Aguilar and R.A. Garcia, *On converse Lyapunov theorems for ISS and iISS switched nonlinear systems*, Systems Control Lett., 42 (2001), pp. 47–53.
[22] E. Panteley and A. Loria, *On global uniform asymptotic stability of nonlinear time-varying systems in cascade*, Systems Control Lett., 33 (1998), pp. 131–138.
[23] L. Rifford, *Existence of Lipschitz and semiconcave control-Lyapunov functions*, SIAM J. Control Optim., 39 (2000), pp. 1043–1064.
[24] L. Rifford, *On the existence of non-smooth control Lyapunov functions in the sense of generalized gradients*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 593–611.
[25] L. Rosier, *Smooth Lyapunov functions for discontinuous stable systems*, Set-Valued Anal., 7 (1999), pp. 375–405.
[26] E.D. Sontag, *A Lyapunov-like characterization of asymptotic controllability*, SIAM J. Control Optim., 21 (1983), pp. 462–471.
[27] E.D. Sontag, *A "universal" construction of Artstein's theorem on nonlinear stabilization*, Systems Control Lett., 13 (1989), pp. 117–123.
[28] E.D. Sontag, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.
[29] E.D. Sontag and Y. Wang, *On characterizations of the input-to-state stability property*, Systems Control Lett., 24 (1995), pp. 351–359.
[30] E.D. Sontag and H.J. Sussmann, *Nonsmooth control-Lyapunov functions*, in Proceedings of the 34th IEEE Conference on Decision and Control, Vol. 3, IEEE Press, Piscataway, NJ, 1995, pp. 2799–2805.
[31] E.D. Sontag. and Y. Wang, *New characterizations of the input-to-state stability*, IEEE Trans. Automat. Control, 41 (1996), pp. 1283–1294.
[32] E.D. Sontag, *Comments on integral variants of ISS*, Systems Control Lett., 34 (1998), pp. 93–100.
[33] E.D. Sontag and Y. Wang, *Lyapunov characterizations of input to output stability*, SIAM J. Control Optim., 39 (2000), pp. 226–249.
[34] A.R. Teel and L. Praly, *A smooth Lyapunov function from a class-KL estimate involving two positive semidefinite functions*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 313–367.
[35] J. Tsinias, *Sufficient Lyapunov like conditions for stabilization*, Math. Control Signals Systems, 2 (1989), pp. 343-357.
[36] J. Tsinias, *Versions of Sontag's input to state stability condition and output feedback stabilization*, J. Math. Systems Estim. Control, 6 (1996), pp. 113–126.
[37] J. Tsinias, *Input to state stability properties of nonlinear systems and applications to bounded feedback stabilization using saturation*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 57–85.
[38] J. Tsinias, *Stochastic input-to-state stability and applications to global feedback stabilization*, Internat. J. Control, 71 (1998), pp. 907–931.
[39] J. Tsinias, *The concept of "exponential ISS" for stochastic systems and applications to feedback stabilization*, Systems Control Lett., 36 (1999), pp. 221–229.
[40] J. Tsinias and I. Karafyllis, *ISS property for time-varying systems and application to partial-static feedback stabilization and asymptotic tracking*, IEEE Trans. Automat. Control, 44 (1999), pp. 2179–2185.
[41] J. Tsinias, *Backstepping design for time-varying nonlinear systems with unknown parameters*, Systems Control Lett., 39 (2000), pp. 219–227.
[42] J. Tsinias, *A converse Lyapunov theorem for non-uniform in time, global exponential robust stability*, Systems Control Lett., 44 (2001), pp. 373–384.
[43] J. Tsinias and J. Spiliotis, *Notions of exponential robust stability, ISS and their Lyapunov characterizations*, Int. J. Robust Nonlinear Control, to appear.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 966–984

RIESZ BASIS PROPERTY OF EVOLUTION EQUATIONS IN HILBERT SPACES AND APPLICATION TO A COUPLED STRING EQUATION*

GEN-QI XU† AND BAO-ZHU GUO‡

Abstract. Suppose that $\{\lambda_n\}$ is the set of zeros of a sine-type generating function of the exponential system $\{e^{i\lambda_n t}\}$ in $L^2(0,T)$ and is separated. Levin and Golovin's classical theorem claims that $\{e^{i\lambda_n t}\}$ forms a Riesz basis for $L^2(0,T)$. In this article, we relate this result with Riesz basis generation of eigenvectors of the system operator of the linear time-invariant evolution equation in Hilbert spaces through its spectrum. A practically favorable necessary and sufficient condition for the separability of zeros of function of sine type is derived. The result is applied to get Riesz basis generation of a coupled string equation with joint dissipative feedback control.

Key words. Riesz basis, function of sine type, string equation

AMS subject classifications. 93C20, 93D15, 35B35, 35P10

DOI. 10.1137/S0363012901400081

1. Introduction.

In a Hilbert space, the most important bases are orthonormal bases. Second in importance are Riesz bases that are bases equivalent to some orthonormal basis. Riesz basis is studied in the context of stabilization of linear infinite dimensional system $\dot{x}(t) = Ax(t)$, in some Hilbert space $\mathbf{H}$, where A is the generator of a C_0-semigroup on $\mathbf{H}$. The system is called a Riesz spectral system [2] if there is a set of eigenvectors of A, which forms a Riesz basis for $\mathbf{H}$. For this kind of system, not only the stability of system is usually determined by the spectrum of the system operator A, which is referred to as the spectrum-determined growth condition, but also the dynamic of the system can be described by eigenpairs under expansion of nonharmonic Fourier series. Riesz basis is also the basis of the so-called method of moment, a powerful method in the study of controllability of hyperbolic systems [19], [1]. A nice recent result on the relation of exact controllability and Riesz basis can be found in [9].

Recently, some progress has been made for the Riesz basis generation of single beam equations under boundary feedback controls [3], [4] and coupled beams under joint dissipative feedback controls [5]. The basic idea is to show that the generalized eigenfunctions of the closed-loop system is quadratically close to that of the free system. This is actually an application of Bari's classical theorem that if $\{\phi_n\}_1^\infty$ is a Riesz basis for a Hilbert space $\mathbf{H}$ and another ω-linearly independent sequence $\{\psi_n\}_1^\infty$ of $\mathbf{H}$ satisfying

$$\sum_{n=1}^{\infty} \|\psi_n - \phi_n\|^2 < \infty, \tag{1}$$

*Received by the editors December 20, 2001; accepted for publication (in revised form) December 23, 2002; published electronically June 25, 2003. This research is supported by the National Natural Science Foundation of China.

http://www.siam.org/journals/sicon/42-3/40008.html

†Department of Mathematics, Shanxi University, Taiyuan 030006, Shanxi, China (gqxu@sxu.edu.cn).

‡Corresponding author. Institute of Systems Science, Academy of Mathematics and System Sciences, Academia Sinica, Beijing 100080, China (bzguo@iss03.iss.ac.cn).

then $\{\psi_n\}_1^\infty$ also forms a Riesz basis itself. The success of this approach to beam equations is attributed to their higher order of eigenfrequencies. In this sense, we can roughly say that the closed-loop system is a "perturbed" system of the free counterpart. In other words, the boundary feedback previously studied for beam equations are "low order" perturbations of the corresponding free systems. Recently, this idea was generally developed for the second order hyperbolic systems with collocated actuator/sensor by [6]. For string equations, however, this is not the case in general. A simple example is the following one dimensional wave equation with boundary feedback control:

$$\begin{cases} y_{tt}(x,t) - y_{xx}(x,t) = 0, & 0 < x < 1,\ t > 0, \\ y(0,t) = 0, \quad y_x(1,t) = u(t), \quad u(t) = -ky_t(1,t). \end{cases} \tag{2}$$

When $k \neq 1$, system (2) is a Riesz spectral system. However, the eigenfunctions of this system are never quadratically close to that of the free system ($k = 0$ in (2)) (see, e.g., [18]). By this reason, we may say that this closed-loop system possesses the same order as the associated free system. Moreover, in some special case such as $k = 1$ in (2), the system is never a Riesz spectral system.

This special property results in many different approaches to deal with Riesz basis generation for string equations. The basic approach is to estimate eigenvalues and eigenfunctions and then find some invertible transformation to transform the set of eigenfunctions to be an orthonormal basis. We refer to [20], [21] as well as many references therein. Mathematically, the general Riesz basis theory is developed in the context of nonharmonic Fourier series, which originated from the works of Paley and Wiener [16] and was developed later by many former Soviet mathematicians. Earlier results are summarized in [24], [15]. A nice summary of the later development can be found in [1]. Among them, a powerful concept—the so-called function of sine type—was introduced by Levin [10]. The main result due to Levin and Golovin says that if $\{\lambda_n\}$ is the set of zeros of a sine-type generating function of the exponential system $\{e^{i\lambda_n t}\}$ in $L^2(0,T)$ and is separated, then $\{e^{i\lambda_n t}\}$ forms a Riesz basis for $L^2(0,T)$.

In this paper, we first relate Levin and Golovin's theorem (generally, Pavlov's theorem) with Riesz basis generation of eigenvectors of the system operator of the time-invariant evolution equation in Hilbert spaces through its spectrum. A remarkable characterization condition is obtained for the separability of zeros of functions of sine type, which is considered practically a hard problem in many applications. The result is then applied to study the Riesz basis property of a coupled string equation jointed by a span dissipative feedback control.

The organization of this paper is as follows: in the next section, we shall first introduce some basic facts about functions of sine type. The main results and some remarks are presented. The proofs of main results are given in section 3. Finally, in section 4, we will check how our string system satisfies the sufficient condition obtained in this article—to be a Riesz spectral system in the energy Hilbert space.

2. Basic notation and main results. To begin with, let us recall some basic facts about functions of sine type. An entire function $f(z)$ is said to be of *exponential type* if the inequality

$$| f(z) | \leq Ae^{B|z|} \tag{3}$$

holds for some positive constants A and B and all complex values of z. The smallest of constants B is said to be the *exponential type* of $f(z)$. For the exponential-type

function f, define a 2π-periodic function on $\mathbb{R}$ by the equality

$$h_f(\phi) = \lim_{r\to\infty} \sup \frac{1}{r} \log |f(re^{i\phi})| \tag{4}$$

as a growth indicator of f. The indicator diagram of f is a convex set G_f such that

$$h_f(\phi) = \sup_{k\in G_f} \mathrm{Re}(ke^{-i\phi}). \tag{5}$$

Furthermore, the entire function f of exponential type is said to be a function of the *Cartwright class* if

$$\int_{\mathbb{R}} \frac{\max\{\log|f|, 0\}}{1+x^2} dx < \infty. \tag{6}$$

In particular, the function f of exponential type satisfying the condition

$$\int_{\mathbb{R}} \frac{|f(x)|^2}{1+x^2} dx < \infty \tag{7}$$

belongs to the *Cartwright class*. The indicator diagram of a *Cartwright class* function is an interval $[i\alpha, i\beta], \alpha \le \beta$, of the imaginary axis. Its length is the width of the indicator diagram [1, pp. 59–60].

An entire function of exponential type with simple zeros $\{\lambda_n\}$ and with the width of the indicator diagram T is called a generating function of exponential family $\{e^{i\lambda_n t}\}_1^\infty$ in $L^2(0,T)$ [1, p. 101]. The following theorem on the basis property of the exponential family $\{e^{i\lambda_n t}\}$ was obtained by Pavlov [17]. (See also Proposition II.3.17 and Theorem II.4.8 of [1].)

THEOREM 1 (Pavlov). *Let $\Lambda := \{\lambda_n\}_1^\infty$ be a countable set of the complex numbers. The family $\{e^{i\lambda_n t}\}_1^\infty$ forms a Riesz basis for $L^2(0,T)$ if and only if the following conditions are satisfied:*

(i) *the sequence $\{\lambda_n\}_1^\infty$ lies in a strip parallel to the real axis:*

$$\sup_{n\ge 1} |\mathrm{Im}\lambda_n| < \infty; \tag{8}$$

(ii) *$\{\lambda_n\}_1^\infty$ is separated in the sense that*

$$\delta(\Lambda) := \inf_{n\ne m} \mid \lambda_n - \lambda_m \mid > 0; \tag{9}$$

(iii) *the generating function of the family $\{e^{i\lambda_n t}\}$ on the interval $(0,T)$ satisfies the Muckenhoupt condition*

$$\sup_{I\in\mathcal{J}} \left\{ \frac{1}{|I|^2} \int_I |f(x+ih)|^2 dx \int_I |f(x+ih)|^{-2} dx \right\} < \infty \tag{10}$$

for some $h \in \mathbb{R}$, where $\mathcal{J}$ is the set of all intervals of the real axis. □

According to Definition II.1.27 of [1], an entire function of exponential type is said to be of sine type if

(a) the zeros of f lie in a strip $\{z \in \mathbb{C} || \mathrm{Im} z| \le H\}$ for some $H > 0$;

(b) there exist $h \in \mathbb{R}$ and positive constants c_1, c_2 such that $c_1 \le |f(x+ih)| \le c_2$ for all $x \in \mathbb{R}$.

It is seen that if the generation function f of $\{e^{i\lambda_n t}\}_1^\infty$ is of sine type, conditions (i) and (iii) of Theorem 1 are always satisfied. In this case, Theorem 1 will reduce to Levin and Golovin's well-known theorem (Proposition II.4.3 of [1]).

In order to apply Pavlov's theorem to get Riesz basis generation of general linear systems in Hilbert spaces, we need to relate eigenvectors of the system operator with the exponential system through its spectrum. Our main result on this point is stated as Theorem 2 below.

Let us recall that for a closed linear operator A in a Hilbert space $\mathbf{H}$, a nonzero $x \in \mathbf{H}$ is called a generalized eigenvector of A, corresponding to an eigenvalue λ of A, if there is a positive integer n such that $(\lambda - A)^n x = 0$. The root subspace of A corresponding to λ is given by

$$\mathbf{H}_\lambda = \{x \mid \exists n \text{ such that } (\lambda - A)^n x = 0\}.$$

The dimension m_λ of $\mathbf{H}_\lambda$ is called the algebraic multiplicity of λ. When $m_\lambda = 1$, we say that λ is algebraically simple. Let $\mathrm{Sp}(A)$, the so-called root subspace of A, be the closed subspace spanned by all generalized eigenvectors of A.

THEOREM 2. *Suppose that a linear (usually unbounded) operator A generates a C_0-semigroup e^{At} on a separable Hilbert space $\mathbf{H}$. Assume that A is a discrete operator (that is, $(\lambda - A)^{-1}$ is compact for some $\lambda \in \rho(A)$) with eigenvalues $\{\lambda_n\}_1^\infty$. Suppose that*

(a) *each eigenvalue λ_n is algebraically simple;*

(b) $\mathrm{Sp}(A) = \mathrm{Sp}(A^*) = \mathbf{H}$;

(c) $\{e^{\lambda_n t}\}_1^\infty$ *forms a Riesz basis for $L^2(0,T)$ for some $T > 0$;*

then there is a set of eigenvectors of A corresponding to $\{\lambda_n\}$, which forms a Riesz basis for $\mathbf{H}$. □

Remark 1. Naturally, one expects that the converse of Theorem 2 is also true, that is, if there is a set of eigenvectors of A that forms a Riesz basis for $\mathbf{H}$, $\{e^{\lambda_n t}\}_1^\infty$ forms a Riesz basis for $L^2(0,T)$ for some $T > 0$. Unfortunately, this is not generally true. In fact, by Theorem 1, a necessary condition for $\{e^{\lambda_n t}\}_1^\infty$ to form a Riesz basis for $L^2(0,T)$ is that $\{\lambda_n\}$ lies in a strip parallel to the imaginary axis. □

By Theorem 2, in order to check whether the system $\dot{x}(t) = Ax(t)$ in the Hilbert space $\mathbf{H}$ is a Riesz spectral system, we need check only conditions (a)–(c). In many applications, the spectrum of A multiplied by $-i$ is the set of zeros of some sine-type function, and such a function can be produced by the characteristic equation for which the spectrum is satisfied. So, the difficulty of checking condition (c) of Theorem 2 lies in the *separability* of $\{\lambda_n\}$ required in Theorem 1. For this reason, finding a characterization condition for the *separability* of zeros of functions of sine type appears to be pressing. Actually, it is already known from Corollary 1 of [24] that if $f(z)$ is a function of sine type with simple separated zeros $\{\lambda_n\}$, then

$$\inf_n \mid f'(\lambda_n) \mid > 0. \tag{11}$$

Unexpectedly, we find that condition (11) also serves a sufficient condition for zeros of $f(z)$ to be separated. This is given in the following theorem.

THEOREM 3. *Let $f(z)$ be an entire function of exponential type satisfying* (3). *Suppose that $f(z)$ is bounded on the real axis and the zeros $\{\lambda_n\}_1^\infty$ of $f(z)$ satisfy*

$$\sup_n \mid \mathrm{Im}\lambda_n \mid = h < \infty, \qquad \sup_n m_n < \infty,$$

where m_n is the multiplicity of λ_n as a zero of $f(z)$.

(i) *If*

$$\inf_n \mid f^{(m_n)}(\lambda_n) \mid > 0, \tag{12}$$

then $\{\lambda_n\}_1^\infty$ *is separated:* $\inf_{n\neq m} \mid \lambda_n - \lambda_m \mid > 0$.

(ii) *Conversely, if* $\{\lambda_n\}_1^\infty$ *is separated and there is a* $y_0 \in \mathbb{R}$ *such that*

$$\inf_{x\in\mathbb{R}} \mid f(x+iy_0) \mid > 0, \tag{13}$$

then (12) *holds true.*

Consequently, the necessary and sufficient condition for zeros $\{\lambda_n\}$ *of a function of sine type to be separated is that condition* (12) *holds true.* □

The second difficulty in applying Theorem 2 is the completeness of the root subspace, which is condition (b) in Theorem 2. To do this, let us recall Lemma 5 of [8] that if A is a discrete operator, then

$$\mathbf{H} = \sigma_\infty(A^*) \oplus \mathrm{Sp}(A),$$

where $\sigma_\infty(A^*) = \{x \in \mathbf{H} | R(\lambda, A^*)x$ is analytic in $\mathbb{C}\}$ [8, Lemma 6, p. 2296]. Hence for a discrete operator A, $\mathrm{Sp}(A) = H$ if and only if $\sigma_\infty(A^*) = \{0\}$.

For an H-valued entire function $f(z)$, one can also define its order ρ_f as the infimum of real number a so that

$$\|f(z)\|_{\mathbf{H}} = \mathcal{O}(e^{|z|^a}) \text{ as } |z| \to \infty;$$

see [23]. We can now state our result on the completeness of the root subspace, which is characterized by the first order resolvent operator of the adjoint operator. Here we use order instead of type as it was used in [23] for operators with resolvent operator being of quotient of entire functions of exponential type. The advantage is that the order is more easily determined than type in applications (see also Theorem 2.6.2 of [13]).

THEOREM 4. *Let* A *be the generator of a* C_0*-semigroup in a Hilbert space* $\mathbf{H}$. *Assume that* A *is discrete (so is* A^**) and for* $\lambda \in \rho(A^*)$, $R(\lambda, A^*)$ *is of the form*

$$R(\lambda, A^*)x = \frac{G(\lambda)x}{F(\lambda)} \qquad \forall x \in \mathbf{H},$$

where for each $x \in \mathbf{H}$, $G(\lambda)x$ *is an H-valued entire function with order less than or equal to* ρ_1 *and* $F(\lambda)$ *is a scalar entire function of order* ρ_2. *Let* $\rho = \max\{\rho_1, \rho_2\} < \infty$ *and an integer* n *so that* $n - 1 \le \rho < n$. *If there are* $n+1$ *rays* γ_j, $j = 0, 1, 2, \ldots, n$, *on the complex plane*

$$\arg\gamma_0 = \frac{\pi}{2} < \arg\gamma_1 < \arg\gamma_2 \cdots < \arg\gamma_n = \frac{3\pi}{2}$$

with

$$\arg\gamma_{j+1} - \arg\gamma_j \le \frac{\pi}{n}, \qquad 0 \le j \le n-1,$$

so that $R(\lambda, A^*)x$ *is bounded on each ray* γ_j, $0 < j < n$, *as* $|\lambda| \to \infty$ *for any* $x \in \mathbf{H}$, *then* $\mathrm{Sp}(A) = \mathrm{Sp}(A^*) = \mathbf{H}$. □

3. Proof of main results. Before proving Theorem 2, we show the following lemma, which is a natural generalization of the well-known fact for $\mathbf{H} = \mathbb{C}$.

LEMMA 1. *Let* $\mathbf{H}$ *be a separable Hilbert space and* $\{e_n(t)\}_1^\infty$ *be a Riesz basis for* $L^2(0,T)$, $T > 0$. *Then for any* $\phi \in L^2(0,T;\mathbf{H})$ *there exists a sequence* $\{\phi_n\}_1^\infty \subset \mathbf{H}$ *such that*

(i)

$$\phi(t) = \sum_{n=1}^{\infty} e_n(t)\phi_n, \tag{14}$$

which holds in $L^2(0,T;\mathbf{H})$, *where* ϕ_n *is uniquely determined by*

$$\phi_n = \int_0^T e_n^*(t)\phi(t)dt \tag{15}$$

and $\{e_n^*(t)\}$ *is the biorthogonal system with respect to* $\{e_n(t)\}$ *in* $L^2(0,T)$ [24, p. 28],

$$\langle e_n^*(\cdot), e_m(\cdot)\rangle = \delta_{nm} \text{ whenever } n, m \geq 1;$$

(ii) *there are constants* $C_i > 0$, $i = 1, 2$, *such that*

$$C_1 \sum_{n=1}^{\infty} \|\phi_n\|_{\mathbf{H}}^2 \leq \|\phi\|_{L^2(0,T;H)}^2 \leq C_2 \sum_{n=1}^{\infty} \|\phi_n\|_{\mathbf{H}}^2. \tag{16}$$

Proof. Our proof is constructed by the following steps. First, take $\{\psi_n\}$ as an orthonormal basis of $\mathbf{H}$. Then for almost every $t \in [0,T]$, one can expand $\phi \in L^2(0,T;\mathbf{H})$ as

$$\phi(t) = \sum_{n=1}^{\infty} \langle \phi(t), \psi_n\rangle_{\mathbf{H}} \psi_n, \qquad t \in [0,T] \text{ a.e.},$$

and so

$$\|\phi(t)\|_{\mathbf{H}}^2 = \sum_{n=1}^{\infty} |\, \langle \phi(t), \psi_n\rangle_{\mathbf{H}} \,|^2 \qquad \forall t \in [0,T] \text{ a.e.} \tag{17}$$

Second, since $\langle \phi(t), \psi_m\rangle_{\mathbf{H}} \in L^2(0,T)$ for every $m \geq 1$, one can write

$$\langle \phi(t), \psi_m\rangle_{\mathbf{H}} = \sum_{n=1}^{\infty} a_n^{(m)} e_n(t) \qquad \forall m \geq 1 \text{ in } L^2(0,T), \tag{18}$$

where the coefficients $a_n^{(m)}$ are determined by

$$a_n^{(m)} = \int_0^T \langle \phi(t), \psi_m\rangle_{\mathbf{H}} e_n^*(t)dt \tag{19}$$

with the property that

$$C_1 \sum_{n=1}^{\infty} |\, a_n^{(m)} \,|^2 \leq \int_0^T |\, \langle \phi(t), \psi_m\rangle_{\mathbf{H}} \,|^2 \, dt \leq C_2 \sum_{n=1}^{\infty} |\, a_n^{(m)} \,|^2 \tag{20}$$

for some constants $C_i > 0$, $i = 1, 2$, which depend only on $\{e_n(t)\}$. Set

$$\phi_n = \sum_{m=1}^{\infty} a_n^{(m)} \psi_m. \tag{21}$$

We show that $\{\phi_n\}$ is the sequence required. Actually, (21) makes sense since by (19) and (17)

$$\sum_{m=1}^{\infty} \mid a_n^{(m)} \mid^2 \leq C \sum_{m=1}^{\infty} \int_0^T \mid \langle \phi(t), \psi_m \rangle_{\mathbf{H}} \mid^2 dt = C \int_0^T \|\phi(t)\|_{\mathbf{H}}^2 dt \tag{22}$$

for some positive constant $C > 0$. Thus $\phi_n \in \mathbf{H}$. Furthermore, by (20) and (22)

$$C_1 \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \mid a_n^{(m)} \mid^2 \leq \sum_{m=1}^{\infty} \int_0^T \mid \langle \phi(t), \psi_m \rangle_{\mathbf{H}} \mid^2 dt = \int_0^T \|\phi(t)\|_{\mathbf{H}}^2 dt. \tag{23}$$

This, together with (21), gives

$$\sum_{n=1}^{\infty} \|\phi_n\|_{\mathbf{H}}^2 = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \mid a_n^{(m)} \mid^2 < \infty. \tag{24}$$

Furthermore, for any integer $N > 1$ and almost any $t \in [0, T]$, it has

$$\begin{aligned} &\left\| \phi(t) - \sum_{n=1}^{N} e_n(t) \phi_n \right\|_{\mathbf{H}}^2 = \left\| \phi(t) - \sum_{n=1}^{N} e_n(t) \sum_{m=1}^{\infty} a_n^{(m)} \psi_m \right\|_{\mathbf{H}}^2 \\ &= \left\| \phi(t) - \sum_{m=1}^{\infty} \left(\sum_{n=1}^{N} a_n^{(m)} e_n(t) \right) \psi_m \right\|_{\mathbf{H}}^2 = \sum_{m=1}^{\infty} \left| \langle \phi(t), \psi_m \rangle - \sum_{n=1}^{N} a_n^{(m)} e_n(t) \right|^2, \end{aligned}$$

and hence there exists a $C_3 > 0$ such that

$$\begin{aligned} \int_0^T \left\| \phi(t) - \sum_{n=1}^{N} e_n(t) \phi_n \right\|_{\mathbf{H}}^2 dt &= \sum_{m=1}^{\infty} \int_0^T \left| \langle \phi(t), \psi_m \rangle - \sum_{n=1}^{N} a_n^{(m)} e_n(t) \right|^2 dt \\ &\leq C_3 \sum_{m=1}^{\infty} \sum_{n=N+1}^{\infty} |a_n^{(m)}|^2. \end{aligned}$$

Letting $N \to 0$, we obtain (14). Finally, by (23) and (24),

$$C_1 \sum_{n=1}^{\infty} \|\phi_n\|_{\mathbf{H}}^2 \leq \int_0^T \|\phi(t)\|_{\mathbf{H}}^2 dt.$$

By (23) and (20),

$$\int_0^T \|\phi(t)\|_{\mathbf{H}}^2 dt = \sum_{m=1}^{\infty} \int_0^T \mid \langle \phi(t), \psi_m \rangle_{\mathbf{H}} \mid^2 dt \leq C_2 \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \mid a_n^{(m)} \mid^2 = C_2 \sum_{n=1}^{\infty} \|\phi_n\|_{\mathbf{H}}^2.$$

The proof is complete. □

Proof of Theorem 2. Let $\{x_n\}_1^\infty$ be eigenvectors of A corresponding to $\{\lambda_n\}$, $\|x_n\| = 1$ for all $n \geq 1$. By assumption, $\{x_n\}_1^\infty$ is complete in $\mathbf{H}$, hence there exists a unique biorthogonal sequence $\{x_n^*\}_1^\infty$ of $\{x_n\}_1^\infty$:

$$\langle x_n, x_m^* \rangle = \delta_{nm}.$$

In order for $\{x_n\}_1^\infty$ to be a Riesz basis for $\mathbf{H}$, it suffices to show that $\{x_n^*\}$ is also complete in $\mathbf{H}$ and for any $\psi \in \mathbf{H}$,

$$\sum_{n=1}^{\infty} |\langle \psi, x_n^* \rangle|^2 < \infty, \qquad \sum_{n=1}^{\infty} |\langle \psi, x_n \rangle|^2 < \infty$$

[24, Thm. 9, p. 32]. Now, for any given $\psi \in \mathbf{H}$, find $\psi_m \to \psi$ as $m \to \infty$:

$$\psi_m = \sum_{n=1}^{m} b_n^{(m)} x_n,$$

where $b_n^{(m)}$ is constant. It is seen that

$$b_n^{(m)} = \langle \psi_m, x_n^* \rangle \to \langle \psi, x_n^* \rangle \text{ as } m \to \infty.$$

The associated solution to the Cauchy problem

$$\dot{x}(t) = Ax(t), x(0) = \psi_m \tag{25}$$

is

$$x(t) = \sum_{n=1}^{m} e^{\lambda_n t} b_n^{(m)} x_n.$$

It follows from the left inequality of (16) that

$$C_1 \sum_{n=1}^{m} |b_n^{(m)}|^2 \leq \int_0^T \|e^{At}\psi_m\|^2 dt \leq \frac{M^2}{2\omega}(e^{2\omega T} - 1)\|\psi_m\|^2$$

for some positive constant C_1, where we assume that the semigroup e^{At} satisfies $\|e^{At}\| \leq Me^{\omega t}$ for some $M, \omega > 0$. Letting $m \to \infty$ above yields

$$\sum_{n=1}^{\infty} |\langle \psi, x_n^* \rangle|^2 < \infty.$$

Finally, since $\{x_n^*\}$ are eigenvectors of A^* corresponding to $\{\overline{\lambda_n}\}$, which is complete on $\mathbf{H}$ by assumption, $\{e^{\overline{\lambda_n} t}\}$ also forms a Riesz basis for $\mathbf{H}$. Repeating the above process to A^*, we obtain

$$\sum_{n=1}^{\infty} |\langle \psi, x_n \rangle|^2 < \infty,$$

which completes the proof. □

To prove Theorem 3, we need the following lemma.

LEMMA 2. *Let $f(z)$ be an analytic function on $\mid z - z_0 \mid < \delta$, $\delta > 1$. Suppose that $\mid f(z) \mid \leq M$ on $\mid z - z_0 \mid \leq 1$ and z_0 is a kth order zero of $f(z)$ with*

$$\frac{f^{(k)}(z_0)}{k!} = a_k \neq 0. \tag{26}$$

Then z_0 is the unique zero point of $f(z)$ in the disc $\mid z - z_0 \mid \leq \frac{|a_k|}{4(|a_k|+M)}$.

Proof. By assumption, we can write the Taylor expansion of $f(z)$ at $z = z_0$ as

$$f(z) = \sum_{n=k}^{\infty} a_k (z - z_0)^k, \qquad \mid z - z_0 \mid \leq 1.$$

In view of the Cauchy inequality (see, e.g., section 2.5 of [22]) and by assumption, we have

$$\mid a_n \mid \leq M \qquad \forall n \geq k.$$

Now for $0 < \mid z - z_0 \mid \leq r = \frac{|a_k|}{4(|a_k|+M)}$, it has

$$\begin{aligned} \mid f(z) \mid &= \mid z - z_0 \mid^k \mid a_k + \textstyle\sum_{n=1}^{\infty} a_{k+n}(z - z_0)^n \mid \\ &\geq \mid z - z_0 \mid^k [\mid a_k \mid - \textstyle\sum_{n=1}^{\infty} \mid a_{k+n} \mid\mid z - z_0 \mid^n] \\ &\geq \mid z - z_0 \mid^k [\mid a_k \mid - M \textstyle\sum_{n=1}^{\infty} r^n] = \mid z - z_0 \mid^k [\mid a_k \mid - M \frac{r}{1-r}] \\ &= \mid z - z_0 \mid^k \frac{3|a_k|}{4(1-r)} > 0, \end{aligned}$$

proving the lemma. □

Proof of Theorem 3. By assumption, we may assume

$$\mid f(x) \mid \leq M_0 \qquad \forall x \in \mathbb{R}.$$

It follows from Theorem 11 of [24] that

$$\mid f(x + iy) \mid \leq M_0 e^{B|y|} \qquad \forall x, y \in \mathbb{R}.$$

In particular, for $H = h + 2$, it has

$$\mid f(z) \mid \leq M_0 e^{BH} = M \qquad \forall \mid \mathrm{Im} z \mid \leq H. \tag{27}$$

Obviously, $|\mathrm{Im} z| < H$ for $\mid z - \lambda_n \mid < 2$; it follows particularly that

$$\mid f(z) \mid \leq M \text{ in } \mid z - \lambda_n \mid \leq 1. \tag{28}$$

Now, since λ_n is an m_nth order zero of $f(z)$,

$$\frac{f^{(m_n)}(\lambda_n)}{m_n!} = a_n \neq 0. \tag{29}$$

Applying Lemma 1 to $f(z)$, we know that $f(z)$ has only one zero point in

$$\mid z - \lambda_n \mid \leq \frac{\mid a_n \mid}{4(\mid a_n \mid + M)} \geq \frac{C}{4(C + M)}, \qquad C = \inf_n \left| \frac{f^{(m_n)}(\lambda_n)}{m_n!} \right| > 0.$$

Consequently, $f(z)$ has only one zero point in $\mid z-\lambda_n \mid \leq \frac{C}{4(C+M)}$. Therefore,

$$\inf_{n\neq m} \mid \lambda_n - \lambda_m \mid > \frac{C}{4(C+M)} > 0.$$

(i) is thus proved.

Suppose (13). We may assume without loss of generality that $\mid y_0 \mid < H_0$, where $h+1 < H_0 <$H. Set

$$\inf_{n\neq m} \mid \lambda_n - \lambda_m \mid > 4\epsilon > 0, \qquad \epsilon < 1.$$

$B_n(\epsilon) = \{z \in \mathbb{C} \mid\mid z-\lambda_n \mid \leq \epsilon\}$. Define

$$\Omega = \{z \in \mathbb{C} \mid\mid \mathrm{Im}z \mid \leq H_0\} - \bigcup_{n=1}^{\infty} B_n(\epsilon). \tag{30}$$

We first show that

$$\mid f(z) \mid \geq \alpha \text{ on } \overline{\Omega} \tag{31}$$

for some $\alpha > 0$, where $\overline{\Omega}$ is the closure of Ω. Since otherwise, there is a sequence $\{z_m\}_1^\infty \subset \overline{\Omega}$ so that

$$\lim_{m\to\infty} f(z_m) = 0.$$

Write $z_m = x_m + iy_m$, $x_m, y_m \in \mathbb{R}$. Since $\mid y_m \mid \leq H_0$, we may assume without loss of generality that

$$y_m \to \eta \text{ as } m \to \infty. \tag{32}$$

For each $m \geq 1$, define

$$g_m(z) = f(z+x_m), \qquad z \in \{z \in \mathbb{C} \mid\mid \mathrm{Im}z \mid < H\}. \tag{33}$$

Then it follows from (27) that

$$\mid g_m(z) \mid \leq M \qquad \forall \mid \mathrm{Im}z \mid < H.$$

By Montel's theorem (Theorem 5.22 in [22]), there exist a subsequence of $\{g_m\}$ still denoted by $\{g_m\}$ and an analytic function $g(z)$ on |Im z|$< H$ such that

$$g_m(z) \to g(z) \text{ uniformly on any compact subset of } \mid \mathrm{Im}z \mid \leq H_0 \text{ as } m \to \infty. \tag{34}$$

Now that

$$g_m(iy_m) = f(z_m) \to 0 \text{ and } \inf_m \mid g_m(iy_0) \mid \geq \inf_{x\in\mathbb{R}} \mid f(x+iy_0) \mid > 0,$$

it follows that $g(i\eta) = 0, \mid g(iy_0) \mid > 0$. That is, g does not vanish identically on |Imz|$\leq H_0$. By Hurwitz's theorem (Theorem 3.45 in [22]) there is an $N > 0$ such that for each $m > N$, $g_m(z)$ has one zero point w_m so that

$$\mid w_m - i\eta \mid \leq \frac{\epsilon}{2} \qquad \forall m > \mathrm{N}. \tag{35}$$

That is,

$$f(x_m + w_m) = 0 \text{ and } \mid w_m - i\eta \mid \leq \frac{\epsilon}{2} \qquad \forall m > \mathrm{N}. \tag{36}$$

Since

$$\mid x_m + w_m - x_m - iy_m \mid = \mid w_m - iy_m \mid \leq \mid w_m - i\eta \mid + \mid iy_m - i\eta \mid \leq \frac{\epsilon}{2} \text{ as } m \to \infty,$$

we see that there exists an $N_0 > N$ such that as $m > N_0$,

$$\mid x_m + w_m - z_m \mid \leq \frac{\epsilon}{2} < \epsilon.$$

This is a contradiction since $x_m + w_m \in \{\lambda_n\}_1^\infty$ and $\{z_m\} \subset \overline{\Omega}$. (31) is thus verified.

Notice that $B_n(\epsilon)$ are nonoverlapping, the function $f(z)/(z-\lambda_n)^{m_n}$ is analytic and free of zero in $B_n(\epsilon)$, and

$$\left| \frac{f(z)}{(z-\lambda_n)^{m_n}} \right| \geq \frac{\alpha}{\epsilon^{m_n}} \text{ on } \mid z - \lambda_n \mid = \epsilon.$$

It follows from Jensen's formula [24, Thm. 2, p. 59] that

$$\log \left| \frac{f^{(m_n)}(\lambda_n)}{m_n!} \right| \geq \log \left(\frac{\alpha}{\epsilon^{m_n}} \right)$$

and hence

$$\left| \frac{f^{(m_n)}(\lambda_n)}{m_n!} \right| \geq \frac{\alpha}{\epsilon^{m_n}}.$$

The proof is complete. □

Proof of Theorem 4. First we show that $\sigma_\infty(A^*) = \{0\}$. Note that for any $x \in \sigma_\infty(A^*)$, $R(\lambda, A^*)x$ is an H-valued entire function of λ with order less than or equal to ρ. Since A^* also generates a C_0-semigroup on $\mathbf{H}$, we may assume without loss of generality that $R(\lambda, A^*)x$ is bounded on the right half complex plane, particularly on the imaginary axis. Set

$$S_j = \{\lambda \in \mathbb{C} \mid \arg \gamma_{j-1} \leq \arg \lambda \leq \arg \gamma_j\}, \qquad j = 1, 2, \ldots, n.$$

By assumption, $R(\lambda, A^*)x$ is bounded on the boundary of S_j, and

$$\|R(\lambda, A^*)x\| = \mathcal{O}(e^{|\lambda|^{\rho+\epsilon}}) \qquad \forall \lambda \in S_j,$$

where $\epsilon > 0$ is chosen so that $\rho + \epsilon < n$. Applying the Phragmén–Lindelöf theorem to $R(\lambda, A^*)x$ in each S_j [24, Thm. 10, p. 80], we know that $R(\lambda, A^*)x$ is uniformly bounded in S_j and so is in the whole complex plane. It follows from the Liouville's theorem that $R(\lambda, A^*)x$ is an constant element of $\mathbf{H}$. Furthermore, by the Hille–Yosida theorem,

$$\lim_{\lambda \to +\infty} R(\lambda, A^*)x = 0$$

and hence $R(\lambda, A^*)x = 0$ or $x = 0$, proving that $\sigma_\infty(A^*) = \{0\}$. Finally, since

$$R(\overline{\lambda}, A)x = \frac{G^*(\lambda)x}{\overline{F(\lambda)}} \qquad \forall x \in \mathbf{H},$$

it holds similarly that $\sigma_\infty(A) = \{0\}$. The proof is complete. □

4. Application to coupled strings. Since the game is almost the same for other types of boundary conditions and joint linear feedback controls, we demonstrate the whole process by considering the following string equation with one end fixed, one end free, and a force stabilizer at the joint d, $0 < d < 1$:

$$\begin{cases} y_{tt}(x,t) - y_{xx}(x,t) = 0, & 0 < x < d, d < x < 1, \\ y(0,t) = y_x(1,t) = 0, \\ y(d^-,t) = y(d^+,t), \\ y_x(d^-,t) - y_x(d^+,t) = -\alpha y_t(d,t), & \alpha > 0. \end{cases} \tag{37}$$

To turn system (37) into a framework of semigroups, we introduce the underlying state Hilbert space $\mathcal{H} = H_E^1(0,1) \times L^2(0,1)$ with the inner product induced norm:

$$\|(u,v)\|^2 = \int_0^1 [\mid u'(x) \mid^2 + \mid v(x) \mid^2] dx,$$

where $H_E^1(0,1) = \{u \mid u \in H^1(0,1), u(0) = 0\}$. System (37) is then written as an evolutionary equation in $\mathcal{H}$:

$$\frac{d}{dt} Y(t) = \mathcal{A} Y(t), \tag{38}$$

where $Y(t) = (y(\cdot,t), y_t(\cdot,t)) \in \mathcal{H}$ and $\mathcal{A}$ is defined by

$$\mathcal{A}(u,v) = (v(x), u''(x)) \tag{39}$$

with

$$\begin{aligned} D(\mathcal{A}) = \{&(u,v) \in H^1(0,1) \times H_E^1(0,1) \mid u(0) = u'(1) = 0, \\ &u \mid_{[0,d]} \in H^2(0,d), u \mid_{[d,1]} \in H^2(d,1), u'(d^-) - u'(d^+) = -\alpha v(d)\}, \end{aligned} \tag{40}$$

where $u \mid_{[a,b]}$ denotes the function u confined to $[a,b]$.

LEMMA 3.

(i) *$\mathcal{A}^{-1}$ exists and is compact on $\mathcal{H}$.*

(ii) *$\mathcal{A}$ is dissipative and hence $\mathcal{A}$ generates a C_0-semigroup of contractions on $\mathcal{H}$.*

(iii) *$\lambda \in \sigma(\mathcal{A})$ if and only if λ is a zero of $g(\lambda)$:*

$$g(\lambda) = 2\alpha^{-1} \cosh \lambda + \sinh \lambda - \sinh \lambda (1 - 2d). \tag{41}$$

(iv) *$\lambda \in \sigma(\mathcal{A})$ is algebraically simple if and only if $g(\lambda) = 0$, $g'(\lambda) \neq 0$.*

Proof. (i) and (ii) are straightforward. In particular, $0 \in \rho(\mathcal{A})$. We prove only (iii) and (iv). It is easy to know that for each $\lambda \in \sigma(\mathcal{A})$, the corresponding eigenfunction takes the form $(\phi, \lambda\phi)$ with $\phi \neq 0$ satisfying

$$\begin{cases} \lambda^2 \phi(x) - \phi''(x) = 0, & 0 < x < d, d < x < 1, \\ \phi(0) = \phi'(1) = 0, \\ \phi(d^-) = \phi(d^+), \\ \phi'(d^-) - \phi'(d^+) = -\alpha\lambda\phi(d), & \alpha > 0. \end{cases} \tag{42}$$

Solving (42) gives

$$\phi(x) = \begin{cases} c_1 \sinh \lambda x, & 0 < x < d, \\ c_2 \cosh \lambda (1 - x), & d < x < 1, \end{cases}$$

where c_1 and c_2 satisfy

$$\begin{cases} c_1 \sinh \lambda d - c_2 \cosh \lambda(1-d) = 0, \\ c_1(\cosh \lambda d + \alpha \sinh \lambda d) + c_2 \sinh \lambda(1-d) = 0. \end{cases}$$

Since c_1, c_2 cannot vanish simultaneously, solving the above equation yields (41).

As for (iv), since it is not the standard problem studied in [13], we give here a direct proof. There are two cases.

Case 1 ($|\sinh \lambda d| + |\cosh \lambda(1-d)| \neq 0$ for any λ). In this case, the eigenfunction corresponding to λ is $(\phi(x,\lambda), \lambda\phi(x,\lambda))$, where

$$\phi(x,\lambda) = \begin{cases} \cosh \lambda(1-d) \sinh \lambda x, & 0 < x < d, \\ \sinh \lambda d \cosh \lambda(1-x), & d < x < 1. \end{cases} \tag{43}$$

Notice that the function $\phi(x,\lambda)$ defined above satisfies

$$\begin{cases} \lambda^2 \phi(x,\lambda) - \phi''(x,\lambda) = 0, & 0 < x < d, d < x < 1, \\ \phi(0,\lambda) = \phi'(1,\lambda) = 0, \\ \phi(d^-,\lambda) = \phi(d^+,\lambda) \end{cases} \tag{44}$$

for all complex numbers λ. Differentiating the above equation with respect to λ, we obtain

$$\begin{cases} \lambda^2 \phi_\lambda(x,\lambda) - \phi''_\lambda(x,\lambda) = -2\lambda\phi(x,\lambda), & 0 < x < d, d < x < 1, \\ \phi_\lambda(0,\lambda) = \phi'_\lambda(1,\lambda) = 0, \\ \phi_\lambda(d^-,\lambda) = \phi_\lambda(d^+,\lambda). \end{cases} \tag{45}$$

Now confining $\lambda \in \sigma(\mathcal{A})$ and solving

$$(\lambda - \mathcal{A})(f,g) = (\phi, \lambda\phi), \tag{46}$$

we have $g = \lambda f - \phi$, and f satisfies

$$\begin{cases} \lambda^2 f(x) - f''(x) = 2\lambda\phi(x,\lambda), & 0 < x < d, d < x < 1, \\ f(0) = f'(1) = 0, \\ f(d^-) = f(d^+), \\ f'(d^-) - f'(d^+) = -\alpha\lambda f(d) + \alpha\phi(d,\lambda). \end{cases} \tag{47}$$

Since λ is geometrically simple, it is algebraically simple if and only if there is no solution to (47). Let

$$z(x) = f(x) + \phi_\lambda(x,\lambda).$$

Then z satisfies

$$\begin{cases} \lambda^2 z(x) - z''(x) = 0, & 0 < x < d, d < x < 1, \\ z(0) = z'(1) = 0, \\ z(d^-) = z(d^+), \\ z'(d^-) - z'(d^+) = -\alpha\lambda z(d) + \beta, \end{cases} \tag{48}$$

where

$$\beta = \alpha\phi(d,\lambda) + \alpha\lambda\phi_\lambda(d,\lambda) + \phi'_\lambda(d^-,\lambda) - \phi'_\lambda(d^+,\lambda). \tag{49}$$

On the one hand, solving (48), one finds that (48) admits a solution (so does (47)) if and only if

$$\frac{2\alpha^{-1}}{\lambda}\beta = g(\lambda) = 0,$$

and on the other hand, computing (49) from (43) directly, one finds

$$\frac{2\alpha^{-1}}{\lambda}\beta = g'(\lambda) = 0.$$

Thus we have proved that λ is algebraically simple if and only if $g(\lambda) = 0$ and $g'(\lambda) \neq 0$.

Case 2 ($\sinh \lambda d = \cosh \lambda(1-d) = 0$ for some λ). In this case, such a λ must be an eigenvalue and the corresponding eigenfunction is $(\phi(x), \lambda\phi(x))$, where

$$\phi(x) = \begin{cases} -\sinh \lambda(1-d) \sinh \lambda x, & 0 < x < d, \\ \cosh \lambda d \cosh \lambda(1-x), & d < x < 1. \end{cases} \tag{50}$$

Solving directly the following equation

$$(\lambda - \mathcal{A})(u, v) = (\phi, \lambda\phi), \tag{51}$$

we find that the solution to (51) must satisfy

$$u(d^-) = -\frac{d}{2}\cosh \lambda d \sinh \lambda(1-d) \neq u(d^+) = -\frac{1-d}{2}\cosh \lambda d \sinh \lambda(1-d).$$

($|\cosh \lambda d \sinh \lambda(1-d)| = 1$ under the assumption.) However, since $(u, v) \in D(\mathcal{A})$, it should have $u(d^-) = u(d^+)$. This contradiction shows that there is no solution to (51), i.e., λ must be algebraically simple. Finally, a simple calculation shows that $g'(\lambda) = 2\alpha^{-1} \sinh \lambda \neq 0$ since under the assumption, $\cosh \lambda = 0$ and so $|\sinh \lambda| = 1$. The proof is complete. □

Let $g(\lambda)$ be defined by (41). Define

$$f(\lambda) = g(i\lambda). \tag{52}$$

The zeros $\{\lambda_n\}$ of $g(\lambda)$ and $\{\mu_n\}$ of $f(\lambda)$ are related through

$$\lambda_n = i\mu_n. \tag{53}$$

Obviously, $f(\lambda)$ is uniformly bounded on the real axis, and hence f belongs to the *Cartwright class*. Thus, its indicator diagram is an interval.

Furthermore, it is easy to show that $g(\lambda)$ is of *exponential type* with type 1 and

$$\begin{cases} Ce^{|x|} \leq |\, g(x+iy)\, | \leq De^{|x|}, & \alpha \neq 2; \\ Ce^{|(1-2d)x|} \leq |\, g(x+iy)\, | \leq De^{|x|}, & \alpha = 2, d \neq 1/2; \\ |g(x+iy)| = e^x, & \alpha = 2, d = 1/2 \end{cases} \tag{54}$$

for some positive constants C, D and all $x \in \mathbb{R}$ whenever x is sufficiently large. Hence as $\alpha = 2, d = 1/2$, $g(\lambda)$ has no zero. In this case, $\sigma(\mathcal{A}) = \emptyset$ and we could not talk about a basis for system (37). For other cases, though, $g(\lambda)$ must have infinite number of zeros [24, pp. 88–89].

In what follows, we always assume that $\alpha \neq 2$ or $\alpha = 2$, $d \neq 1/2$. In both cases, (54) shows that f is of sine type and

$$h_f(\phi) = \begin{cases} |\sin\phi|, & \alpha \neq 2, \\ \begin{cases} \sin\phi, & \phi \in [0,\pi], \\ -|1-2d|\sin\phi, & \phi \in [-\pi,0], \end{cases} & \alpha = 2, d \neq 1/2, \end{cases} \tag{55}$$

$$G_f = \begin{cases} [-i,i], & \alpha \neq 2, \\ [-|1-2d|i, i], & \alpha = 2, d \neq 1/2. \end{cases} \tag{56}$$

Therefore, the width of the indicator diagram of f equals

$$T = \begin{cases} 2, & \alpha \neq 2, \\ 1+|1-2d|, & \alpha = 2, d \neq 1/2. \end{cases} \tag{57}$$

LEMMA 4. *If $\alpha \neq 2$, then both the root subspaces of $\mathcal{A}$ and $\mathcal{A}^*$ are complete in $\mathcal{H}$: $\mathrm{Sp}(\mathcal{A}) = \mathrm{Sp}(\mathcal{A}^*) = \mathcal{H}$.*

Proof. We apply Theorem 4 for the proof. To do this, we need the adjoint $\mathcal{A}^*$ of $\mathcal{A}$, which can be found using

$$\begin{cases} \mathcal{A}^*(u,v) = (-v(x), -u''(x)), \\ D(\mathcal{A}^*) = \{(u,v) \in H^1(0,1) \times H_E^1(0,1) \mid u(0) = u'(1) = 0, \\ \qquad\qquad u\,|_{[0,d]} \in H^2(0,d), u\,|_{[d,1]} \in H^2(d,1), u'(d^-) - u'(d^+) = \alpha v(d)\}. \end{cases} \tag{58}$$

Let $\mathcal{A}_0$ be the operator $\mathcal{A}$ with $\alpha = 0$. Then $\mathcal{A}_0$ is a skew-adjoint operator in $\mathcal{H}$: $\mathcal{A}_0^* = -\mathcal{A}_0$. $\mathcal{A}_0^*$ generates a unitary-group and hence

$$\|R(\lambda, \mathcal{A}_0^*)\| \leq \frac{1}{|\lambda|} \qquad \forall \lambda \in \mathbb{R}.$$

Now, for any $(u,v) \in \mathcal{H}, \lambda \in \rho(\mathcal{A}), \lambda < 0$, set

$$(\phi,\psi) = R(\lambda, \mathcal{A}_0^*)(u,v), (p,q) = R(\lambda, \mathcal{A}^*)(u,v) - (\phi,\psi).$$

Then $q = -\lambda p$ and p satisfies

$$\begin{cases} \lambda^2 p(x) - p''(x) = 0, & 0 < x < d, d < x < 1, \\ p(0) = p'(1) = 0, \\ p(d^-) = p(d^+), \\ p'(d^-) - p'(d^+) = -\alpha\lambda p(d) - \alpha\lambda\phi(d) + \alpha u(d). \end{cases} \tag{59}$$

Solving (59) gives

$$\begin{cases} p'(x) = \dfrac{-\alpha\lambda\phi(d) + \alpha u(d)}{\alpha/2g(\lambda)} \begin{cases} \cosh\lambda(1-d)\cosh\lambda x, & 0 < x < d, \\ -\sinh\lambda d \sinh\lambda(1-x), & d < x < 1, \end{cases} \\ q(x) = \dfrac{-\alpha\lambda\phi(d) + \alpha u(d)}{\alpha/2g(\lambda)} \begin{cases} \cosh\lambda(1-d)\sinh\lambda x, & 0 < x < d, \\ \sinh\lambda d\cosh\lambda(1-x), & d < x < 1. \end{cases} \end{cases} \tag{60}$$

Notice the following facts:

(a) $|\lambda\phi(d)| \leq |\lambda|\sqrt{d}\|\phi'\|_{H^1} \leq \sqrt{d}|\lambda|\|R(\lambda, \mathcal{A}_0^*)(u,v)\| \leq \sqrt{d}\|(u,v)\|$, $|u(d)| \leq \sqrt{d}\|(u,v)\|$;

(b) $g(\lambda) = e^{-\lambda}[\alpha^{-1} - 1/2 + o(1)]$ as $\lambda \to -\infty$;

(c) $\lim_{|\lambda|\to\infty} \|R(\lambda, \mathcal{A}_0^*)(u,v)\| = 0$.

We can easily check by a direct computation from (60) that $\|(p,q)\| = \|(p',q)\|_{L^2\times L^2}$ is uniformly bounded as $\lambda \to -\infty$. Since

$$\|R(\lambda,\mathcal{A}^*)(u,v)\| \le \|(p,q)\| + \|R(\lambda,\mathcal{A}_0^*)(u,v)\|,$$

we see that $\|R(\lambda,\mathcal{A}^*)(u,v)\|$ is also uniformly bounded as $\lambda \to -\infty$. Then, it is easily found that

$$\phi(x) = -\frac{\sinh\lambda x \int_0^1 \cosh\lambda(1-s)(v-\lambda u)ds}{\lambda\cosh\lambda} + \int_0^x \frac{\sinh\lambda(x-s)}{\lambda}(v-\lambda u)ds;$$

hence we can write

$$R(\lambda,\mathcal{A}^*)(u,v) = (p,q) + (\phi, u-\lambda\phi) = \frac{G(\lambda;u,v)}{F(\lambda)} \qquad \forall (u,v)\in\mathcal{H}, \tag{61}$$

where $G(\lambda;u,v)$ is an $\mathcal{H}$-valued entire function with order less than or equal to 1 and $F(\lambda)$ is a scalar entire function of order 1. Therefore, all conditions of Theorem 4 are satisfied with $\rho = 1, n = 2, \gamma_1 = \{\lambda|\arg\lambda = \pi\}$. The result follows. □

LEMMA 5. *Suppose $\alpha \neq 2$ or $\alpha = 2, d \neq 1/2$.*

(i) *When d is irrational or $\alpha^{-2} < d(1-d)$, all zeros of $g(\lambda)$ are simple and separated.*

(ii) *While $d = q/p$ is rational and $\alpha^{-2} \ge d(1-d)$, if*

$$\eta_0 = \pm\sqrt{\frac{\alpha^{-2}}{d(1-d)} - 1} \tag{62}$$

does not satisfy

$$\left(\frac{\eta_0 \pm (1-2d)\frac{\alpha^{-1}}{\sqrt{d(1-d)}}}{2\alpha^{-1}+1}\right)^{1-2d} = \eta_0 \pm \frac{\alpha^{-1}}{\sqrt{d(1-d)}}, \tag{63}$$

then all zeros of $g(\lambda)$ are simple and separated.

Proof. Suppose that $g(\lambda)$ has zeros $\{\lambda_n\}_1^\infty$. In view of Theorem 3, it suffices to show that

$$\inf_n |\, g'(\lambda_n)\,| > 0. \tag{64}$$

If λ is a zero point of $g(\lambda)$, then there exists an η such that

$$\begin{cases} 2\alpha^{-1}\cosh\lambda + \sinh\lambda = \eta, \\ \sinh\lambda(1-2d) = \eta. \end{cases} \tag{65}$$

Solving (65) yields

$$e^\lambda = \frac{\eta \pm \sqrt{\eta^2 - 4\alpha^{-2}+1}}{2\alpha^{-1}+1}, \qquad e^{\lambda(1-2d)} = \eta \pm \sqrt{\eta^2+1}. \tag{66}$$

Hence the necessary and sufficient condition for λ to be a zero of $g(\lambda)$ is that

$$\left(\frac{\eta \pm \sqrt{\eta^2 - 4\alpha^{-2}+1}}{2\alpha^{-1}+1}\right)^{1-2d} = \eta \pm \sqrt{\eta^2+1}. \tag{67}$$

In this case, λ can be found through the first or the second equality of (66). In what follows, we always relate λ to η via this method.

When η solves (67), one has

$$\begin{aligned} g'(\lambda) = g(\lambda) + g'(\lambda) &= (2\alpha^{-1}+1)e^{\lambda} + 2d\cosh\lambda(1-2d) - e^{\lambda(1-2d)} \\ &= \eta \pm \sqrt{\eta^2 - 4\alpha^{-2}+1} - \left(\eta \pm \sqrt{\eta^2+1}\right) + 2d\cosh\lambda(1-2d) \\ &= \pm\left[\sqrt{\eta^2-4\alpha^{-2}+1} - (1-2d)\sqrt{\eta^2+1}\right]. \end{aligned} \tag{68}$$

It is seen that $g(\lambda) = g'(\lambda) = 0$ if and only if the solution of (67), $\eta = \eta_0$, satisfies

$$\sqrt{\eta_0^2 - 4\alpha^{-2}+1} = (1-2d)\sqrt{\eta_0^2+1},$$

that is, η_0 satisfies (62). When $\eta = \eta_0$, (67) becomes (63). We have two cases.

Case 1 (η_0 does not satisfy (63)). Let

$$2\alpha^{-1}\cosh\lambda_n + \sinh\lambda_n = \eta_n. \tag{69}$$

Suppose that $\mid g'(\lambda_{n_k}) \mid \to 0$ as $k \to \infty$ for some subsequence $\{\lambda_{n_k}\}$ of $\{\lambda_n\}$. Since all $\{\lambda_{n_k}\}$ lie on the strip $-M < \mathrm{Re}\lambda_{n_k} \le 0$ for some $M > 0$, the corresponding $\{\eta_{n_k}\}$ are uniformly bounded: $\mid \eta_{n_k} \mid \le C$ for all n and some $C > 0$. Let η be an accumulation point of $\{\eta_{n_k}\}$. We may assume without loss of generality that

$$\eta_{n_k} \to \eta \text{ as } k \to \infty.$$

So η satisfies (67). On the other hand,

$$\begin{aligned} g'(\lambda_{n_k}) &= \pm\left[\sqrt{\eta_{n_k}^2 - 4\alpha^{-2}+1} - (1-2d)\sqrt{\eta_{n_k}^2+1}\right] \\ &\to \pm\left[\sqrt{\eta^2-4\alpha^{-2}+1} - (1-2d)\sqrt{\eta^2+1}\right] = 0 \text{ as } k \to \infty. \end{aligned}$$

Hence $\eta = \eta_0$. That is, η_0 satisfies (67) or (63), contradicting the assumption. Therefore,

$$\inf_n \mid g'(\lambda_n) \mid > 0.$$

By Theorem 3, $\{\lambda_n\}$ is separated.

Case 2 (η_0 does satisfy (63)). In this case, there is an λ_0 such that $g(\lambda_0) = g(\lambda_0) + g'(\lambda_0) = 0$ and

$$\begin{cases} 2\alpha^{-1}\cosh\lambda_0 + \sinh\lambda_0 = \eta_0, \\ \sinh\lambda_0(1-2d) = \eta_0. \end{cases} \tag{70}$$

Hence

$$\begin{aligned} 2\alpha^{-1}\sinh\lambda_0 + \cosh\lambda_0 &= (2\alpha^{-1}+1)(\sinh\lambda_0 + \cosh\lambda_0) - \eta_0 \\ &= \pm\sqrt{\eta_0^2 - 4\alpha^{-2}+1} = \pm(1-2d)\alpha^{-1}\sqrt{\frac{1}{d(1-d)}} \end{aligned} \tag{71}$$

and

$$\cosh\lambda_0(1-2d) = \pm\alpha^{-1}\sqrt{\frac{1}{d(1-d)}}. \tag{72}$$

Hence $\cosh \lambda_0(1-2d)$ is real and so is $\sinh \lambda_0(1-2d)$. (Its imaginary part equals zero; in particular, η_0 is real.) Let $\lambda_0 = x_0 + iy_0, x_0, y_0 \in \mathbb{R}$. Then since the right hands of (71) and (72) and

$$2\alpha^{-1} \cosh \lambda_0 + \sinh \lambda_0 = \sinh \lambda_0(1-2d) \tag{73}$$

are real numbers, comparing the imaginary parts of (71)–(73) gives

$$\begin{cases} \sinh x_0(1-2d) \sin y_0(1-2d) = 0, \\ (2\alpha^{-1} \cosh x_0 + \sinh x_0) \sin y_0 = 0, \\ (2\alpha^{-1} \sinh x_0 + \cosh x_0) \sin y_0 = 0. \end{cases} \tag{74}$$

Hence $\sin y_0(1-2d) = \sin y_0 = 0$. That is, d is rational. Furthermore, since η_0 solving (67) implies that η_0 is real, it must have

$$\alpha^{-1} \geq d(1-d).$$

The proof is complete. □

By virtue of Lemmas 3–5 and Theorem 2, we obtain the Riesz basis property for the system (37).

THEOREM 5. *Suppose that $\alpha \neq 2$. Assume that one of the following conditions is satisfied:*

(a) *d is irrational or d is rational but $\alpha^{-2} < d(1-d)$.*

(b) *d is rational and $\alpha^{-2} \geq d(1-d)$ but η_0 defined by (62) does not satisfy (63).*

Then each eigenvalue of $\mathcal{A}$ is algebraically simple and there is a set of eigenfunctions of $\mathcal{A}$ which forms a Riesz basis for the energy space $\mathcal{H}$.

Remark 2. Numerical simulation by MATLAB shows that some solution of (62) does satisfy (63), and in this case, $g(\lambda)$ may have multiple zeros with multiplicity at most 2. For the multiple zero case, we shall discuss it in a separate paper.

From Theorem 5, we know that when the system (37) is a Riesz spectral system, the spectrum-determined growth condition holds, which was obtained specifically by Liu in 1986 [11]. The general case can be found in [14] and [12]. We do not touch the stability here because once we have the spectrum-determined growth condition, simple spectral analysis can be made to get the stability of system (37). We refer readers to [7] for this.

REFERENCES

[1] S.A. AVDONIN AND S.A. IVANOV, *Families of Exponentials: The Method of Moments in Controllability Problems for Distributed Parameter Systems*, Cambridge University Press, Cambridge, UK, 1995.

[2] R.F. CURTAIN AND H. ZWART, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Texts Appl. Math. 21, Springer-Verlag, New York, 1995.

[3] B.-Z. GUO, *Riesz basis approach to the stabilization of a flexible beam with a tip mass*, SIAM J. Control Optim., 39 (2001), pp. 1736–1747.

[4] B.Z. GUO AND R. YU, *The Riesz basis property of discrete operators and application to an Euler-Bernoulli beam equation with boundary linear feedback control*, IMA J. Math. Control Inform., 18 (2001), pp. 241–251.

[5] B.Z. GUO AND K.Y. CHAN, *Riesz basis generation, eigenvalues distribution, and exponential stability for an Euler-Bernoulli beam with joint feedback control*, Rev. Mat. Complut., 14 (2001), pp. 205–229.

[6] B.Z. GUO AND Y.H. LUO, *Riesz basis property of a second order hyperbolic system with collocated scalar input/output*, IEEE Trans. Automat. Control, 47 (2002), pp. 693–698.

[7] B.Z. GUO AND W.D. ZHU, *On the energy decay of two coupled strings through a joint damper*, J. Sound Vibration, 203 (1997), pp. 447–455.

[8] N. Dunford and J.T. Schwartz, *Linear Operators*, Part III, Wiley-Interscience, New York, 1971.
[9] B. Jacob and H. Zwart, *Exact controllability of C_0-groups with one-dimensional input operators*, in Advances in Mathematical Systems Theory, F. Colonius et al., eds., Birkhäuser, Boston, 2000, pp. 221–242.
[10] B.J. Levin, *On bases of exponential functions in L^2*, Zap. Har'kov. Gos. Univ. i Har'kov. Mat. Obsc., 27 (1961), pp. 39–48.
[11] K.S. Liu, *Energy decay problems in the design of a point stabilizer for coupled string vibrating systems*, SIAM J. Control Optim., 26 (1988), pp. 1348–1356.
[12] Z.H. Luo, B.Z. Guo, and O. Morgul, *Stability and Stabilization of Infinite Dimensional Systems with Applications*, Springer-Verlag, London, 1999.
[13] J. Locker, *Spectral Theory of Non-Self-Adjoint Two-Point Differential Operators*, Math. Surveys Monogr. 73, American Mathematical Society, Providence, RI, 2000.
[14] A.F. Neves, H.D.S. Ribeiro, and O. Lopes, *On the spectrum of evolution operators generated by hyperbolic systems*, J. Funct. Anal., 67 (1986), pp. 320–344.
[15] N.K. Nikol'skii, *A Treatise on the Shift Operator*, Spinger-Verlag, Berlin, 1986.
[16] R.E.A.C. Paley and N. Wiener, *Fourier Transform in the Complex Domain*, Amer. Math. Soc. Colloq. Publ. 19, New York, 1934.
[17] B.S. Pavlov, *Basicity of an exponential systems and Muckenhoupt's condition*, Soviet Math. Dokl., 20 (1979), pp. 655–659.
[18] P. Rideau, *Contrôle d'un Assemblage de Poutres Flexibles par des Capteurs-Actionneurs Ponctuels: Étude du Spectre du Système*, Thèse, Ecole Nationale Supérieure des Mines de Paris, Sophia-Antipolis, France, 1985.
[19] D.L. Russell, *Nonharmonic Fourier series in control theory of distributed systems*, J. Math. Anal. Appl., 18 (1967), pp. 542–559.
[20] M.A. Shubov, *Basis property of eigenfunctions of nonselfadjoint operator pencils generated by the equation of nonhomogeneous damped string*, Integral Equations Operator Theory, 25 (1996), pp. 289–328.
[21] M.A. Shubov, *Spectral operators generated by damped hyperbolic equations*, Integral Equations Operator Theory, 28 (1997), pp. 358–372.
[22] E.C. Titchmarsh, *The Theory of Functions*, Oxford University Press, London, 1952.
[23] S.M. Verduyn Lunel, *The closure of the generalized eigenspace of a class of infinitesimal generators*, Proc. Roy. Soc. Edinburgh Sect. A, 117 (1991), pp. 171–192.
[24] R.M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, London, 1980.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 985–1001

ON THE POISSON EQUATION FOR PIECEWISE-DETERMINISTIC MARKOV PROCESSES*

OSWALDO L. V. COSTA† AND FRANÇOIS DUFOUR‡

Abstract. In this paper we study the problem of the existence of a solution for the Poisson equation (PE) associated to a piecewise-deterministic Markov process (PDP). It is well known that the long run average cost of a stochastic process can be obtained through a solution of the PE associated with the process. Our first result will show that the existence of a solution for the PE of a PDP is equivalent to the existence of a solution for the PE of an embedded discrete- time Markov chain associated with the PDP. It is important to point out that, due to the possibility of jumps from the boundary, the differential formula for the PDPs has a special form, so that general results for the PE of continuous- time stochastic processes cannot be directly applied. Usually discrete-time conditions for the existence of a solution of a PE of a Markov chain are easier to apply than the continuous-time counterpart. We follow this approach to derive our second result, which establishes sufficient conditions for the existence of a PE to the embedded Markov chain, and consequently for the PE of the PDP. The condition is illustrated with an application to the capacity expansion model.

Key words. piecewise-deterministic Markov process, Poisson equation, Markov chain

AMS subject classifications. 60J25, 60J10

DOI. 10.1137/S0363012901393523

1. Introduction.

Piecewise-deterministic Markov processes (PDPs) have been introduced in the literature by Davis [1] as a general class of stochastic models. PDPs are a family of Markov processes involving deterministic motion punctuated by random jumps. The motion of the PDP X_t depends on three local characteristics, namely the flow, the jump rate, and the transition measure Q, which specifies the postjump location. Starting from x the motion of the process follows the flow $\phi(t,x)$ until the first jump time T_1, which occurs either spontaneously in a Poisson-like fashion with rate $\lambda(t,x)$ or when the flow $\phi(t,x)$ hits the boundary of the state-space. In either case the location of the process at the jump time T_1 is selected by the transition measure $Q((T_1,x),.)$ and the motion restarts from this new point as before. A suitable choice of the state space and the local characteristics ϕ, λ, and Q provides stochastic models covering a great number of problems of operations research [1].

In this paper we are interested in studying the problem of the existence of a solution for a Poisson equation (PE) associated to a PDP. It is well known that the long run average cost associated with a process can be obtained by solving a PE. Several general results on PEs can be found in the literature, and we can mention [4, 5, 6] for a general overview on the subject. However, as we shall see below, due to

*Received by the editors August 6, 2001; accepted for publication (in revised form) November 24, 2002; published electronically June 25, 2003. This work has been supported by a PICS Franco-Brazilian grant (ref. 921).

http://www.siam.org/journals/sicon/42-3/39352.html

†Departamento de Engenharia de Telecomunicações e Controle, Escola Politécnica da Universidade de São Paulo, CEP: 05508 900-São Paulo, Brazil (oswaldo@lac.usp.br). This author received financial support from FAPESP (Research Council of the State of São Paulo), grant 97/04668-1, CNPq (Brazilian National Research Council), grant 305173/88, PRONEX, grant 015/98, and IMAGIMB.

‡Corresponding author. Mathématiques Appliquées de Bordeaux, Université Bordeaux I, 351 cours de la Libération, 33405 Talence Cedex, France (dufour@math.u-bordeaux.fr). The other affiliation of this author is GRAPE, Université Bordeaux IV, France.

the special features of PDPs (mixture of deterministic and random jumps), general results cannot be directly applied to PDPs (see Remark 2.3).

As shown in [2], associated with the PDP we can define an embedded Markov chain Y_n, such that the existence of an invariant probability measure for the PDP X_t is equivalent to the existence of an invariant σ-finite measure for the embedded Markov chain Y_n. In this paper we somewhat extend this result and show that the existence of a solution of the PE for the PDP X_t is equivalent to the existence of a solution of the PE for the embedded Markov chain Y_n. The advantage of doing this is that conditions for the existence of a PE for discrete-time Markov chains are in general easier to check than those for continuous-time Markov processes. Following this approach we present conditions for the existence of a solution for a PE of the embedded Markov chain Y_n, which in turns implies the existence of a solution for a PE of the PDP X_t.

The paper is organized as follows. In section 2 we give the main definitions and assumptions. In section 3 we obtain the equivalence results linking the existence of a solution of a PE for the PDP and the embedded Markov chain. We also show that, as expected, when an invariant probability measure for the PDP exists (and consequently there exists a σ-finite measure for the embedded Markov chain), the stationary value costs obtained from these invariant measures coincide. In section 4 we provide a sufficient condition for the existence of a solution for the PE associated with the embedded Markov chain. This condition is written, as in [4], as a Lyapunov function criterion and also provides explicit bounds on the solution to the PE. We conclude the paper in section 5 by presenting an application to a capacity expansion model.

2. Notation and preliminary definitions.

2.1. Preliminaries. In this subsection we present some standard notation and some basic definitions related to the motion of a PDP (for further details, see [1]).

Let X be a metric space, $\mathcal{B}(X)$ the Borel σ-field of X, $\mathcal{P}(X)$ the set of probability measures on $(X, \mathcal{B}(X))$, and $\mathbb{B}(X)$ the set of all Borel measurable functions from X into R. For any $f \in \mathbb{B}(X)$ and $\mu \in \mathcal{P}(X)$, $\mu(f)$ denotes $\int_X f(x)\mu(dx)$.

Let X and Y be two metric spaces. Then a kernel $V(.,.)$ defined on $X \times \mathcal{Y}$ is a map from $X \times \mathcal{B}(Y)$ into $[0,1]$ such that for each $A \in \mathcal{B}(Y)$, $V(.,A)$ is a nonnegative bounded measurable function on X and for each $x \in X$, $V(x,.)$ is a bounded measure on $\mathcal{B}(Y)$. Consequently, for any measurable function $f \in \mathbb{B}(Y)$, $Vf(.)$ defined as

$$Vf(x) \doteq \int_Y f(y)V(x,dy) \tag{1}$$

belongs to $\mathbb{B}(X)$.

Let E^0 be an open nonempty subset of $\mathbb{R}^d$ and ∂E^0 its boundary. We consider $\phi(t,x)$ as being the flow of a Lipschitz continuous vector field $\mathcal{X}$. Define

$$\partial^{\pm} E^0 = \Big\{ z \in \partial E^0; z = \phi(\pm t, x) \text{ for some } x \in E^0, \text{ for some } t \geq 0 \Big\}$$

and $\partial_1 E^0 = \partial^- E^0 - \partial^+ E^0$ (where for sets A, B, $A - B = A \cap B^c$). Set $E = E^0 \cup \partial_1 E^0$, $\partial^* E = \partial^+ E^0$, and for each $x \in E$, write

$$t^*(x) = \inf\{t > 0; \phi(t,x) \in \partial^* E\},$$

where $\inf\{\emptyset\} := \infty$, and define $t^*(z) = 0$ for $z \in \partial^* E$. Note that the points in E are such that starting from $x \in E$, we have that $\phi(t,x) \in E$ for all t sufficiently small, that is, the flow stays inside E for all t sufficiently small. On the other hand the points in $\partial^* E$ are such that starting from $z \in \partial^* E$, we have that $\phi(t,z) \notin E$ for all t sufficiently small, that is, the flow leaves the set $E^0 \cup \partial E^0$.

We consider the following parameters for our problem:

(a) $\lambda(.) : E \mapsto R_+$ is a Borel measurable function;

(b) $Q(.,.)$ is a kernel defined on $E \cup \partial^* E \times \mathcal{B}(E)$ such that (for all $x \in E \cup \partial^* E$), $Q(x,.) \in \mathcal{P}(E)$.

Let D denote the space of right-continuous functions $\omega(.)$ on R_+ taking values in E such that the left limit exists for all $t > 0$. Denote by x_t the coordinate function $x_t(\omega) = \omega(t)$ for all $\omega \in D$. Let $\mathcal{F}_t^0 = \sigma\{x_s; 0 \le s \le t\}$ and $\mathcal{F}^0 = \vee_{t \ge 0} \mathcal{F}_t^0$. For $\omega \in \Omega$, set for $k = 0, 1, 2, \ldots$

$$
\begin{aligned}
&T_0(\omega) = 0,\\
&T_k(\omega) = \begin{cases} \inf\Big\{t > T_{k-1}(\omega); x_t(\omega) \ne x_{t-}(\omega)\Big\} & \text{if } T_{k-1}(\omega) < \infty,\\ \infty & \text{otherwise,} \end{cases}\\
&T_\infty(\omega) = \lim_{k \to \infty} T_k(\omega),\\
&Z_k(\omega) = \begin{cases} x_{T_k(\omega)}(\omega) & \text{if } T_k(\omega) < \infty,\\ \Delta & \text{if } T_k(\omega) = \infty, \end{cases}
\end{aligned}
$$

where Δ represents a cemetery state.

Let us consider $\Omega \subset D$ such that $\omega \in \Omega$ if

$$x_t(\omega) = \phi(t, Z_0(\omega)) I_{\{0 \le t < T_1(\omega)\}} + \sum_{k=2}^{\infty} I_{\{T_{k-1}(\omega) \le t < T_k(\omega)\}} \phi(t - T_{k-1}(\omega), Z_{k-1}(\omega)).$$

We define the motion of the process $\{X_t\}$ starting from a point $x \in E$ in the following way. Take a random variable T_1 such that

$$P_x(T_1 > t) = \begin{cases} \exp -\Lambda(t,x) & \text{for } t < t^*(x),\\ 0 & \text{for } t \ge t^*(x), \end{cases}$$

where

$$\Lambda(t,x) = \int_0^t \lambda(\phi(s,x)) ds. \tag{2}$$

If T_1 generated according to the above probability is equal to infinity, then for $t \in \mathbb{R}_+$, $X_t = \phi(t,x)$. Otherwise select independently an E-valued random variable having distribution $Q(\phi(T_1,x),.)$. The trajectory of $\{X_t\}$ starting at x, for $t \le T_1$, is given by

$$X_t = \begin{cases} \phi(t,x) & \text{for } t < T_1,\\ Z_1 & \text{for } t = T_1. \end{cases}$$

Starting from $X_{T_1} = Z_1$, we now select the next interjump time $T_2 - T_1$ and postjump location $X_{T_2} = Z_2$ in a similar way. This gives a piecewise-deterministic trajectory

for the process $\{X_t\}$ with jump times $T_1, T_2, \ldots$, and postjump location $Z_1, Z_2, \ldots$. It can be shown [1, pp. 62–66] that $\{X_t\}$ is a strong Markov process. The procedure above defines a family of measures $\{P_x; x \in E\}$ on $(\Omega, \mathcal{F}^0)$. The final assumption is that for every $x \in E$ and $t \in \mathbb{R}_+$

$$E_x[N_t] < \infty, \tag{3}$$

where

$$N_t := \sum_{k=1}^{\infty} I_{\{T_k \le t\}}. \tag{4}$$

In particular, (3) implies that $T_k \to \infty$ as $k \to \infty$ almost surely.

For any $\mu \in \mathcal{P}(E)$, define P_μ on $(\Omega, \mathcal{F}^0)$ as

$$P_\mu(A) = \int_A P_x(A)\mu(dx).$$

Let $\mathcal{F}_t^\mu$ be the completion of $\mathcal{F}_t^0$ with respect to all P_μ-null sets of $\mathcal{F}^0$, and

$$\mathcal{F}_t = \bigcap_{\mu \in \mathcal{P}(E)} \mathcal{F}_t^\mu.$$

By the same arguments as Theorem 25.3 in [1, p. 63], it follows that $\mathcal{F}_t$ is right-continuous.

2.2. The PDP differential formula. In this subsection we present the PDP differential formula, which will be used in what follows. In order to do this we need to make some initial definitions.

Define the operator $\mathfrak{C} : \mathbb{B}(\partial^* E) \to R$ as follows:

$$(\forall z \in \partial^* E)\ (\forall g \in \mathbb{B}(\partial^* E)) \quad \mathfrak{C}g(z) := Qg(z) - g(z). \tag{5}$$

Define also

$$\mathfrak{B}f(t, x, \omega) := f(x) - f(X_{t-}(\omega)).$$

It was shown in [1, p. 83] that $\mathfrak{B}h$ is in $\mathcal{L}_1^{loc}(p)$ if h satisfies the identity

$$E_x\left[\sum_i \left|h(X_{T_i \wedge \tau_n}) - h(X_{(T_i \wedge \tau_n)-})\right|\right] < \infty$$

for some sequence of stopping times $\tau_n \uparrow \infty$. In particular, this always holds if (see [1, p. 83]) for each $t \ge 0$,

$$E_x\left[\sum_{T_i \le t} \left|h(X_{T_i}) - h(X_{T_i-})\right|\right] < \infty. \tag{6}$$

Clearly from (3), if h is bounded, then (6) will hold.

Define

$$p^*(t) = \sum_{i=1}^{\infty} I_{\{t \ge T_i\}} I_{\{X_{T_i-} \in \partial^* E\}},$$

which is a measure that counts the number of jumps from the boundary, and for $A \in \mathcal{B}(E)$

$$p(t,A) := \sum_{k=1}^{\infty} I_{\{T_k \le t\}} I_{\{X_{T_k} \in A\}},$$

$$\widetilde{p}(t,A) := \int_0^t Q(X_s,A)(\lambda(X_s))ds + \int_0^t Q(X_{s-},A)dp^*(s),$$

and

$$q(t,A) := p(t,A) - \widetilde{p}(t,A).$$

Associated with the PDP we can define a multivalued operator (see chapter 1 in [3]) $\widetilde{\mathcal{A}}$. This multivalued operator $\widetilde{\mathcal{A}}$ is a subset of $\mathbb{B}(E) \times \mathbb{B}(E)$ with domain $\mathcal{D}(\widetilde{\mathcal{A}}) \subset \mathbb{B}(E)$ defined as follows: $f \in \mathcal{D}(\widetilde{\mathcal{A}})$ if the following conditions are satisfied:

(a) $(\exists \mathcal{X}f \in \mathbb{B}(E))$ such that

$$(\forall x \in E)\,(\forall t \in [0, t^*(x)))) \quad f(\phi(t,x)) = f(x) + \int_0^t \mathcal{X}f(\phi(s,x))ds.$$

(b) $\lim_{t \uparrow t^*(x)} f(\phi(t,x))$ exists whenever $t^*(x) < \infty$.

The range of $\widetilde{\mathcal{A}}$ is given by

$$\mathcal{R}(\widetilde{\mathcal{A}}) = \Big\{ g \in \mathbb{B}(E) : (\exists f \in \mathcal{D}(\widetilde{\mathcal{A}})) \quad \text{such that} \quad g = \mathcal{X}f + (Qf - f)\lambda \Big\}.$$

For $f \in \mathcal{D}(\widetilde{\mathcal{A}})$, $\mathcal{A}f$ will denote a function in $\mathcal{R}(\widetilde{\mathcal{A}})$ such that $(f, \mathcal{A}f) \in \widetilde{\mathcal{A}}$. Moreover, for $f \in \mathcal{D}(\widetilde{\mathcal{A}})$, we define

$$f(z) = \lim_{t \uparrow t^*(x)} f(\phi(t,x)) \qquad \forall z = \lim_{t \uparrow t^*(x)} \phi(t,x) \in \partial^* E.$$

Notice that the limit exists from (b) of the definition of $\mathcal{D}(\widetilde{\mathcal{A}})$.

Remark 2.1. (a) Note that $\widetilde{A}$ is a linear operator in the sense that $\alpha f + \beta g \in \mathcal{D}(\widetilde{\mathcal{A}})$ whenever $f \in \mathcal{D}(\widetilde{\mathcal{A}})$, $g \in \mathcal{D}(\widetilde{\mathcal{A}})$, and $(\alpha, \beta) \in \mathbb{R}^2$.

(b) We had to introduce here the notion of multivalued operator, which is different from the extended generator defined by Davis (see Theorem 26.14 in [1]), to ensure that the function $\mathcal{X}f(x)$ belongs to $\mathbb{B}(E)$.

We can now state the PDP differential formula.

THEOREM 2.2 (PDP differential formula; see [1, Thm. 31.3, p. 83]). *Suppose that $f \in \mathcal{D}(\widetilde{\mathcal{A}})$ and satisfies the condition in* (6). *Then for each $t \ge 0$*

$$f(X_t) - f(X_0) = \int_0^t \mathcal{A}f(X_s)ds + \int_0^t \int_E \mathfrak{B}f(s,x,.)q(ds,dx) + \int_0^t \mathfrak{C}f(X_{s-})dp^*(s). \tag{7}$$

Moreover, the stochastic integral term (the one with respect to $q(ds,dx)$) is a martingale.

Remark 2.3. The previous equation must be compared to the differential formula (see (11)) presented in [4]. In formula (7) there is an extra term, associated with the jumps from the boundary, given by $\int_0^t \mathfrak{C}f(X_{s-})dp^*(s)$. The expectation of this term cannot be expressed as an integral term with respect to the Lebesgue measure. Consequently, due to this special feature, more specific results have to be developed.

2.3. The embedded Markov chain and the PE. Associated with the PDP defined in section 2.1, the following stochastic kernels can be introduced (recall the definition of Λ in (2)): for all $x \in E \cup \partial^* E$ and $A \in \mathcal{B}(E)$

$$L(x,A) \doteq \int_0^{t^*(x)} e^{-s-\Lambda(s,x)} I_A(\phi(s,x))ds, \tag{8}$$

$$K(x,A) \doteq \int_0^{t^*(x)} e^{-s-\Lambda(s,x)} \lambda(\phi(s,x))Q(\phi(s,x),A)ds$$

$$+ e^{-t^*(x)-\Lambda(t^*(x),x)} Q(\phi(t^*(x),x),A), \tag{9}$$

$$G(x,A) \doteq L(x,A) + K(x,A). \tag{10}$$

It will be useful in what follows to define the function $\mathcal{L}(x)$ as follows:

$$\mathcal{L}(x) := L(x,E). \tag{11}$$

Note that for every $x \in E$, $0 < L(x,E) < 1$, and for every $x \in E \cup \partial^* E$, $G(x,E) = 1$. Thus $G(.,.)$ is a stochastic kernel of an embedded Markov chain associated with the PDP $\{X_t\}$, which we shall denote by $\{Y_n\}$.

In order to introduce the PE for the embedded Markov chain, we need to define the following kernel acting on the boundary: for all $x \in E \cup \partial^* E$ and $A \in \mathcal{B}(\partial^* E)$,

$$H(x,A) \doteq e^{-t^*(x)-\Lambda(t^*(x),x)} I_A(\phi(t^*(x),x)). \tag{12}$$

Moreover, it is easy to see from the definitions of the kernels L, H, and G (see (8), (12), (10)) that for $z \in \partial^* E$ and any functions $f \in \mathbb{B}(E)$ and $r \in \mathbb{B}(\partial^* E)$, we have

$$Lf(z) = 0, \qquad Hr(z) = r(z), \quad \text{and} \quad Gf(z) = Qf(z). \tag{13}$$

Next we define the PE for PDPs which, as we will see in Theorem 2.5, is associated with the long run average cost.

DEFINITION 2.4 (the PE for the PDP). *Let $f \in \mathbb{B}(E)$ and $r \in \mathbb{B}(\partial^* E)$. A pair of functions $(w,h) \in \mathcal{D}(\widetilde{\mathcal{A}}) \times \mathcal{D}(\widetilde{\mathcal{A}})$ is a solution for the PE for the PDP with charge (f,r) if*

(i) $(\forall x \in E) \quad \mathcal{A}w(x) = 0 \qquad$ *and* $\qquad (\forall z \in \partial^* E) \quad \mathfrak{C}w(z) = 0,$

(ii) $(\forall x \in E) \quad \mathcal{A}h(x) = w(x) - f(x) \quad$ *and* $\quad (\forall z \in \partial^* E) \quad \mathfrak{C}h(z) = -r(z).$

The next result shows the connection between the PE for the PDP and the long run average cost associated with cost functions f and r.

THEOREM 2.5. *For $f \in \mathbb{B}(E)$ and $r \in \mathbb{B}(\partial^* E)$, suppose that (w,h) is a solution for the PE for the PDP with charges (f,r) such that*

$$\liminf_{t\to\infty} \frac{1}{t} E_x[h(X_t)] = 0, \tag{14}$$

and assume that w and h satisfy the condition in (6). *Then for all $x \in E$*

$$w(x) = \liminf_{t\to\infty} \frac{1}{t} E_x \left[\int_0^t f(X_s)ds + \int_0^t r(X_{s-})dp^*(s) \right].$$

Proof. By the PDP differential formula (see Theorem 2.2) and from (i) in Definition 2.4,

$$\begin{aligned} w(X_t) - w(x) &= \int_0^t \mathcal{A}w(X_s)ds + \int_0^t \int_E \mathfrak{B}w(s,x,.)q(ds,dx) + \int_0^t \mathfrak{C}w(X_{s-})dp^*(s) \\ &= \int_0^t \int_E \mathfrak{B}w(s,x,.)q(ds,dx) \end{aligned}$$

and the stochastic integral term is a martingale. Thus for all $t \in R_+$,

$$E_x[w(X_t)] - w(x) = 0. \tag{15}$$

Again by the PDP differential formula (see Theorem 2.2) and from (ii) in Definition 2.4,

$$\begin{aligned} h(X_t) - h(x) &= \int_0^t \mathcal{A}h(X_s)ds + \int_0^t \int_E \mathfrak{B}h(s,x,.)q(ds,dx) + \int_0^t \mathfrak{C}h(X_{s-})dp^*(s) \\ &= \int_0^t (w(X_s) - f(X_s))ds + \int_0^t \int_E \mathfrak{B}h(s,x,.)q(ds,dx) \\ &\quad - \int_0^t r(X_{s-})dp^*(s) \end{aligned}$$

and the stochastic integral term is a martingale. From (15) we have for all $t \in R_+$

$$\begin{aligned} E_x[h(X_t)] - h(x) &= \int_0^t E_x[w(X_s)]ds - \int_0^t E_x[f(X_s)]ds - E_x\left[\int_0^t r(X_{s-})dp^*(s)\right] \\ &= w(x)t - \int_0^t E_x[f(X_s)]ds - E_x\left[\int_0^t r(X_{s-})dp^*(s)\right]. \end{aligned}$$

Therefore,

$$w(x) = \frac{1}{t}\left(E_x\left[\int_0^t f(X_s)ds + \int_0^t r(X_{s-})dp^*(s)\right]\right) + \frac{1}{t}\left(E_x\left[h(X_t)\right] - h(x)\right),$$

and taking the limit as $t \to \infty$ we have from (14) the desired result. □

We will show in the next section that the solution for the PE for the (continuous-time) PDP is closely related to the solution for the PE for the (discrete-time) embedded Markov chain with kernel G, defined next.

DEFINITION 2.6 (the PE for G). *Let $f \in \mathbb{B}(E)$ and $r \in \mathbb{B}(\partial^* E)$. A pair of functions $(w,h) \in \mathbb{B}(E \cup \partial^* E) \times \mathbb{B}(E \cup \partial^* E)$ is a solution for the PE for G with charge $Lf + Hr$ if for all $x \in E \cup \partial^* E$*

$$\begin{aligned} &\text{(i)} \quad Gw(x) = w(x), \\ &\text{(ii)} \quad Gh(x) = h(x) + Lw(x) - Lf(x) - Hr(x). \end{aligned}$$

3. Equivalence results. In this section we present some connections between the PE for the PDP and the Markov chain $\{Y_n\}$ associated with G.

The first result presents some relations between the kernel G and the operator $\mathcal{A}$.

According to the notation introduced in (1), if $V(.,.)$ is any stochastic kernel defined on $E \cup \partial^* E \times \mathcal{B}(E)$, then for any function $h \in \mathcal{D}(\widetilde{\mathcal{A}})$, $V\mathcal{A}h(x)$ denotes the

function in $\mathbb{B}(E \cup \partial^* E)$ defined by $\int_E \mathcal{A}h(y)\, V(x, dy)$. If V is a kernel such that $Vf \in \mathcal{D}(\widetilde{\mathcal{A}})$, then $\mathcal{A}Vf$ will denote a function in $\mathcal{R}(\widetilde{\mathcal{A}})$ such that $(Vf, \mathcal{A}Vf) \in \widetilde{\mathcal{A}}$, according to the definition of $\mathcal{A}$ presented in section 2.2.

THEOREM 3.1. *Assume that $h \in \mathcal{D}(\widetilde{\mathcal{A}})$. Then*

$$(\forall x \in E \cup \partial^* E) \qquad L\mathcal{A}h(x) = Gh(x) - h(x) - H(Q - I)h(x). \tag{16}$$

Assume that $h \in \mathbb{B}(E)$ and $r \in \mathbb{B}(\partial^* E)$. Then $Gh \in \mathcal{D}(\widetilde{\mathcal{A}})$, $Lh \in \mathcal{D}(\widetilde{\mathcal{A}})$, $Hr \in \mathcal{D}(\widetilde{\mathcal{A}})$ and

$$(\forall x \in E) \qquad \mathcal{A}Gh(x) = (I + \lambda(x)Q)(G - I)h(x), \tag{17}$$

$$\mathcal{A}Lh(x) = Lh(x) - h(x) + \lambda(x)QLh(x), \tag{18}$$

$$\mathcal{A}Hr(x) = (1 + \lambda(x))Hr(x) + \lambda(x)(Q - I)Hr(x). \tag{19}$$

Proof. First note that using the definition of L and $\mathcal{A}$, we have that

$$\begin{aligned} L\mathcal{A}h(x) = &\int_0^{t^*(x)} e^{-s-\Lambda(s,x)} \mathcal{X}h(\phi(s,x))ds \\ &+ \int_0^{t^*(x)} e^{-s-\Lambda(s,x)} \lambda(\phi(s,x))[Q - I]h(\phi(s,x))ds. \end{aligned} \tag{20}$$

Now subtracting $Lh(x)$ from both sides of (20), we obtain

$$\begin{aligned} L\mathcal{A}h(x) - Lh(x) = &\int_0^{t^*(x)} e^{-s-\Lambda(s,x)} [\mathcal{X}h(\phi(s,x)) - (1 + \lambda(\phi(s,x)))h(\phi(s,x))]ds \\ &+ \int_0^{t^*(x)} e^{-s-\Lambda(s,x)} \lambda(\phi(s,x))Qh(\phi(s,x))ds. \end{aligned}$$

Using the definition of $\mathcal{X}h$, it follows that

$$\begin{aligned} L\mathcal{A}h(x) - Lh(x) &= \left[e^{-s-\Lambda(s,x)} h(\phi(s,x))\right]_0^{t^*(x)} \\ &\quad + \int_0^{t^*(x)} e^{-s-\Lambda(s,x)} \lambda(\phi(s,x))Qh(\phi(s,x))ds \\ &= Hh(x) - h(x) + Kh(x) - HQh(x), \end{aligned} \tag{21}$$

and from the definition of G (see (10)), we obtain that (21) gives (16).

For (17), note that a short calculation using the fact that $t^*(\phi(t,x)) = t^*(x) - t$ for $t < t^*(x)$ shows that

$$\begin{aligned} Gh(\phi(t,x)) = e^{t+\Lambda(t,x)} \Bigg[&Gh(x) \\ &- e^{t+\Lambda(t,x)} \int_0^t e^{-s-\Lambda(s,x)} [h(\phi(s,x)) + \lambda(\phi(s,x))Qh(\phi(s,x))]ds \Bigg]. \end{aligned} \tag{22}$$

From (22), it is easy to deduce that $Gh \in \mathcal{D}(\widetilde{\mathcal{A}})$ and also that

$$\mathcal{X}Gh(x) = [1 + \lambda(x)]Gh(x) - h(x) - \lambda(x)Qh(x). \tag{23}$$

Consequently, from (23),

$$\mathcal{A}Gh(x) = [1 + \lambda(x)]Gh(x) - h(x) - \lambda(x)Qh(x) + \lambda(x)[Q - I]Gh(x) \tag{24}$$

and (17) follows from (24).

Now, using similar arguments as above, it follows that $Lh \in \mathcal{D}(\widetilde{\mathcal{A}})$, $Hr \in \mathcal{D}(\widetilde{\mathcal{A}})$. The last two identities, (18) and (19), are obtained similarly, yielding the desired results. □

The next result shows the connection between the PE for the PDP and the PE for G.

THEOREM 3.2. *Let $f \in \mathbb{B}(E)$ and $r \in \mathbb{B}(\partial^* E)$. Then (w, h) is a solution for the PE for the PDP with charges (f, r) if and only if (w, h) is a solution for the PE for G with charge $Lf + Hr$.*

Proof. Suppose that $(w, h) \in \mathbb{B}(E \cup \partial^* E) \times \mathbb{B}(E \cup \partial^* E)$ is a solution for the PE for the kernel G with charges $Lf + Hr$. From Definition 2.6, for all $x \in E \cup \partial^* E$,

$$Gw(x) = w(x), \tag{25}$$

$$Gh(x) = h(x) + Lw(x) - Lf(x) - Hr(x). \tag{26}$$

From Theorem 3.1, (17), and (25), it follows that $w = Gw \in \mathcal{D}(\widetilde{\mathcal{A}})$ and

$$\begin{aligned} \mathcal{A}w &= \mathcal{A}Gw \\ &= (I + \lambda Q)(G - I)w \\ &= (I + \lambda Q)(Gw - w) \\ &= 0. \end{aligned}$$

From (13) and (25), we have for all $z \in \partial^* E$, (see (5)) $\mathfrak{C}w(z) = Qw(z) - w(z) = Gw(z) - w(z) = 0$. From (26), we have for all $z \in \partial^* E$, $Gh(z) = Lw(z) - Lf(z) + h(z) - Hr(z)$. Using (13), we have

$$Gh(z) = Qh(z) = h(z) - r(z), \tag{27}$$

implying that $\mathfrak{C}h(z) = -r(z)$ for all $z \in \partial^* E$. From (26), it follows that $h \in \mathcal{D}(\widetilde{\mathcal{A}})$, by combining Theorem 3.1 and Remark 2.1. Moreover, applying the three identities (17), (18), (19) of Theorem 3.1 to the functions $\mathcal{A}Gh$, $\mathcal{A}L(w-f)$, and $\mathcal{A}Hr$, it follows from (26) that

$$\begin{aligned} \mathcal{A}h(x) &= \mathcal{A}\{Gh(x) - L(w - f)(x) + Hr(x)\} \\ &= (I + \lambda(x)Q)(G - I)h(x) \\ &\quad - L(w - f)(x) - (w - f)(x) + \lambda(x)QL(w - f)(x) \\ &\quad + (1 + \lambda(x))Hr(x) + \lambda(x)(Q - I)Hr(x) \\ &= \lambda(x)Q\{Gh - L(w - f) + Hr\}(x) \\ &\quad + \{Gh - L(w - f) + Hr\}(x) \\ &\quad + (1 + \lambda(x))(Hr(x) - Hr(x)) \\ &\quad + (w - f)(x) \\ &= (w - f)(x). \end{aligned}$$

Suppose now that (w, h) is a solution for the PE for the PDP with charges (f, r), so that according to Definition 2.4, for $x \in E, z \in \partial^* E$,

$$\mathcal{A}w(x) = 0 \text{ and } \mathfrak{C}w(z) = 0, \tag{28}$$

$$\mathcal{A}h(x) = w(x) - f(x) \text{ and } \mathfrak{C}h(z) = -r(z). \tag{29}$$

From (28) and (16) in Theorem 3.1, we obtain that

$$\begin{aligned} 0 &= L\mathcal{A}w(x) \\ &= Gw(x) - w(x) - H(Q-I)w(x) \\ &= Gw(x) - w(x) \end{aligned}$$

since from (28)

$$\begin{aligned} H(Q-I)w(x) &= e^{-t^*(x)-\Lambda(t^*(x),x)}(Qw(\phi(t^*(x),x)) - w(\phi(t^*(x),x))) \\ &= e^{-t^*(x)-\Lambda(t^*(x),x)}\mathfrak{C}w(\phi(t^*(x),x)) \\ &= 0. \end{aligned}$$

Now from (16) in Theorem 3.1 and (29), it follows that

$$\begin{aligned} L\mathcal{A}h(x) &= L(w-f)(x) \\ &= Gh(x) - h(x) - H(Q-I)h(x) \\ &= Gh(x) - h(x) + Hr(x) \end{aligned} \tag{30}$$

since from (29),

$$\begin{aligned} HQh(x) - Hh(x) &= e^{-t^*(x)-\Lambda(t^*(x),x)}(Qw(\phi(t^*(x),x)) - w(\phi(t^*(x),x))) \\ &= e^{-t^*(x)-\Lambda(t^*(x),x)}\mathfrak{C}h(\phi(t^*(x),x)) \\ &= -e^{-t^*(x)-\Lambda(t^*(x),x)}r(\phi(t^*(x),x)) \\ &= -Hr(x). \end{aligned}$$

Thus from (30), $L(w-f) = Gh - h + Hr$, which yields the desired result. □

Next we show that, if there exists an invariant probability measure for $\{X_t\}$ (equivalently a σ-finite invariant measure π for G; see [2]), then, as expected, the stationary cost values for the PDP and the Markov chain $\{Y_n\}$ are the same.

PROPOSITION 3.3. *Assume that an invariant probability measure for $\{X_t\}$, labeled μ, exists and satisfies (see* (4)*) $E_\mu[N_t] < \infty$. Denote by σ the associated boundary measure (see equation* (34.18) *in* [1]*) and by π the corresponding invariant σ-finite measure for $\{Y_n\}$. Then the following identity holds:*

$$\pi(Lf + Hr) = \int_E f(x)\mu(dx) + \int_{\partial^* E} r(z)\sigma(dz),$$

provided that one of the members exists.

Proof. Assume that f and r are positive real-valued functions. Let us define r_n and f_n as follows: $r_n(x) = r(x) \wedge n$, $f_n(x) = f(x) \wedge n$. Then by repeating the same arguments as in Proposition 34.13 and Theorem 34.15 of Davis [1, pp. 116–117], we obtain that

$$E_\mu\left[\int_0^t e^{-s} r_n(X_{s-})dp^*(s)\right] = (1 - e^{-t})\sigma(r_n).$$

From the monotone convergence theorem, it follows that

$$E_\mu\left[\int_0^\infty e^{-s} r_n(X_{s-})dp^*(s)\right] = \sigma(r_n).$$

Now using Fubini's theorem and the fact that μ is an invariant probability measure for $\{X_t\}$, we obtain that

$$E_\mu\left[\int_0^\infty e^{-s} f_n(X_s)ds + \int_0^\infty e^{-s} r_n(X_{s-})dp^*(s)\right] = \mu(f_n) + \sigma(r_n).$$

On the other hand, from Proposition 32.34 of [1], it is easy to show that

$$E_\mu\left[\int_0^\infty e^{-s} f_n(X_s)ds + \int_0^\infty e^{-s} r_n(X_{s-})dp^*(s)\right] = \mu \sum_{i=0}^{\infty} K^i(Lf_n + Hr_n).$$

Again using the monotone convergence theorem, it follows that

$$\mu(f) + \sigma(r) = \mu \sum_{i=0}^{\infty} K^i(Lf + Hr).$$

However, from the proof of Theorem 3.5 in [2], we have that $\mu \sum_{i=0}^{\infty} K^i = \pi$ and therefore

$$\mu(f) + \sigma(r) = \pi(Lf + Hr).$$

Now, for the general case, we consider the following classical decompositions,

$$f = (0 \wedge f) - (0 \wedge -f) \qquad \text{and} \qquad r = (0 \wedge r) - (0 \wedge -r),$$

so that we can apply the previous identity to obtain the final result. □

Remark 3.4. Combining the two previous results we have that $(\mu(f) + \sigma(r), h)$ is a solution for the PE for the PDP with charges (f, r) if and only if $(\pi(Lf + Hr), h)$ is a solution for the PE for G with charge $Lf + Hr$.

We conclude this section with a result about the uniqueness of the solution of the PE for PDPs.

PROPOSITION 3.5. *Assume that the process $\{X_t\}$ is irreducible and has an invariant probability measure, labeled μ, such that $E_\mu[N_t] < \infty$. Denote by σ the associated boundary measure. Suppose that $(\mu(f) + \sigma(r), h_1)$ and $(\mu(f) + \sigma(r), h_2)$ are two solutions for the PE for the PDP with charges (f, r) such that $\mu(|h_1| + |h_2|) < \infty$ and h_1, h_2 satisfy the condition in* (6). *Then for some constant c, $h_1(x) = h_2(x) + c$ for μ-almost every $x \in E$.*

Proof. Set $h = h_1 - h_2$, so that for all $x \in E$, $\mathcal{A}h(x) = 0$ and for all $z \in \partial^* E$, $\mathfrak{C}h(z) = 0$. By using Theorem 2.2 and the fact that h satisfies the condition in (6), we obtain that for all $t \in \mathbb{R}_+$, $h(x) = E_x[h(X_t)]$. Consequently, we obtain that $h(x) = Uh(x)$, where U denotes the resolvent kernel associated with $\{X_t\}$. Since U is positive recurrent, it follows that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} U^k h(x) = \mu(h) \quad \mu\text{-a.s.}$$

Therefore,

$$h(x) = \mu(h) \quad \mu\text{-a.s.},$$

showing the desired result. □

4. Sufficient condition. In this section we present a sufficient condition for the existence of a solution for the PE for G. In order to do that, we recall now some classical definitions related to Markov chains. For a complete exposition on the subject the reader is referred to the book of Meyn and Tweedie [6].

In order to simplify the notation, let us define

$$X \doteq E \cup \partial^* E.$$

Let us denote by τ_A the first hitting time of the set $A \in \mathcal{B}(X)$ not including time 0:

$$\tau_A \doteq \inf\{n \geq 1 : Y_n \in A\}.$$

DEFINITION 4.1. *$\{Y_n\}$ is φ-irreducible, if there exists a measure φ defined on $(X, \mathcal{B}(X))$ such that, whenever $\varphi(A) > 0$, it follows that $P_x(\tau_A < \infty) > 0$ for all $x \in X$.*

If $\{Y_n\}$ is φ-irreducible, then there exists a maximal irreducibility measure (see [6, Proposition 4.2.2, p. 88]). In the following, we will use ψ to denote a maximal irreducibility measure for $\{Y_n\}$. Associated with a maximal irreducibility measure for $\{Y_n\}$ is the definition

$$\mathcal{B}(X)^+ \doteq \{A \in \mathcal{B}(X) : \psi(A) > 0\}.$$

DEFINITION 4.2. *$\{Y_n\}$ is called recurrent if it is ψ-irreducible and $\sum_{n=1}^{\infty} G^n(x, A) = \infty$ for all $x \in X$ and $A \in \mathcal{B}(X)^+$.*

DEFINITION 4.3. *A set $C \in \mathcal{B}(X)$ is called ν_a-petite if there exists a nontrivial measure ν_a on $(X, \mathcal{B}(X))$ such that*

$$(\forall x \in C)\,(\forall B \in \mathcal{B}(X)) \quad \sum_{n=1}^{\infty} a(n) G^n(x, B) \geq \nu_a(B),$$

where the positive sequence $\{a(n)\}$ satisfies $\sum_{n=0}^{\infty} a(n) = 1$.

The resolvent K_{a_ϵ} of the Markov chain G is defined for $0 < \epsilon < 1$ by

$$(\forall x \in X)\,(\forall A \in \mathcal{B}(X)) \quad K_{a_\epsilon}(x, A) \doteq (1 - \epsilon) \sum_{n=1}^{\infty} \epsilon^n G^n(x, B).$$

The following assumptions will be used in what follows:

(A.1) The Markov kernel G is recurrent.

(A.2) For a petite set $C \in \mathcal{B}(X)^+$ the Markov kernel G satisfies the following inequality for all $x \in X$,

$$GV(x) \leq V(x) - v(x) + bI_C(x),$$

for a function $v \geq \mathcal{L}$ and $V \geq 0$ everywhere finite and bounded in C.

The following results follow from (A.1) and (A.2) (recall the definition of $\mathcal{L}$ in (11)).

PROPOSITION 4.4. *Suppose that* (A.1) *and* (A.2) *hold. Then*

(i) *there exists a unique (up to constant multiples) σ-finite invariant measure π for G with $0 < \pi(C) < \infty$ and $\pi(\mathcal{L}) < \infty$;*

(ii) *$\pi(v) < \infty$.*

Proof. Since $v(x) \geq \mathcal{L}(x)$, it follows from (A.2) that $GV(x) \leq V(x) - \mathcal{L}(x) + bI_C(x)$, and from this result and (A.1) we have from Corollary 4.5 in [2] that (i) holds. From Theorem 2.2(iii) in [4] we obtain that for any $A \in \mathcal{B}(X)^+$ there exists $c(A)$ such that

$$E_x\left[\sum_{k=0}^{\tau_A-1} v(Y_k)\right] \leq V(x) + c(A).$$

Therefore, from Theorems 10.0.1 and 10.4.9 in [6], we have that

$$\begin{aligned}\pi(v) &= \int_C \pi(dx) E_x\left[\sum_{k=1}^{\tau_C} v(Y_k)\right] \\ &= \int_C \pi(dx) E_x\left[\sum_{k=0}^{\tau_C-1} v(Y_k)\right] \\ &\leq \left[\sup_{x\in C} V(x) + c(C)\right]\ \pi(C) < \infty,\end{aligned}$$

which gives (ii). □

There is no loss of generality in assuming that $\pi(\mathcal{L}) = 1$.

THEOREM 4.5. *Suppose that* (A.1) *and* (A.2) *hold. Then for any function* $g \in \mathbb{B}(X)$ *such that* $\sup_{x\in X} \frac{|g(x)|}{v(x)} < \infty$, *there exists* $h \in \mathbb{B}(X)$ *such that*

$$Gh(x) = h(x) - [g(x) - \pi(g)\mathcal{L}(x)], \tag{31}$$

and for some $R > 0$,

$$|h(x)| \leq R(V(x) + 1). \tag{32}$$

Proof. Since $|g(x)| \leq \beta v(x)$ for a constant $\beta > 0$, it follows from Proposition 4.4 (ii) that $\pi(|g|) < \infty$. Define $\bar{g}(x) \doteq g(x) - \pi(g)\mathcal{L}(x)$. Using the fact that $\mathcal{L}(x) \leq v(x)$, it is easy to obtain that $\sup_{x\in X} \frac{|\bar{g}(x)|}{v(x)} < \infty$.

In the case where the Markov chain G is strongly aperiodic, the proof of Theorem 2.3 in [4] can be used to show the existence of the function h satisfying (31) and (32).

In the general case, the K_{a_ϵ}-chain can be considered, and following the proof of Theorem 2.3 in [4] we have that

$$K_{a_\epsilon} V_\epsilon \leq V_\epsilon - v + b' K_{a_\epsilon} I_C, \tag{33}$$

where the function V_ϵ equals a constant multiple of V and b' is a strictly positive constant.

However, now we cannot follow the proof of Glynn and Meyn (see Theorem 2.3 in [4]) because the function v may be less than 1 and consequently we cannot adopt the proof of Theorem 14.2.9 in [6] to get the (V3) condition for the K_{a_ϵ}-chain. (For the definition of the (V3) condition, see page 337 in the book of Meyn and Tweedie [6].)

From now on we follow a different line from [4]. We adopt the notation of Meyn and Tweedie for the splitting technique of a Markov chain (see p. 103 in [6]).

Since $C \in \mathcal{B}(X)^+$ and using Theorem 5.2.2 in [6], there exist a set D, a probability ν, and a real $\delta > 0$ such that $D \subset C$, $D \in \mathcal{B}(X)^+$, $\nu(D) = 1$, and

$$(\forall x \in X)\ (\forall A \in \mathcal{B}(X)) \qquad K_{a_\epsilon}(x, A) \geq \delta I_D(x)\nu(A). \tag{34}$$

Note that V_ϵ is bounded on D since by hypothesis V is bounded on C. Since K_{a_ϵ} satisfies the minorization condition (34), we can introduce the split chain $\check{K}_{a_\epsilon}$ defined on $(\check{X}, \mathcal{B}(\check{X}))$. Let us define the function $\check{V}$ on $\check{X}$ by

$$(\forall x_0 \in X_0) \quad \check{V}(x_0) = V_\epsilon(x), \tag{35}$$

$$(\forall x_1 \in D_1) \quad \check{V}(x_1) = V_\epsilon(x), \tag{36}$$

$$(\forall x_1 \in X_1 - D_1) \quad \check{V}(x_1) = v(x_1) + [\delta\{\nu(V_\epsilon) - b'\} \vee 0]. \tag{37}$$

Using the fact that $\check{V}$ is bounded on $D_0 \cup D_1$ and (35)–(37), it can be shown that there exists a strictly positive number d such that

$$\check{K}_{a_\epsilon}\check{V} \le \check{V} - \check{v} + b'\check{K}_{a_\epsilon} I_{C_0 \cup C_1} + d I_{D_0 \cup D_1}, \tag{38}$$

where $\check{v}(x_i) = v(x)$ for $i = 0, 1$.

Let us introduce the following notation: $\check{H}_c \doteq \sum_{i=0}^{\infty} c(i)\check{K}^i_{a_\epsilon}$, where c is a distribution on $\mathbb{Z}_+$.

Since $C_0 \cup C_1$ is a petite set for the split chain $\check{K}_{a_\epsilon}$, it follows that for any set B in $\mathcal{B}(\check{X})^+$ we can find a measure Ψ_{a_ϵ} such that $\Psi_{a_\epsilon}(B) > 0$ and satisfies for $i = 0, 1$

$$I_{C_0 \cup C_1}(x_i) \le \Psi^{-1}_{a_\epsilon}(B)\check{H}_{a_\epsilon}(x_i, B). \tag{39}$$

Combining (38) and (39), we obtain that there exists a real $\beta > 0$ such that

$$\begin{aligned} \check{K}_{a_\epsilon}\check{V}(x_i) &\le \check{V}(x_i) - \check{v}(x_i) + \beta\Psi_a^{-1}(B)\frac{1}{2}\left[I + \check{K}_{a_\epsilon}\right]\check{H}_{a_\epsilon}(x_i, B) \\ &\le \check{V}(x_i) - \check{v}(x_i) + \beta\Psi_a^{-1}(B)\check{H}_{a * a_\epsilon}(x_i, B), \end{aligned} \tag{40}$$

where a is the distribution defined by $a(0) = a(1) = \frac{1}{2}$.

Now we can go back to the proof of Glynn and Meyn (see Theorem 2.2 in [4]). Indeed, by noting that $\sum_{i=0}^{\infty} i\, a * a_\epsilon(i) < \infty$, we can conclude from (40) that

$$\check{E}_{x_i}\left[\sum_{k=0}^{\tau_{\check{\alpha}}-1} f(\check{Y}_k)\right] \le \check{V}(x_i) + c, \tag{41}$$

where the Markov chain $\{\check{Y}_k\}$ is generated by $\check{K}_{a_\epsilon}$.

Consequently, we can claim that there exists a solution (labeled h_ϵ) for the PE $K_{a_\epsilon}h_\epsilon(x) = h_\epsilon(x) - [g(x) - \pi(g)\mathcal{L}(x)]$ satisfying $|h_\epsilon(x)| \le R_\epsilon(V_\epsilon(x) + 1)$ for some $R_\epsilon > 0$. Finally, the existence of a function h satisfying (31) and (32) follows from Theorem 2.2 in [4]. □

We present now a sufficient condition for the existence of a solution for the PE for G (and thus for the PDP).

COROLLARY 4.6. *Suppose that* (A.1) *and* (A.2) *hold. Then for all functions* $f \in \mathbb{B}(E)$ *and* $r \in \mathbb{B}(\partial^* E)$ *satisfying* $\sup_{x \in X} \frac{|Lf(x) + Hr(x)|}{v(x)} < \infty$, *there exists a function* $h \in \mathbb{B}(X)$ *such that*

$$Gh(x) = h(x) + Lw(x) - Lf(x) - Hr(x),$$

where $w(x) \doteq \pi(Lf + Hr)$.

Moreover, for some $R \in \mathbb{R}_+$, $|h(x)| \le R(V(x) + 1)$.

Proof. By noting that $Lw(x) = \mathcal{L}(x)\pi(Lf + Hr)$, the result follows as a direct consequence of Theorem 4.5. □

Remark 4.7. Notice that condition (A.2) in Theorem 4.5 does not imply that $\pi(E) < \infty$ (the invariant measure for G can be just σ-finite), since we are just requiring that $v \geq \mathcal{L}$. This condition (A.2) is very much similar to condition (V3) of [6], which has been used in several recent references to obtain f-regularity of a sampled process (see, for instance, [4], [6], [7]). However, note that since v may be less than 1, condition (A.2) implies neither f-regularity nor positive Harris recurrence for the Markov kernel G.

Remark 4.8. Suppose that for a petite set $C \in \mathcal{B}(X)^+$ there exists a function $V \in \mathcal{D}(\widetilde{\mathcal{A}})$ positive, everywhere finite and bounded in C such that for some functions $\alpha(x) \geq 1$, $\gamma(z) \geq 1$, and every $x \in C^c$,

$$\mathcal{A}V(\phi(s,x)) \leq -\alpha(\phi(s,x)) \tag{42}$$

for every $0 \leq s < t^*(x)$, and

$$QV(z) - V(z) \leq -\gamma(z) \tag{43}$$

for every $z = \phi(t^*(x), x)$. From (16) it follows that for every $x \in C^c$,

$$\begin{aligned} GV(x) - V(x) &= L\mathcal{A}V(x) + H(Q - I)V(x) \\ &\leq -(L\alpha(x) + H\gamma(x)) \\ &\leq -\mathcal{L}(x) \end{aligned}$$

so that condition (A.2) will be satisfied with $v(x) = L\alpha(x) + H\gamma(x)$ for $x \in C^c$. Thus whenever $f \in \mathbb{B}(E)$ and $r \in \mathbb{B}(\partial^* E)$ are such that $f(\phi(t,x))$ and $r(z)$ are bounded for every $x \in C$, $0 \leq t < t^*(x)$, $z = \phi(t^*(x), x)$, and $|f(\phi(t,x))| \leq a\alpha(\phi(t,x))$, $|r(z)| \leq a\gamma(z)$ for some $a > 0$ and every $x \in C^c$, $0 \leq t < t^*(x)$, $z = \phi(t^*(x), x)$, the PE with charges f and r will have a solution. Condition (42), obtained from the discrete-time condition (A.2), resembles the continuous-time condition in equation (15) of [4]. Note, however, that we need the extra boundary condition (43), and that (42) was defined over the petite set C for the discrete-time process $\{Y_n\}$. So in this sense the conditions in (42) and (43) can be seen as a continuous-time condition using the structure of petite sets for the discrete-time Markov process $\{Y_n\}$, usually easier to be checked.

5. Example. In this section we apply the results of the previous section to a capacity expansion model (see sections 21.13 and 34.45 in [1] for a complete and more detailed description of the model) to illustrate the relevance of our approach. The demand for some utility is modeled as a random point process, i.e., it increases by one unit at random times. This demand is met by consecutive construction of identical expansion projects. Each project meets K units of demand when completed. We assume that if there is an excess demand of at least N units then the construction of a new project is started at a rate of k_i per unit of time, where i corresponds to the present level of excess demand, which is completed after a lead time of p units of time; if the excess demand is less than N, then no construction takes place. New demands occur with rate $\lambda_i(\zeta)$, where ζ represents the elapsed construction time of the present project. We assume that $\lambda_i(\zeta)$ is a continuous function in $\zeta \in [0, p]$ and that there exists $\theta > 0$ such that $k_i \geq \theta$ for all $i \geq N$. This problem can be modeled

as a PDP with state space

$$E = \Big\{\{N-K,\dots,N-1\}\times\{0\}\Big\} \cup \Big\{\mathbb{N}_{\mathrm{N}}\times[0,p)\Big\},$$

where $\mathbb{N}_j$ denotes the set of integers greater than or equal to j.

We show that for this example the condition given by Corollary 4.6, through Remark 4.8, can be applied.

Suppose that there exists a sequence of positive numbers $\{c_i\}$ and $c>0$ such that

$$\limsup_{i\to\infty}\frac{c_i}{k_i}\max_{\zeta\in[0,p]}\lambda_i(\zeta) < \frac{c}{p}, \tag{44}$$

and for some i_0,

$$\sum_{j=0}^{K-1} c_{i-K+j} \geq c \tag{45}$$

for all $i\geq i_0$. In view of hypothesis (44), there exist $\varepsilon>0$ and $M>N, M>i_0$, such that for all $\zeta\in[0,p]$ and $i>M$ we have that

$$\frac{c_i}{k_i}\lambda_i(\zeta) \leq \frac{c}{p}-\varepsilon$$

and thus

$$\theta\varepsilon \leq k_i\varepsilon \leq \frac{c}{p}k_i - c_i\lambda_i(\zeta). \tag{46}$$

It is easy to see that $\{Y_n\}$ is irreducible. Set $a_{i+1}=a_i+c_i$, $i=N-K,N-K+1,\dots$, and define the function V as

$$(\forall x=(i,\zeta)\in E)\qquad V(x)=\begin{cases} a_i+\dfrac{c}{p}(p-\zeta), & i\geq N,\\ a_i+c, & i<N.\end{cases}$$

We have that for $x=(i,\zeta)\in E$, $i\geq N$,

$$\mathcal{A}V(x) = -\frac{c}{p}k_i + c_i\lambda_i(\zeta)$$

and for $z=(i,p)$,

$$QV(z)-V(z) = c - \sum_{j=0}^{K-1} c_{i-K+j},$$

so that from (45) and (46) we have that (42) and (43) are satisfied on the compact set $C:=\{x=(i,\zeta): N-K\leq i\leq M, 0\leq\zeta\leq p\}$. Set $\alpha(x)=\frac{c}{p}k_i-c_i\lambda_i(\zeta)$ for $z=(i,\zeta)\in E$, $i\geq N$, and $\gamma(z)=\sum_{j=0}^{K-1}c_{i-K+j}-c$ for $x=(i,p)$, $i\geq N$. From (ii) in Theorem 6.0.1 in [6], it follows that C is a petite set and thus from Remark 4.8 there exists a solution of the PE for the PDP with charge (f,r) whenever $f\in\mathbb{B}(E)$ and $r\in\mathbb{B}(\partial^* E)$ are bounded in C, and $|f(x)|\leq a\alpha(x)$, $|r(z)|\leq a\gamma(z)$ for some $a>0$, and every $x=(i,\zeta)\in C^c, z=(i,p)$.

Acknowledgment. The authors are grateful to anonymous referees for their suggestions which have greatly improved the presentation of the paper.

REFERENCES

[1] M. H. A. Davis, *Markov Models and Optimization*, Chapman and Hall, London, 1993.
[2] F. Dufour and O. L. V. Costa, *Stability of piecewise-deterministic Markov processes*, SIAM J. Control Optim., 37 (1999), pp. 1483–1502.
[3] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.
[4] P. Glynn and S. P. Meyn, *A Lyapunov bound for solutions of the Poisson equation*, Ann. Probab., 24 (1996), pp. 916–931.
[5] O. Hernandez-Lerma and J.-B. Lasserre, *Discrete-Time Markov Control Processes*, Springer-Verlag, Berlin, 1996.
[6] S. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, Berlin, 1993.
[7] S. Meyn and R. Tweedie, *State-dependent criteria for convergence of Markov chains*, Ann. Appl. Probab., 4 (1994), pp. 149–168.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 1002–1012

A NESTED COMPUTATIONAL APPROACH TO THE DISCRETE-TIME FINITE-HORIZON LQ CONTROL PROBLEM*

GIOVANNI MARRO†, DOMENICO PRATTICHIZZO‡, AND ELENA ZATTONI†

Abstract. A new algorithmic setting is proposed for the discrete-time finite-horizon linear quadratic (LQ) optimal control problem with constrained or unconstrained final state, no matter whether the problem is cheap, singular, or regular. The proposed solution, based on matrix pseudoinversion, is completed and made practically implementable by a nesting procedure for welding optimal subarcs that enables arbitrary enlargement of the control time interval.

Key words. linear quadratic problems, algorithms, discrete-time systems, pseudoinversion

AMS subject classifications. 49N05, 49N10, 93C05, 93C62

DOI. 10.1137/S0363012901384429

1. Introduction. The linear quadratic (LQ) and linear quadratic regulator (LQR) optimal control problems have been extensively studied and are now well settled in the literature. In particular, the regular LQ and LQR problems are exhaustively treated in many well-known specialist books and textbooks, such as, for instance, those by Bryson and Ho [3], Kwakernaak and Sivan [12], Anderson and Moore [1], Chen and Francis [5], and Lewis and Syrmos [13]. Due to increasing interest in H_2 and H_∞ control problems, more recently much effort has been made to extend the solutions of LQ and LQR problems also to cheap and singular cases. In general, these new contributions concern the infinite-horizon problems, where optimality jointly with internal stability is required. The basic properties and algorithms referring to cheap and singular optimal control have been investigated by Arnold and Laub in [2], Van Dooren in [23], Ionescu, Oară, and Weiss in [11, 10, 9], Geerts in [6], Saberi, Sannuti, Chen, and Stoorvogel in [17, 18, 4, 22], and Soethoudt, Trentelman and Ran in [21, 16]. Extension to cheap and singular problems is obtained by using matrix pencils or linear matrix inequalities.

The above contributions mainly refer to continuous- or discrete-time *infinite-horizon* LQR problems. Conversely, this work focuses on the *finite-horizon* case for discrete-time systems. It applies to cheap, singular, or regular problems. Considering a *sharp* constraint on the final state, this paper extends the results given in [16], where the infinite-horizon case with *asymptotic* endpoint constraints is investigated. However, in the present paper, the requirement that the cost functions are positive semidefinite is still present, while in [16] it is not.

From the algorithmic viewpoint, the present work differs from those of the abovementioned papers and is inspired by the system structure algorithm introduced by Silverman in his well-known contribution [20]; see also [19, 8]. In the present work, the solution is achieved by the straightforward technique of pseudoinverting a particular system matrix as in [20], but the time interval considered can be arbitrarily

*Received by the editors February 6, 2001; accepted for publication (in revised form) November 28, 2002; published electronically June 25, 2003. This work was partially supported by the Italian Ministry of University and Research.

http://www.siam.org/journals/sicon/42-3/38442.html

†Dipartimento di Elettronica, Informatica e Sistemistica, Università di Bologna, viale Risorgimento, 2, 40136 Bologna, Italy (gmarro@deis.unibo.it, ezattoni@deis.unibo.it).

‡Dipartimento di Ingegneria dell'Informazione, Università di Siena, via Roma, 77, 53100 Siena, Italy (prattichizzo@dii.unisi.it).

enlarged through a suitable multilevel computational scheme; this enables overcoming any dimensionality constraint thus achieving a certain interest with respect to computational practice.

The following notation will be used: $\mathbb{R}^n$ stands for the set of all n-tuples of real numbers, A' and $A^{\#}$ are used for the transpose and pseudoinverse of matrix A, respectively, and I stands for the identity matrix with a suitable dimension.

2. Statement of the problem. Consider the linear discrete-time-invariant dynamical system

$$\begin{aligned} x(k+1) &= A\,x(k) + B\,u(k), \quad x(0) = x_0, \\ e(k) &= C\,x(k) + D\,u(k), \end{aligned} \tag{1}$$

with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^p$, $e \in \mathbb{R}^q$, $k \in [0, N-1]$, the cost function

$$J := \sum_{k=0}^{N-1} e(k)'e(k) + x(N)'Z'Z\,x(N), \tag{2}$$

and the constraint on the final state

$$G\,x(N) = y_f, \quad y_f \in \mathbb{R}^r. \tag{3}$$

Constraint (3) is assumed to be *feasible*; i.e., the initial state x_0, the constraint on the final state (G, y_f), and the number of steps N of the control time interval are assumed to be such that at least one N-step state trajectory from x_0 to an $x(N)$ satisfying (3) exists. It is worth noting that the final state can be completely assigned by assuming $G = I$ in (3); in this case, the terminal cost becomes irrelevant to the problem solution, and in (2) matrix Z can be assumed to be zero. On the other hand, assumptions $G = 0$ and $y_f = 0$ correspond to consider the final state as completely free. The discrete-time finite-horizon LQ optimal control problem with constrained final state can be stated as follows.

PROBLEM 1. *Consider the dynamic system* (1)*, and find a control sequence*[1] $u(k)$ $(k = 0, 1, \ldots, N-1)$ *such that the cost function* (2) *is minimized under the constraint* (3).

The solution of the above problem is given with no distinctions among cheap $(D'D = 0)$, singular $(\det(D'D) = 0)$, and regular cases $(\det(D'D) \neq 0)$. To this aim, it is convenient to introduce the following notation for the sequences of controls and of extended outputs (i.e., of the outputs completed with the square root of the terminal cost):

$$u_N := \begin{bmatrix} u(0) \\ u(1) \\ \vdots \\ u(N-1) \end{bmatrix}, \quad e_N := \begin{bmatrix} e(0) \\ e(1) \\ \vdots \\ e(N-1) \\ Z\,x(N) \end{bmatrix}. \tag{4}$$

Note that, owing to the above definition of e_N, the cost function (2) can also be given as

$$J = \|e_N\|_2^2, \tag{5}$$

[1]The control sequence corresponding to the minimum value of cost may be nonunique, particularly if the system (1) is not left-invertible.

where $\|e_N\|_2$ is the 2-norm of vector e_N. Furthermore, let us express the final state $x(N)$ as a function of the initial state x_0 and the control sequence u_N,

$$x(N) = A^N x_0 + A^{N-1} B\, u(0) + \cdots + B\, u(N-1) = A^N x_0 + L_N\, u_N,$$

with

$$L_N := \left[\begin{array}{cccc} A^{N-1}B & A^{N-2}B & \cdots & B \end{array} \right], \tag{6}$$

so that the constraint (3) can be written as

$$y_f - GA^N x_0 = GL_N u_N. \tag{7}$$

According to (4), (5), and (7), Problem 1 is restated as follows.

PROBLEM 1a. *Referring to the dynamic system* (1)*, find a vector* u_N *such that* $\|e_N\|_2$ *is minimized under the constraint* (7)*.*

An algorithmic solution of Problem 1 is derived through a constructive proof of the following theorem.

THEOREM 1. *Consider the dynamic system* (1)*. Assume that the constraint* (3) *is feasible. Let* K *be a matrix, the columns of which form a basis for* $\ker(GL_N)$*. Let*

$$T_N := -(I - K\,(B_N K)^{\#}\, B_N)\,(GL_N)^{\#}\, GA^N - K\,(B_N K)^{\#}\, A_N, \tag{8}$$

$$V_N := (I - K\,(B_N K)^{\#}\, B_N)\,(GL_N)^{\#}, \tag{9}$$

$$C_N := (I - B_N K\,(B_N K)^{\#})\,(A_N - B_N\,(GL_N)^{\#}\, GA^N), \tag{10}$$

$$D_N := (I - B_N K\,(B_N K)^{\#})\, B_N\,(GL_N)^{\#}, \tag{11}$$

with

$$A_N := \left[\begin{array}{c} C \\ CA \\ \vdots \\ CA^{N-1} \\ ZA^N \end{array} \right], \quad B_N := \left[\begin{array}{cccc} D & 0 & \cdots & 0 \\ CB & D & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{N-2}B & CA^{N-3}B & \cdots & D \\ ZA^{N-1}B & ZA^{N-2}B & \cdots & ZB \end{array} \right]. \tag{12}$$

Then

$$u_N^o := T_N\, x_0 + V_N\, y_f \tag{13}$$

is the optimal solution of Problem 1 *with the feasible constraint* (3)*, and vector*

$$e_N^o := C_N\, x_0 + D_N\, y_f \tag{14}$$

is the optimal extended output sequence providing the optimal cost $J^o = \|e_N^o\|_2^2$*.*

Proof. First, let us consider the plainer problem defined only by (1) and (2), i.e., with no constraints on the final state. The sequence of outputs $e(0), e(1), \ldots, e(N-1)$ and the square root $Z\,x(N)$ of the terminal cost are related to the input sequence $u(0), u(1), \ldots, u(N-1)$ by the set of equations

$$\begin{aligned} e(0) &= C\, x_0 + D\, u(0), \\ e(1) &= C\,(A\, x_0 + B\, u(0)) + D\, u(1), \\ &\vdots \\ e(N-1) &= C\,(A^{N-1} x_0 + A^{N-2} B\, u(0) + \cdots + B\, u(N-2)) + D\, u(N-1), \\ Z\, x(N) &= Z\,(A^N x_0 + A^{N-1} B\, u(0) + \cdots + B\, u(N-1)), \end{aligned}$$

which can be written in the compact form

$$e_N = A_N\, x_0 + B_N\, u_N. \tag{15}$$

The sequence of controls u_N minimizing $\|e_N\|_2$ is obtained by applying the pseudoinverse property to (15). Hence

$$u_N^o = -B_N^{\#} A_N\, x_0. \tag{16}$$

It is worth noting that, in the case of an unconstrained final state, the optimal control sequence depends solely on the initial state x_0. From (15) and (16), it follows that

$$e_N^o = (I - B_N B_N^{\#})\, A_N\, x_0.$$

Hence, in the absence of constraints on the terminal state, the theorem is proven with

$$T_N := -B_N^{\#} A_N,$$
$$C_N := (I - B_N B_N^{\#})\, A_N.$$

In this case, of course, V_N and D_N are not defined.

The complete optimal control problem, including the constraint on the final state, can easily be solved by using Lemma 1 in the appendix. In fact, (15) and (7) are taken from (23) and (24), respectively, under the correspondences

$$\zeta = e_N, \quad \Phi = B_N, \quad \lambda = u_N, \quad \bar{\rho} = A_N\, x_0, \quad \Gamma = GL_N, \quad \bar{\sigma} = y_f - GA^N x_0. \tag{17}$$

The optimal control sequence u_N^o, as expressed in (13), can then be derived from (25) by taking into account the correspondences in (17). Similarly, the optimal extended output sequence e_N^o, as expressed in (14), can be derived from (26). □

REMARK 1. *Note that the optimal cost J^o can be expressed in terms of the matrices C_N and D_N defined in* (10) *and* (11) *as the quadratic form*

$$J^o = \begin{bmatrix} x_0 \\ y_f \end{bmatrix}' \begin{bmatrix} C_N' C_N & C_N' D_N \\ D_N' C_N & D_N' D_N \end{bmatrix} \begin{bmatrix} x_0 \\ y_f \end{bmatrix}. \tag{18}$$

The results of Theorem 1 apply whether the system (A, B) is stable or not, or stabilizable or not. However, stability of system (1) is usually required to avoid numerical problems concerning powers of dynamic matrix A arising in (12). A deep discussion on the ill-conditioning of powers of matrices with eigenvalues greater than one is provided in [7]. In the case of stabilizable systems, the following remark applies to overcome the ill-conditioning of matrices involved in Theorem 1.

REMARK 2. *If the system* (1) *is stabilizable, a preliminary pole placement is performed through feedback $u(k) = \bar{u}(k) + Hx(k)$ such that $\bar{A} := A + BH$ is stable and powers of $\bar{A}$ can be robustly computed. Now, let $\bar{u}^o(k)$ be the input in* (13) *which solves the optimal control problem for $(\bar{A}, B, C + DH, D)$. Then the solution for the original system (A, B, C, D) is simply computed as $u^o(k) = \bar{u}^o(k) + H\, x^o(k)$, where $x^o(k)$ is the optimal state time history, which is the same for both systems.*

REMARK 3. *Let P be a basis matrix of* $\ker B_N \cap \ker GL_N$*; see* (12) *and* (6). *From* (25) *of Lemma* 1 *in the appendix, it follows that all the solutions of Problem* 1 *lie on the linear variety*

$$u_N^o(\gamma) = T_N\, x_o + V_N\, y_f + P\, \gamma.$$

According to Remark 5 *in the appendix, the minimum* 2*-norm vector solving Problem* 1 *is*

$$\hat{u}_N^o = \hat{T}_N\, x_o + \hat{V}_N\, y_f,$$

with

$$\begin{aligned} \hat{T}_N &= (I - PP^{\#})\, T_N, \\ \hat{V}_N &= (I - PP^{\#})\, V_N. \end{aligned}$$

It is worth noting that, if the final state is unconstrained, i.e., $G = 0$ and $y_f = 0$ in (3), the technique herein devised proves to be an effective alternative to the solution of the difference (possibly generalized) Riccati equation. The main difference with respect to the Riccati equation is that the pseudoinversion procedure is not recursive in nature and, as already mentioned, is more general since it also works for constrained final state problems.

LQ optimal control problems with hard constraints on the final states are usually approached by means of dynamic programming techniques or gradient methods, which, although formalized in a wide nonlinear setting, always provide solutions through recursive procedures.

The results of Theorem 1 and Remarks 2 and 3 are summarized in the following algorithm solving Problem 1.

ALGORITHM 1.

Step 1. *Check the stability of matrix* A*, and stabilize it by a state feedback according to Remark* 2 *if necessary.*

Step 2. *Evaluate the matrices* A_N*,* B_N*, and* L_N *as defined in* (12) *and* (6)*. Note that if the final state* $x(N)$ *is completely assigned, the computation of the last row of* A_N *and* B_N *is skipped.*

Step 3. *Evaluate a basis matrix* K *for* $\ker(GL_N)$ *and the matrices* T_N*,* V_N*,* C_N*, and* D_N *as defined by* (8)*,* (9)*,* (10)*, and* (11)*. Compute the basis matrix* P *and the matrices* $\hat{T}_N$ *and* $\hat{V}_N$ *defined in Remark* 3 *if* u_N^o *is required to be minimum* 2*-norm.*

Step 4. *Evaluate the optimal control sequence* u_N^o *as defined by* (13) *and the optimal cost* J^o *as defined by* (18)*.*

3. The problem nesting procedure. In the proposed solution, the optimal control sequence is computed as a function of x_0 and y_f by pseudoinverting suitably defined matrices. Since the dimensions of the matrices to be pseudoinverted are proportional to the number of steps N of the control time interval, this technique is subject to a dimensionality constraint: N cannot be greater than a maximum which depends on the available computational capability. However, this drawback can be overcome by resorting to the contrivance of nesting problems of the same type at different levels, as described below for a two-level nesting procedure.

In what follows, extension of (4), (6), (12), (10), (11) is used to denote the variables defined on a generic N_1-step (or N_2-step) control time interval; i.e., the variable subscript is intended for denoting the control interval length.

Refer to Problem 1. For the sake of simplicity, assume that $N_1, N_2 \in \mathbb{Z}^+$ exist, satisfying the pseudoinversion dimensionality constraint and such that $N_1 N_2 = N$. Then the following corollary of Theorem 1 can be stated.

COROLLARY 1. *The solution of Problem* 1 *can be obtained through* N_2 *optimal control problems with* N_1 *steps and the final state sharply assigned* (first level problems) *and one optimal control problem with* N_2 *steps and the final state weighted*

and/or constrained as in the original problem (second level problem). *Let* R_{N_1} *be a basis matrix of the subspace* $\operatorname{im} L_{N_1}$. *Denote the initial and final states of the* j*th* $(j=1,\dots,N_2-1)$ *first level problem, respectively, by* $\tilde{x}(j)$ *and* $\tilde{x}(j+1)$, *and let them belong to the optimal state trajectory of the second level optimal control problem defined by the system*

$$\begin{aligned} \tilde{x}(j+1) &= \tilde{A}\tilde{x}(j)+\tilde{B}\alpha(j), \quad \tilde{x}(0)=x_0, \\ \tilde{e}(j) &= \tilde{C}\tilde{x}(j)+\tilde{D}\alpha(j), \end{aligned} \tag{19}$$

with

$$\begin{aligned} \tilde{A} &:= A^{N_1}, \\ \tilde{B} &:= R_{N_1}, \\ \tilde{C} &:= C_{N_1}+D_{N_1}A^{N_1}, \\ \tilde{D} &:= D_{N_1}R_{N_1}, \end{aligned} \tag{20}$$

and $\tilde{x}$, α, $\tilde{e}$ *vectors of suitable dimension, by the cost function*

$$\tilde{J}=\sum_{j=0}^{N_2-1}\tilde{e}(j)'\,\tilde{e}(j)+\tilde{x}(N_2)'\,Z'\,Z\,\tilde{x}(N_2) \tag{21}$$

and by the constraint

$$G\,\tilde{x}(N_2)=y_f. \tag{22}$$

Proof. Note that the optimal cost J_j of the jth first level N_1 steps problem is equal to the jth contribution to the optimal cost of the second level problem (this latter also equal to the original problem optimal cost). In fact,

$$\begin{aligned} J_j &= \begin{bmatrix} \tilde{x}(j) \\ \tilde{x}(j+1) \end{bmatrix}' \begin{bmatrix} C_{N_1}'C_{N_1} & C_{N_1}'D_{N_1} \\ D_{N_1}'C_{N_1} & D_{N_1}'D_{N_1} \end{bmatrix} \begin{bmatrix} \tilde{x}(j) \\ \tilde{x}(j+1) \end{bmatrix} \\ &= (C_{N_1}\tilde{x}(j)+D_{N_1}\tilde{x}(j+1))'\,(C_{N_1}\tilde{x}(j)+D_{N_1}\tilde{x}(j+1)), \end{aligned}$$

and

$$C_{N_1}\tilde{x}(j)+D_{N_1}\tilde{x}(j+1)=C_{N_1}\tilde{x}(j)+D_{N_1}\left(A^{N_1}\tilde{x}(j)+R_{N_1}\,\alpha(j)\right)=\tilde{C}\tilde{x}(j)+\tilde{D}\alpha(j)=\tilde{e}(j).$$

Hence the thesis is an immediate consequence of Theorem 1. □

For the nesting procedure to be convenient from the computational viewpoint, the row dimension of matrices $\tilde{C}$ and $\tilde{D}$ appearing in (19), equal to the row dimension of matrices C_{N_1} and D_{N_1}, should be reduced according to the following property.

PROPERTY 1. *Matrices* $\bar{C}_{N_1}\in\mathbb{R}^{(r\times n)}$ *and* $\bar{D}_{N_1}\in\mathbb{R}^{(r\times n)}$ *exist such that*

$$\begin{bmatrix} \bar{C}_{N_1}' \\ \bar{D}_{N_1}' \end{bmatrix} \begin{bmatrix} \bar{C}_{N_1} & \bar{D}_{N_1} \end{bmatrix} = \begin{bmatrix} C_{N_1}' \\ D_{N_1}' \end{bmatrix} \begin{bmatrix} C_{N_1} & D_{N_1} \end{bmatrix},$$

and whose row dimension is $r=\operatorname{rank}\,[C_{N_1}\ D_{N_1}]\le 2n$.

Matrices $\bar{C}_{N_1}$ and $\bar{D}_{N_1}$ can easily be evaluated through a standard singular value decomposition procedure. Note that their row dimension is not directly related to the value of N_1 and is $2n$ at most, while that of C_{N_1} and D_{N_1} is $q\,(N_1+1)$. On the other hand, substitution does not affect the optimal control inputs and the optimal cost for

the finite-horizon N_1N_2-step LQ problem. This is due to the way matrices C_N and D_N enter the optimal cost (18).

REMARK 4. *The assumption of $N = N_1N_2$ does not affect the generality of the proposed approach because, if this is not the case, one can divide the interval N into two parts of lengths N_a and $N_b = N_1N_2$ not violating the computational capability limits. Find the cost and the optimal control matrices on the two intervals, and then simply weld the N_a and N_b intervals by optimizing the overall cost with respect to the unknown state variable $x(N_a)$. Note that $x(N_a)$ must belong to the linear variety reachable in N_a steps from x_0.*

The following algorithm, based on Corollary 1, solves Problem 1 in a nested computational framework.

ALGORITHM 2.

Step 1. *Divide the control time interval into N_2 parts of N_1 steps each, where N_1 and N_2 are such that the maximum computational capability condition is met.*

Step 2. *Consider any of the N_2 subintervals and the corresponding N_1-step optimal control problem with the final state completely assigned and zero weighting matrix Z, referred to as the first level LQ problem. Note that, at this stage, both the initial and final states are assumed to be fixed, but they are, in general, unknown. Compute the resolvent matrices T_{N_1}, V_{N_1}, C_{N_1}, and D_{N_1} for the first level problem as described in steps* 1 *and* 2 *of Algorithm* 1 *and, if convenient, reduce the row dimension of C_{N_1} and D_{N_1} according to Property* 1. *These matrices are common to all the N_1-step optimal control problems.*

Step 3. *Solve the second level problem, defined by* (19), (21), *and* (22), *with Algorithm* 1, *thus obtaining the sequence of the intermediate states (i.e., the initial/terminal state of each pair of subsequent first level subintervals) as $\tilde{x}(j)$, $j \in [0, N_2]$.*

Step 4. *Use the jth pair of initial/terminal states $\tilde{x}(j), \tilde{x}(j+1)$ to solve the jth first level N_1-step problem as described in step* 3 *of Algorithm* 1. *Note that the matrices in* (13) *and* (15) *are equal for all the N_1-step problems and have been computed at step* 1.

Step 5. *The optimal control sequence and the optimal cost for the original control problem are simply obtained by grouping in a unique vector the optimal control sequences and summing up the optimal costs of all the N_1-step optimal control subproblems.*

The nesting idea enables extending the pseudoinversion procedure to finite-horizon LQ optimal control problems with a large value of N. However, the nesting procedure is subject to a dimensionality constraint, as well, which depends on the available computational capability and limits the admissible value of N_1 and N_2.

3.1. The multilevel nesting algorithm. This section shows how to overcome the drawback on the size of N_1 and N_2 for the two-level nesting procedure by means of a multi-level nesting algorithm whose innermost procedure is represented by the simple nesting procedure of Corollary 1.

The multilevel nesting algorithm reduces the computational effort needed to solve a finite-horizon LQ problem. The multilevel procedure basically corresponds to iteratively solving the second level N_2-step optimal control problem, stated in Step 2, by means of a further nesting procedure.

It is an easy matter to verify that the simple implementation of the iteration idea leads to increasing at each level the dimension of the matrices needed to compute the optimal cost $\|e_N^o\|_2^2$. Property 1 enables overcoming the dimensionality drawback

while iterating Algorithm 2 in the multilevel case. The multilevel nesting algorithm solves the N-step LQ optimal control problem as follows.

ALGORITHM 3.

Step 1. *Set* $k=1$ *and* $N_k=N$.

Step 2. *Divide the control time interval into* $N_k=N_{k,1}\,N_{k,2}$ *steps each, where* $N_{k,1}$ *is such that the computational complexity needed to evaluate the resolvent matrices* $T_{N_{k,1}}$, $V_{N_{k,1}}$, $C_{N_{k,1}}$, *and* $D_{N_{k,1}}$ *is met. Compute these matrices.*

Step 3. *Compute the reduced-size cost matrices* $\bar{C}_{N_{k,1}}\in\mathbb{R}^{(r_k\times n)}$ *and* $\bar{D}_{N_{k,1}}\in\mathbb{R}^{(r_k\times n)}$ *as described in Property* 1.

Step 4. *State the second-level* $N_{k,2}$*-step problem as defined in* (19), (21), *and* (22) *with these reduced-size cost matrices.*

Step 5. *Solve the* $N_{k,2}$ *step problem if this is computationally possible. If not, set* $k=k+1$ *and* $N_k=N_{k,2}$, *and go back to step* 1.

Step 6. *Move back over all the nested levels, and assess the optimal control input and the cost for the original* N*-step problem.*

Table 1 illustrates a four-level nesting algorithm. At the second and third levels, $k=2$ and $k=3$, only basic matrices for the innermost problems are evaluated. The two-level nested problem is completely solved only at the last nesting level, i.e., $k=4$. It is worth noting that if matrix factorization of Property 1 were not applied, the dimensions of the matrices $C_{N_{k,1}}$ and $D_{N_{k,1}}$ would significantly increase with k. In fact, at the nesting level $k+1$, both would be $q\prod_{j=1}^{k}N_{j,1}\times n$, while factorization implies the maximum value $r_{k+1}\times n$ with $r_{k+1}\le 2n$ at every step.

TABLE 1

A four-level nested procedure. Matrices are evaluated only for problems with $N_{1,1}$, $N_{2,1}$, $N_{3,1}$, *and* $N_{3,2}$ *steps.*

$k=1$	$N=N_{1,1}N_{1,2}$			
$k=2$	$\mathbf{N_1=N_{1,1}}$	$N_{1,2}=N_{2,1}N_{2,2}$		
$k=3$		$\mathbf{N_1=N_{2,1}}$	$N_{2,2}=N_{3,1}N_{3,2}$	
$k=4$			$\mathbf{N_1=N_{3,1}}$	$\mathbf{N_2=N_{3,2}}$

4. An example. As an illustrative example, let us consider the following constrained LQ optimal control problem for the system:

$$A=\begin{bmatrix} 0.5 & 1 & -0.4 & 0 \\ 0.1 & 0.7 & 0 & -0.5 \\ 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0.6 \end{bmatrix},\quad B=\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix},\quad C=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},\quad D=\begin{bmatrix} 1 & 0 \\ 1 & 0.5 \end{bmatrix},$$

with initial condition $x_0=[1,\ 2,\ 3,\ 4]'$ and constrained final state and weighting matrix as

$$y_f=\begin{bmatrix} 1 \\ 1 \end{bmatrix}=\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}x(N),\quad Z=\begin{bmatrix} 1 & 0 & 2 & 1 \\ 0 & 0 & 3 & 1 \end{bmatrix}.$$

Problem 1 was first solved according to Algorithm 1 implemented in Matlab© 5.3 running on a 350 MhZ Pentium© II. The corresponding CPU time was 77.72 s. Instead, the CPU time corresponding to the implementation of Algorithm 2, with $N_1=25$ and $N_2=8$, was 0.676 s. Note that the three-level nesting procedure, with $N_{1,1}=8$, $N_{2,1}=5$, and $N_{2,2}=5$, yields a dramatic reduction of the CPU time: 0.16 s. The results for CPU times are summarized in Table 2.

TABLE 2
CPU time is compared for different nesting level algorithms.

Simple pseudoinversion	$N = 200$	CPU time $= 77.72\,s$
Two-level nesting	$N_1 = 8;\ N_2 = 25$	CPU time $= 0.676\,s$
Three-level nesting	$N_{1,1} = 8;\ N_{2,1} = 5;\ N_{2,2} = 5$	CPU time $= 0.16\,s$

It can be seen that the multilevel nesting algorithm greatly reduces the computational burden for the finite-horizon LQ optimal control problem.

The optimal cost is

$$\min_{u_{200}} J = 0.687,$$

and the final state is

$$x(200) = \begin{bmatrix} -0.4821 \\ 1.4821 \\ -0.5109 \\ 1.5109 \end{bmatrix}.$$

5. Concluding remarks. The problem considered in this paper is out of the standard scheme of LQ optimal control problems considered in the literature, not only because of the completely general approach which makes it possible to deal with cheap, singular, and regular problems with no distinction, but also because the state of the system can be sharply assigned at both the initial and final time instants. Furthermore, the proposed nesting procedure is very easily implementable thus making feasible the pseudoinversion solution of optimal control problems on a time interval of finite, but arbitrary, length.

A practical application of the algorithms described herein are the computations of convolution profiles for H_2-optimal tracking and H_2-optimal rejection of an N-step previewed signal in the nonminimum phase case. These problems were described by the authors in [14, 15], where geometric-type conditions for perfect or almost perfect tracking or rejection with stability were derived.

The possibility of using the algorithmic setting described in this paper for the solution of the standard discrete-time infinite-horizon LQR problem is under investigation.

Appendix. This appendix briefly presents the manipulations used in the proof of Theorem 1.

LEMMA 1. *The problem of finding a vector λ which minimizes the 2-norm of the vector*

$$\zeta := \Phi\,\lambda + \bar{\rho} \tag{23}$$

and satisfies the constraint

$$\Gamma\,\lambda = \bar{\sigma}, \tag{24}$$

where the matrices Φ, Γ and the column vectors $\bar{\rho}$, $\bar{\sigma}$ are given and the constraint (24) *is assumed to be feasible (i.e., satisfying $\bar{\sigma} \in \operatorname{im} \Gamma$), admits as the set of all solutions the linear variety*

$$\lambda^o(\gamma) := \left(I - K\,(\Phi K)^{\#}\,\Phi\right)\Gamma^{\#}\bar{\sigma} - K\,(\Phi K)^{\#}\,\bar{\rho} + P\,\gamma, \tag{25}$$

where K is a basis matrix for $\ker \Gamma$*, P is a basis matrix for* $\operatorname{im} K \cap \ker \Phi$*, and* γ *is a free vector parameterizing the solution* λ *in the column space of* P*.*

The corresponding ζ *vector with minimum* 2*-norm is given by*

$$\zeta^o := \left(I - \Phi K \left(\Phi K\right)^{\#}\right) \left(\Phi \Gamma^{\#} \bar{\sigma} + \bar{\rho}\right). \tag{26}$$

Proof. The constraint (24) can be solved with respect to λ as

$$\lambda = \Gamma^{\#} \bar{\sigma} + K \nu, \tag{27}$$

where ν parameterizes the solution in $\ker \Gamma$ whose basis matrix is K. Then the following expression for vector ζ can be obtained by substituting (27) into (23):

$$\zeta = \Omega \nu + \bar{\eta}, \tag{28}$$

where

$$\Omega := \Phi K, \qquad \bar{\eta} := \Phi \Gamma^{\#} \bar{\sigma} + \bar{\rho}. \tag{29}$$

The expression for ν minimizing the 2-norm of ζ is simply obtained as

$$\nu = -\Omega^{\#} \bar{\eta} + H \gamma, \tag{30}$$

where vector γ parameterizes the solution in $\ker \Omega$ whose basis matrix is H. The expression

$$\zeta^o = \left(I - \Omega \Omega^{\#}\right) \bar{\eta} \tag{31}$$

of the minimum norm vector ζ can then be obtained by substituting (30) into (28). Expression (26) for ζ^o is derived from (31) and (29). By substituting (30) into (27), it follows that

$$\lambda = \Gamma^{\#} \bar{\sigma} - K \Omega^{\#} \bar{\eta} + K H \gamma. \tag{32}$$

Moreover, since

$$\operatorname{im} H := \ker \left(\Phi K\right) = \ker K + \operatorname{im} Y,$$

with Y such that

$$\operatorname{im} \left(K Y\right) = \operatorname{im} K \cap \ker \Phi,$$

(32) can also be written as

$$\lambda = \Gamma^{\#} \bar{\sigma} - K \Omega^{\#} \bar{\eta} + P \gamma, \tag{33}$$

where P has been defined in the statement of the lemma. Finally, by substituting (29) into (33), linear variety (25) is obtained. □

Remark 5. *The minimum* 2*-norm vector* λ *minimizing the cost function* (23) *is*

$$\lambda^o = \left(I - P P^{\#}\right) \left(\left(I - K \left(\Phi K\right)^{\#} \Phi\right) \Gamma^{\#} \bar{\sigma} - K \left(\Phi K\right)^{\#} \bar{\rho}\right),$$

obtained by choosing

$$\gamma = -P^{\#} \left(\left(I - K \left(\Phi K\right)^{\#} \Phi\right) \Gamma^{\#} \bar{\sigma} - K \left(\Phi K\right)^{\#} \bar{\rho}\right);$$

see (25)*. If the minimum norm is not required, a plainer expression for* λ *follows by assuming that* $\gamma = 0$*.*

REFERENCES

[1] B. Anderson and J. Moore, *Optimal Control: Linear Quadratic Methods*, Prentice–Hall International, London, 1989.

[2] W. Arnold and A. Laub, *Generalized eigenproblem algorithms and software for algebraic Riccati equations*, Proc. IEEE, 72 (1984), pp. 1746–1754.

[3] A. Bryson and Y. Ho, *Applied Optimal Control*, Blaisdell Publishing Company, Waltham, MA, 1969.

[4] B. Chen, *H_∞ Control and Its Applications*, Lecture Notes in Control and Inform. Sci. 235, Springer-Verlag, New York, 1999.

[5] T. Chen and B. Francis, *Optimal Sampled-Data Control Systems*, Springer-Verlag, London, 1995.

[6] T. Geerts, *The algebraic Riccati equation and singular optimal control: The discrete-time case*, in Proceedings of the International Symposium on Systems and Networks: Mathematical Theory and Applications (MTNS '93), Vol. 2, Regensburg, Germany, 1993, pp. 129–134.

[7] G. Golub and C. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[8] M. Hautus and L. Silverman, *System structure and singular control*, Linear Algebra Appl., 50 (1983), pp. 369–402.

[9] C. Ionescu, V. Oară, and M. Weiss, *General matrix pencil techniques for the solution of algebraic Riccati equation: A unified approach*, IEEE Trans. Automat. Control, 42 (1997), pp. 1085–1097.

[10] V. Ionescu and C. Oară, *Generalized continuous-time Riccati theory*, Linear Algebra Appl., 232 (1996), pp. 111–130.

[11] V. Ionescu and M. Weiss, *On computing the stabilizing solution of the discrete-time Riccati equation*, Linear Algebra Appl., 174 (1992), pp. 229–238.

[12] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*, John Wiley & Sons, New York, 1972.

[13] F. Lewis and V. Syrmos, *Optimal Control*, John Wiley & Sons, New York, 1995.

[14] G. Marro, D. Prattichizzo, and E. Zattoni, *Convolution profiles for right-inversion of multivariable non-minimum phase discrete-time systems*, Automatica J. IFAC, 38 (2002), pp. 1695–1703.

[15] G. Marro, D. Prattichizzo, and E. Zattoni, *A unified algorithmic setting for signal-decoupling compensators and unknown input observers*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia, 2000.

[16] A. C. M. Ran and H. L. Trentelman, *Linear quadratic problems with indefinite cost for discrete time systems*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 776–797.

[17] A. Saberi and P. Sannuti, *Cheap and singular controls for linear quadratic regulators*, IEEE Trans. Automat. Control, 32 (1987), pp. 208–219.

[18] A. Saberi, P. Sannuti, and B. Chen, *H_2 Optimal Control*, System and Control Engineering, Prentice–Hall International, London, 1995.

[19] L. Silverman, *Inversion of multivariable linear systems*, IEEE Trans. Automat. Control, 14 (1969), pp. 270–276.

[20] L. Silverman, *Discrete Riccati equations: Alternative algorithms, asymptotic properties, and system theory interpretations*, in Control and Dynamic Systems, C. Leondes, ed., Academic Press, New York, 1976, pp. 313–386.

[21] J. Soethoudt and H. Trentelman, *The regular indefinite linear-quadratic problem with linear endpoint constraints*, Systems Control Lett., 12 (1989), pp. 23–31.

[22] A. Stoorvogel and A. Saberi, *The discrete-time algebraic Riccati equation and linear matrix inequality*, Linear Algebra Appl., 274 (1998), pp. 317–365.

[23] P. Van Dooren, *A generalized eigenvalue approach for solving Riccati equations*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 121–135.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 1013–1032

NULL CONTROLLABILITY WITH VANISHING ENERGY*

ENRICO PRIOLA† AND JERZY ZABCZYK‡

Abstract. Linear, null controllable systems, for which an arbitrary initial state can be transferred to the origin with arbitrarily small energy, are characterized. Theorems are stated in terms of an associated algebraic Riccati equation and in terms of the spectrum of the linear part of the system. The results so obtained allow us to determine Ornstein–Uhlenbeck operators for which the Liouville theorem about bounded harmonic functions holds.

Key words. null controllability, minimal energy, algebraic Riccati equation

AMS subject classifications. 93B05, 93C05, 60H15

DOI. 10.1137/S0363012902409970

1. Introduction.

In this paper we are concerned with the following infinite dimensional linear control system:

$$\frac{dy}{dt} = Ay(t) + Bu(t), \quad y(0) = x \in H, \tag{1.1}$$

$t \geq 0$, where H denotes a *complex* Hilbert space, with inner product $\langle \cdot, \cdot \rangle$ and norm $|\cdot|$, and A is a possibly unbounded linear operator on H, which generates a $\mathcal{C}_0$-semigroup e^{tA} on H. Moreover, B is a bounded linear operator from another Hilbert space U into H, i.e., $B \in \mathcal{L}(U, H)$, and $u : [0, +\infty) \to U$ is a locally square integrable function, which represents the control on the system. For any control function u, there exists a weak solution $y^{x,u}(t)$ of (1.1), which is given by

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

see, for instance, [10]. The system (1.1) is called *null controllable* if it is null controllable at some universal time $\tilde{T} > 0$. This means that for any state $a \in H$, we can find a control u such that $y^{a,u}(\tilde{T}) = 0$ (the control u transfers a to 0 at time $\tilde{T}$); see [28], [1], [2], [5], [17], [18], [25]. Throughout the paper we will assume that the system (1.1) is null controllable. Taking into account that 0 is an equilibrium point for our system, the following definition has a clear control-theoretic meaning. We say that the system (1.1) is *null controllable with vanishing energy*, *an NCVE system* for short, if it is null controllable and for any $x \in H$ there exists a sequence of controls (u_n) and of times (T_n) such that $y^{x,u_n}(T_n) = 0$ for any $n \in \mathbf{N}$, and

$$\lim_{n\to\infty} \int_0^{T_n} |u_n(s)|^2 ds = 0. \tag{1.2}$$

*Received by the editors June 20, 2002; accepted for publication (in revised form) January 29, 2003; published electronically July 8, 2003.

http://www.siam.org/journals/sicon/42-3/40997.html

†Dipartimento di Matematica, Università di Torino, via Carlo Alberto 10, 10123, Torino, Italy (priola@dm.unito.it). This author was partially supported by the Italian national project MURST "Analisi e controllo di equazioni di evoluzione deterministiche e stocastiche."

‡Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland (zabczyk@impan.gov.pl). Current address: Interdisciplinary Centre for Mathematical & Computational Modelling, Warsaw University, Warsaw, Poland. This author was supported by the KBN grant "Rownania paraboliczne w przestrzeniach Hilberta" 2 P03A 037 16.

The purpose of this paper is to characterize NCVE systems. This characterization will be given in terms of spectral properties of A as well as in terms of *maximal solutions* to an associated algebraic Riccati equation. To the best of our knowledge, the concept of an NCVE system is formulated for the first time in this paper. However, important examples of NCVE systems have been already investigated; see, for instance, [14] and references therein.

Let us recall that the *spectral bound* $s(A)$ of the operator A is given by the formula $s(A) = \sup\{\mathrm{Re}(\lambda) : \lambda \in \sigma(A)\}$, where $\sigma(A)$ denotes the spectrum of A. (If $\sigma(A) = \emptyset$, then $\mathrm{s}(A) = -\infty$.) A vector $x \neq 0$ is called a *generalized eigenvector* of the operator A if there exists $\lambda \in \mathbf{C}$ and a natural number k such that $(\lambda - A)^k x = 0$. Note that necessarily $x \in D(A^k)$ and if $k = 1$, then x is an eigenvector of A. A $\mathcal{C}_0$-semigroup e^{tA} is said to be *exponentially stable* if there exist $M > 0$, $\omega < 0$ such that for all $t \geq 0$, $\|e^{tA}\|_{\mathcal{L}(H)} \leq Me^{\omega t}$. To formulate our first main result we need the following assumptions.

Hypothesis 1.1.

(i) There exists a sequence $(\lambda_n) \subset \sigma(A)$ such that each λ_n is isolated in $\sigma(A)$ and $\lim_{n\to\infty} \mathrm{Re}(\lambda_n) = s(A)$.

(ii) There exist closed linear subspaces H_s and H_u invariant for e^{tA}, $t \geq 0$, such that their direct sum is H, i.e., $H = H_s \oplus H_u$ and, moreover,
   (a) the semigroup e^{tA} restricted to H_s is exponentially stable on H_s,
   (b) the set of all generalized eigenvectors of A contained in H_u is linearly dense in H_u.

THEOREM 1.1. *Let the system* (1.1) *be null controllable and assume that Hypothesis* 1.1 *holds. Then the following statements are equivalent:*

(a) *the system* (1.1) *is NCVE;*

(b) $s(A) = \sup\{\mathrm{Re}(\lambda) : \lambda \in \sigma(A)\} \leq 0$.

What seems surprising, even in the finite dimensional case, is that the solutions of the uncontrolled linear system may have a polynomial growth, but nevertheless any state can be transferred to 0 with arbitrarily small energy. As we shall see in section 5.2, the index $s(A)$ cannot be replaced by $\omega(A)$, the growth bound of the semigroup e^{tA}, $\omega(A) = \inf\{\omega \in \mathbf{R} : \|e^{tA}\|_{\mathcal{L}(H)} \leq Me^{\omega t}$ for some M and all $t \geq 0\}$. We will comment on Hypothesis 1.1 in section 5.3.

Let us consider a smaller class of the so-called exactly controllable systems. The system (1.1) is called *exactly controllable* if it is exactly controllable at some $\tilde{T} > 0$, i.e., for any $a, b \in H$, there exists a control $\hat{u} = u_{a,b}$ such that $y^{a,\hat{u}}(\tilde{T}) = b$. We say that (1.1) is *exactly controllable with vanishing energy, an ECVE system* for short, if it is exactly controllable and for any a, $b \in H$, there exists a sequence of controls (u_n) and of times (T_n) such that $y^{a,u_n}(T_n) = b$, $n \in \mathbf{N}$, and (1.2) holds. An analogue of Theorem 1.1 for exactly controllable systems is stated below.

THEOREM 1.2. *Assume the system* (1.1) *is exactly controllable, the operator A generates a $\mathcal{C}_0$-group on H, and Hypothesis* 1.1 *holds. Then the following statements are equivalent:*

(a) *the system* (1.1) *is ECVE;*

(b) $\sigma(A) \subset \{i\lambda : \lambda \in \mathbf{R}\}$.

Applications of these theorems to delay systems and hyperbolic and parabolic equations are given in section 6. In order to prove Theorems 1.1 and 1.2, we combine arguments from both control theory and spectral theory. Moreover, we will apply our third main result, which requires only the null controllability of (1.1).

THEOREM 1.3. *A null controllable system* (1.1) *is NCVE if and only if the algebraic Riccati equation*

$$PA + A^*P - PBB^*P = 0, \quad P \geq 0, \tag{1.3}$$

has a unique bounded nonnegative solution $P = 0$.

This theorem will be a consequence of a stronger result concerning a more general algebraic Riccati equation:

$$PA + A^*P - PBB^*P + R = 0, \tag{1.4}$$

where R is a symmetric, bounded, nonnegative operator on H. By a *solution* P of this equation we mean a bounded nonnegative operator P such that for any $x, y \in D(A)$,

$$\langle Px, Ay\rangle + \langle PAx, y\rangle - \langle PBB^*Px, y\rangle + \langle Rx, y\rangle = 0. \tag{1.5}$$

A solution $\hat{P}$ of (1.5) is called *maximal* if for any solution S of (1.5), we have $S \leq \hat{P}$, i.e., $\langle Sx, x\rangle \leq \langle \hat{P}x, x\rangle$, $x \in H$; see [3], [16], [2]. In the literature results close to the next theorem are available; see, for instance, [2, p. 283], [17], [18], and [9]. However, since we could not find the exact statement we needed, we decided to include it in the present paper.

THEOREM 1.4. *If the system* (1.1) *is null controllable, then for an arbitrary linear, bounded operator* $R \geq 0$ *there exists a maximal solution* $\hat{P} \geq 0$ *of the equation* (1.4)*, and it is given by the formula*

$$\langle \hat{P}x, x\rangle = \inf_{t \geq \tilde{T}} \inf_{u \in L^2(0,t;U),\ y^{x,u}(t)=0} \left\{ \int_0^t (\langle Ry^{x,u}(s), y^{x,u}(s)\rangle + |u(s)|^2) ds \right\}. \tag{1.6}$$

We conjecture that results similar to Theorems 1.1 and 1.2 could be true even for control systems in Banach spaces; however, the proofs should be different from ours. Indeed, a satisfactory theory of the linear quadratic problem involving the differential Riccati equation is not available in general Banach spaces, even in the reflexive case. It would be also interesting to extend our characterizations to systems with unbounded control operators B, in particular to boundary control systems described by linear PDEs; see [26], [1], [2], [17], [18]. This extension will be a subject of our future research.

There are close links between the results of the present paper and the theory of bounded harmonic functions for the so-called Ornstein–Uhlenbeck operator. Consider an H-valued Ornstein–Uhlenbeck process X satisfying the following equation:

$$dX(t) = AX(t)dt + BdW(t), \quad X(0) = x \in H,\ t \geq 0,$$

where W is a U-valued Wiener process. The generator $\mathcal{L}$ of the process X is called an Ornstein–Uhlenbeck operator. A function h such that $\mathcal{L}h = 0$ is called harmonic for $\mathcal{L}$. The problem of existence of bounded harmonic functions for differential operators is called the Liouville problem and is intensively studied; see, e.g., [20] and [23]. Roughly speaking, in our special situation, there exist nonconstant, bounded, harmonic functions for $\mathcal{L}$ if and only if the associated control system is not NCVE. We study this topic in a separate paper [21]. We also mention that a preliminary expanded version of our results is given in [22], which contains additional details.

Section 2 is devoted to preliminaries from spectral theory and linear control theory. Section 3 is on the proof of Theorems 1.4 and 1.3. In section 4 we will give a proof of Theorems 1.1 and 1.2. Comments and possible extensions are discussed in section 5. In section 6 we collect some examples which illustrate our results.

2. Preliminaries.

2.1. Basic concepts from spectral theory. Let us introduce some basic facts from spectral theory. Let $A : D(A) \subset H \to H$ be a closed operator with dense domain and denote by $\sigma(A)$ its spectrum. An element $\mu \in \sigma(A)$ is called *isolated* if there exists a neighborhood U of μ such that $U \cap \sigma(A) = \{\mu\}$.

Let $\lambda \in \mathbf{C}$ be an eigenvalue of A. The dimension of $\mathrm{Ker}(\lambda - A)$ is called the geometric multiplicity of λ. Remark that $\mathrm{Ker}(\lambda - A)^k \subset \mathrm{Ker}(\lambda - A)^{k+1}$, $k \in \mathbf{N}$. Hence we can say that v is a generalized eigenvector of A if

$$v \in \bigcup_{k \geq 1} \mathrm{Ker}(\lambda - A)^k.$$

If there exists p such that $\mathrm{Ker}(\lambda - A)^p = \mathrm{Ker}(\lambda - A)^{p+1}$ and $\dim(\mathrm{Ker}(\lambda - A)^p) = m < \infty$, then λ is named an eigenvalue with finite algebraic multiplicity m. (The integer p is named the index of λ.) We refer to [11] for more details on this subject.

Let $A : D(A) \subset H \to H$ be the generator of a $\mathcal{C}_0$-semigroup e^{tA}. A closed subspace $K \subset H$ is called *invariant* for e^{tA} if $e^{tA}(K) \subset K$, $t \geq 0$. In this case we have that $A(D(A) \cap K) \subset K$ and the restriction of e^{tA} to K is still a $\mathcal{C}_0$-semigroup on K with generator $A^K : (D(A) \cap K) \subset K \to K$, $A^K v = Av$, $v \in D(A) \cap K$.

In what follows we need the following theorem which provides the existence of useful invariant subspaces for e^{tA}; see Lemma 2.5.7 in [5] and [10, p. 245].

Assume that $\sigma(A) = \sigma_0 \cup \sigma_1$, where σ_0 and σ_1 are two disjoint closed subsets of $\mathbf{C}$ and, in addition, σ_0 is bounded. Since the distance between σ_0 and σ_1 is positive, there exists a bounded open set Ω containing σ_0 such that its closure is disjoint from σ_1. We may assume that the boundary γ of Ω consists of a finite number of rectifiable, closed simple Jordan paths oriented in the usual positive direction. Let us introduce the spectral Riesz projection P_0:

$$P_0 x = \frac{1}{2\pi i} \int_\gamma (w - A)^{-1} x \, dw, \quad x \in H. \tag{2.1}$$

Define $P_0 H = E_0$. (Note that E_0 can be equal to H.) We get the *spectral decomposition*

$$H = E_0 \oplus E_1, \quad E_1 = P_1 H, \ \text{ where } \ (I - P_0) = P_1. \tag{2.2}$$

The closed subspaces E_0 and E_1 are invariant for e^{tA} and, moreover, $E_0 \subset D(A)$. The restrictions A_i of A to E_i, $i = 0, 1$, satisfy $\sigma(A_i) = \sigma_i$. We have

$$A_0 : E_0 \to E_0, \quad A_1 : (D(A) \cap E_1) \subset E_1 \to E_1. \tag{2.3}$$

The operator A_0 generates a uniformly continuous group e^{tA_0} on E_0, and A_1 generates a $\mathcal{C}_0$-semigroup e^{tA_1} on E_1. The restrictions of e^{tA} to E_0 and E_1 coincide with e^{tA_0} and e^{tA_1}, respectively.

If H, K are two Hilbert spaces, we denote by $\mathcal{L}(H, K)$ the Banach space of all bounded linear operators from H into K endowed with the uniform norm $\|\cdot\|_{\mathcal{L}(H,K)}$. In particular, when $H = K$, we set $\mathcal{L}(H, H) = \mathcal{L}(H)$ and $\|\cdot\|_{\mathcal{L}(H)} = \|\cdot\|$. Let us consider the cone $\mathcal{K}_+(H)$ of $\mathcal{L}(H)$ consisting of all symmetric nonnegative bounded linear operators on H. On $\mathcal{K}_+(H)$, we will have the following order:

$$T \leq S \quad \Leftrightarrow \quad \langle Tx, x\rangle \leq \langle Sx, x\rangle, \quad x \in H, \ T, S \in \mathcal{K}_+(H).$$

Note that $\langle Tx, x\rangle = \langle Sx, x\rangle$, $x \in H$, implies that $T = S$. We will use the following result: if a family $(P_t) \subset \mathcal{K}_+(H)$, $t > 0$, is decreasing as t tends to 0^+, then there exists $P \in \mathcal{K}_+(H)$ such that P_t converges strongly to P (see [5, p. 606]), i.e.,

$$\lim_{t\to 0^+} P_t x = Px, \quad x \in H, \quad \text{or equivalently} \quad \lim_{t\to 0^+} \langle P_t x, x\rangle = \langle Px, x\rangle, \quad x \in H. \tag{2.4}$$

2.2. Control-theoretic concepts. Let us collect here some preliminaries from control theory which will be used throughout the paper; for more details, see [28] or [5]. The system (1.1) will often be denoted by the pair (A, B).

Set $Q = BB^*$ and introduce the *controllability operator* Q_t,

$$Q_t x = \int_0^t e^{sA}\, BB^*\, e^{sA^*} x\, ds \;=\; \int_0^t e^{sA} Q e^{sA^*} x\, ds, \quad t > 0,\ x \in H, \tag{2.5}$$

where B^* denotes the adjoint operator of B and the integral is in the Bochner sense. We remark that the null controllability of (1.1) at time $\tilde{T} > 0$ implies the null controllability at any $t \geq \tilde{T}$. It is known that the null controllability of (1.1) at $\tilde{T}$ is equivalent to the following condition:

$$e^{\tilde{T}A}(H) \subset Q_{\tilde{T}}^{1/2}(H). \tag{2.6}$$

By the closed graph theorem the operator $\Gamma_s = Q_s^{-1/2} e^{sA}$ is a well-defined bounded linear operator on H for any $s \geq \tilde{T}$. Fix $a \in H$; among the controls u steering a to 0 at time $t \geq \tilde{T}$, there exists a unique optimal control $\hat{u}$ which minimizes the energy functional $u \mapsto \int_0^t |u(s)|^2 ds$, and, moreover,

$$\int_0^t |\hat{u}(s)|^2 ds = |\Gamma_t a|^2 = \langle \Gamma_t^* \Gamma_t a, a\rangle, \quad \text{where } \Gamma_t a = Q_t^{-1/2} e^{tA} a. \tag{2.7}$$

Thus $|Q_t^{-1/2} e^{tA} a|^2$ is the minimal energy to attain 0 from a at time t, and Γ_t is sometimes called the *minimal energy operator*.

Since the map: $t \mapsto |\Gamma_t x|^2$ is decreasing, we get that the system (1.1) is NCVE *if and only if*

$$\lim_{t\to+\infty} |\Gamma_t x|^2 = 0, \quad x \in H.$$

Now let us deal with *exactly controllable systems*. It is known that the system (A, B) is *exactly controllable* at $t > 0$ if and only if Q_t is one to one and onto on H. When (A, B) is exactly controllable at $t > 0$, one can define the minimal energy $E_t(a, b)$, which is needed to transfer the state a into b at time $t > 0$,

$$E_t(a,b) = E_t^{(A,B)}(a,b) = \inf_{u\in L^2(0,t;U)} \left\{ \int_0^t |u(s)|^2 ds \,:\, y^{a,u}(t) = b \right\}, \quad a,\ b \in H, \tag{2.8}$$

where $y^{a,u}$ denotes the solution to (1.1) such that $y^{a,u}(0) = a$. Note that $|\Gamma_t a|^2 = E_t(a, 0)$, $a \in H$; see (2.7). It is known (see, e.g., [28]) that $E_t(a, b) = |Q_t^{-1/2}(e^{tA}a - b)|^2$ and that there exists an optimal control $\hat{u}$ such that $E_t(a, b) = \int_0^t |\hat{u}(s)|^2 ds$. It is clear that (1.1) is an ECVE system *if and only if* $\lim_{t\to+\infty} E_t(a, b) = 0$ for any a, $b \in H$.

We remark that if (A, B) is null controllable at $t > 0$ and, in addition, A generates a $\mathcal{C}_0$-group on H, then (A, B) is exactly controllable at $t > 0$. Moreover, if $H = \mathbf{R}^n$,

$U = \mathbf{R}^m$, and (A, B) is null controllable at some $t_0 > 0$, then (A, B) is exactly controllable at any $t > 0$.

Let us consider the *Lyapunov equation*:

$$\langle Ax, Ky\rangle + \langle Kx, Ay\rangle = -\langle Sx, y\rangle, \quad x,\, y \in D(A), \tag{2.9}$$

where $S \in \mathcal{K}_+(H)$ and A generates an exponentially stable semigroup e^{tA} on H, i.e., $\omega(A) < 0$. Then there exists a unique solution $K \in \mathcal{K}_+(H)$ to (2.9) which is given by

$$Kx = \int_0^\infty e^{sA} S e^{sA^*} x ds, \;\; x \in H.$$

Finally recall that a solution $P \in \mathcal{K}_+(H)$ to the *algebraic Riccati equation*

$$\langle Px, Ay\rangle + \langle PAx, y\rangle - \langle PBB^*Px, y\rangle + \langle Rx, y\rangle = 0, \;\; x,\, y \in D(A),$$

has the properties that $P(D(A)) \subset D(A^*)$ and, in addition, that $PAx + A^*Px - PBB^*Px + Rx = 0$ for any $x \in D(A)$.

3. Proofs of Theorems 1.4 and 1.3. Throughout this section we assume only that the system (1.1) is *null controllable at some* $\tilde{T} > 0$. Let $R,\, S \in \mathcal{K}_+(H)$, $Q = BB^*$. We introduce the infinite dimensional *differential Riccati equation*:

$$\frac{d}{dt}P_t = A^*P_t + P_tA + R - P_tQP_t, \;\; t > 0, \quad P_0 = S. \tag{3.1}$$

We say that a map $t \mapsto P_t \in \mathcal{L}(H)$, $t \geq 0$, such that $P_0 = S$ is a (global) solution of (3.1) if for any $h, k \in D(A)$ the real map $t \to \langle P_t h, k\rangle$ is absolutely continuous on $[0, +\infty)$ and it satisfies

$$\frac{d}{dt}\langle P_t h, k\rangle = \langle P_t h, Ak\rangle + \langle P_t Ah, k\rangle + \langle Rh, k\rangle - \langle P_t Q P_t h, k\rangle \tag{3.2}$$

for almost all $t > 0$. It is well known (see, e.g., [28]) that for any $S \in \mathcal{K}_+(H)$ there exists a unique solution $P_t = P_t^S \in \mathcal{L}(H)$ to (3.1). Moreover, $P_t^S \in \mathcal{K}_+(H)$ and

$$\langle P_t^S x, x\rangle = \inf_{u \in L^2(0,t;U)} \left\{ \int_0^t (|u(s)|^2 + \langle Ry^{x,u}(s), y^{x,u}(s)\rangle) ds + \langle Sy^{x,u}(t), y^{x,u}(t)\rangle \right\}, \tag{3.3}$$

$t \geq 0$, where $y^{x,u}$ is the solution of (1.1), $y^{x,u}(0) = x$.

Let us introduce the quantity $\Gamma_t^R(x)$ generalizing $|\Gamma_t x|^2$ as

$$\begin{aligned} \Gamma_t^R(x) &= \inf_{u \in L^2(0,t;U):\; y^{x,u}(t) = 0} J_t(u), \quad \text{where} \\ J_t(u) &= J_t^x(u) = \int_0^t (\langle Ry^{x,u}(s), y^{x,u}(s)\rangle + |u(s)|^2) ds, \quad t \geq \tilde{T}. \end{aligned} \tag{3.4}$$

Of course, choosing $R = 0$, we get $\Gamma_t = \Gamma_t^0$. We remark that the map $t \mapsto \Gamma_t^R(x)$ is decreasing on $[\tilde{T}, +\infty)$ for any $x \in H$.

In order to prove Theorem 1.4 we proceed as follows. First we show that $\Gamma_t^R(\cdot)$ is a quadratic functional on H. Then we prove that the map $t \mapsto (\Gamma_t^R)^*\Gamma_t^R \in \mathcal{L}(H)$ satisfies the differential Riccati equation (3.2). Finally we obtain that $(\Gamma_t^R)^*\Gamma_t^R$ *converges strongly* to the *maximal* solution $\hat{P}$ of the algebraic Riccati equation (1.4) as $t \to +\infty$.

We start to establish some properties of Γ_t^R (see also [4, Theorem 1.2]), where a related result is proved.

LEMMA 3.1. *For any $t \geq \tilde{T}$, the following statements hold:*

(i) *there exists a unique control $\hat{u}$ minimizing the functional J_t in* (3.4)*;*

(ii) *the map Γ_t^R is a quadratic functional of x; i.e., we have $\Gamma_t^R(x) = \langle \Gamma_t^R x, x\rangle$, where $\Gamma_t^R \in \mathcal{K}_+(H)$.*

Proof. (i) Inserting the explicit formula for $y^{x,u}$ into the functional J_t and using the linear operator $L_t : L^2(0,t;U) \to L^2(0,t;H)$, $L_t u(s) = \int_0^s e^{(s-r)A} Bu(r)dr$, $s \in (0,t)$, $u \in L^2(0,t;U)$, we get

$$J_t(u) = a + \langle v, u\rangle_{L^2(0,t;U)} + \langle Wu, u\rangle_{L^2(0,t;U)}, \tag{3.5}$$

where $a = \int_0^t \langle R(e^{sA}x), e^{sA}x\rangle ds$, $v = 2L_t^* R\,(e^{(\cdot)\,A}x)$, $Wu = u + L_t^* R(L_t u(\cdot))$. Noting that W is a coercive symmetric operator on $L^2(0,t;U)$ we obtain that J_t is a continuous, strictly convex, and coercive functional on $L^2(0,t;U)$. Hence the infimum of J_t over the nonempty affine set of all $u \in L^2(0,t;U)$ such that

$$\int_0^t e^{(t-r)A} Bu(r)dr = -e^{tA}x$$

is attained at exactly one point $\hat{u}$. This proves (i).

(ii) Let us define the cost functionals $J_t^\epsilon : L^2(0,t;U) \to \mathbf{R}$:

$$J_t^\epsilon(u) = \int_0^t (\langle Ry^{x,u}(s), y^{x,u}(s)\rangle + |u(s)|^2)ds + \frac{1}{\epsilon}\langle y^{x,u}(t), y^{x,u}(t)\rangle, \quad \epsilon > 0,$$

$u \in L^2(0,t;U)$. We know that the infimum of J_t^ϵ over all controls u is $\langle P_t^\epsilon x, x\rangle$, where P_t^ϵ verifies

$$\frac{d}{dt}P_t^\epsilon = A^* P_t^\epsilon + P_t^\epsilon A + R - P_t^\epsilon Q P_t^\epsilon, \quad t > 0, \quad P_0^\epsilon = \frac{1}{\epsilon} I.$$

By the definition of Γ_t^R it is clear that $P_t^\epsilon \leq \Gamma_t^R$ for any $\epsilon > 0$. Note that P_t^ϵ is increasing as ϵ tends to 0. Hence $\Gamma_t^R - P_t^\epsilon$ is decreasing in $\mathcal{K}_+(H)$. It follows that P_t^ϵ converges strongly to some $P_t \in \mathcal{K}_+(H)$ as $\epsilon \to 0^+$, i.e.,

$$\lim_{\epsilon \to 0^+} \langle P_t^\epsilon x, x\rangle = \langle P_t x, x\rangle, \quad x \in H,\ t > 0.$$

Our next goal will be to verify that $\Gamma_t^R = P_t$. This will give the assertion (ii). For this purpose, we show that

$$\lim_{\epsilon \to 0^+} \langle P_t^\epsilon x, x\rangle = \lim_{\epsilon \to 0^+} J_t^\epsilon(u_\epsilon) = J_t(\hat{u}), \quad t \geq \tilde{T}, \tag{3.6}$$

where u_ϵ are the controls which minimize J_t^ϵ and $\hat{u}$ is the control which minimizes J_t among all controls steering x to 0 in time t. We will also obtain that u_ϵ converges weakly to $\hat{u}$ as ϵ tends to 0.

First recall that, by (3.5), we have

$$\begin{aligned} J_t^\epsilon(u) &= a + \langle v, u\rangle_{L^2(0,t;U)} + \langle Wu, u\rangle_{L^2(0,t;U)} + \frac{1}{\epsilon}|e^{tA}x + \mathcal{L}_t u|^2 \\ &= J_t(u) + \frac{1}{\epsilon}|e^{tA}x + \mathcal{L}_t u|^2, \quad u \in L^2(0,t;U), \end{aligned}$$

where $\mathcal{L}_t : L^2(0,t;U) \to H$, $\mathcal{L}_t u = \int_0^t e^{(t-s)A} B u(s) ds$. It follows that

$$J_t^\epsilon(u_\epsilon) = J_t(u_\epsilon) + \frac{1}{\epsilon}|e^{tA}x + \mathcal{L}_t u_\epsilon|^2 \le J_t^\epsilon(\hat{u}) = J_t(\hat{u}), \quad \epsilon > 0.$$

Since W is coercive, we deduce that the set (u_ϵ), $\epsilon > 0$, is uniformly bounded in $L^2(0,t;U)$. Thus it is possible to extract from (u_ϵ) a sequence, still denoted by u_ϵ, such that u_ϵ converges weakly to $u_0 \in L^2(0,t;U)$. We prove that u_0 transfers x to 0 at time t, i.e.,

$$e^{tA}x + \mathcal{L}_t u_0 = 0. \tag{3.7}$$

Computing the directional derivative of J_t^ϵ in u_ϵ along any direction $v \in L^2(0,t;U)$, we get

$$0 = DJ_t^\epsilon(u_\epsilon, v) = \langle v, u_\epsilon \rangle_{L^2(0,t;U)} + 2\langle W u_\epsilon, v \rangle_{L^2(0,t;U)} + \frac{2}{\epsilon}\langle e^{tA}x + \mathcal{L}_t u_\epsilon, \mathcal{L}_t v \rangle, \ \epsilon > 0.$$

Since u_ϵ converges weakly to u_0, it is clear that the above identities can be true only if $\langle e^{tA}x + \mathcal{L}_t u_\epsilon, \mathcal{L}_t v \rangle$ converges to 0 as $\epsilon \to 0^+$. We have

$$\langle e^{tA}x + \mathcal{L}_t u_0, \mathcal{L}_t v \rangle = 0, \quad v \in L^2(0,t;U). \tag{3.8}$$

Now we can choose a control v_1 such that $\mathcal{L}_t v_1 = e^{tA}x$ because the system (1.1) is null controllable at time t. Taking the control $v = u_0 + v_1$ in (3.8), we obtain $|e^{tA}x + \mathcal{L}_t u_0|^2 = 0$ and so (3.7) is verified.

Using that J_t is weakly lower semicontinuous and $J_t(\hat{u}) \ge J_t^\epsilon(u_\epsilon)$, we get

$$J_t(\hat{u}) \ge \liminf_{\epsilon \to 0} \left(J_t(u_\epsilon) + \frac{1}{\epsilon}|e^{tA}x + \mathcal{L}_t u_\epsilon|^2 \right) \ge \liminf_{\epsilon \to 0} J_t(u_\epsilon) \ge J_t(u_0). \tag{3.9}$$

Since $y^{x,\hat{u}}(t) = y^{x,u_0}(t) = 0$ and $\hat{u}$ was the minimum point for J_t, we get that $\hat{u} = u_0$. Moreover, by (3.9) we achieve $\liminf_{\epsilon \to 0} J_t^\epsilon(u_\epsilon) = J_t(u_0)$. Using that $J_t^\epsilon(u_\epsilon)$ is increasing we finally get

$$\lim_{\epsilon \to 0^+} J_t^\epsilon(u_\epsilon) = J_t(\hat{u}). \qquad \square$$

LEMMA 3.2. *The following assertions hold:*

(i) *The map* $t \mapsto P_t = (\Gamma_t^R)^* \Gamma_t^R \in \mathcal{L}(H)$ *solves*

$$\frac{d}{dt}P_t = A^* P_t + P_t A + R - P_t Q P_t \ \text{ for any } \ t \ge \tilde{T}. \tag{3.10}$$

(ii) *For any* $S \in \mathcal{K}_+(H)$, *denoting by* P_t^S, $t \ge 0$, *the unique solution of* (3.1), *one has* $P_t^S \le (\Gamma_t^R)^* \Gamma_t^R$, $t \ge \tilde{T}$.

Proof. For simplicity of notation, we will write $\hat{\Gamma}_t$ instead of Γ_t^R.

(i) Fix any $t \ge \tilde{T}$. We prove the assertion by showing that

$$P_s^{\hat{\Gamma}_t^* \hat{\Gamma}_t} = \hat{\Gamma}_{t+s}^* \hat{\Gamma}_{t+s}, \quad s \ge 0. \tag{3.11}$$

Fix $x \in H$ and let $\hat{u}$ be the control which transfers x to 0 in time $t+s$ with the minimal energy. Then we have

$$\begin{aligned}
\langle \hat{\Gamma}_{t+s}x, \hat{\Gamma}_{t+s}x \rangle &= \int_0^{t+s} (\langle R y^{x,\hat{u}}(r), y^{x,\hat{u}}(r) \rangle + |\hat{u}(r)|^2) dr \\
&= \int_0^s (\langle R y^{x,\hat{u}}(r), y^{x,\hat{u}}(r) \rangle + |\hat{u}(r)|^2) dr + \int_s^{s+t} (\langle R y^{x,\hat{u}}(r), y^{x,\hat{u}}(r) \rangle + |\hat{u}(r)|^2) dr \\
&\ge \int_0^s (\langle R y^{x,\hat{u}}(r), y^{x,\hat{u}}(r) \rangle + |\hat{u}(r)|^2) dr + \langle \hat{\Gamma}_t y^{x,\hat{u}}(s), \hat{\Gamma}_t y^{x,\hat{u}}(s) \rangle \ge \langle P_s^{\hat{\Gamma}_t^* \hat{\Gamma}_t} x, x \rangle.
\end{aligned}$$

On the other hand, for any arbitrary control u, writing $z(s) = y^{x,u}(s)$, there results

$$\begin{aligned}
&\int_0^s (\langle Ry^{x,u}(r), y^{x,u}(r)\rangle + |u(r)|^2)dr + \langle \hat\Gamma_t^*\hat\Gamma_t y^{x,u}(s), y^{x,u}(s)\rangle \\
&= \int_0^s (\langle Ry^{x,u}(r), y^{x,u}(r)\rangle + |u(r)|^2)dr + \int_0^t (\langle Ry^{z(s),v}(r), y^{z(s),v}(r)\rangle + |v(r)|^2)dr \\
&= \int_0^s (\langle Ry^{x,u}(r), y^{x,u}(r)\rangle + |u(r)|^2)dr \\
&\quad + \int_s^{t+s} (\langle Ry^{z(s),v}(w-s), y^{z(s),v}(w-s)\rangle + |v(w-s)|^2)dw,
\end{aligned}$$

where v is the optimal control which transfers the state $z(s) = y^{x,u}(s)$ to 0 in time t with the minimal energy. Consequently the control $\tilde u$,

$$\tilde u(r) = u(r), \quad r \in [0,s], \quad \tilde u(r) = v(r-s), \quad r \in [s, s+t],$$

transfers x to 0 in time $t+s$. Therefore

$$\int_0^s (\langle Ry^{x,u}(r), y^{x,u}(r)\rangle + |u(r)|^2)dr + \langle \hat\Gamma_t^*\hat\Gamma_t y^{x,u}(s), y^{x,u}(s)\rangle \ \geq \ \langle \hat\Gamma_{t+s}^*\hat\Gamma_{t+s}x, x\rangle.$$

Taking the infimum over the controls u, it follows that $\langle P_s^{\hat\Gamma_t^*\hat\Gamma_t}x, x\rangle \geq \langle \hat\Gamma_{t+s}^*\hat\Gamma_{t+s}x, x\rangle$. This gives (3.11). The proof of (i) is complete.

(ii) Fix $x \in H$ and let $\hat u$ be the control which transfers x to 0 in time t with the minimal energy. Since $y^{x,\hat u}(t) = 0$, we have

$$\begin{aligned}
&\langle \hat\Gamma_t x, \hat\Gamma_t x\rangle = \int_0^t (\langle Ry^{x,\hat u}(s), y^{x,\hat u}(s)\rangle + |\hat u(s)|^2)ds \\
&= \int_0^t (\langle Ry^{x,\hat u}(s), y^{x,\hat u}(s)\rangle + |\hat u(s)|^2)ds + \langle Sy^{x,\hat u}(t), y^{x,\hat u}(t)\rangle \geq \langle P_t^S x, x\rangle, \quad t \geq \tilde T. \qquad \square
\end{aligned}$$

Proof of Theorem 1.4. Since $(\Gamma_t^R)^*\Gamma_t^R$ is decreasing for $t \geq \tilde T$, formula (1.6) is equivalent to the fact that there exists

$$\lim_{t\to+\infty} \langle (\Gamma_t^R)^*\Gamma_t^R x, x\rangle = \langle \hat P x, x\rangle, \quad x \in H, \tag{3.12}$$

where $\hat P$ is the maximal solution to (1.4). To prove this assertion, the proof is split up into two parts.

Step I. We show that the map $t \mapsto (\Gamma_t^R)^*\Gamma_t^R$ converges strongly to a solution of the algebraic Riccati equation (1.4) as $t \to +\infty$. Since $(\Gamma_t^R)^*\Gamma_t^R$ is decreasing on $(\tilde T, +\infty)$, applying (2.4), we find that there exists $\hat P \in \mathcal{K}_+(H)$ such that $(\Gamma_t^R)^*\Gamma_t^R$ converges strongly to $\hat P$. By Lemma 3.2, we know that $P_t = (\Gamma_t^R)^*\Gamma_t^R$ solves the differential Riccati equation

$$\frac{d}{dt}\langle P_t h, k\rangle = \langle P_t h, Ak\rangle + \langle P_t Ah, k\rangle + \langle Rh, k\rangle - \langle P_t Q P_t h, k\rangle, \quad t \geq \tilde T,$$

for any h, $k \in D(A)$. Passing to the limit as $t \to +\infty$, by elementary arguments we find that

$$\frac{d}{dt}\langle (\Gamma_t^R)^*\Gamma_t^R h, k\rangle \ \to \ 0 \ \text{ as } \ t \to +\infty.$$

This yields that $\hat P$ solves the algebraic Riccati equation.

Step II. We show that $\hat{P}$ is maximal. If $S \geq 0$ is a solution of the algebraic Riccati equation, then $P_t^S = S$, $t \geq 0$, where P_t^S solves (3.1). Thanks to (ii) in Lemma 3.2, we get $P_t^S \leq (\Gamma_t^R)^* \Gamma_t^R$, $t \geq \tilde{T}$. Letting $t \to +\infty$, we infer that $S \leq \hat{P}$, and the proof is complete. □

Proof of Theorem 1.3. Recall that the system (1.1) is NCVE if and only if the operators $\Gamma_t = \Gamma_t^0$ (see (2.7)) converge strongly to 0 as $t \to +\infty$. The assertion follows from Theorem 1.4 □

4. Proofs of Theorems 1.1 and 1.2.

Proof of Theorem 1.1. We will use hypothesis (i) in (1.1) in order to show that (a) implies (b) and hypothesis (ii) in (1.1) in order to prove the converse.

(a) ⇒ (b) We know that $\Gamma_t^* \Gamma_t \to 0$ strongly as $t \to +\infty$. Assume by contradiction that $s(A) > 0$. By hypothesis (i) in (1.1), this implies that there exists an *isolated* element $\mu \in \sigma(A)$ such that $\mathrm{Re}(\mu) > 0$. Let us introduce the spectral Riesz projection P_0 associated with μ (see the notation in (2.1)),

$$P_0 x = \frac{1}{2\pi i} \int_\gamma (w - A)^{-1} x \, dw, \quad x \in H, \tag{4.1}$$

where γ is a circle enclosing μ in its interior and $\sigma(A) \backslash \{\mu\}$ in its exterior. We get the decomposition $H = E_0 \oplus E_1$, where $P_0 H = E_0$, $E_1 = P_1 H$, $(I - P_0) = P_1$.

The subspaces E_0 and E_1 are both invariant for e^{tA}; moreover, $E_0 \subset D(A)$, $A(E_0) \subset E_0$, and $A(E_1 \cap D(A)) \subset E_1$. Consider the restrictions A_i of A to E_i, $i = 0, 1$, and define

$$B_0 = P_0 B, \quad B_1 = P_1 B;$$

we have $B_i \in \mathcal{L}(U, E_i)$, $i = 0, 1$. We know that A_0 generates a uniformly continuous group e^{tA_0} on E_0, $\sigma(A_0) = \{\mu\}$, and A_1 generates a $\mathcal{C}_0$-semigroup e^{tA_1} on E_1. Let us split (1.1) into two systems: (A_0, B_0) on E_0 and (A_1, B_1) on E_1.

$$\begin{cases} \text{(i)} \;\; y_0'(t) = A_0 y_0(t) + B_0 u(t), \;\; y_0(0) = x_0 \in E_0, \\ \text{(ii)} \;\; y_1'(t) = A_1 y_1(t) + B_1 u(t), \;\; y_1(0) = x_1 \in E_1. \end{cases} \tag{4.2}$$

Since (4.2) is null controllable at $t \geq \tilde{T}$, also (A_0, B_0) and (A_1, B_1) are *null controllable* at $t \geq \tilde{T}$. In addition, e^{tA_0} is a uniformly continuous group, and therefore we have that the system $(-A_0, B_0)$, i.e.,

$$\frac{dy_0}{dt} = -A_0 y_0(t) + B_0 u(t), \;\; y_0(0) = x_0 \in E_0,$$

is exactly controllable at $t \geq \tilde{T}$. Consequently, for any $t \geq \tilde{T}$, the controllability operator $R_t \in \mathcal{L}(E_0)$,

$$R_t x_0 = \int_0^t e^{-sA_0} B_0 B_0^* e^{-sA_0^*} x_0 ds, \;\; x_0 \in E_0, \tag{4.3}$$

is one to one and onto E_0. Since $\sigma(-A_0) = \{-\mu\}$, we also have that e^{-tA_0} is exponentially stable. Thus the bounded linear operator R,

$$R x_0 = \int_0^\infty e^{-sA_0} B_0 B_0^* e^{-sA_0^*} x_0 ds, \;\; x_0 \in E_0,$$

is well defined and one has $\lim_{t\to+\infty}\langle R_t x_0, x_0\rangle = \langle Rx_0, x_0\rangle$, $x_0 \in E_0$. We remark that (R_t) is an increasing family of symmetric bounded operators on E_0 and that (4.3) is equivalent to $\langle R_t x, x\rangle \geq C|x|^2$ for any $t \geq \tilde{T}$, $C > 0$. It follows that R is coercive on E_0. Moreover, since $-A_0$ is stable, the operator R is the unique symmetric nonnegative solution of the Lyapunov equation:

$$(-A_0)R + R(-A_0)^* = -B_0B_0^*.$$

Using that R is an isomorphism, we achieve

$$R^{-1}A_0 + A_0^*R^{-1} = R^{-1}B_0B_0^*R^{-1} \text{ on } E_0.$$

Now consider the minimal energy operator $\Gamma_t^{E_0} : E_0 \to E_0$ associated with the exactly controllable system (A_0, B_0); i.e.,

$$|\Gamma_t^{E_0}x_0|^2 = \inf_{u\in L^2(0,t;U)}\left\{\int_0^t |u(s)|^2 ds \ : \ y_0^{x_0,u}(t) = 0\right\}, \ \ t > 0, \ x_0 \in E_0,$$

where $y_0^{x_0,u}(t) = e^{tA_0}x_0 + \int_0^t e^{(t-s)A_0}B_0u(s)ds$. It is easy to verify that

$$|\Gamma_t x_0| \geq |\Gamma_t^{E_0}x_0|, \ \ t \geq \tilde{T}.$$

It follows that $|\Gamma_t^{E_0}x_0|$ converges to 0 as $t \to +\infty$ for any $x_0 \in E_0$. By Theorem 1.3, this contradicts the fact that $P = R^{-1}$ is a nonzero solution to the algebraic Riccati equation $PA_0 + A_0^*P = PB_0B_0^*P$ on E_0. This finishes the proof.

(a) $\Leftarrow$ (b) Here we assume that $s(A) \leq 0$. The proof uses hypothesis (ii) in (1.1) and Theorem 1.3. In order to prove that $\Gamma_t^*\Gamma_t$ converges strongly to 0 as $t \to +\infty$, we show that if $P \geq 0$ is a solution of

$$PA + A^*P = PBB^*P, \tag{4.4}$$

then P is identically 0. Since $H = H_s \oplus H_u$, we will separately verify that $P = 0$ on H_s and on H_u.

We first prove that $Px = 0$ for any $x \in H_s$. To this end, let us denote by e^{tA_s} the restriction of e^{tA} to the subspace H_s. We have that $P = P_t^P$, $t \geq 0$, where P_t^P is the unique solution of the differential Riccati equation (3.1) with $R = 0$ such that $P_0 = P$. It follows that

$$\langle Px, x\rangle = \inf_{u\in L^2_{loc}}\left\{\int_0^t |u(s)|^2 ds \ + \ \langle Py^{x,u}(t), y^{x,u}(t)\rangle\right\}, \ \ t > 0, \ x \in H. \tag{4.5}$$

Fix any $x \in H_s$ and take the control $u = 0$; we obtain for any $t > 0$,

$$|\sqrt{P}\,x|^2 = \langle Px, x\rangle \leq \langle Pe^{tA_s}x, e^{tA_s}x\rangle. \tag{4.6}$$

Since e^{tA_s} is exponentially stable on H_s, letting $t \to +\infty$, we infer $Px = 0$ for any $x \in H_s$.

In order to show that $Px = 0$ for any $x \in H_u$, it is enough to check that $Pv = 0$ for any generalized eigenvector v of A. Indeed, we assume that there exists a set of generalized eigenvectors which spans a linear dense subspace of H_u. Now the proof is split up into two steps.

Step I. *We prove that if v is an eigenvector of A, then $Pv = 0$.*
Let $Av = \lambda v$, $\lambda \in \mathbf{C}$, $\mathrm{Re}(\lambda) \leq 0$. We have

$$\langle A^*Pv, v\rangle + \langle PAv, v\rangle = |B^*Pv|^2 \;\; \Rightarrow \;\; 2\mathrm{Re}(\lambda)\langle Pv, v\rangle = |B^*Pv|^2. \tag{4.7}$$

Since $\mathrm{Re}(\lambda) \leq 0$, we get that $B^*Pv = 0$. By (4.4), we find

$$A^*Pv + PAv = A^*Pv + \lambda Pv = PBB^*Pv = 0 \;\; \Rightarrow \;\; A^*Pv = -\lambda Pv.$$

Let us recall that the null controllability at $t > 0$ is equivalent to the fact that there exists $C_t > 0$ such that

$$\int_0^t |B^*e^{sA^*}x|^2 ds \geq C_t |e^{tA^*}x|^2, \quad x \in H. \tag{4.8}$$

Because of $B^*Pv = 0$ and (4.8), one has

$$0 = \int_0^t |e^{-\lambda s}B^*Pv|^2 ds = \int_0^t |B^*e^{sA^*}Pv|^2 ds \;\geq\; C_t e^{-2Re(\lambda)t}|Pv|^2, \tag{4.9}$$

$t \geq \tilde{T}$. Hence $Pv = 0$, and the assertion is proved.

Step II. *We check that if w is a generalized eigenvector of A, then $Pw = 0$.*

We know that there exists an eigenvalue $\lambda \in \mathbf{C}$, $\mathrm{Re}(\lambda) \leq 0$, and $m \in \mathbf{N}$ such that $w \in \mathrm{Ker}(\lambda - A)^m$. We will use induction on k to show that P is identically 0 on each $N_k = \mathrm{Ker}(\lambda - A)^k$, $k \in \mathbf{N}$. This will imply $Pw = 0$.

By the previous step, P is identically 0 on N_1. Let us assume that P is identically 0 on N_k and prove that the same is true for N_{k+1}. If $(\lambda - A)^{k+1}u = 0$, then

$$(\lambda - A)u \in N_k.$$

By induction we get $P(\lambda - A)u = 0$, i.e., $PAu = \lambda Pu$. It follows that

$$2\mathrm{Re}\langle PAu, u\rangle = 2\mathrm{Re}(\lambda)\langle Pu, u\rangle = |B^*Pu|^2,$$

which yields $B^*Pu = 0$. Arguing as before, it follows that $Pu = 0$. This completes the proof. □

Proof of Theorem 1.2. We are assuming that (1.1) is exactly controllable and that the operator A generates a $\mathcal{C}_0$-group e^{tA}. (We remark that A generates a $\mathcal{C}_0$-group e^{tA} on H if and only if both A and $-A$ generate $\mathcal{C}_0$-semigroups on H; see [10].) Recall the definition of the minimal energy $E_t^{(A,B)}(a, b) = E_t(a, b)$, which is needed to transfer the state a into b at time $t \geq \tilde{T}$; see (2.8).

The assertion is equivalent to the following one:

$$\lim_{t\to+\infty} E_t^{(A,B)}(a, b) = 0 \;\; \text{for all} \;\; a,\, b \in H \;\; \Longleftrightarrow \;\; \sigma(A) \subset \{i\lambda \,:\, \lambda \in \mathbf{R}\}. \tag{4.10}$$

First we need to check that the system $(-A, -B)$ is exactly controllable at $t \geq \tilde{T}$. To this end, fix a, $b \in H$. Since (A, B) is exactly controllable at t, there exists a control u such that $y^{b,u}(t) = a$. Define $z(s) = y(t - s)$, $s \in [0, t]$. One verifies that

$$z'(s) = -Az(s) - Bv(s) \quad \text{with the control} \quad v(s) = u(t - s), \;\; s \in [0, t].$$

Hence z is a solution of the system $(-A,-B)$; we have $z = z^{a,v}$, $z(0) = a$, and $z(t) = b$. This proves that $(-A,-B)$ is exactly controllable at t. Moreover, we find

$$\int_0^t |u(s)|^2 ds = \int_0^t |v(s)|^2 ds.$$

It follows that $E_t^{(A,B)}(a,b) = E_t^{(-A,-B)}(b,a)$ for any a, $b \in H$. ($E_t^{(-A,-B)}$ denotes the minimal energy with respect to $(-A,-B)$.)

Now we prove $\Longrightarrow$ in (4.10). We know in particular that $E_t^{(A,B)}(a,0)$ tends to 0 as $t \to +\infty$. Thanks to Theorem 1.1, we deduce that $s(A) \le 0$. Moreover, $E_t^{(A,B)}(0,b) = E_t^{(-A,-B)}(b,0)$, which tends to 0 as $t \to +\infty$ for any $b \in H$. It follows that $s(-A) \le 0$ and the claim is proved.

We verify $\Longleftarrow$ in (4.10). Set $E_t = E_t^{(A,B)}$ and fix any $t \ge 2\tilde{T}$; we prove that

$$E_t(a,b) \le E_{t/2}(a,0) + E_{t/2}(0,b), \quad a,\, b \in H. \tag{4.11}$$

To this end, choose two optimal controls u, v such that $\int_0^{t/2} |u(s)|^2 ds = E_{t/2}(a,0)$, $\int_0^{t/2} |v(s)|^2 ds = E_{t/2}(0,b)$. Define the control $\hat{u}$ in $L^2(0,t;U)$:

$$\hat{u}(r) = u(r), \quad r \in [0,t/2]; \quad \hat{u}(r) = v(r-t/2), \quad r \in (t/2,t].$$

We have

$$\begin{aligned} y^{a,\hat{u}}(t) &= e^{tA}a + \int_0^t e^{(t-s)A} B\hat{u}(s) ds \\ &= e^{t/2A}\left(e^{t/2A}a + \int_0^{t/2} e^{(t/2-s)A} Bu(s) ds \right) + \int_0^{t/2} e^{(t/2-r)A} Bv(r) dr \\ &= e^{t/2A} y^{a,u}(t/2) \,+\, y^{0,v}(t/2) = b. \end{aligned}$$

Since $E_t(a,b) \le \int_0^t |\hat{u}(s)|^2 ds = \int_0^{t/2} |u(s)|^2 ds + \int_0^{t/2} |v(s)|^2 ds$, we obtain formula (4.11). We remark that

$$E_{t/2}(0,b) = E_{t/2}^{(A,B)}(0,b) = E_{t/2}^{(-A,-B)}(b,0).$$

Applying Theorem 1.1, we deduce that both (A,B) and $(-A,-B)$ are NCVE systems. Thus $E_{t/2}^{(-A,-B)}(b,0)$ tends to 0 as $t \to +\infty$. The assertion follows letting $t \to +\infty$ in (4.11). □

5. Possible extensions and comments.

5.1. Strongly stable semigroups. A $\mathcal{C}_0$-semigroup e^{tA} on H is called *strongly stable* if $\lim_{t\to+\infty} e^{tA}x = 0$ for any $x \in H$. Arguing as in (4.5) and (4.6), it is possible to prove the following.

PROPOSITION 5.1. *Let the system* (1.1) *be null controllable and assume that* e^{tA} *is strongly stable. Then* (1.1) *is an NCVE system.*

This result can be used to show that a null controllable system (1.1) is NCVE when the operator A is of the following type:

(i) A is self-adjoint, $\langle Ax, x\rangle \le 0$, $x \in D(A)$, and $\lambda = 0$ is not an eigenvalue of A. (Note that in this case e^{tA} is strongly stable; see [10, p. 324].)

(ii) A generates a *relatively strongly compact semigroup* e^{tA}; i.e., e^{tA} is relatively compact with respect to the strong topology in $\mathcal{L}(H)$, $t \geq 0$; see [10, p. 317]. These semigroups are uniformly bounded, i.e., $\|e^{tA}\| \leq M$, $t \geq 0$, and yield the following decomposition for H: $H = H_s \oplus H_r$, where

$$H_s = \{x \in H : \lim_{t \to +\infty} e^{tA}x = 0\}, \quad H_r = \overline{\text{lin}[\{x \in D(A) : Ax = i\alpha x,\ \alpha \in \mathbf{R}\}]}.$$

5.2. Growth bound. We show now that there exist NCVE systems (A, B) for which Hypothesis 1.1 holds but the growth bound $\omega(A)$ of the semigroup e^{tA} is *positive*.

We recall that the *growth bound* $\omega(A)$ of a $\mathcal{C}_0$-semigroup e^{tA} is equal to

$$\omega(A) = \lim_{t \to +\infty} \frac{1}{t} \log \|e^{tA}\|.$$

Following [27] (see also [28, pp. 224–225]), for any sequence $(\lambda_m) \subset \mathbf{R}$ such that $|\lambda_m| \to +\infty$ as $m \to \infty$, we can construct a semigroup e^{tA} on $H = l^2_{\mathbf{C}}$ such that

$$\|e^{tA}\| = e^t, \quad t \geq 0, \quad \sigma(A) = \{i\lambda_m, \quad m \in \mathbf{N}\}.$$

Moreover, each $i\lambda_m$ is an eigenvalue of A with finite algebraic multiplicity, and the system of all generalized eigenvectors of A is dense in H. Thus A satisfies Hypothesis 1.1 with $H_s = \{0\}$, $H_u = H$, and $s(A) = 0$.

The required semigroup e^{tA} is defined as follows. Regarding each $x \in H$ as an infinite column and writing x in the form (x^m), where $x^m \in \mathbf{C}^m$, i.e., $H = \{(x^m) : \sum_{m=1}^{+\infty} |x^m|^2 < +\infty\}$, we set

$$e^{tA} = \bigoplus_{m \in \mathbf{N}} e^{i\lambda_m t} e^{tA^m}, \ \text{ i.e., } \ e^{tA}x = (e^{i\lambda_m t} e^{tA^m} x^m),$$

where each $A^m = A^m_{ij}$ is a nilpotent matrix of order m on $\mathbf{C}^m$ such that $A^m_{ij} = 1$ if $j = i + 1$ $(i = 1, \dots, m-1)$, $A^m_{ij} = 0$ for the remaining (i, j). We get

$$(5.1) \quad D(A) = \left\{(x^m) \in H : \sum_{m=1}^{+\infty} |\lambda_m|^2 |x^m|^2 < +\infty\right\}, \ \ A(x^m) = ((\lambda_m i + A^m)x^m),$$

$x = (x^m) \in D(A)$. Let us consider any null controllable system (A, B) on H, where A is given in (5.1). For instance, take any $B \in \mathcal{L}(U, H)$ invertible such that $B^{-1} \in \mathcal{L}(H, U)$. (Note that for any $x \in H$, the control $u(s) = -\frac{1}{t} B^{-1} e^{sA} x$, $s \in [0, t]$, transfers x into 0 at time $t > 0$.) By Theorem 1.1, we get that the system (A, B) is NCVE.

5.3. Comments on Hypothesis 1.1. This hypothesis seems to be rather technical. However, compare it with assumptions $(\mathcal{P})$ in [2, p. 272]. Note that Hypothesis 1.1 holds in the following important cases:

(i) H is finite dimensional;

(ii) the semigroup e^{tA} is eventually compact (see below);

(iii) the system (1.1) is null controllable and the operator $B \in \mathcal{L}(U, H)$ is compact.

Case (ii) includes, in particular, delay systems. Case (iii) follows by a general result on stabilizable systems; see [10] and also [8] and [5].

We stress that in Hypothesis 1.1 it is possible that the subspace H_s or H_u is equal to $\{0\}$. In particular, if all the generalized eigenvectors of A are linearly dense in H, we have $H_u = H$ and we set $H_s = \{0\}$; see (5.1).

Eventually compact semigroups. A semigroup e^{tA} on H is called *eventually compact* if there exists $t_0 > 0$ such that $e^{t_0 A}$ is a compact operator. (It follows that e^{tA} is compact for any $t \geq t_0$.) For such semigroups, it is known that $\sigma(A)$ is discrete and at most countable; it consists entirely of eigenvalues of finite multiplicity; see [10, pp. 330 and 247]. Moreover, for any $r \in \mathbf{R}$, the set

$$\{\mu \in \sigma(A) \,:\, \mathrm{Re}(\mu) \geq r\} \text{ is finite.} \tag{5.2}$$

For any eventually compact semigroup e^{tA}, one has $\omega(A) = s(A)$ (i.e., any eventually compact semigroup satisfies the spectral determining growth condition).

PROPOSITION 5.2. *If the operator A in* (1.1) *generates an eventually compact semigroup e^{tA}, then Hypothesis* 1.1 *holds.*

Proof. According to (5.2), we denote by σ_0 the finite set of all eigenvalues of A: $\mu_1, \ldots, \mu_n$, such that $\mathrm{Re}(\mu_i) \geq 0$ and set $\sigma_1 = \sigma(A) \backslash \sigma_0$. Using the spectral decomposition (2.2), we get two closed e^{tA}-invariant subspaces E_0 and E_1, associated with σ_0 and σ_1, where E_0 is finite dimensional and spanned by all generalized eigenvectors associated with μ_i, $i = 1, \ldots, n$. Moreover, e^{tA_1}, the restriction of e^{tA} to E_1, is exponentially stable. (Indeed, note that $e^{t_0 A_1}$ is compact in $\mathcal{L}(E_1)$ since $e^{t_0 A}$ is compact in $\mathcal{L}(H)$, and so $\omega(A_1) = s(A_1)$.) □

We finally mention a useful result: if a semigroup e^{tA} is analytic on H and with compact resolvent, then e^{tA} is compact for any $t > 0$.

Completeness of generalized eigenvectors. An important problem related to condition (ii) in Hypothesis (1.1) is to establish when the system G of all generalized eigenvectors of A is *complete*, i.e., G spans a dense linear subspace of H. This is, of course, the case when A is self-adjoint and with compact resolvent. However, a deep result proved by Dunford and Schwartz states that G is complete also when A is not necessarily self-adjoint but has a Hilbert–Schmidt resolvent and the uniform norm of its resolvent satisfies a particular growth condition; see [11]. This is true, in particular, when A generates an analytic semigroup and has Hilbert–Schmidt resolvent. Other theorems on the completeness of generalized eigenvectors are known in special situations; see [11] and the references therein. Moreover, we mention [5, Theorem 2.5.10], which gives conditions on delay systems in order that the system of all generalized eigenvectors is complete.

Stabilizable systems. A system (A, B) is called *exponentially stabilizable* if there exists an operator $F \in \mathcal{L}(H, U)$ such that for some $\beta < 0,\ C > 0$,

$$\|e^{t(A+BF)}\| \leq Ce^{\beta t}, \quad t \geq 0.$$

It is known that if a system is null controllable, then it is also exponentially stabilizable. By [10, Theorem 8.24] (see also [8] and [5, Theorem 5.2.6] for the case in which B is of finite rank), we get the following proposition.

PROPOSITION 5.3. *If the system* (1.1) *is exponentially stabilizable and the operator $B \in \mathcal{L}(U, H)$ is compact, then Hypothesis* 1.1 *holds.*

Note that if the system (1.1) is exponentially stabilizable, then there always exists the maximal solution $\hat{P}$ to the algebraic Riccati equation (1.4); see [2, p. 283]. (We remark that this maximal solution in general does not exist; see [2, p. 285].) Moreover, by [2, Theorem III.4.1] one has

$$\langle \hat{P}x, x\rangle = \inf\left\{ \int_0^\infty (\langle Ry^{x,u}(s), y^{x,u}(s)\rangle + |u(s)|^2) ds \,:\, u \in L^2(0,\infty;U),\ \ y^{x,u} \in L^2(0,\infty;H) \right\}.$$

6. Examples.

6.1. Delay systems. Let us consider the following controlled discrete delay system; see [1], [5], and [13] for more details:

$$\begin{cases} y'(t) = A_0 y(t) + \sum_{i=1}^{p} A_i y(t-h_i) + Bu(t), \ \ t \geq 0, \\ y(0) = r \in \mathbf{C}^n, \qquad y(s) = f(s), \ \ -h_p \leq s < 0, \end{cases} \tag{6.1}$$

where $0 < h_1 < \cdots < h_p$ represent the point delays, $y(t) \in \mathbf{C}^n$, $A_i \in \mathcal{L}(\mathbf{C}^n)$, $i = 0, \dots, p$; the map $f \in L^2([-h_p, 0]; \mathbf{C}^n)$, $B \in \mathcal{L}(\mathbf{C}^m, \mathbf{C}^n)$, $u \in L^2([0,T]; \mathbf{C}^m)$ for any $T > 0$. System (6.1) is said to be *null controllable* in time $\tilde{T} > 0$ if for any $r \in \mathbf{C}^n$ and any $f \in L^2([-h_p, 0]; \mathbf{C}^n)$, there exists $u \in L^2([0,T]; \mathbf{C}^m)$ such that for the corresponding solution $y^{(r,f),u}$,

$$y^{(r,f),u}(t) = 0 \ \text{ for all } \ t \in [\tilde{T} - h_p, \tilde{T}].$$

It is known that if (6.1) is null controllable at some time $\tilde{T}$, then $\tilde{T} > h_p$. Moreover, (6.1) is null controllable at any $t > h_p$ if and only if

$$\operatorname{rank}\left[\lambda I - \sum_{i=1}^{p} A_i e^{-\lambda h_i}, \ B\right] = n \tag{6.2}$$

for any $\lambda \in \mathbf{C}$; see [19]. A direct sufficient condition for the null controllability of (6.1) is

$$\operatorname{rank}[B, A_0 B, \dots, A_0^{n-1} B] = n \ \text{ and } \ \operatorname{Im}(A_j) \subset \operatorname{Im}(B), \ \ j = 1, \dots, p;$$

see [7, p. 202]. We now deduce from Theorem 1.1 the following result.

PROPOSITION 6.1. *The null controllable system* (6.1) *is NCVE if and only if*

$$\sup\left\{ \operatorname{Re}(\lambda) \ : \ \lambda \in \mathbf{C}, \ \det\left[\lambda I - A_0 - \sum_{i=1}^{p} A_i e^{-\lambda h_i}\right] = 0 \right\} \ \leq \ 0.$$

Proof. It is well known (see, e.g., [5]) that the delay system (6.1) can be reformulated as a linear system (1.1). To this end we introduce the Hilbert space $H = M^2([-h_p, 0]) = \mathbf{C}^n \oplus L^2([-h_p, 0]; \mathbf{C}^n)$ and the family of operators T_t, $T_t \in \mathcal{L}(H)$,

$$T_t(r, f) = (y(t), y(t + \cdot)), \quad (r, f) \in H, \ \ t \geq 0,$$

where $y(t)$ is the solution of (6.1) with $B = 0$ and $y(s) = f(s)$, $s \in [-h_p, 0]$. The operators T_t form a $\mathcal{C}_0$-semigroup on H with generator $\mathcal{A}$, $\mathcal{A} : D(\mathcal{A}) \subset H \to H$,

$$D(\mathcal{A}) = \{(f(0), f) \in H, \ \ f \in W^{1,2}([-h_p, 0]; \mathbf{C}^n)\},$$

$\mathcal{A}(f(0), f) = (A_0 f(0) + \sum_{i=1}^{p} A_i f(-h_i), \ \ f')$ for any $(f(0), f) \in D(\mathcal{A})$; here f' stands for the derivative of f. The spectrum of $\mathcal{A}$ consists entirely of eigenvalues of finite multiplicity, and further

$$\sigma(\mathcal{A}) = \{\lambda \in \mathbf{C} \ : \ \det(N(\lambda)) = 0\}, \ \ N(\lambda) = \left[\lambda I - A_0 - \sum_{i=1}^{p} A_i e^{-\lambda h_i}\right].$$

We reformulate (6.1) as

$$Y'(t) = \mathcal{A}Y(t) + \mathcal{B}u(t),$$

where $Y(t) = (y(t), y(t+\cdot)) \in H$, $Y(0) = (r, f)$, $t \geq 0$, and $y(t)$ denotes the solution of (6.1). Moreover, $u : [0, +\infty) \to \mathbf{C}^m$ and $\mathcal{B} \in \mathcal{L}(\mathbf{C}^m, H)$, $\mathcal{B}a = (Ba, 0)$ for any $a \in \mathbf{C}^m$. The resolvent operator of $\mathcal{A}$ is compact for any $\lambda \notin \sigma(\mathcal{A})$; see [5, Corollary 2.4.7]. Moreover, T_t is differentiable in (h_p, ∞) (i.e., the map $t \mapsto T_t \in \mathcal{L}(H)$ is differentiable on (h_p, ∞)); see [1, p. 60]. In particular, it follows that T_t is compact for any $t > h_p$; see [10, Lemma 4.28]. Thus the operator $\mathcal{A}$ satisfies Hypothesis 1.1, and the proof of the proposition is complete. □

6.2. Commuting systems. The following example shows that Theorem 1.1 may hold even if Hypothesis 1.1 is *not* satisfied.

Let us consider $H = L^2(\mathbf{R}, \mu)$, where μ is a locally finite measure on $\mathbf{R}$, and introduce the multiplication operator A,

$$Ax(s) = sx(s),\ s \in \mathbf{R},\ x \in H,\ D(A) = \{x \in H \text{ such that } s \mapsto sx(s) \in H\}. \tag{6.3}$$

The operator A is self-adjoint on H. (Recall that by the spectral theorem each self-adjoint operator on a given Hilbert space is unitarily isomorphic to a multiplication operator on some L^2-space.) We assume that A generates a $\mathcal{C}_0$-semigroup on H. This is equivalent to supposing that the measure μ is concentrated on $(-\infty, a]$ for some $a \in \mathbf{R}$. (This implies that the essential range of the identity map $s \mapsto s$ is bounded from above; see for instance [10].) Let us introduce a bounded operator B on H,

$$Bx(s) = b(s)x(s), \quad s \in \mathbf{R},\ x \in H, \tag{6.4}$$

where b is a real measurable map on $\mathbf{R}$ which is assumed to be essentially bounded, i.e., the essential range of b, $b_{ess}(\mathbf{R}) = \{r \in \mathbf{R},\ \mu(h \in \mathbf{R} : |b(h) - r| < \epsilon) > 0 \text{ for all } \epsilon > 0\}$, is bounded. We consider (1.1), where A and B are the operators which we have just introduced above. Let us compute the controllability operator Q_t:

$$Q_t x(s) = \frac{e^{2ts} - 1}{2s} b^2(s)\, x(s), \quad s \in \mathbf{R},\ x \in H.$$

The null controllability of the system is equivalent to the fact that the map $s \mapsto \frac{e^{ts}}{|b(s)|}\sqrt{\frac{2|s|}{|e^{2ts}-1|}}$ is essentially bounded. We assume that $0 \notin b_{ess}(\mathbf{R})$; this yields the null controllability of (1.1). By the theory of multiplication operators, we have $\sigma(A) = (-\infty, a]$. In general, *A does not satisfy* Hypothesis 1.1. (For instance, it may have no point spectrum.) However, we can prove the following result.

PROPOSITION 6.2. *The null controllable system (A, B), where A, B are of the form* (6.3), (6.4)*, respectively, is NCVE if and only if $s(A) = a \leq 0$ or if and only if $\mu(0, \infty) = 0$.*

Proof. Let us assume that $a > 0$. This means that there exist $0 < p < q$ such that $\mu([p, q]) > 0$. Let us compute the minimal energy $|\Gamma_t x|^2$ which is needed to transfer x into 0 at time t. We have

$$|\Gamma_t x|^2 = 2\int_{-\infty}^{a} \frac{e^{2ts}}{b^2(s)} \frac{|s|}{|e^{2ts} - 1|} x^2(s)\mu(ds), \quad x \in H.$$

Choosing $\hat{x} = I_{[p,q]}$, we find

$$|\Gamma_t \hat{x}|^2 = 2\int_{p}^{q} \frac{e^{2ts}}{b^2(s)} \frac{s}{e^{2ts} - 1} \mu(ds).$$

Fix $T>0$. Since for any $s\in[p,q]$, $t\geq T$, one has $\frac{e^{2ts}}{e^{2ts}-1}\geq 1$, we obtain

$$\lim_{t\to\infty}|\Gamma_t\hat{x}|^2\geq 2\int_p^q\frac{|s|}{b^2(s)}\mu(ds)>0,$$

and the system is not NCVE. Now let $\mu(0,\infty)=0$. We have

$$|\Gamma_t x|^2=2\int_{-\infty}^0\frac{|s|}{b^2(s)}\frac{e^{2sT}}{1-e^{2Ts}}x^2(s)\,k(t,s)\mu(ds),$$

where $0\leq k(t,s)=\frac{1-e^{2sT}}{1-e^{2ts}}\,\frac{e^{2st}}{e^{2sT}}\leq 1$, $s<0$, $t>T>0$. Since $\lim_{t\to\infty}k(t,s)=0$, we get $\lim_{t\to\infty}|\Gamma_t x|^2=0$ by the Lebesgue theorem. The proof is complete. □

6.3. Hyperbolic systems. Let us consider the following generalized damped wave equation:

$$\left\{\begin{array}{l}\dfrac{d^2y}{dt^2}(t)+\Lambda y(t)=\rho\dfrac{dy}{dt}(t)+u(t),\quad t>0,\quad \rho\in\mathbf{R},\\ y(0)=x\in D(\sqrt{\Lambda}),\quad \dfrac{dy}{dt}(0)=x_1\in K,\end{array}\right. \tag{6.5}$$

where Λ is a positive definite self-adjoint operator on a Hilbert space K, $\Lambda:D(\Lambda)\subset K\to K$, $y,u:[0,+\infty)\to K$. We assume that Λ has a compact resolvent. This implies that $\sigma(\Lambda)$ consists of a sequence of eigenvalues $0<\lambda_1\leq\lambda_2\leq\cdots\leq\lambda_n\leq\cdots\to+\infty$. We denote by (e_k) the corresponding orthonormal basis of eigenvectors, i.e., $\Lambda e_k=\lambda_k e_k$, $k\in\mathbf{N}$. For a specific application, consider a bounded open subset Ω of $\mathbf{R}^n$ with regular boundary $\partial\Omega$ and introduce

$$K=L^2(\Omega),\quad \Lambda=-\triangle,\quad D(\Lambda)=H^2(\Omega)\cap H_0^1(\Omega),\ D(\sqrt{\Lambda})=H_0^1(\Omega),$$

$$\left\{\begin{array}{l}\dfrac{\partial^2y}{dt^2}(t,\xi)=\triangle_\xi y(t,\xi)+\rho\dfrac{\partial y}{\partial t}(t,\xi)+u(t,\xi),\quad t>0,\ \xi\in\Omega,\\ y(t,\xi)=0,\quad t>0,\ \xi\in\partial\Omega;\quad y(0,\xi)=x(\xi),\quad \dfrac{\partial y}{dt}(0,\xi)=x_1(\xi),\end{array}\right. \tag{6.6}$$

$x\in H_0^1(\Omega)$, $x_1\in L^2(\Omega)$. Equation (6.5) can be reformulated as a system like (1.1). To this end, let us introduce the Hilbert space H, $H=D(\sqrt{\Lambda})\oplus K$, with inner product

$$\langle X,Y\rangle_H=\langle\sqrt{\Lambda}x,\sqrt{\Lambda}y\rangle_K+\langle x_1,y_1\rangle_K,\quad \text{where } X=\begin{pmatrix}x\\x_1\end{pmatrix},\ Y=\begin{pmatrix}y\\y_1\end{pmatrix}\in H.$$

Define the operators $A_\rho:D(A_\rho)=D(\Lambda)\oplus D(\sqrt{\Lambda})\subset H\to H$ and $B\in\mathcal{L}(K,H)$,

$$A_\rho\begin{pmatrix}x\\x_1\end{pmatrix}=\begin{pmatrix}0&I\\-\Lambda&\rho I\end{pmatrix}\begin{pmatrix}x\\x_1\end{pmatrix},\quad Bu=\begin{pmatrix}0\\u\end{pmatrix},$$

$x\in D(\Lambda)$, $x_1\in D(\sqrt{\Lambda})$, $u\in K$. It is not difficult to prove that A_ρ generates a $\mathcal{C}_0$-group e^{tA_ρ} on H, $\rho\in\mathbf{R}$. We can write (6.5) as $Y'(t)=A_\rho Y(t)+Bu(t)$, $Y(0)=Y_0\in H$, where $Y(t)=\binom{y(t)}{y_1(t)}$ and $Y_0=\binom{x}{x_1}$. Note that if $\rho=0$, then $A=A_0$ is skew-adjoint (i.e., $A^*=-A$) and generates a unitary group. The spectrum of A_ρ consists entirely of eigenvalues and can be easily computed:

$$\sigma(A_\rho)=\left\{\frac{\rho\pm\sqrt{\rho^2-4\lambda_k}}{2}\ :\ \lambda_k\in\sigma(\Lambda)\right\}.$$

It is clear that $s(A_\rho) \le 0$ if and only if $\rho \le 0$. In the case $\rho = 0$, there exists an orthonormal basis $(E_k^\pm)$ of eigenvectors of A:

$$E_k^\pm = \frac{1}{\sqrt{2}} \begin{pmatrix} \frac{e_k}{\sqrt{\lambda_k}} \\ \pm i\, e_k \end{pmatrix}, \quad k \in \mathbf{N}, \quad AE_k^\pm = \pm i\sqrt{\lambda_k} E_k^\pm.$$

It is known (see [6] or [2]) that (A_ρ, B) is *exactly controllable* at any $t > 0$ for any $\rho \in \mathbf{R}$. Denote by $E_t(X, Y)$ the minimal energy which is needed to transfer the state X into Y at time t; see (2.8). Thanks to Theorems 1.1 and 1.2, we obtain the following result.

PROPOSITION 6.3. *The system (A_ρ, B) is NCVE if and only if $\rho \le 0$. Moreover, (A_ρ, B) is ECVE if and only if $\rho = 0$.*

Proof. Recall that $s(A_\rho) \le 0$ if and only if $\rho \le 0$. We consider three different cases.

(a) $\rho > 0$. Since (i) in Hypothesis 1.1 is verified, we can conclude that (A_ρ, B) is not NCVE.

(b) $\rho = 0$. In this case there exists the basis $(E_k^\pm)$ on H consisting of eigenvectors of A. It follows that (ii) in Hypothesis 1.1 holds. (We can take $H_s = \{0\}$, $H_u = H$.) Moreover, $\sigma(A) \subset \{i\mathbf{R}\}$. Hence the system is ECVE.

(c) $\rho < 0$. Following [6], we find that e^{tA_ρ} is exponentially stable. (Indeed, one verifies that the map $F(x, x_1) = |x_1|^2 - \rho \mathrm{Re}\langle x_1, x\rangle + \frac{\rho^2}{2}|x|^2 + \langle\sqrt{\Lambda}x, \sqrt{\Lambda}x\rangle$, $x \in D(\sqrt{\Lambda})$, $x_1 \in K$, is a Lyapunov function for (6.5) with $u = 0$.) Thus even in this case condition (ii) in Hypothesis 1.1 holds and the system is NCVE. However, by Theorem 1.2, the system is not ECVE. □

6.4. Parabolic systems. Important examples of null controllable systems are *parabolic systems*. In this case we assume that the operator A in (1.1) generates an analytic semigroup on H. (For instance, as A we can take a $-\Lambda$, where Λ is a positive definite self-adjoint, unbounded operator on H; see section 6.3.)

If, in addition, we take $U = H$ and $B = I$, then it is well known that the parabolic system (A, I) with A unbounded is null controllable at any $t > 0$ but not exactly controllable; see [6], [2], [17], [18]. Note that parabolic systems with a compact resolvent satisfy Hypothesis 1.1.

Remark 6.1. It is of some interest to establish the rate with which the minimal energy tends to 0 as the horizon t of the control tends to $+\infty$. For wave equations, this problem was recently studied in [14]. (We also mention [15, p. 171], which considers the case when BB^* is invertible and e^{tA} is uniformly bounded.)

The behavior of the minimal energy $|\Gamma_t x|$ as $t \to 0^+$ is completely characterized in [24] for finite dimensional systems. On the contrary, in infinite dimensions such a characterization is not possible and the behavior of Γ_t as $t \to 0^+$ becomes much more complicated; see [12] which also contains useful references on this subject.

Acknowledgments. The authors wish to thank Professors G. Da Prato and L. Pandolfi for useful discussions and suggestions. The first author wishes to thank the Mathematics Research Center of the Warwick University and the Institute of Mathematics of the Polish Academy of Sciences in Warsaw, where parts of this paper were discussed, for their warm hospitality. Finally, the authors thank the referees for their careful reports.

REFERENCES

[1] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter, *Representation and Control of Infinite Dimensional Systems,* Vol. I, Birkhäuser Boston, Boston, 1993.

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

[3] S. Bittanti, J. L. Alan, and J. C. Willems, *The Riccati Equation*, Comm. Control Engrg. Ser., Springer-Verlag, Berlin, 1991.

[4] S. Chen and I. Lasiecka, *Feedback exact null controllability for unbounded control problems in Hilbert space,* J. Optim. Theory Appl., 74 (1992), pp. 191–219.

[5] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Texts Appl. Math. 21, Springer-Verlag, New York, 1995.

[6] G. Da Prato, A. J. Pritchard, and J. Zabczyk, *On minimum energy problems*, SIAM J. Control Optim., 29 (1991), pp. 209–221.

[7] G. Da Prato and J. Zabczyk, *Ergodicity for Infinite-Dimensional Systems,* London Math. Soc. Lecture Note Ser. 229, Cambridge University Press, Cambridge, UK, 1996.

[8] W. Desch and W. Schappacher, *Spectral properties of finite-dimensional perturbed linear semigroups,* J. Differential Equations, 59 (1985), pp. 80–102.

[9] Z. Emirsajlow, *Feedback approximation of the minimum energy control for linear infinite-dimensional systems with zero terminal state,* J. Optim. Theory Appl., 78 (1993), pp. 337–363.

[10] K. Engel and R. Nagel, *One-parameter Semigroups for Linear Evolution Equations,* Grad. Texts in Math. 194, Springer-Verlag, New York, 2000.

[11] I. Gohberg, S. Goldberg, and M. A. Kaashoek, *Classes of Linear Operators,* Vol. I, Oper. Theory Adv. Appl. 49, Birkhäuser-Verlag, Basel, 1990.

[12] F. Gozzi and P. Loreti, *Regularity of the minimum time function and minimum energy problems: The linear case*, SIAM J. Control Optim., 37 (1999), pp. 1195–1221.

[13] J. K. Hale, *Theory of Functional Differential Equations,* Springer-Verlag, New York, 1977.

[14] S. Ivanov, *Control norms for large control times*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 405–418.

[15] N. V. Krylov, M. Röckner, and J. Zabczyk, *Stochastic PDE's and Kolmogorov Equations in Infinite Dimensions*, Lecture Notes in Math. 1715, G. Da Prato, ed., Springer-Verlag, Berlin, 1999.

[16] H. Langer, A. C. M. Ran, and D. Temme, *Nonnegative solutions of algebraic Riccati equations,* Linear Algebra Appl., 261 (1997), pp. 317–352.

[17] I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations: Continuous and Approximation Theories.* I. *Abstract Parabolic Systems*, Encyclopedia Math. Appl. 74, Cambridge University Press, Cambridge, UK, 2000.

[18] I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations: Continuous and Approximation Theories.* II. *Abstract Hyperbolic-Like Systems over a Finite Time Horizon,* Encyclopedia Math. Appl. 75, Cambridge University Press, Cambridge, UK, 2000.

[19] A. W. Olbrot and L. Pandolfi, *Null controllability of a class of functional differential systems*, Internat. J. Control, 47 (1988), pp. 193–208.

[20] R. G. Pinsky, *Positive Harmonic Functions and Diffusion*, Cambridge Stud. Adv. Math. 45, Cambridge University Press, Cambridge, UK, 1995.

[21] E. Priola and J. Zabczyk, *Liouville Theorems in Finite and Infinite Dimensions*, Preprint 9, Scuola Normale Superiore di Pisa, Pisa, Italy, 2003.

[22] E. Priola and J. Zabczyk, *Null Controllability with Vanishing Energy*, Preprint 10/2002, Department of Mathematics, University of Turin, Turin, Italy, 2002; also available online from http://www2.dm.unito.it/paginepersonali/priola/index.htm.

[23] K.-I. Sato, *Levy Processes and Infinite Divisible Distributions*, Cambridge University Press, Cambridge, UK, 1999.

[24] T. Seidman, *How violent are fast controls?*, Math. Control Signals Systems, 1 (1988), pp. 89–95.

[25] M. Sirbu and G. Tessitore, *Null controllability of an infinite dimensional SDE with state and control-dependent noise*, Systems Control Lett., 44 (2001), pp. 385–394.

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

[27] J. Zabczyk, *A note on C_0-semigroups*, Bull. Acad. Polon. Sci. Sér. Math. Astronom. Phys., 23 (1975), pp. 895–898.

[28] J. Zabczyk, *Mathematical Control Theory: An Introduction*, Systems Control Found. Appl., Birkhäuser-Verlag, Basel, 1992.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 1033–1043

BOUNDARY FEEDBACK STABILIZATION OF AN UNSTABLE HEAT EQUATION*

WEIJIU LIU†

Abstract. In this paper we study the problem of boundary feedback stabilization for the unstable heat equation

$$u_t(x,t) = u_{xx}(x,t) + a(x)u(x,t).$$

This equation can be viewed as a model of a heat conducting rod in which not only is the heat being diffused (mathematically due to the diffusive term u_{xx}) but also the destabilizing heat is generating (mathematically due to the term au with $a > 0$). We show that for *any* given continuously differentiable function a and *any* given positive constant λ we can *explicitly* construct a boundary feedback control law such that the solution of the equation with the control law converges to zero exponentially at the rate of λ. This is a continuation of the recent work of Boskovic, Krstic, and Liu [*IEEE Trans. Automat. Control*, 46 (2001), pp. 2022–2028] and Balogh and Krstic [*European J. Control*, 8 (2002), pp. 165–176].

Key words. heat equation, boundary control, stabilization

AMS subject classifications. 35K05, 93D15

DOI. 10.1137/S0363012902402414

1. Introduction.

In this paper we continue the study of boundary feedback control of an unstable heat equation

$$u_t(x,t) = u_{xx}(x,t) + \mu u(x,t) \quad \text{in } (0,1)\times(0,\infty).$$

Hereafter, the subscripts denote the derivatives. This equation can be viewed as a model of a heat conducting rod in which not only is the heat being diffused (mathematically due to the term u_{xx}) but also the destabilizing heat is generating (mathematically due to the term μu with $\mu > 0$). This feedback control problem was recently addressed by Boskovic, Krstic, and Liu in [5], and it was shown that the unstable rod can be exponentially stabilized by a boundary feedback control law if the constant $\mu < 3\pi^2/4$; that is, the destabilizing heat generation is not very big. More recently, Balogh and Krstic [3, 4] removed the condition $\mu < 3\pi^2/4$ and replaced μ by an arbitrarily large function $a(x)$:

$$u_t(x,t) = u_{xx}(x,t) + a(x)u(x,t) \quad \text{in } (0,1)\times(0,\infty). \tag{1.1}$$

They used a backstepping method for the finite difference semidiscretized approximation of the above equation to derive a Dirichlet boundary feedback control law that makes the closed-loop system stable with an arbitrary prescribed stability margin. They showed that the integral kernel in the control law is bounded. However, some problems like the smoothness of the kernel and Neumann boundary control (usually more difficult than the Dirichlet one) were left open. Using a different method, we completely solve these problems by solving a partial differential equation of the kernel

*Received by the editors February 8, 2002; accepted for publication (in revised form) February 24, 2003; published electronically July 8, 2003. This work was done while the author was with the University of Cincinnati and was supported by the Taft Memorial Fund.

http://www.siam.org/journals/sicon/42-3/40241.html

†Department of Mechanical Engineering, Room 3-455C, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139 (weiliu@mit.edu).

with strange boundary conditions (see (2.1) below). This strange boundary value problem has stood open since the work of [5] was started in 1998. We also derive Neumann boundary feedback control laws which seemingly cannot be achieved in [4]. From the proof of Lemma 2.2 below it can been seen that the feedback law is constructed explicitly and can be calculated numerically via a scheme of successive approximation. This makes its implementation possible in real problems.

The problem of boundary feedback control that we address here is not new. Some of the results on feedback stabilization of parabolic equations include the work of Amamm [2], Burns and Rubio [6], Burns, Rubio, and King [7], Day [8], Lasiecka and Triggiani [10, 11, 12, 13], and Triggiani [15]. For a detailed review of these references, we refer to [4] and [5]. In comparison with the existing literature, the novelty of the paper is the explicit construction of the feedback laws and the complete solving of the strange boundary value problem mentioned above.

The paper is organized as follows. Section 2 is devoted to the stabilization of unstable Dirichlet boundary value problems and section 3 to the stabilization of unstable Neumann boundary value problems. We raise an open problem in section 4.

2. Dirichlet boundary conditions. In what follows, we denote by $H^s(0,1)$ the usual Sobolev space (see, e.g., [1, 14]) for any $s \in \mathbf{R}$. For $s \geq 0$, $H_0^s(0,1)$ denotes the completion of $C_0^\infty(0,1)$ in $H^s(0,1)$, where $C_0^\infty(0,1)$ denotes the space of all infinitely differentiable functions on $(0,1)$ with compact support in $(0,1)$. We denote by $\|\cdot\|$ the norm of $L^2(0,1)$. $C^n[0,1]$ denotes the space of all n times continuously differentiable functions on $[0,1]$.

It is well known that the Dirichlet boundary value problem

$$\begin{cases} u_t(x,t) = u_{xx}(x,t) + a(x)u(x,t) & \text{in } (0,1)\times(0,\infty), \\ u(0,t) = u(1,t) = 0 & \text{in } (0,\infty) \end{cases}$$

is unstable if a is positive and large. To design a boundary feedback law to stabilize it for any function $a \in C^1[0,1]$, we consider the problem

$$\begin{cases} k_{xx}(x,y) - k_{yy}(x,y) = (a(y)+\lambda)k(x,y), & 0 \leq y \leq x \leq 1, \\ k(x,0) = 0, & 0 \leq x \leq 1, \\ k_x(x,x) + k_y(x,x) + \frac{d}{dx}(k(x,x)) = a(x) + \lambda, & 0 \leq x \leq 1, \end{cases} \tag{2.1}$$

where λ is any constant. From the proof of Lemma 2.4 below we will see why we want to consider this problem. For the moment, let us assume this problem has a unique solution k for $a \in C^1[0,1]$. (This will be proved in Lemma 2.2 below.) Using the solution k, we then obtain Dirichlet boundary feedback law

$$u(1,t) = -\int_0^1 k(1,y)u(y,t)dy \qquad \text{in } (0,\infty) \tag{2.2}$$

and Neumann boundary feedback law

$$u_x(1,t) = -k(1,1)u(1,t) - \int_0^1 k_x(1,y)u(y,t)dy \qquad \text{in } (0,\infty). \tag{2.3}$$

With one of the boundary feedback laws, the system

$$\begin{cases} u_t(x,t) = u_{xx}(x,t) + a(x)u(x,t) & \text{in } (0,1)\times(0,\infty), \\ u(0,t) = 0 & \text{in } (0,\infty), \\ u(x,0) = u^0(x) & \text{in } (0,1) \end{cases} \tag{2.4}$$

is exponentially stable. In this controlled system, the left-hand end of a rod is insulated while the temperature or the heat flux at the other end is adjusted according to the measurement of k-weighted averaged temperature over the whole rod. Physically, if the destabilizing heat is generating inside the rod, then we cool the right end of the rod so that it is not overheated. To state this result, we introduce the compatible conditions for the initial data:

$$u^0(0) = 0, \quad u^0(1) = -\int_0^1 k(1,y)u^0(y)dy, \tag{2.5}$$

$$u^0(0) = 0, \quad u_x^0(1) = -k(1,1)u^0(1) - \int_0^1 k_x(1,y)u^0(y)dy. \tag{2.6}$$

THEOREM 2.1. *Assume that $\lambda > 0$ is any positive constant and $a \in C^1[0,1]$ is any function. For arbitrary initial data $u^0(x) \in H^1(0,1)$ with compatible condition* (2.5) *or* (2.6)*, equation* (2.4) *with either* (2.2) *or* (2.3) *has a unique solution that satisfies*

$$\|u(t)\|_{H^1} \le M\|u^0\|_{H^1}e^{-\lambda t} \quad \forall t > 0, \tag{2.7}$$

where M is a positive constant independent of u^0.

The idea of proving the theorem is to carefully construct a transformation

$$w(x,t) = u(x,t) + \int_0^x k(x,y)u(y,t)dy$$

to convert the system (2.4) with either (2.2) or (2.3) into the exponentially stable system

$$\begin{cases} w_t = w_{xx} - \lambda w & \text{in } (0,1)\times(0,\infty), \\ w(0,t) = w(1,t) = 0 & \text{in } (0,\infty), \\ w(x,0) = w^0(x) & \text{in } (0,1) \end{cases} \tag{2.8}$$

or

$$\begin{cases} w_t = w_{xx} - \lambda w & \text{in } (0,1)\times(0,\infty), \\ w(0,t) = w_x(1,t) = 0 & \text{in } (0,\infty), \\ w(x,0) = w^0(x) & \text{in } (0,1), \end{cases} \tag{2.9}$$

where $w^0(x) = u^0(x) + \int_0^x k(x,y)u^0(y)dy$. This will be achieved in the following lemmas.

LEMMA 2.2. *Suppose that $a \in C^1[0,1]$. Then problem* (2.1) *has a unique solution which is twice continuously differentiable in $0 \le y \le x \le 1$.*

Proof. Using the variable changes

$$\xi = x + y, \quad \eta = x - y$$

and denoting

$$G(\xi,\eta) = k(x,y) = k\left(\frac{\xi+\eta}{2}, \frac{\xi-\eta}{2}\right),$$

problem (2.1) is transformed to

$$\begin{cases} G_{\xi\eta}(\xi,\eta) = \frac{1}{4}\left(a\left(\frac{\xi-\eta}{2}\right) + \lambda\right)G(\xi,\eta), & 0 \le \eta \le \xi \le 2, \\ G(\xi,\xi) = 0, & 0 \le \xi \le 2, \\ \frac{\partial}{\partial\xi}(G(\xi,0)) = \frac{1}{4}\left(a\left(\frac{\xi}{2}\right) + \lambda\right), & 0 \le \xi \le 2, \end{cases} \tag{2.10}$$

which is equivalent to the following integral equation:

$$G(\xi,\eta)=\frac{1}{4}\int_{\eta}^{\xi}\left(a\left(\frac{\tau}{2}\right)+\lambda\right)d\tau+\frac{1}{4}\int_{\eta}^{\xi}\int_{0}^{\eta}\left(a\left(\frac{\tau-s}{2}\right)+\lambda\right)G(\tau,s)dsd\tau. \tag{2.11}$$

By the method of successive approximations we can show that this equation has a unique continuous solution. In fact, set

$$G_0(\xi,\eta)=\frac{1}{4}\int_{\eta}^{\xi}\left(a\left(\frac{\tau}{2}\right)+\lambda\right)d\tau,$$
$$G_n(\xi,\eta)=\frac{1}{4}\int_{\eta}^{\xi}\int_{0}^{\eta}\left(a\left(\frac{\tau-s}{2}\right)+\lambda\right)G_{n-1}(\tau,s)dsd\tau$$

and denote $M=\sup_{0\le x\le 1}|a(x)+\lambda|$. Then one can readily show that

$$|G_0(\xi,\eta)|\le\frac{1}{4}M(\xi-\eta)\le M,$$
$$|G_1(\xi,\eta)|\le M^2\xi\eta,$$
$$|G_2(\xi,\eta)|\le\frac{M^3}{(2!)^2}\xi^2\eta^2,$$

and, by induction,

$$|G_n(\xi,\eta)|\le\frac{M^{n+1}}{(n!)^2}\xi^n\eta^n.$$

These estimates show that the series

$$G(\xi,\eta)=\sum_{n=0}^{\infty}G_n(\xi,\eta)$$

converges absolutely and uniformly in $0\le\eta\le\xi\le 2$, and then its sum is a continuous solution of (2.11). Moreover, it follows from (2.11) that G is twice continuously differentiable because $a\in C^1[0,1]$. Indeed, differentiating (2.11) with respect to ξ gives

$$\frac{\partial G(\xi,\eta)}{\partial\xi}=\frac{1}{4}\left(a\left(\frac{\xi}{2}\right)+\lambda\right)+\frac{1}{4}\int_{0}^{\eta}\left(a\left(\frac{\xi-s}{2}\right)+\lambda\right)G(\xi,s)ds,$$

which implies that $\frac{\partial G(\xi,\eta)}{\partial\xi}$ is continuous since $G(\xi,\eta)$ is continuous. By analogy, we can show that other derivatives of G are continuous. □

Remark 2.3. The proof of Lemma 2.2 provides a numeric computation scheme of successive approximation to compute the kernel function k in our feedback laws (2.2) and (2.3). This makes the feedback laws (2.2) and (2.3) implementable in real problems.

LEMMA 2.4. *Let $k(x,y)$ be the solution of problem* (2.1) *and define the linear bounded operator $K:H^i(0,1)\to H^i(0,1)$ $(i=0,1,2)$ by*

$$w(x)=(Ku)(x)=u(x)+\int_{0}^{x}k(x,y)u(y)dy\quad\text{for } u\in H^i(0,1). \tag{2.12}$$

Then

1. *K has a linear bounded inverse $K^{-1} : H^i(0,1) \to H^i(0,1)$ $(i = 0, 1, 2)$, and*
2. *K converts the system* (2.2) *and* (2.4) *and the system* (2.3) *and* (2.4) *into* (2.8) *and* (2.9)*, respectively.*

Proof. To prove that (2.12) has a bounded inverse, we set

$$v(x) = \int_0^x k(x,y)u(y)dy$$

and then

$$w(x) = u(x) + v(x).$$

Hence we have

$$\begin{aligned} v(x) &= \int_0^x k(x,y)[w(y) - v(y)]dy \\ &= \int_0^x k(x,y)w(y)dy - \int_0^x k(x,y)v(y)dy. \end{aligned} \tag{2.13}$$

To show that this equation has a unique continuous solution, we set

$$\begin{aligned} v_0(x) &= \int_0^x k(x,y)w(y)dy, \\ v_n(x) &= -\int_0^x k(x,y)v_{n-1}(y)dy \end{aligned}$$

and denote $M = \sup_{0 \le y \le x \le 1} |k(x,y)|$. Then

$$\begin{aligned} |v_0(x)| &\le M\|w\|, \\ |v_1(x)| &\le M^2\|w\|x, \\ |v_2(x)| &\le \frac{M^3\|w\|}{2!}x^2, \end{aligned}$$

and, by induction,

$$|v_n(x)| \le \frac{M^{n+1}\|w\|}{n!}x^n.$$

These estimates show that the series

$$v(x) = \sum_{n=0}^{\infty} v_n(x)$$

converges absolutely and uniformly in $0 \le x \le 1$ and that its sum is a continuous solution of (2.13). Moreover, there exists a constant $C > 0$ such that

$$\|v\| \le C\|w\|. \tag{2.14}$$

This implies that there exists a bounded linear operator $\Phi : L^2(0,1) \to L^2(0,1)$ such that

$$v(x) = (\Phi w)(x)$$

and then

$$u(x) = w(x) - v(x) = ((I - \Phi)w)(x) = (K^{-1}w)(x). \tag{2.15}$$

It is clear that $K^{-1} : L^2(0,1) \to L^2(0,1)$ is bounded. To show that $K^{-1} : H^1(0,1) \to H^1(0,1)$ is bounded, we take the derivative in (2.13) and obtain

$$v_x(x) = k(x,x)w(x) + \int_0^x k_x(x,y)w(y)dy - k(x,x)v(x) - \int_0^x k_x(x,y)v(y)dy,$$

which, combined with (2.14), implies that there exists constant $C > 0$ such that

$$\|v_x\| \le C\|w\|$$

and then by (2.15)

$$\|u\|_{H^1} \le \|w\|_{H^1} + \|v\|_{H^1} \le C\|w\|_{H^1}.$$

By analogy, we can show that $K^{-1} : H^2(0,1) \to H^2(0,1)$ is bounded.

To prove that the transformation (2.12) converts the system (2.2) and (2.4) and the system (2.3) and (2.4) into (2.8) and (2.9), respectively, we compute as follows:

$$\begin{aligned} w_t(x,t) &= u_t(x,t) + \int_0^x k(x,y)u_t(y,t)dy \\ &= u_t(x,t) + \int_0^x k(x,y)[u_{yy}(y,t) + a(y)u(y,t)]dy \\ &= u_t(x,t) + k(x,x)u_x(x,t) - k(x,0)u_x(0,t) \\ &\quad - k_y(x,x)u(x,t) + k_y(x,0)u(0,t) \\ &\quad + \int_0^x [k_{yy}(x,y)u(y,t) + k(x,y)a(y)u(y,t)]dy, \end{aligned} \tag{2.16}$$

$$w_x(x,t) = u_x(x,t) + k(x,x)u(x,t) + \int_0^x k_x(x,y)u(y,t)dy, \tag{2.17}$$

$$\begin{aligned} w_{xx}(x,t) &= u_{xx}(x,t) + \frac{d}{dx}(k(x,x))u(x,t) + k(x,x)u_x(x,t) \\ &\quad + k_x(x,x)u(x,t) + \int_0^x k_{xx}(x,y)u(y,t)dy. \end{aligned} \tag{2.18}$$

It then follows from (2.1) and (2.4) that

(2.19)

$$\begin{aligned} w_t - w_{xx} + \lambda w = {} & u_t(x,t) + k(x,x)u_x(x,t) - k(x,0)u_x(0,t) \\ & -k_y(x,x)u(x,t) + k_y(x,0)u(0,t) \\ & + \int_0^x [k_{yy}(x,y)u(y,t) + k(x,y)a(y)u(y,t)]dy \\ & -u_{xx}(x,t) - \frac{d}{dx}(k(x,x))u(x,t) - k(x,x,t)u_x(x,t) \\ & -k_x(x,x)u(x,t) - \int_0^x k_{xx}(x,y)u(y,t)dy \\ & +\lambda u(x,t) + \lambda \int_0^x k(x,y)u(y,t)dy \end{aligned}$$

$$
\begin{aligned}
&= \left(a(x) - k_x(x,x) - k_y(x,x) - \frac{d}{dx}(k(x,x)) + \lambda\right) u(x,t) \\
&\quad + k_y(x,0)u(0,t) - k(x,0)u_x(0,t) \\
&\quad + \int_0^x \left[k_{yy}(x,y) - k_{xx}(x,y,t) + (a(y)+\lambda)k(x,y,t)\right] u(y,t)dy \\
&= 0.
\end{aligned}
$$

By the boundary condition of (2.4), we deduce that $w(0,t) = 0$. Using feedback law (2.2) or (2.3), we obtain

$$w(1,t) = u(1,t) + \int_0^1 k(1,y)u(y,t)dy = 0$$

or

$$w_x(1,t) = u_x(1,t) + k(1,1)u(1,t) + \int_0^1 k_x(1,y)u(y,t)dy = 0. \qquad \square$$

We are now ready to prove Theorem 2.1.

Proof of Theorem 2.1. We first note that problem (2.4) with either (2.2) or (2.3) is well posed since, by Lemma 2.4, they can be transformed to the problem (2.8) or (2.9) via the isomorphism defined by (2.12), and the problem (2.8) or (2.9) is well posed (see, e.g., [9, Chap. IV]). Moreover, there exists a positive constant $C > 0$ such that

$$
\begin{aligned}
\|u(t)\|_{H^1} &\le C\|w(t)\|_{H^1}, \\
\|w^0\|_{H^1} &\le C\|u^0\|_{H^1}.
\end{aligned}
$$

Therefore, it is sufficient to prove (2.7) for the solution w of (2.8) or (2.9). We do so only for problem (2.8) since the situation for problem (2.9) is similar.

We define the energy

$$E(t) = \frac{1}{2}\int_0^1 w(x,t)^2 dx.$$

Multiplying the first equation of (2.8) by w and integrating from 0 to 1 by parts we get

$$
\begin{aligned}
\dot{E}(t) &= w_x w\Big|_0^1 - \int_0^1 w_x(x,t)^2 dx - \lambda\int_0^1 w(x,t)^2 dx \\
&= -\int_0^1 w_x(x,t)^2 dx - \lambda\int_0^1 w(x,t)^2 dx \\
&\le -2\lambda E(t),
\end{aligned}
$$

which implies

$$E(t) \le E(0)e^{-2\lambda t} \quad \text{for } t \ge 0.$$

Set

$$V(t) = \int_0^1 w_x(x,t)^2 dx.$$

Multiplying the first equation of (2.8) by w_{xx} and integrating from 0 to 1 by parts we obtain

$$\begin{aligned}\dot{V}(t) &= -2\int_0^1 w_{xx}^2 dx + 2\lambda\int_0^1 w w_{xx} dx \\ &= -2\int_0^1 w_{xx}^2 dx - 2\lambda\int_0^1 w_x^2 dx \\ &\le -2\lambda V(t),\end{aligned}$$

which implies that

$$V(t) \le V(0)e^{-2\lambda t}.$$

This shows that (2.7) holds. □

3. Neumann boundary conditions. To stabilize the Neumann boundary value problem

$$\begin{cases} u_t(x,t) = u_{xx}(x,t) + a(x)u(x,t) & \text{in } (0,1)\times(0,\infty), \\ u_x(0,t) = u_x(1,t) = 0 & \text{in } (0,\infty), \end{cases}$$

we consider the problem

$$\begin{cases} k_{xx}(x,y) - k_{yy}(x,y) = (a(y)+\lambda)k(x,y), & 0 \le y \le x \le 1, \\ k_y(x,0) = 0, & 0 \le x \le 1, \\ k_x(x,x) + k_y(x,x) + \frac{d}{dx}(k(x,x)) = a(x) + \lambda, & 0 \le x \le 1, \\ k(0,0) = 0, \end{cases} \tag{3.1}$$

where λ is any constant. Using the solution k, we then obtain Dirichlet boundary feedback law

$$u(1,t) = -\int_0^1 k(1,y)u(y,t)dy \qquad \text{in } (0,\infty) \tag{3.2}$$

and Neumann boundary feedback law

$$u_x(1,t) = -k(1,1)u(1,t) - \int_0^1 k_x(1,y)u(y,t)dy \qquad \text{in } (0,\infty). \tag{3.3}$$

With one of the boundary feedback laws, the system

$$\begin{cases} u_t(x,t) = u_{xx}(x,t) + a(x)u(x,t) & \text{in } (0,1)\times(0,\infty), \\ u_x(0,t) = 0 & \text{in } (0,\infty), \\ u(x,0) = u^0(x) & \text{in } (0,1) \end{cases} \tag{3.4}$$

is exponentially stable. To state this result, we introduce the compatible conditions for the initial data

$$u_x^0(0) = 0, \quad u^0(1) = -\int_0^1 k(1,y)u^0(y)dy, \tag{3.5}$$

$$u_x^0(0) = 0, \quad u_x^0(1) = -k(1,1)u^0(1) - \int_0^1 k_x(1,y)u^0(y)dy. \tag{3.6}$$

THEOREM 3.1. *Assume that $\lambda > 0$ is any positive constant and $a \in C^1[0,1]$ is any function. For arbitrary initial data $u^0(x) \in H^1(0,1)$ with the compatible condition* (3.5) *or* (3.6)*, equation* (3.4) *with either* (3.2) *or* (3.3) *has a unique solution that satisfies*

$$\|u(t)\|_{H^1} \le M\|u^0\|_{H^1}e^{-\lambda t},$$

where M is a positive constant independent of u^0.

Proof. The proof is the same as that of Theorem 2.1. The only thing we need to do is to show that problem (3.1) has a unique solution. This is given in Lemma 3.2 below. □

LEMMA 3.2. *Suppose that $a \in C^1[0,1]$. Then problem* (3.1) *has a unique solution which is twice continuously differentiable in $0 \le y \le x \le 1$.*

Proof. Using the variable changes

$$\xi = x + y, \quad \eta = x - y$$

and denoting

$$G(\xi,\eta) = k(x,y) = k\left(\frac{\xi+\eta}{2}, \frac{\xi-\eta}{2}\right),$$

problem (3.1) is transformed into

$$\text{(3.7)} \qquad \begin{cases} G_{\xi\eta}(\xi,\eta) = \frac{1}{4}\left(a\left(\frac{\xi-\eta}{2}\right)+\lambda\right)G(\xi,\eta), & 0 \le \eta \le \xi \le 2, \\ G_\xi(\xi,\xi) = G_\eta(\xi,\xi), & 0 \le \xi \le 2, \\ \frac{\partial}{\partial\xi}(G(\xi,0)) = \frac{1}{4}\left(a\left(\frac{\xi}{2}\right)+\lambda\right), & 0 \le \xi \le 2, \\ G(0,0) = 0. \end{cases}$$

Integrating the first equation of (3.7) with respect to η from 0 to ξ gives

$$\begin{aligned} G_\xi(\xi,\xi) &= G_\xi(\xi,0) + \frac{1}{4}\int_0^\xi \left(a\left(\frac{\xi-s}{2}\right)+\lambda\right)G(\xi,s)ds \\ &= \frac{1}{4}\left(a\left(\frac{\xi}{2}\right)+\lambda\right) + \frac{1}{4}\int_0^\xi \left(a\left(\frac{\xi-s}{2}\right)+\lambda\right)G(\xi,s)ds. \end{aligned}$$

It then follows from the second equation of (3.7) that

$$\begin{aligned} \frac{d}{d\xi}[G(\xi,\xi)] &= G_\xi(\xi,\xi) + G_\eta(\xi,\xi) \\ &= 2G_\xi(\xi,\xi) \\ &= \frac{1}{2}\left(a\left(\frac{\xi}{2}\right)+\lambda\right) + \frac{1}{2}\int_0^\xi \left(a\left(\frac{\xi-s}{2}\right)+\lambda\right)G(\xi,s)ds. \end{aligned}$$

Integrating from 0 to ξ and using the fourth equation of (3.7) gives

$$\text{(3.8)} \qquad G(\xi,\xi) = \frac{1}{2}\int_0^\xi \left(a\left(\frac{\tau}{2}\right)+\lambda\right)d\tau + \frac{1}{2}\int_0^\xi\int_0^\tau \left(a\left(\frac{\tau-s}{2}\right)+\lambda\right)G(\tau,s)dsd\tau.$$

Integrating twice the first equation of (3.7) first with respect to η from 0 to η and second with respect to ξ from η to ξ and using (3.8), we obtain the following integral equation:

$$G(\xi,\eta) = \frac{1}{2}\int_0^\eta \left(a\left(\frac{\tau}{2}\right)+\lambda\right)d\tau + \frac{1}{2}\int_0^\eta\int_0^\tau \left(a\left(\frac{\tau-s}{2}\right)+\lambda\right)G(\tau,s)dsd\tau + \frac{1}{4}\int_\eta^\xi \left(a\left(\frac{\tau}{2}\right)+\lambda\right)d\tau + \frac{1}{4}\int_\eta^\xi\int_0^\eta \left(a\left(\frac{\tau-s}{2}\right)+\lambda\right)G(\tau,s)dsd\tau. \tag{3.9}$$

As in the proof of Lemma 2.2, by the method of successive approximations we can show that this equation has a unique continuous solution. Moreover, it follows from (3.9) that G is twice continuously differentiable because $a \in C^1[0,1]$. □

Similar to Lemma 2.4, we have the following lemma.

LEMMA 3.3. *Let $k(x,y)$ be the solution of problem* (3.1) *and define the linear bounded operator $K : H^i(0,1) \to H^i(0,1)$ $(i = 0,1,2)$ by*

$$w(x) = (Ku)(x) = u(x) + \int_0^x k(x,y)u(y)dy \quad \text{for } u \in H^i(0,1).$$

Then

1. *K has a linear bounded inverse $K^{-1} : H^i(0,1) \to H^i(0,1)$ $(i = 0,1,2)$, and*
2. *K converts the system* (3.2) *and* (3.4) *and the system* (3.3) *and* (3.4) *into*

$$\begin{cases} w_t = w_{xx} - \lambda w & \text{in } (0,1)\times(0,\infty), \\ w_x(0,t) = w(1,t) = 0 & \text{in } (0,\infty), \\ w(x,0) = w^0(x) & \text{in } (0,1) \end{cases}$$

or

$$\begin{cases} w_t = w_{xx} - \lambda w & \text{in } (0,1)\times(0,\infty), \\ w_x(0,t) = w_x(1,t) = 0 & \text{in } (0,\infty), \\ w(x,0) = w^0(x) & \text{in } (0,1), \end{cases}$$

respectively, where $w^0(x) = u^0(x) + \int_0^x k(x,y)u^0(y)dy$.

4. Remarks. An interesting problem is to stabilize the problem

$$u_t(x,t) = u_{xx}(x,t) + a(x,t)u(x,t),$$

where the function a depends on t. To address the problem, it can been seen from the computations in (2.16)–(2.19) that we have to consider the problem

$$\begin{cases} k_{xx}(x,y,t) - k_{yy}(x,y,t) - k_t(x,y,t) = (a(y,t)+\lambda)k(x,y,t), & 0 \le y \le x \le 1, \\ k_y(x,0,t) = 0, & 0 \le x \le 1, \\ k_x(x,x,t) + k_y(x,x,t) + \frac{\partial}{\partial x}(k(x,x,t)) = a(x,t) + \lambda, & 0 \le x \le 1, \end{cases}$$

where λ is any constant. But we do not know if this problem has a solution. Once we can show that this problem has a solution, all the results in sections 2 and 3 hold immediately.

Acknowledgment. The author thanks the referee for bringing Balogh and Krstic [3, 4] to his attention.

REFERENCES

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

[2] H. Amamm, *Feedback stabilization of linear and semilinear parabolic systems*, in Semigroup Theory and Applications, P. Clement, S. Invernizzi, E. Mitidieri, and I. I. Vrabie, eds., Lecture Notes in Pure and Appl. Math. 116, Marcel Dekker, New York, 1989, pp. 21–57.

[3] A. Balogh and M. Krstic, *Infinite-step backstepping for a heat equation-like PDE with arbitrarily many unstable eigenvalues*, in Proceedings of the 2001 American Control Conference, Vol. 3, IEEE Press, Piscataway, NJ, 2001, pp. 2480–2485.

[4] A. Balogh and M. Krstic, *Infinite dimensional backstepping—style feedback transformations for a heat equation with an arbitrary level of instability*, European J. Control, 8 (2002), pp. 165–176.

[5] D. M. Boskovic, M. Krstic, and W. J. Liu, *Boundary control of an unstable heat equation via measurement of domain-averaged temperature*, IEEE Trans. Automat. Control, 46 (2001), pp. 2022–2028.

[6] J. A. Burns and D. Rubio, *A distributed parameter control approach to sensor location for optimal feedback control of thermal processes*, in Proceedings of the 36th IEEE Conference on Decision and Control, Vol. 3, IEEE Press, Piscataway, NJ, 1998, pp. 2243–2247.

[7] J. A. Burns, D. Rubio, and B. B. King, *Regularity of feedback operators for boundary control of thermal processes*, in Proceedings of the First International Conference on Nonlinear Problems in Aviation and Aerospace, Daytona Beach, FL, 1996.

[8] W. A. Day, *A decreasing property of solutions of parabolic equations with applications to thermoelasticity*, Quart. Appl. Math., 40 (1982/83), pp. 468–475.

[9] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasi-linear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.

[10] I. Lasiecka and R. Triggiani, *Stabilization of Neumann boundary feedback of parabolic equations: The case of trace in the feedback loop*, Appl. Math. Optim., 10 (1983), pp. 307–350.

[11] I. Lasiecka and R. Triggiani, *Stabilization and structural assignment of Dirichlet boundary feedback parabolic equations*, SIAM J. Control Optim., 21 (1983), pp. 766–803.

[12] I. Lasiecka and R. Triggiani, *The regulator problem for parabolic equations with Dirichlet boundary control.* I. *Riccati's feedback synthesis and regularity of optimal solution*, Appl. Math. Optim., 16 (1987), pp. 147–168.

[13] I. Lasiecka and R. Triggiani, *The regulator problem for parabolic equations with Dirichlet boundary control.* II. *Galerkin approximation*, Appl. Math. Optim., 16 (1987), pp. 187–216.

[14] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. I, Springer-Verlag, Berlin, Heidelberg, New York, 1972.

[15] R. Triggiani, *Boundary feedback stabilization of parabolic equations*, Appl. Math. Optim., 6 (1980), pp. 201–220.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 1044–1055

ON REGULARITY OF TRANSPORT DENSITY IN THE MONGE–KANTOROVICH PROBLEM*

GIUSEPPE BUTTAZZO† AND EUGENE STEPANOV†

Abstract. We show that the optimal regularity result for the transport density in the classical Monge–Kantorovich optimal mass transport problem, with the measures having summable densities, is a Sobolev differentiability along transport rays.

Key words. Monge–Kantorovich transport problem, transport density, regularity

AMS subject classifications. 49Q20, 49N60, 49Q10

DOI. 10.1137/S036301290138568X

1. Introduction. Let φ^+, φ^- be some given finite positive Borel measures in $\mathbf{R}^n$ satisfying $\varphi^+(\mathbf{R}^n) = \varphi^-(\mathbf{R}^n)$. We assume $\operatorname{supp}\varphi^\pm = X^\pm \subset \Omega$ with $\Omega \subset \mathbf{R}^n$ some open bounded convex set and let $||\cdot||$ stand for the Euclidean norm in $\mathbf{R}^n$. The classical Monge–Kantorovich problem consists of finding a *transport map* ψ_{opt}: $X^+ \to X^-$ minimizing the functional

$$MK(\psi) := \int_{X^+} ||x - \psi(x)||\, d\varphi^+(x)$$

over all Borel measurable maps ψ: $X^+ \to X^-$ satisfying $\varphi^- = \psi_\# \varphi^+$, where $\psi_\#$ denotes the push forward operator acting on every Borel measure α according to the formula

$$\psi_\#\alpha(B) := \alpha(\psi^{-1}(B)) \text{ for all Borel } B \subset \Omega.$$

Although this problem may have no solutions, its relaxed setting always has ones. The latter setting consists of finding a Borel measure γ over $\Omega \times \Omega$ (called *optimal plan*) satisfying $\pi^\pm_\#\gamma = \varphi^\pm$, $\pi^\pm$: $\Omega \times \Omega \to \Omega$ being the projections on the first and second factor, respectively (i.e., $\pi^\pm(x^+, x^-) := x^\pm$), which minimizes the functional

$$MK_0(\gamma) := \int_{\Omega\times\Omega} ||x - y||\, d\gamma(x, y).$$

It is well known that an important role in the Monge–Kantorovich problem is played by the transport density μ, which is the measure defined by

$$\mu(B) := \int_{\Omega\times\Omega} \mathcal{H}^1((x, y) \cap B)\, d\gamma(x, y) \tag{1}$$

for all Borel $B \subset \Omega$, where γ is an optimal plan and $\mathcal{H}^k$ stands for the k-dimensional Hausdorff measure. Its physical meaning is the work for transporting the mass through the set B. It has been proven in [1] and in [11] (see also [5]) that when $\varphi^\pm \ll dx$, dx

*Received by the editors February 28, 2001; accepted for publication (in revised form) January 28, 2003; published electronically July 8, 2003.
http://www.siam.org/journals/sicon/42-3/38568.html

†Dipartimento di Matematica, Università di Pisa, via Buonarroti 2, 56127 Pisa, Italy (buttazzo @dm.unipi.it, e.stepanov@sns.it). The work of the first author is part of the European Research Training Network "Homogenization and Multiple Scales" (contract HPRN-2000-00109).

standing for the Lebesgue measure, then also $\mu \ll dx$, and what is more important, μ is unique; that is, it depends only on $\varphi^{\pm}$ but not on a particular choice of the optimal plan γ (which on the contrary might not be unique). In this case we will denote by $f^{\pm}$ the density of $\varphi^{\pm}$ with respect to the Lebesgue measure and by a the density of μ with respect to the Lebesgue measure.

In this paper we study the higher regularity (namely, differentiability) of the transport density in the Monge–Kantorovich problem. We show that unless the transport problem is one-dimensional, then the optimal result one can get is a Sobolev regularity of restrictions of the density to the transport directions.

2. Preliminaries: dual problem and transport rays. It is well known that the dual setting for the Monge–Kantorovich problem consists of finding a $u \in W^{1,\infty}(\Omega)$ (called *transport potential*) which maximizes the functional

$$MK'(v) := \int_{X^+} v \, d\varphi^+(x) - \int_{X^-} v \, d\varphi^-(x)$$

over all $v \in \mathrm{Lip}_1(\Omega)$, where $\mathrm{Lip}_1(\Omega)$ stands for the set of Lipschitz continuous functions on Ω with Lipschitz constant one. We call a *transport ray* the maximal open (i.e., not containing the endpoints) interval $(x, y) := \{z \in \Omega \,:\, z = tx + (1-t)y,\ 0 < t < 1\}$ which satisfies

$$|u(x) - u(y)| = ||x - y||.$$

In other words, a transport ray is a maximal line segment along which the transport potential is decreasing at a maximal rate equal to one. Given a transport potential u, we denote by T the transport set (the union of all the transport rays without ray ends).

We recall the basic construction of the proof of existence of the optimal transport map from [4]. If u is a transport potential and R is a transport ray, then for every $p \in R$ the set

$$\{x \in \Omega \,:\, u(x) = p \text{ and } \exists \nabla u(x) \neq 0\}$$

admits a Borel covering $\{S_{pi}\}_{i=1}^{\infty}$ such that there exist Lipschitz coordinates $U\colon \mathbf{R}^n \to \mathbf{R}^{n-1}$ and $V\colon \mathbf{R}^{n-1} \to \mathbf{R}^n$ satisfying $V(U(x)) = x$ for all $x \in S_{pi}$. Enumerating the triples $(p, i, j) \in \mathbf{Q} \times \mathbf{N}^2$ and ordering them in the order of enumeration, we can define the countable sequence of disjoint clusters of transport rays T_{pij} as follows:

$$T_{pij} := \bigcup \{R = (a, b) \subset \Omega \,:\, R \cap S_{pi} = \{z\}, \min\{||z - a||, ||z - b||\} > 1/j\} \setminus \bigcup_{(p'i'j') < (p,i,j)} T_{p'i'j'}.$$

Each cluster T_{pij} admits coordinates $G := G_{pij}\colon T_{pij} \to \mathbf{R}^{n-1} \times \mathbf{R}$ (with Lipschitz inverse F), which are Lipschitz over

$$T^{\sigma}_{pij} := \{x \in T_{pij} \,:\, \min\{||z - a||, ||z - b||\} \geq \sigma > 0\},$$

where a and b stand for the ends of the unique transport ray R_x passing through x. Moreover, one has

$$G(x) = (U(z), u(x) - u(z)), \text{ where } \{z\} := R_x \cap S_{pi}.$$

In what follows we will make use of the following lemmata.

LEMMA 1. *Let $N \subset \mathbf{R}^n$ be a set with Lebesgue measure zero. Then for a.e. $x \in T$ one has $\mathcal{H}^1(R_x \cap N) = 0$.*

Proof. Observe that $\mathcal{H}^n(G(T_{pij}^{1/l} \cap N)) = 0$ for every $(p,i,j) \in \mathbf{Q} \times \mathbf{N}^2$ and $l \in \mathbf{N}$; hence by the Fubini theorem $\mathcal{H}^1(G(R_x \cap T_{pij}^{1/l}) \cap G(N)) = 0$ for a.e. $x \in T_{pij}$. Making the countable union over all $l \in \mathbf{N}$ we obtain $\mathcal{H}^1(G(R_x) \cap G(N)) = 0$ and hence $\mathcal{H}^1(R_x \cap N) = 0$ for a.e. $x \in T_{pij}$, which concludes the proof. □

LEMMA 2. *Let $\rho \in L^p_{loc}(\Omega)$. Then for a.e. $x \in T$ one has $\rho \in L^p_{loc}(R_x)$.*

Proof. Let $\rho \in L^p_{loc}(\Omega)$. The area formula gives

$$\int_{T_{pij}} \rho^p(x)\,dx = \int_{G(T_{pij})} \rho^p(F(x'))J_nF(x')\,dx',$$

where J_nF stands for the Jacobian of F. Since over $G(T^{\sigma}_{pij})$ one has $J_nF \geq c > 0$, then $\rho(F(\cdot)) \in L^p_{loc}(G(T_{pij}))$, and hence by the Fubini theorem

$$\rho(F(z,\cdot)) \in L^p_{loc}(G(T_{pij}) \cap (\{z\} \times \mathbf{R}))$$

for $\mathcal{H}^{n-1}$-a.e. $z \in \mathbf{R}^{n-1}$. This means $\rho \in L^p_{loc}(R_z)$ for $\mathcal{H}^{n-1}$-a.e. $z \in S_{pi}$, which implies the result. □

LEMMA 3. *Let $D \subset \Omega \times \Omega$ be a Borel set and γ_D stand for the restriction to D of the optimal transport plan γ. Then γ_D is an optimal plan for the Monge–Kantorovich problem of transporting the measure $\varphi^+_D := \pi^+_{\#}\gamma_D$ to $\varphi^-_D := \pi^-_{\#}\gamma_D$.*

Proof. Supposing the contrary we would have the existence of a better plan γ'_D with the same marginals (i.e., $\pi^{\pm}_{\#}\gamma'_D = \varphi^{\pm}_D$), which would then give rise to the better plan γ' for the original problem, according to the relationship

$$\gamma'(e) := \gamma'_D(e \cap D) + \gamma(e \setminus D)$$

for all Borel $e \subset \Omega \times \Omega$, which contradicts the optimality of γ. □

3. Nondifferentiability. It is a simple exercise to show that for one-dimensional transport (i.e., $n = 1$) the transport density has certain regularity properties. Namely, one has the following simple result.

PROPOSITION 1. *Let $\Omega \subset \mathbf{R}$ be a bounded interval. Then $f^{\pm} \in L^p(\Omega)$ implies $a \in W^{1,p}(\Omega)$. Moreover, $f^{\pm} \in W^{k,p}(\Omega)$ implies $a \in W^{k+1,p}(R)$ for every transport ray $R \subset \Omega$.*

Proof. According to [3], the pair (a,ν), where $a > 0$ with support in Ω is the transport density and ν is the unit direction of transport ray, solves the system

$$\begin{cases} -(a\nu)' &= f(x) \text{ in } \Omega, \\ |\nu(x)| &\leq 1 \text{ in } \Omega, \\ |\nu(x)| &= 1 \text{ a.e. where } a(x) > 0, \end{cases}$$

where $f := f^+ - f^-$. Let F stand for the primitive of f which has support in Ω. Clearly then, $a(x) := |F(x)|$, $\nu(x) := -\operatorname{sign} F(x)$ satisfies the above system, thus proving the first claim. To prove the second claim, observe that over each transport ray R (which is in this case just an open interval), either $\nu(x) \equiv 1$ or $\nu(x) \equiv -1$, and hence either $a(x) = F(x)$ or $a(x) = -F(x)$. □

Remark. It is clear from the proof that in the one-dimensional case one always has $a = 0$ in the ray ends.

Let us now show that, generally speaking, the result of Proposition 1 cannot hold for $n \geq 2$.

Example 1. Let $n = 2$ and

$$X^+ = \{(x, y) : 0 \leq x \leq g(y), 0 \leq y \leq L\}$$

for some bounded positive Borel function g, while $f^-(x, y) = f^+(x - b, y)$ (assuming b to be sufficiently large so that $X^+ \cap X^- = \emptyset$), so that

$$X^- = \{(x, y) : b \leq x \leq g(y) + b, 0 \leq y \leq L\}.$$

In this case the Monge–Kantorovich problem admits an exact solution $\psi'(x, y) = (x + b, y)$. In fact, setting $u(x, y) = -x$ in the dual problem, we obtain that the supremum of the latter is less than or equal to $b \int_{X^+} f^+(z)\, dz$, that is, the value $MK(\psi')$, whence the optimality follows. Now we can calculate the transport density a as

$$a(x, y) = \begin{cases} \displaystyle\int_0^x f^+(\xi, y)\, d\xi, & 0 \leq x \leq g(y), \\ \displaystyle\int_0^{g(y)} f^+(\xi, y)\, d\xi, & g(y) < x \leq b, \\ \displaystyle\int_{x-b}^{g(y)} f^+(\xi, y)\, d\xi, & b \leq x \leq g(y) + b \end{cases} = \int_{\max(x-b,0)}^{\min(x,g(y))} f^+(\xi, y)\, d\xi;$$

in particular,

$$a(x, y) = \int_0^{g(y)} f^+(\xi, y)\, d\xi \text{ in } \mathbf{R}^2 \setminus (X^+ \cup X^-).$$

Note that the above formula is valid, generally speaking, for a.e. $y \in [0, L]$. If one chooses now $g(y) = \text{const}$ and f^+ independent of x, namely $f^+ = f^+(y)$ such that $f^+ \in L^p(0, L)$, but $f^+ \notin L^{p+\alpha}(0, L)$, for all $\alpha > 0$, then clearly $a \notin L^{p+\alpha}(\Omega)$. In particular, this means, in view of the Sobolev imbedding theorem, that $a \notin W^{1,p}(\Omega)$, and hence Proposition 1 is not valid.

We remark that the same example but with nonconstant and discontinuous g shows that, generally speaking, even $f^\pm \in C^\infty(X^\pm)$ does not imply $a \in C(\Omega)$ or $a \in W^{1,1}(\Omega)$. Here the regularity of the solution is spoiled by the geometry of the problem (say, by the discontinuity of g).

Note that it has been shown in [5] that when $f^\pm \in L^p(X^\pm)$, $1 \leq p \leq +\infty$, while X^+ and X^- are disjoint, then $\mu \ll dx$ and for the respective density one has $a \in L^p(\Omega)$. The above example shows also that such a result cannot be improved.

4. Differentiability along transport rays. In spite of the above discouraging example, one can still claim that some regularization occurs in the direction of the transport rays. In fact, in [8] it has been proven that when $f^\pm$ are Lipschitz continuous with disjoint supports (and with some extra technical condition on the supports), then also the transport density is Lipschitz continuous along transport rays. We have the following more general result for the case of just summable $f^\pm$ without any extra conditions on supports.

THEOREM 1. *Let $f^\pm \in L^p(X^\pm)$. Then for a.e. $x \in T$ one has $a \in W^{1,p}_{loc}(R_x)$, where R_x is the transport ray passing through x.*

Proof. The result will be obtained in several steps.

Step 1. For each triple $(p,i,j) \in \mathbf{Q} \times \mathbf{N}^2$ we consider the restriction γ_{pij} of the optimal transport plan γ to the ray cluster $T_{pij} \times \Omega$, where T_{pij} is a ray cluster. It is clearly optimal for the Monge–Kantorovich problem of transporting the measure $\varphi^+_{pij} := \pi^+_{\#}\gamma_{pij}$ to the measure $\varphi^-_{pij} := \pi^-_{\#}\gamma_{pij}$. Note that the measures $\varphi^\pm_{pij}$ are just the restrictions of $\varphi^\pm$ to T_{pij}. Then the transport set for each such new problem is T_{pij}, while the respective transport density a_{pij} satisfies

$$-\operatorname{div} a_{pij}\nu = f_{pij}$$

weakly in Ω, where $f_{pij} := f^+_{pij} - f^-_{pij}$, $f^\pm_{pij}$ being the restrictions of $f^\pm$ to T_{pij}, and ν is the unit direction of the transport ray (both for the original and for the new problem). Note that $a = a_{pij}$ over T_{pij}; hence it is enough to prove the assertion of the theorem for each T_{pij}, but since in the latter problem the vector field ν varies Lipschitz continuously, it is enough to assume that in the original problem ν is Lipschitz continuous.

Step 2. In view of the Lemma 1, we may assume that $\mathcal{H}^1$-a.e. $z \in R_x$ is a Lebesgue point of the density a (this is true for a.e. $x \in T$). We also assume that the first Cartesian coordinate axis is directed along R_x (in the direction of transport), and, moreover, the origin of the coordinate system coincides with the upper end of R_x. Hence, the lower end of R_x is le_1, where e_1 is the unit vector of the first coordinate direction and $l > 0$ is the length of R_x. For $z \in R_x$ we denote by $B'_\varepsilon(z) \subset \mathbf{R}^{n-1}$ the $(n-1)$-dimensional ball in the plane perpendicular to R_x, which has radius $\varepsilon > 0$ and is centered at z. Let $\phi\colon \mathbf{R}^{n-1} \to \mathbf{R}$ be a smooth positive function with

$$\int_{\mathbf{R}^{n-1}} \phi(z')\,dz' = 1, \qquad B'_{1/2}(0) \subset \operatorname{supp}\phi \subset\subset B'_1(0).$$

Denote then

$$\begin{aligned} a_\varepsilon(z_1) &:= \frac{1}{2\varepsilon}\int_{z_1-\varepsilon}^{z_1+\varepsilon}\int_{B'_\varepsilon(0)} a(z_1,z')\phi_\varepsilon(z')\,d\mathcal{H}^{n-1}(z')d\mathcal{H}^1(z_1),\\ f_\varepsilon(z_1) &:= \frac{1}{2\varepsilon}\int_{z_1-\varepsilon}^{z_1+\varepsilon}\int_{B'_\varepsilon(0)} f(z_1,z')\phi_\varepsilon(z')\,d\mathcal{H}^{n-1}(z')d\mathcal{H}^1(z_1), \end{aligned}$$

where

$$\phi_\varepsilon(z') := \frac{1}{\varepsilon^n}\phi\left(\frac{z'}{\varepsilon}\right) \text{ and } f := f^+ - f^-.$$

Now we follow the lines of Steps 2–3 of Proposition 6.1 from [8]. In fact, using $\phi_\varepsilon\psi$, where $\psi\colon [\sigma, l-\sigma] \to \mathbf{R}$ is a positive Lipschitz function with compact support and $\sigma > 0$ is sufficiently small, as a test function to the equation

$$-\operatorname{div} a\nu = f \text{ in } \Omega,$$

where ν is the unit direction of the transport ray, we obtain

$$\begin{aligned} \int_0^{l-\sigma}\psi(z_1)\tilde f_\varepsilon(z_1)d\mathcal{H}^1(z_1) &= -\int_\sigma^{l-\sigma}\psi'(z_1)\tilde a_\varepsilon(z_1)d\mathcal{H}^1(z_1)\\ &+\int_\sigma^{l-\sigma}\psi'(z_1)\int_{B'_\varepsilon(0)} a(z_1,z')(\nu(z_1,z)+e_1)\cdot e_1\phi_\varepsilon(z')\,d\mathcal{H}^{n-1}(z')d\mathcal{H}^1(z_1)\\ &+\int_\sigma^{l-\sigma}\psi(z_1)\int_{B'_\varepsilon(0)} a(z_1,z')\nu(z_1,z)\cdot\nabla\phi_\varepsilon(z')\phi_\varepsilon(z')\,d\mathcal{H}^{n-1}(z')d\mathcal{H}^1(z_1), \end{aligned} \tag{2}$$

where

$$\begin{aligned}\tilde{f}_\varepsilon(z_1) &:= \int_{B'_\varepsilon(0)} f(z_1,z')\phi_\varepsilon(z')\,d\mathcal{H}^{n-1}(z'),\\ \tilde{a}_\varepsilon(z_1) &:= \int_{B'_\varepsilon(0)} a(z_1,z')\phi_\varepsilon(z')\,d\mathcal{H}^{n-1}(z').\end{aligned}$$

Since on Step 1 of the proof we supposed that ν is Lipschitz continuous, one can estimate

$$||\nu(z_1,z')+e_1|| = ||\nu(z_1,z')+\nu(z_1,0)|| \le C||z'|| \le C\varepsilon$$

for small $\varepsilon>0$, where $C=C(\sigma)$. This means that the second term on the right-hand side of (2) can be estimated as

$$\begin{aligned}&\left|\int_\sigma^{l-\sigma}\psi'(z_1)\int_{B'_\varepsilon(0)} a(z_1,z')(\nu(z_1,z)+e_1)\cdot e_1\phi_\varepsilon(z')\,d\mathcal{H}^{n-1}(z')d\mathcal{H}^1(z_1)\right|\\ &\qquad\le C\varepsilon\int_\sigma^{l-\sigma}\psi'(z_1)\tilde{a}_\varepsilon(z_1)\,d\mathcal{H}^1(z_1).\end{aligned}$$

Analogously, the third term on the right-hand side of (2) can be estimated as

$$\begin{aligned}&\left|\int_\sigma^{l-\sigma}\psi(z_1)\int_{B'_\varepsilon(0)} a(z_1,z')\nu(z_1,z)\cdot\nabla\phi_\varepsilon(z')\phi_\varepsilon(z')\,d\mathcal{H}^{n-1}(z')d\mathcal{H}^1(z_1)\right|\\ &\qquad\le C\int_\sigma^{l-\sigma}\psi(z_1)\tilde{a}_{2\varepsilon}(z_1)\,d\mathcal{H}^1(z_1).\end{aligned}$$

Combining (2) with the above estimates and choosing ψ as in Step 4 of Proposition 6.1 from [8], we arrive at the relationship

$$\left|a_\varepsilon(s)-a_\varepsilon(t)+\int_s^t f_\varepsilon(s)\,d\mathcal{H}^1(s)\right| \le C\int_s^t a_{2\varepsilon}(s)\,d\mathcal{H}^1(s)+C\varepsilon \tag{3}$$

with some positive constant $C>0$ for all sufficiently small $\varepsilon>0$, whenever $\sigma<t<s<l-\sigma$.

Step 3. For an $m>0$ let $X_m^+ := \{f^+\le m\}$ and

$$f_m^+(x) := \begin{cases} f^+(x), & x\in X_m^+,\\ 0, & \text{elsewhere.}\end{cases}$$

If γ is an optimal plan, denote by γ_m the restriction of γ to $X_m^+\times X^-$, and set $\varphi_m^- := \pi_\#^-\gamma_m$, f_m^- denoting the density of the latter with respect to the Lebesgue measure. According to Lemma 3, γ_m is optimal for the mass optimization problem (MK_m) with respect to the measures $\varphi_m^\pm := f_m^\pm\,dx$. For each of such problems denote $a_m\in L^1(\Omega)$ the respective transport density. In the same way let f_{mk}^- stand for the restriction of f_m^- to the set $X_{mk}^- := \{f_m^-\le k\}$, $k>0$. Letting γ_{mk} stand for the restriction of γ_m to $X_m^+\times X_{mk}^-$, f_{mk}^+ stand for the density of $\varphi_{mk}^+ := \pi_\#^-\gamma_{mk}$ with respect to the Lebesgue measure, and $a_{mk}\in L^1(\Omega)$ stand for the transport density for the Monge–Kantorovich problem (MK_{mk}) with respect to the measures $\varphi_{mk}^\pm := f_{mk}^\pm\,dx$ (for which in view of Lemma 3 γ_{mk} is optimal), we see that the directions of the transport rays for each problem (MK_{mk}) coincide with those of the original problem;

namely, for each transport ray R^{mk} for the former there is a ray R for the latter such that $R^{mk} \subset R$. Hence one concludes that

$$-\mathrm{div}\,(a_{mk}\nu) = f_{mk} \text{ in } \Omega,$$

where $f_{mk} := f^+_{mk} - f^-_{mk}$. From (3) we derive

$$\left| a_{mk,\varepsilon}(s) - a_{mk,\varepsilon}(t) + \int_s^t f_{mk,\varepsilon}(s)\, d\mathcal{H}^1(s) \right| \le C \int_s^t a_{mk,2\varepsilon}(s)\, d\mathcal{H}^1(s) + C\varepsilon. \tag{4}$$

According to our choice of ϕ_ε we have that

$$a_{mk,\varepsilon}(s) \to a_{mk}(s,0) \text{ and } f_{mk,\varepsilon}(s) \to f_{mk}(s,0)$$

as $\varepsilon \to 0$. Moreover, the assertion (ii) of Lemma 4 shows that $a_{mk} \in L^\infty(\Omega)$; hence one has

$$a_{mk,\varepsilon}(s) \le C', \qquad f_{mk,\varepsilon}(s) \le C',$$

where the constant $C' = C'(m,k) > 0$ is independent of $\varepsilon > 0$. Thus one can pass to the limit in (4) as $\varepsilon \to 0$ obtaining

$$\left| a_{mk}(s,0) - a_{mk}(t,0) + \int_s^t f_{mk}(\xi,0)\, d\mathcal{H}^1(\xi) \right| \le C \int_s^t a_{mk}(\xi,0)\, d\mathcal{H}^1(\xi). \tag{5}$$

Step 4. Note that $a_{mk}(\cdot)$ is nondecreasing with respect to k, a_m is also nondecreasing, and $a_{mk} \le a_m \le a$. Moreover, since as $k \to \infty$, $m \to \infty$, the sets $X^+_{mk} \uparrow \Omega$, then $a_{mk} \to a$ a.e. on Ω. The same is true also for the sequence $\{f_{mk}\}$. We may assume that the ray R_x is chosen so that the latter convergences occur also $\mathcal{H}^1$-a.e. on R_x. Now, since $f \in L^1(\Omega)$ and $a \in L^1(\Omega)$ by the assertion (i) of Lemma 4, then one can pass to the consecutive limits as $m \to +\infty$ and as $k \to +\infty$ in (5), arriving at the inequality

$$\left| a(s,0) - a(t,0) + \int_s^t f(\xi,0)\, d\mathcal{H}^1(\xi) \right| \le C \int_s^t a(\xi,0)\, d\mathcal{H}^1(\xi).$$

The latter expression means

$$|a'_s(s,0) + f(s,0)| \le C a(s,0)$$

whenever $\sigma \le s \le l - \sigma$. Recall now that $a \in L^1(\Omega)$ by assertion (i) of Lemma 4, while $f \in L^p(\Omega)$ according to the conditions of the theorem being proven. Since according to Lemma 2 we may assume that $a(\cdot,0) \in L^1(\sigma, l-\sigma)$ and $f(\cdot,0) \in L^p(\sigma, l-\sigma)$, then the above inequality implies $a'_s(\cdot,0) \in L^1(\sigma, l-\sigma)$; hence $a(\cdot,0) \in L^\infty(\sigma, l-\sigma)$. Again using the same inequality one concludes $a'_s(\cdot,0) \in L^p(\sigma, l-\sigma)$, which shows the statement. □

LEMMA 4. *In the Monge–Kantorovich problem one has that*

(i) *if either $f^+ \in L^1(X^+)$ or $f^- \in L^1(X^-)$, then $a \in L^1(\Omega)$;*
(ii) *if $f^\pm \in L^\infty(X^\pm)$, then $a \in L^\infty(\Omega)$;*
(iii) *if, moreover, $X^+ \cap X^- = \emptyset$, then $f^+ \in L^\infty(X^+)$ (resp., $f^- \in L^\infty(X^-)$) implies $a \in L^\infty_{loc}(\Omega \setminus X^-)$ (resp., $a \in L^\infty_{loc}(\Omega \setminus X^+)$).*

The proof of this assertion can be found in [5].

We can also refine the above result in the following way.

THEOREM 2. *Let $f^+ \in L^{p^+}(X^+)$ (resp., $f^- \in L^{p^-}(X^-)$) and $X^+ \cap X^- = \emptyset$. Then for a.e. $x \in T$ one has $a \in W^{1,p^+}_{loc}(R_x \backslash X^-)$ (resp., $a \in W^{1,p^-}_{loc}(R_x \backslash X^+)$), where R_x is the transport ray passing through x.*

The proof just follows the scheme of that of Theorem 2 but uses the assertion (iii) of Lemma 4 rather than the assertion (ii).

5. Behavior of density at the ends of transport rays. It is also interesting to study the behavior of the transport density near the ray ends. In fact, the result on vanishing of transport density at the ends of transport rays for Lipschitz continuous $f^{\pm}$ was extensively used in [8] in the proof of existence of optimal transport maps via the ODE approach, as well as in [7, 10, 9] for deriving the law of evolution of a sandpile shape. In general the following result holds.

THEOREM 3. *Let $f^+ \in L^\infty(X^+)$ (resp., $f^- \in L^\infty(X^-)$). Then for a.e. $x \in T$ one has $a(z) \to 0$ when $z \to l^+$ (resp., $z \to l^-$) while $z \in R_x$, where l^+ and l^- are the upper and lower ends, respectively, of the transport ray R_x passing through x.*

Proof. As in the proof of Theorem 1, it is enough to consider the restrictions of a to each of the ray clusters T_{pij}. These restrictions in fact are transport densities for the Monge–Kantorovich problem with respect to the restrictions of $\varphi^{\pm}$ to T_{pij}. In the remaining part of the proof we will assume a ray cluster T_{pij} to be chosen and will be dealing with the respective restrictions without explicitly referring to the cluster.

Let $x \in T$ be such a point that both a and $f^{\pm}$ are continuous in the generalized (Lebesgue) sense $\mathcal{H}^1$-a.e. on R_x. (A.e. $x \in T$ will suit that purpose according to Lemma 1). Suppose $f^+ \in L^\infty(X^+)$. For every $z \in T$ denote by $d(z)$ the distance from z to the upper end of R_z. Suppose now $z \in R_x$, $d(z) = \theta$. Since d is an upper semicontinuous function according to Lemma 24 from [4], then for each sufficiently small $\varepsilon > 0$ one has $d(y) \le 2\theta$ for all $y \in B_\varepsilon(z) \cap T$, where $B_\varepsilon(z) \subset \mathbf{R}^n$ stands for the ball with radius ε centered at z.

We also remark that there exists a $\delta > 0$ such that

$$||l^+ - L^-|| \ge \delta \tag{6}$$

for all lower ray ends L^-. In fact, since we are working within the ray cluster T_{pij}, we have $u(l^+) \le p - 1/j$. If (6) were false, there would exist a sequence of lower ray ends $\{L^-_\nu\}$ such that $L^-_\nu \to l^+$. But $u(L^-_\nu) \ge p + 1/j$ by construction of the ray cluster, and hence, in view of continuity of u, one would have $u(l^+) \ge p + 1/j$ leading to a contradiction.

We use now the following slightly refined version of Lemma 4.6 from [5].

LEMMA 5. *Let l^+ and l^- be the upper and lower ends of the transport ray R_x, respectively, $z \in R_x$, and $\varepsilon > 0$ be sufficiently small. If (6) holds for some $\delta > 0$ and for all lower ray ends L^-, then the upper end of every transport ray R_y which intersects $B_\varepsilon(z)$ belongs to the cylinder with axis R_x and cross-sectional area $C\varepsilon^{n-1}/\delta^{n-1}$, where $C > 0$ is a constant depending only on the geometry of the problem.*

Applying the above lemma to our situation, one has that for $\varepsilon > 0$ sufficiently small, all the upper ends of transport rays R_y which intersect $B_\varepsilon(z)$ necessarily belong to the piece $C_{\theta,\varepsilon}$ of the cylinder mentioned in the lemma, with the length 4θ and with z being the middle point of its axis. From (1) one derives

$$\mu(B_\varepsilon(z)) \le 2\varepsilon \int_{C_{\theta,\varepsilon} \times \Omega} d\gamma = 2\varepsilon \varphi^+(C_{\theta,\varepsilon}).$$

Dividing the above relationship by ε^n and taking the limit as $\varepsilon \to 0$, we obtain $a(z) \le C'\theta$ for some $C' > 0$, which shows the statement for $f^+ \in L^\infty(X^+)$. The case $f^- \in L^\infty(X^-)$ is symmetric. □

Let us remark that the transport density may not vanish at the end of the transport ray when $f^+ \in L^p(X^+)$ with $1 \le p < \infty$, as the following example shows.

Example 2. Let $X^+ \subset \mathbf{R}^n$ be the unit ball centered at the origin, $f^+(x) := |x|^\alpha$ with $-n/p < \alpha < -1$. Then $f^+ \in L^p(X^+)$. Let also $X^- := \{x \in \mathbf{R}^n : l \le |x| \le L\}$ with $l > 1$ and

$$f^-(x) := \frac{n}{(n+\alpha)(L^n - l^n)} \text{ in } X^-.$$

Then one has $u(x) := -|x|$ and the transport is radial. In particular, for $x \in X^+$ one has

$$a(x) = \frac{1}{n+\alpha}|x|^{\alpha+1},$$

and hence $a(x) \not\to 0$ as $|x| \to 0$.

6. Application to *p*-Laplacian equations. In [6] it was proved that if u_p is a solution to the equation

$$-\text{div}\,(|\nabla u_p|^{p-2}\nabla u_p) = -\text{div}\,(|F|^{p-2}F) \text{ in } \Omega,$$

where $F \in L^s_{loc}(\Omega;\mathbf{R}^n)$, then $\nabla u_p \in L^s_{loc}(\Omega;\mathbf{R}^n)$, and, moreover, for $\Omega_0 \subset\subset \Omega$ one has

$$||\nabla u_p||^s_{s;\Omega_0} \le C^{1/p}(||F||^s_{s;\Omega_0} + ||u_p||^p_{p;\Omega_0}), \tag{7}$$

where $C = C(\Omega_0, p, s)$. (Further, we will omit the reference to Ω_0.) It is, however, important to know whether there is such an estimate with C bounded when $p \to +\infty$, while, say, $1 \le s/p \le \theta < +\infty$. In fact, similar estimates (called "uniform in p") on the gradient of the solution to the p-Laplacian equation play an important role in the study of limiting behavior of such solution as $p \to \infty$ (see, for instance, [2]). Unfortunately, the above discussion plus a known relation between the p-Laplacian equations and the Monge–Kantorovich problem shows that the desired estimates cannot be found.

THEOREM 4. *There is no estimate of the type* (7) *uniform in p (i.e., with C bounded), when* $p \to +\infty$.

Proof. The result will be shown in two steps.

Step 1. Consider $\varphi^\pm \ll dx$ be as in the previous discussion. Let $\{u_p\}_{p\ge 2}$ be the sequence of unique solutions to the problems

$$(\Lambda_p) \qquad \begin{cases} -\text{div}\,(|\nabla u_p|^{p-2}\nabla u_p) = f \text{ in } \Omega, \\ u_p = 0 \text{ on } \partial\Omega, \end{cases}$$

where $f := f^+ - f^-$, and suppose $f \in L^r(\Omega)$, $1 < r \le +\infty$. It is a matter of simple though slightly technical calculations to show that up to subsequences (not relabeled)

(A) $u_p \rightharpoonup u$ weakly in $W^{1,s}_0(\Omega)$ for every $s < +\infty$, while $u \in W^{1,\infty}(\Omega)$, and, moreover, u is a transport potential for the problem of transporting φ^+ to φ^-;

(B) $A_p := |\nabla u_p|^{p-2}\nabla u_p \to \alpha$, where α is a vector measure, $*$-weakly in the sense of measures.

In fact, to show (A), using u_p as a test function to (Λ_p) and applying the Hölder inequality, we get

$$\int_\Omega |\nabla u_p|^p \, dx = \int_\Omega f u_p \, dx \le ||f||_r \cdot ||u_p||_{r'}.$$

By the Sobolev inequality we have

$$||u_p||_{r'} \le C(r', n)||\nabla u_p||_q,$$

where $q = q(r', n)$ is such that $nq/(n-q) = r'$, and $C(r', n) = r'n/(n-1)$. Combining the above estimates and again using the Hölder inequality one arrives at

$$||\nabla u_p||_p^{p-1} \le C(r', n)||f||_r \cdot |\Omega|^{1/q(r',n)-1/p} \tag{8}$$

for p sufficiently large. Fix now an arbitrary $s > 2$. Since by Hölder inequality

$$||\nabla u_p||_s \le ||\nabla u_p||_p \cdot |\Omega|^{1/s-1/p}$$

for $p > s$, and in view of (8), one has that up to a subsequence $u_p \rightharpoonup u$ weakly in $W_0^{1,s}(\Omega)$. Note that also $u \in W_0^{1,t}(\Omega)$ for all $t > s$ and, moreover, using (8) we estimate

$$||\nabla u||_t \le \liminf_p ||\nabla u_p||_t \le |\Omega|^{1/t}.$$

The latter actually means $u \in W^{1,\infty}(\Omega)$ and $||\nabla u||_\infty \le 1$. It remains to note that u_p are minima of the functionals

$$I_p(v) := \frac{1}{p}\int_\Omega |\nabla v|^p \, dx - \int_\Omega v f \, dx$$

defined over some $W_0^{1,s}(\Omega)$ with $s > 1$ fixed. Since the sequence I_p Γ-converge to the functional

$$I(u) := \begin{cases} -\int_\Omega v f \, dx, & |\nabla v| \le 1 \text{ a.e. on } \Omega, \\ +\infty & \text{otherwise,} \end{cases}$$

defined over the same $W_0^{1,s}(\Omega)$, hence u is the minimum of the latter; in other words, it is a transport potential.

To show (B), we estimate by the Hölder inequality

$$||A_p||_1 = ||\nabla u_p||_{p-1}^{p-1} \le ||\nabla u_p||_p^{p-1} \cdot |\Omega|^{1/p}$$

to see that, in view of (8), the sequence $\{A_p\}$ is bounded in $L^1(\Omega; \mathbf{R}^n)$ and hence converges up to a subsequence $*$-weakly in the sense of measures.

Let $\mu := |\alpha|$. We show now $\alpha = \mu \nabla_\mu u$, and, moreover, the pair (μ, u) then satisfies

$$(\Lambda_\infty) \qquad \begin{cases} -\operatorname{div}(\mu \nabla_\mu u) &= f \text{ in } \Omega, \\ |\nabla_\mu u| &= 1 \quad \mu\text{-a.e. in } \Omega, \end{cases}$$

where ∇_μ is the tangential gradient operator with respect to μ defined in [3]. In other words, μ is the transport density of the Monge–Kantorovich problem of transporting

$\varphi^+ := f^+\,dx$ to $\varphi^- := f^-\,dx$. For this purpose recall that for all $\phi \in C_0^\infty(\Omega)$ one has $\int_\Omega A_p \nabla\phi\,dx = \int_\Omega f\phi\,dx$ and $A_p \rightharpoonup \alpha$ $*$-weakly in the sense of measures, which implies

$$\int_\Omega \nabla\phi\,d\alpha = \int_\Omega f\phi\,dx.$$

Consider now the sequence $\{\phi_\nu\} \subset C_0^\infty$ such that $\nabla\phi_\nu \to \nabla_\mu u$ in $L^2(\Omega,\mu;\mathbf{R}^n)$ and $\phi_\nu \to u$ uniformly. Passing to the limit in the above relationship, we have

$$\int_\Omega \nabla_\mu u\,d\alpha = \int_\Omega fu\,dx. \tag{9}$$

Let us estimate $\mu(\Omega) = |\alpha|(\Omega)$. We have

$$\mu(\Omega) \le \liminf_p ||A_p||_1 = \liminf_p ||\nabla u_p||_{p-1}^{p-1} \le |\Omega|^{1/p} \left(\int_\Omega |\nabla u_p|^p\,dx\right)^{(p-1)/p},$$

where to obtain the last estimate the Hölder inequality is used. From (Λ_p) using the convergence of $\{u_p\}$, we obtain

$$\int_\Omega |\nabla u_p|^p\,dx = \int_\Omega fu_p\,dx \to \int_\Omega fu\,dx.$$

Therefore, $\mu(\Omega) \le \int_\Omega fu\,dx$. Then from (9) we get

$$\mu(\Omega) \le \int_\Omega \nabla_\mu u\,d\alpha, \tag{10}$$

while on the other hand, since $||\nabla u||_\infty \le 1$, hence $||\nabla_\mu u||_\infty \le 1$, one has

$$\int_\Omega \nabla_\mu u\,d\alpha \le \mu(\Omega). \tag{11}$$

Thus in (10) and (11) in fact the equality holds, namely,

$$\mu(\Omega) = |\alpha|(\Omega) = \int_\Omega \nabla_\mu u\,d\alpha,$$

which implies $|\alpha| = \alpha \cdot \nabla_\mu u$, that is, $\alpha = |\alpha|\nabla_\mu u = \mu\nabla_\mu u$. Clearly, $|\nabla_\mu u| = 1$ μ-a.e.

Step 2. Since $\varphi^\pm \ll dx$, then $\mu \ll dx$ and (Λ_∞) assumes the form

$$(\Lambda'_\infty) \qquad \begin{cases} -\operatorname{div}(a\nabla u) &= f \text{ in } \Omega, \\ |\nabla u| &= 1 \text{ a.e. where } a \ne 0. \end{cases}$$

Consider now a function v such that

$$-\Delta v = f \text{ in } \Omega.$$

Clearly, since $f \in L^r(\Omega)$, such a function exists and the elliptic regularity theory implies $u \in W^{2,r}(\Omega)$. Define the vector field F over Ω by

$$F(x) := |\nabla v(x)|^{1/(p-1)} \frac{\nabla v(x)}{|\nabla v(x)|}$$

with $F(x) = 0$ whenever $\nabla v(x) = 0$. Since by the Sobolev embedding theorem $\nabla v \in L^{r^*}(\Omega; \mathbf{R}^n)$, r^* is the critical Sobolev exponent with respect to r, one has $F \in L^{r^*(p-1)}(\Omega; \mathbf{R}^n)$ and

$$-\mathrm{div}\,(|F|^{p-2}F) = -\Delta v = f \text{ in } \Omega,$$

while

$$|| \, |F|^{p-1} \, ||_{r^*} = ||\nabla v||_{r^*} \leq c(r)||f||_r. \tag{12}$$

Suppose now that the estimate (7) is uniform on p when $p \to +\infty$ (i.e., $C(p,s)$ is bounded, provided $1 \leq s/p \leq \theta \leq +\infty$). Then setting $s := t(p-1)$, where $r < t < r^*$, from (12) we conclude that $|\nabla u_p|^{p-2}\nabla u_p$ are bounded in $L^t(\Omega_0; \mathbf{R}^n)$ whenever $\Omega_0 \subset\subset \Omega$. Hence we would have $a \in L^t_{loc}(\Omega)$ with some $t > r$, and it is enough to refer to Example 1 to see the contradiction. □

REFERENCES

[1] L. Ambrosio, *Lecture Notes on Optimal Transport Problems*, Preprint 32, Scuola Normale Superiore, Pisa, Italy, 2000.

[2] T. Bhattacharya, E. Di Benedetto, and J. Manfredi, *Limits as $p \to \infty$ of $\Delta_p u_p = f$ and related extremal problems*, Rend. Sem. Mat. Univ. Politec. Torino, special issue (1989), pp. 15–67.

[3] G. Buttazzo and G. Bouchitté, *Characterization of optimal shapes and masses through Monge-Kantorovich equation*, J. Eur. Math. Soc. (JEMS), 3 (2001), pp. 139–168.

[4] L. Caffarelli, M. Feldman, and R. McCann, *Constructing optimal maps for Monge's transport problem as a limit of strictly convex costs*, J. Amer. Math. Soc., 15 (2002), pp. 1–206.

[5] L. De Pascale and A. Pratelli, *Regularity properties for Monge transport density and for solutions of some shape optimization problems*, Calc. Var. Partial Differential Equations, 14 (2002), pp. 249–274.

[6] E. Di Benedetto and J. Manfredi, *On the higher integrability of the gradient of weak solutions of certain degenerate elliptic problems*, Amer. J. Math., 115 (1993), pp. 1107–1134.

[7] L. Evans, M. Feldman, and R. F. Gariepy, *Fast/slow diffusion and collapsing sandpiles*, J. Differential Equations, 137 (1997), pp. 166–209.

[8] L. C. Evans and W. Gangbo, *Differential equations methods for the Monge-Kantorovich mass transfer problem*, Mem. Amer. Math. Soc., 137 (1999), no. 653.

[9] M. Feldman, *Growth of a sandpile around an obstacle*, in Monge Ampère Equation: Application to Geometry and Optimization (Deerfield Beach, FL, 1997), Contemp. Math. 226, L. Caffarelli and M. Milman, eds., AMS, Providence, RI, 1999, pp. 55–78.

[10] M. Feldman, *Variational evolution problems and nonlocal geometric motion*, Arch. Ration. Mech. Anal., 146 (1999), pp. 221–274.

[11] M. Feldman and R. McCann, *Uniqueness and transport density in Monge's mass transportation problem*, Calc. Var. Partial Differential Equations, 15 (2002), pp. 81–113.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 1056–1070

BOUNDARY CONTROLLABILITY BETWEEN SUB- AND SUPERCRITICAL FLOW*

MARTIN GUGAT†

Abstract. There are several studies of the boundary controllability of quasi-linear hyperbolic systems where it is assumed that the eigenvalues of the system matrix do not change their signs during the controlled process.

In this paper we consider the flow through a frictionless horizontal rectangular channel that is governed by de St. Venant equations and show that the state can be controlled in finite time from a stationary initial state to a given stationary terminal state in such a way that during this transition, the state stays in the class of C^1 functions, so in particular no shocks occur. There is no restriction on the initial and terminal state, so in some cases it is necessary that one or both eigenvalues of the system matrix change the sign during the process. Various different cases occur: control between subcritical states, control between supercritical states, transition from a subcritical to a supercritical state, and transition from a supercritical to a subcritical state. In the last two cases of a control between states of a different type, one eigenvalue of the system matrix changes its sign during the process. When this happens at a boundary point during the process, it is necessary to switch the type of boundary conditions. We show how to construct controls where at each boundary at most one such switching is necessary.

Key words. de St. Venant equations, subcritical states, supercritical states, global controllability

AMS subject classifications. 35L45, 35L50, 35L65, 93C20

DOI. 10.1137/S0363012902409660

1. Introduction.

In the study of boundary controllability of quasi-linear hyperbolic systems it is common practice to assume that the eigenvalues of the system matrix do not change their signs during the controlled process (see [3], [15], [14], [6], [5], [10], [7], [9], [12], [8]). In this paper we present a controllability result of a more general type where the number of positive eigenvalues is allowed to change during the control process; this is necessary if this is a different number for the initial state and for the target state.

To be specific, we consider the flow through a frictionless horizontal rectangular channel that is governed by de St. Venant equations. The system is controlled at the boundary points of the channel. For this system, the assumption that the number of positive eigenvalues and the number of negative eigenvalues of the system matrix remain constant during the control process means that the state is assumed to remain subcritical if the initial state is subcritical and to stay supercritical if the initial state is supercritical; in the supercritical case, also the direction of the flow has to be preserved. In some cases, however, for example, during a strong rainfall, the system goes from a subcritical to a supercritical state. The de St. Venant system is of practical interest on account of the trend to operate combined sewer systems and other channel networks, for example, in hydropower optimization, using model predictive control (see, for example, [1], [11]).

*Received by the editors June 17, 2002; accepted for publication (in revised form) February 24, 2003; published electronically July 8, 2003. This work was supported by the DFG research cluster "Real-time optimization of complex systems," grant Le595/13-1.

http://www.siam.org/journals/sicon/42-3/40966.html

†Technische Universität Darmstadt, Fachbereich Mathematik, AG 10, Schlossgartenstrasse 7, 64289 Darmstadt, Germany (gugat@mathematik.tu-darmstadt.de).

In this paper, we present a result on global controllability of de St. Venant system with no restriction on the stationary initial state and the target state. Although classical solutions generally break down in finite time as a result of collision of characteristic curves, we show that starting from a stationary initial state we can reach any other stationary state in finite time with boundary controls that can be chosen in such a way that the state remains continuously differentiable, that is, the state is a classical solution of our system. In order to achieve this, it is essential to change the state sufficiently slowly. To make this possible, the control period has to be chosen sufficiently long.

Apart from its mathematical interest, this result is relevant for applications where shocks have to be avoided. Our proofs are based upon the characteristic form of the equations. In the proofs, we show how the boundary control functions can be obtained. As these controls are not unique, there is some flexibility with respect to optimization. Our result is global in the sense that it holds for initial and terminal states that can be arbitrarily far away from each other.

If the type of the flow is not changed by the controls, that is, if the flow remains either sub- or supercritical, the nature of the boundary conditions also remains unchanged. In the case of a subcritical flow, one variable is controlled at each end of the channel whereas in the supercritical case, both variables are controlled at the end of the channel where inflow occurs. If the flow is controlled from a subcritical to a supercritical state, during this process the type of the boundary conditions changes: While we start with the control of one variable at each end of the channel, we switch to the control of both variables at the end of the channel where inflow occurs. We show that it is possible to do this with only one switching time for each end of the channel. This means that at each end of the channel, one eigenvalue of the system matrix is zero exactly at this switching time. The switching times for both ends can be different.

If the flow is controlled from a supercritical to a subcritical regime, the situation is similar. Also in this case it is possible to control the system with only one switching time for each end of the channel; this means that for each end there is exactly one point in time where one eigenvalue of the system matrix vanishes and the type of the boundary conditions changes.

Our paper has the following structure: In [7], we have shown that boundary control between arbitrary subcritical states is possible. Using similar methods, namely virtual prolongation of the channel in order to be able to work with Cauchy problems, we show that control between supercritical states with the same flow direction is possible. For our analysis it is necessary to show that also control between a subcritical and a supercritical state that are sufficiently close together is possible. The two cases are as follows: Control from a sub- to a supercritical regime and control from a super- to a subcritical state are treated separately. In the analysis of these two cases, also critical initial and target states are included.

Together, the four separate controllability results yield the desired general theorem about controllability without restriction on the stationary initial and target states.

2. De St. Venant equations. We consider the flow through a horizontal rectangular channel without friction. Let $b > 0$ denote the constant width of the channel. The channel is parametrized lengthwise by $x \in [0, L]$. Let $A(x, t)$ denote the wetted area at x at time t, and let $Q(x, t)$ denote the corresponding flow rate. Then the

conservation of mass yields the equation

$$\frac{d}{dt}A(x,t) + \frac{d}{dx}Q(x,t) = 0.$$

Let $U = Q/A$ denote the average velocity over the cross section of the channel and $h = A/b$ the average water height. The conservation of energy implies the equation

$$\frac{d}{dt}U(x,t) + \frac{d}{dx}\left[U(x,t)^2/2 + gh(x,t)\right] = 0.$$

In terms of the functions U and h, the quasi-linear system can be written as

$$\partial_t \begin{pmatrix} h \\ U \end{pmatrix} + \begin{pmatrix} U & h \\ g & U \end{pmatrix} \partial_x \begin{pmatrix} h \\ U \end{pmatrix} = 0.$$

The eigenvalues of the system matrix are $U + \sqrt{gh}$ and $U - \sqrt{gh}$. Let $c(x,t) = \sqrt{gh(x,t)}$ denote the corresponding wave celerity. In terms of the functions U and c, the system equation is

$$\partial_t \begin{pmatrix} c \\ U \end{pmatrix} + \begin{pmatrix} U & c/2 \\ 2c & U \end{pmatrix} \partial_x \begin{pmatrix} c \\ U \end{pmatrix} = 0.$$

With the Riemann invariants $R^+ = U + 2c$, $R^- = U - 2c$ and the diagonal matrix

$$A(R^+, R^-) := \begin{pmatrix} \frac{3}{4}R^+ + \frac{1}{4}R^- & 0 \\ 0 & \frac{1}{4}R^+ + \frac{3}{4}R^- \end{pmatrix},$$

de St. Venant equations can be written in the diagonal form

$$\partial_t \begin{pmatrix} R^+ \\ R^- \end{pmatrix} + A(R^+, R^-)\, \partial_x \begin{pmatrix} R^+ \\ R^- \end{pmatrix} = 0. \tag{2.1}$$

3. Controllability with phase transitions. In this section we present our central controllability result where phase transitions are admitted in the sense that the type of the state is allowed to change between sub- and supercritical. Also critical states are admitted. Note that for the system that we consider here, the stationary states are exactly the constant states.

THEOREM 3.1. *Consider the de St. Venant system with boundary controls on an interval $[0, L]$. Let a stationary initial state A and a stationary target state B be given. Then we can find boundary controls that steer the system in finite time from A to B in such a way that the state remains continuously differentiable.*

Moreover, this can be done in such a way that the absolute values of the first partial derivatives of the state remain below any given upper bound.

In the following four sections, we present partial controllability results that we combine in section 8 to prove Theorem 3.1.

4. Controllability between subcritical states. A state is called subcritical, if one eigenvalue of the system matrix is greater than zero and the other eigenvalue is less than zero. The question of controllability between subcritical states has been answered in [7], where the following theorem is proved.

THEOREM 4.1. *From a constant subcritical state (R^+, R^-) ($3/4R^+ + 1/4R^- > 0$ and $1/4R^+ + 3/4R^- < 0$), the system can be steered to any other constant subcritical state by boundary controls in finite time with a continuously differentiable state.*

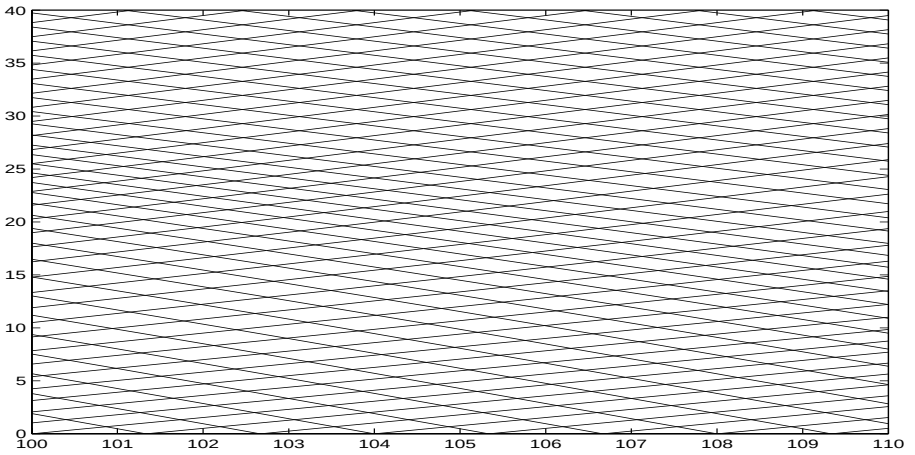


FIG. 1.

Moreover, for every given positive constant, the system can be steered in such a way that the maximum norms of the derivatives of the state and of the controls remain below this bound.

In the proof of the theorem, the considered channel of finite length is prolonged to an infinite channel, where a Cauchy problem can be considered. Under certain assumptions, this Cauchy problem has a classical, that is, continuously differentiable, solution; the essential assumption is that the maximum norm of the derivatives of the initial state is sufficiently small.

The controls are then obtained by cutting a strip whose width is the length of the channel from the x-t plane; the values on the boundary of the strip for positive times yield the desired boundary controls.

Example 1. Consider the subcritical initial state with $U = 0.3$ and $c = 1$, that is, $R_1^+ = 2.3$, $R_1^- = -1.7$, and the subcritical target state with $U = 0$ and $c = 1$, that is, $R_2^+ = 2$, $R_2^- = -2$. Note that for the target state, the water is at rest.

Figure 1 shows the characteristic curves (i.e., the curves whose slopes are the eigenvalues of the system matrix) for a control from the initial state (R_1^+, R_1^-) to the target state (R_2^+, R_2^-) where the water is still. For the target states, the absolute value of the slope of the leftgoing characteristic curves is the same as for the characteristic curves that go from left to right. The channel interval $[0, L]$ is shifted to $[100, 110]$.

At time $T = 40$, the system has reached the desired target state. So we see that our results imply that by boundary controls at both boundaries the water flow in a horizontal channel can be steered to a standstill.

5. Controllability between supercritical states. A state is called supercritical if both eigenvalues of the system matrix are greater than zero or both eigenvalues are less than zero. In this section we study the control between supercritical stationary flows with the same direction. In our result, we consider the problem of steering the system from an initial state for which the corresponding system matrix has two positive eigenvalues to a target state that has also two positive eigenvalues. The problem of steering the system from an initial state with two negative eigenvalues to a target state with two negative eigenvalues can be transformed to the first problem by changing the orientation of the channel. Controllability between supercritical states

has been studied in [9] in a more general framework with source terms on tree-shaped channel networks.

5.1. A local result. We start with a result about local controllability. In this result a compact set of initial states is considered. We show that a certain radius and a control time exist, such that for each of these initial states, it is possible to reach all target states in balls with this radius around the initial state within the control time. So this is a result about uniform local controllability. For the result it is essential that a uniform lower bound for the eigenvalues of the system matrix is valid.

THEOREM 5.1. *Consider the de St. Venant system* (2.1) *with boundary conditions of the form*

$$R^+(0,t) = g_1(t) \quad \textit{and } R^-(0,t) = g_2(t).$$

Let a nonempty compact rectangular set $\Omega = [a_+, b_+] \times [a_-, b_-] \subset R^2$ *be given such that for all* $(R^+, R^-) \in \Omega$, *we have* $3/4R^+ + 1/4R^- > 0$ *and* $1/4R^+ + 3/4R^- > 0$, *that is,* Ω *contains only supercritical states with the same flow direction. Let* $\delta > 0$ *be given. Define*

$$T = \max_{(R^+,R^-)\in\Omega} \max\left\{ \frac{L+\delta}{|(3/4)R^+ + (1/4)R^-|}, \frac{L+\delta}{|(1/4)R^+ + (3/4)R^-|} \right\}. \tag{5.1}$$

Then there exists a number $\alpha > 0$ *such that for all constant initial states* $(R_1^+, R_1^-) \in \Omega$ *and for all constant terminal states* $(R_2^+, R_2^-) \in \Omega$ *with*

$$\max\{|R_1^+ - R_2^+|,\ |R_1^- - R_2^-|\} \le \alpha, \tag{5.2}$$

the boundary controls g_1, g_2 *can be chosen such that in time* T *the system has reached the terminal state* (R_2^+, R_2^-) *and the corresponding solution is continuously differentiable.*

Moreover, by choosing α *sufficiently small, the maximal absolute values of the derivatives*

$$\partial_x R^+(x,t),\ \partial_x R^-(x,t),\ \partial_t R^+(x,t),\ \partial_t R^-(x,t),\ g_1'(t),\ g_2'(t)$$

for $(x,t) \in [0,L] \times [0,T]$ *can be made arbitrarily small.*

Proof. The proof is similar to the proof of the corresponding result for the subcritical case in [7]. We define an initial state $\varphi : R \to R^2$ on the real line in the following way. For $x \in [0,\infty)$, let $\varphi(x) = (R_1^+, R_1^-)$. For $x \in (-\infty, -\delta]$, let $\varphi(x) = (R_2^+, R_2^-)$. For $x \in (-\delta, 0)$, define φ in $C^1(R)$ such that the components φ_1 and φ_2 of φ are monotone functions on $(-\delta, 0)$ and

$$\max_{x\in R}\{|\varphi_1'(x)|,\ |\varphi_2'(x)|\} \le (3/\delta)\max\{|R_1^+ - R_2^+|,\ |R_1^- - R_2^-|\}. \tag{5.3}$$

Since the set Ω is compact, we can choose a number $c_0 > 0$ such that for all $(R^+, R^-) \in \Omega$ we have $|(R^+, R^-)| \le c_0$. Then we can choose a number $k \ge 1$ such that for all (R^+, R^-) with $|(R^+, R^-)| \le 2c_0$ we have $|A(R^+, R^-)| \le k$. Define $a = -kT$ and $b = L + kT$. Then $(A, 0) \in \Sigma([a,b] \times [0,T], 2, 2c_0, 1/2, k)$, where the class Σ is defined in the introduction of [4]. Theorem 3.IV from Cirina [4] implies the existence of a number $\beta > 0$ such that if

$$\max_{x\in\mathbb{R}}\{|\varphi_1'(x)|,\ |\varphi_2'(x)|\} \le \beta, \tag{5.4}$$

the de St. Venant system with initial state φ on $[a,b]$ has a unique continuously

differentiable solution on the convex hull of the two sets $[a,b]\times\{0\}$ and $[0,L]\times[0,T]$. Let τ denote this region.

Choose $\alpha \le \beta\delta/3$. Then (5.2) and (5.3) imply (5.4). Therefore by Theorem 3.IV from Cirina [4], a C^1 solution $(R^+(x,t), R^-(x,t))$ with initial state φ on $[a,b]$ exists on the set τ.

Moreover, Theorem 3.IV guarantees the existence of a constant $N>0$ such that for all $(x,t)\in[0,L]\times[0,T]$ we have

$$\max\{|\partial_x R^+(x,t)|,\ |\partial_x R^-(x,t)|\} \le N \max_{x\in\mathbb{R}}\{|\varphi_1'(x)|,\ |\varphi_2'(x)|\} \le 3N\alpha/\delta.$$

This implies that by choosing α sufficiently small, we can make the absolute values of the derivatives $\partial_x R^+(x,t)$, $\partial_x R^-(x,t)$ arbitrarily small.

Since the initial state φ_1 has values only between R_1^+ and R_2^+, this is also the case for the first component R^+ of the solution on $[0,T]$. Similarly, R^- can attain only values between R_1^- and R_2^-.

Define

$$M = \max_{(x,t)\in[0,L]\times[0,T]}\{|(3/4)R^+ + (1/4)R^-|,\ |(1/4)R^+ + (3/4)R^-|\}.$$

The system equation implies that

$$\max_{(x,t)\in[0,L]\times[0,T]}\{|\partial_t R^+|,\ |\partial_t R^-|\} \le M \max_{(x,t)\in[0,L]\times[0,T]}\{|\partial_x R^+|,\ |\partial_x R^-|\} \le 3MN\alpha/\delta.$$

So we can also make the absolute values of the time derivatives of (R^+, R^-) arbitrarily small by choosing α sufficiently small.

The slopes of the characteristic curves are given by $dx_+/dt = (3/4)R^+ + (1/4)R^-$ and $dx_-/dt = (1/4)R^+ + (3/4)R^-$, respectively. In the area of points above both of these characteristics starting at the point $(-\delta, 0)$, the component R^+ of the solution has the value R_2^+ and R^- has the value R_2^-. By the definition of T and on account of the rectangular shape of Ω, this implies that at time T the system has reached the terminal state (R_2^+, R_2^-) on the interval $[0,L]$.

This can be seen as follows. Since the set Ω contains only supercritical states, both the x_+ and the x_- characteristic curves have positive slope. Hence at time T both characteristic curves coming from the point $(x,t) = (-\delta, 0)$ have reached the point (L,T).

We now define the functions g_1, g_2 by setting $g_1(t) := R_+(0,t)$ and $g_2(t) := R_-(0,t)$ and have thus constructed the required boundary controls. □

5.2. A global result. Now we apply the same globalization technique as in [7] to our local result: The uniformity in the local result implies that boundary control between all elements of the set Ω is possible in finite time.

THEOREM 5.2. *Let a nonempty compact rectangular set Ω as in Theorem 5.1 be given.*

Then there exist boundary controls g_1, g_2 that steer the de St. Venant system in finite time from any constant initial state in the set Ω to any constant terminal state in Ω in such a way that the corresponding solution is continuously differentiable.

Moreover, this can be done in such a way that the absolute values of the derivatives of the state remain smaller than any given upper bound.

The proof is exactly as in [7]: The idea is to introduce a finite number of intermediate states on the line connecting the initial state with the target state and to control the system successively from one intermediate state to the next.

Now we can prove that it is possible to steer the system from any constant supercritical state to any other constant supercritical state with the same flow direction in finite time with a continuously differentiable state and bounds for the absolute values of the derivatives.

THEOREM 5.3. *The de St. Venant system can be steered from a constant supercritical state to any other constant supercritical state with the same flow direction in finite time with a continuously differentiable state by boundary controls at one end.*

Moreover, for every given positive constant the system can be steered in such a way that the maximum norms of the state derivatives and of the control derivatives remain below this bound.

Proof. Let two constant supercritical states (R_A^+, R_A^-) and (R_B^+, R_B^-) be given.

For a natural number n and $k \in \{0, \dots, n\}$ define $\lambda_{k,n} = k/n$ and let

$$R_{k,n}^+ = (1 - \lambda_{k,n})R_A^+ + \lambda_{k,n}R_B^+, \quad R_{k,n}^- = (1 - \lambda_{k,n})R_A^- + \lambda_{k,n}R_B^-.$$

Then $R_{0,n}^+ = R_A^+$, $R_{0,n}^- = R_A^-$, $R_{n,n}^+ = R_B^+$, $R_{n,n}^- = R_B^-$.

If the number n is chosen sufficiently large, for all $k \in \{1, \dots, n\}$ the nonempty sets

$$\Omega_k = [\min\{R_{k-1}^+, R_k^+\},\ \max\{R_{k-1}^+, R_k^+\}] \times [\min\{R_{k-1}^-, R_k^-\},\ \max\{R_{k-1}^-, R_k^-\}]$$

contain only supercritical states with the same flow direction; hence for all $k \in \{1, \dots, n\}$ Theorem 5.2 is applicable to Ω_k. So we can steer the system from the initial state $(R_A^+, R_A^-) \in \Omega_1$ to $(R_{1,n}^+, R_{1,n}^-) \in \Omega_1$, from $(R_{1,n}^+, R_{1,n}^-) \in \Omega_2$ to $(R_{2,n}^+, R_{2,n}^-) \in \Omega_2$, and so forth and finally from $(R_{n-1,n}^+, R_{n-1,n}^-) \in \Omega_n$ to the desired terminal state $(R_B^+, R_B^-) \in \Omega_n$ that we reach after n such steps.

The assertion about the maximum norms of the derivatives and the controls also follows from Theorem 5.2. ∎

Example 2. Consider the supercritical initial state with $U = 2$ and $c = 1$, that is, $R_1^+ = 4$, $R_1^- = 0$, and the supercritical target state with $U = 1.9$ and $c = 1$, that is, $R_2^+ = 3.9$, $R_2^- = -0.1$.

Define $p(a, \cdot)$ as the cubic polynomial that is determined by the four conditions $p(a, -\delta) = a$, $p'(a, -\delta) = p(a, 0) = p'(a, 0) = 0$, that is,

$$p(a, x) = \frac{6a}{\delta^3}\left(\frac{x^3}{3} + \delta\frac{x^2}{2}\right).$$

Note that $p(a, \cdot)$ is monotone on the interval $[-\delta, 0]$.

Define the function φ with the components

$$\varphi_1(x) = \begin{cases} R_2^+, & x \in (-\infty, -\delta), \\ R_1^+ + p(R_2^+ - R_1^+, x), & x \in [-\delta, 0), \\ R_1^+, & x \in [0, \infty), \end{cases}$$

and

$$\varphi_2(x) = \begin{cases} R_2^-, & x \in (-\infty, -\delta), \\ R_1^- + p(R_2^- - R_1^-, x), & x \in [-\delta, 0), \\ R_1^-, & x \in [0, \infty). \end{cases}$$

Then the function $\varphi = (\varphi_1, \varphi_2)$ satisfies the requirements of the proof of Theorem 5.1. Figure 2 shows the characteristic curves for the solution of the Cauchy problem with the initial data given by φ. The point zero is shifted to 180 and the point $-\delta$ is shifted

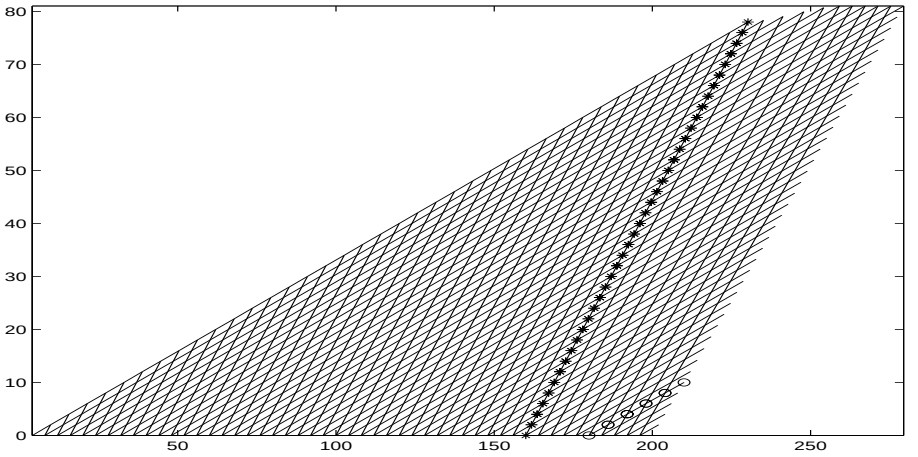


FIG. 2.

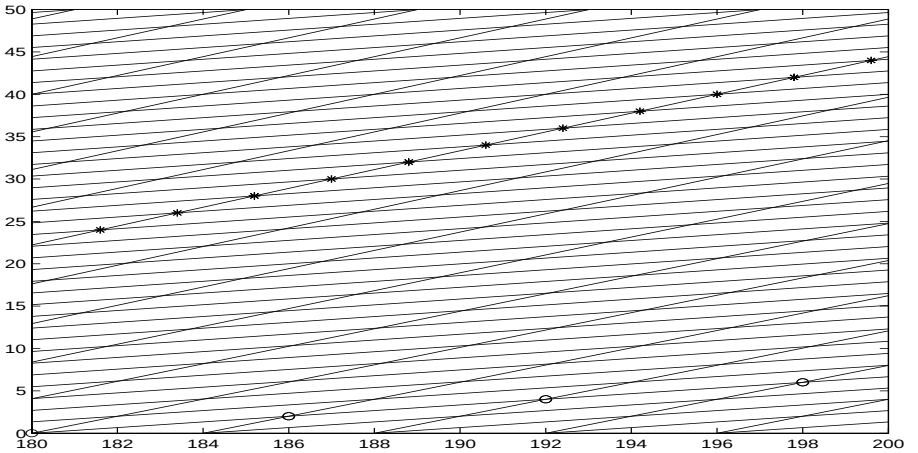


FIG. 3.

to 160. The channel goes from $x = 180$ to $x = 200$. Above the characteristic curve marked by asterisks the state is equal to the target state (R_2^+, R_2^-). Below the line marked by o's, the state equals the initial state (R_1^+, R_1^-).

The control is obtained by cutting the rectangle $[0, L] \times [0, T]$ from the x-t plane; the boundary values yield the control functions. In Figure 3, the rectangle is shifted to $[180, 200] \times [0, 50]$. At time $T = 50$, the system has reached the desired target state.

6. Controllability from super- to subcritical states. In this section we show how the system can be controlled from a given supercritical initial state to a subcritical state with the same value of the Riemann invariant R^-. We also admit critical initial or target states. A state is called critical if one eigenvalue of the corresponding system matrix is zero.

THEOREM 6.1. *Let a supercritical or critical initial state (R_1^+, R_1^-) with positive eigenvalues (that is, $3R_1^+ + R_1^- > 0$, $R_1^+ + 3R_1^- \geq 0$) and a subcritical or critical target state (R_2^+, R_1^-), where the second Riemann invariant has the same value as for the initial state and $R_2^+ > -R_1^-/3$, be given. Then there exists a boundary control*

that steers the system in finite time from the initial state to the target state with a continuously differentiable solution. During the process, the nature of the boundary conditions changes at each boundary at most once.

If the number $|R_2^+ - R_1^+|$ is sufficiently small, for every positive upper bound the system can be steered in such a way that the maximum norm of the derivatives of the state remains below this bound.

Proof. Let $R_2^- = R_1^-$. The assumptions imply that $(3/4)R_2^+ + (1/4)R_2^- > 0$. Since the terminal state is subcritical or critical, this implies that $(1/4)R_2^+ + (3/4)R_2^- \leq 0$. Hence $R_2^+ \leq -3R_2^- \leq R_1^+$, since the initial state is supercritical or critical.

Let $\delta > 0$ be given. Define the function φ_1 by $\varphi_1(x) = R_2^+$ if $x \in (-\infty, -\delta]$, $\varphi_1(x) = R_1^+$ if $x \in [0, \infty)$, and extend φ_1 to an increasing continuously differentiable function on the whole real axis. If $R_2^+ = R_1^+$, the initial state equals the terminal state and there is nothing to prove. So in what follows, we can assume that $R_2^+ < R_1^+$, and we can choose φ_1 in such a way that it is strictly increasing on the interval $[-\delta, 0]$. Define φ_2 by $\varphi_2(x) = R_1^-$ for all $x \in (-\infty, \infty)$.

Now we consider the Cauchy problem for our system (2.1) with the initial data given by the function φ. Since the components of the function φ are both increasing, Theorem 2.1 in [13] implies the existence of a unique continuously differentiable solution of the Cauchy problem for all $t \geq 0$ (see also Remark 2.4 in [13]).

Let (S^+, S^-) denote the solution of this Cauchy problem.

The slope of the characteristic curve with positive slope that starts at the point $(x, t) = (-\delta, 0)$ is at least $(3/4)R_2^+ + (1/4)R_2^- > 0$, since φ_1 is increasing. Hence after the time $T = (L+\delta)/((3/4)R_2^+ + (1/4)R_2^-)$, this curve has reached the x-value L, that is, the end L of the channel. In the area above this characteristic curve the Riemann invariants have the values (R_2^+, R_2^-); hence the state equals the desired target state in this area. This implies that the state in the points in $[0, L] \times [T, \infty)$ is the desired target state, i.e., after the time T the channel state has reached the target.

The initial values of the Riemann invariants are given by the increasing function φ_1 and the constant function φ_2, and the values of the Riemann invariants are constant on the characteristic curves. Therefore on the line (L, t), $t \geq 0$, the eigenvalue $(1/4)S^+ + (3/4)S^-$ is decreasing with time. Since $(1/4)R_1^+ + (3/4)R_1^- \geq 0$ and $(1/4)R_2^+ + (3/4)R_2^- \leq 0$, and $R_2^+ < R_1^+$, there exists an interval $[t_L^1, t_L^2]$ that consists exactly of the points where the eigenvalue $(1/4)S^+ + (3/4)S^-$ is equal to zero. If the initial state is critical, that is, if initially the eigenvalue is zero, that is, $(1/4)R_1^+ + (3/4)R_1^- = 0$, we have $t_L^1 = 0$. If the target state is critical, that is, if for the target state the eigenvalue is zero, that is, $(1/4)R_2^+ + (3/4)R_2^- = 0$, we have $t_L^2 = T$. If both the initial and the terminal state are not critical, the interval $[t_L^1, t_L^2]$ reduces to a single point t_L since φ_1 is strictly increasing on the interesting interval $[-\delta, 0]$; so in this case there is exactly one time t_L where this eigenvalue is zero at the right channel boundary.

Analogously, we see that on the line $(0, t)$, $t \geq 0$, the eigenvalue $(1/4)S^+ + (3/4)S^-$ is decreasing with time and vanishes exactly on an interval $[t_0^1, t_0^2]$. If the target state is critical, $t_0^2 = T$. If the initial state is critical, $t_0^1 = t_0^2 = 0$. If both the initial and the terminal state are not critical, the interval $[t_0^1, t_0^2]$ reduces to a single point t_0, that is, in this case there is exactly one time t_0 where this eigenvalue is zero at the left channel boundary.

For $t \in [0, \infty)$, define $g_1(t) = S^+(0, t)$, $g_2(t) = S^-(0, t) = R_1^-$, and $g_3(t) = S^-(L, t) = R_1^-$. The boundary conditions for the problem of flow control for the channel are then defined in the following way.

At the left channel–boundary 0, $R^+(0,t) = g_1(t)$ and if $g_1(t)+3R_1^- > 0$, $R^-(0,t) = g_2(t) = R_1^-$. Note that $g_1(t) + 3R_1^- = R^+(0,t) + 3R^-(0,t)$.

At the right channel–boundary L, if $R^+(L,t) + 3R_1^- < 0$, $R^-(L,t) = g_3(t) = R_1^-$.

Since the continuously differentiable function (S^+, S^-) satisfies the conditions defined above, it is clear that a classical solution of the initial-boundary problem on $[0, L]$, $t \geq 0$, with the above boundary conditions exists. Moreover, since the boundary conditions are such that the value of R^- is always R_1^- and the value of R^+ is prescribed at the left boundary for all times $t \geq 0$ to be the same as for S^+, the solution must in fact be equal to the restriction of (S^+, S^-) to $[0, L] \times [0, \infty)$. So we see that the system can be controlled from the initial state to the target state in finite time with a continuously differentiable solution and with only one switching time at each end.

The last part of the assertion follows from the a priori bound for the derivatives in Theorem 3.IV in [4], since for any given δ if $|R_2^+ - R_1^+|$ is sufficiently small, we can make the derivative φ_1' arbitrarily small. □

Remark 1. Note that throughout the solution the value of the Riemann invariant R^- remains constant. Our boundary conditions work in the following way: The values of the Riemann invariant R^+ are prescribed at the left boundary 0 at all times $t \geq 0$. The values of the other Riemann invariant R^- are at the beginning also described on the left boundary 0 to have the constant value R_1^-; then as time proceeds, by the change of the value of R^+ through the boundary conditions the slope $(1/4)R^+ + (3/4)R^-$ of the characteristic curves where R^- is constant becomes negative. In fact, the eigenvalue $(1/4)R^+ + (3/4)R^-$ is zero on the characteristic curve that travels from left to right with slope $(3/4)R^+ + (1/4)R^-$ where $R^+ = -3R_1^-$. If the initial state is supercritical (not critical) and the target state is subcritical (also not critical), the switching time for the boundary conditions is the time where this characteristic curve arrives at the right end.

Note that the boundary conditions for the control problem defined above are of similar character as the boundary conditions considered in [2] for scalar quasi-linear equations.

Example 3. Consider the supercritical initial state with $R_1^+ = 5.7$, $R_1^- = -1.5$ and the subcritical target state with $R_2^+ = 3.3$, $R_1^- = -1.5$.

As in Example 2, define the functions φ_1 and φ_2 by cubic polynomials on the interval $[-\delta, 0]$.

Figure 4 shows the characteristic curves for the solution of the Cauchy problem with the initial data given by φ. The point zero is shifted to 140 and the point $-\delta$ is shifted to 40. The channel goes from $x = 140$ to $x = 180$. Above the characteristic curve marked by asterisks, the state is equal to the target state (R_2^+, R_2^-). Below the line marked by o's, the state equals the initial state (R_1^+, R_1^-). The control is obtained by cutting the rectangle $[0, L] \times [0, T]$ from the x-t plane; the boundary values yield the control functions. In fact, only the control values for $g_1(t) = S^+(0,t)$ need to be obtained in this way. In Figure 5 the rectangle is shifted to $[140, 180] \times [0, 67.5]$. At time $T = 67.5$, the desired target state has been reached.

7. Controllability from sub- to supercritical states. To complete our analysis, in this section we study the control of the system between the sub- and the supercritical phase. Again we also admit critical target or initial states. It turns out that this case is more complicated than the situation treated in section 6 because for the corresponding Cauchy problem we cannot expect that the solution exists for all $t \geq 0$, since after a finite time shocks can occur. We present controls where this

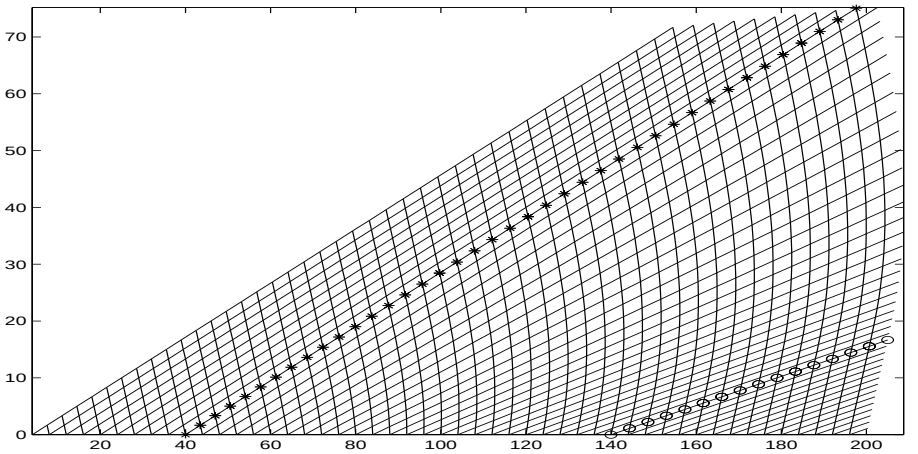


FIG. 4.

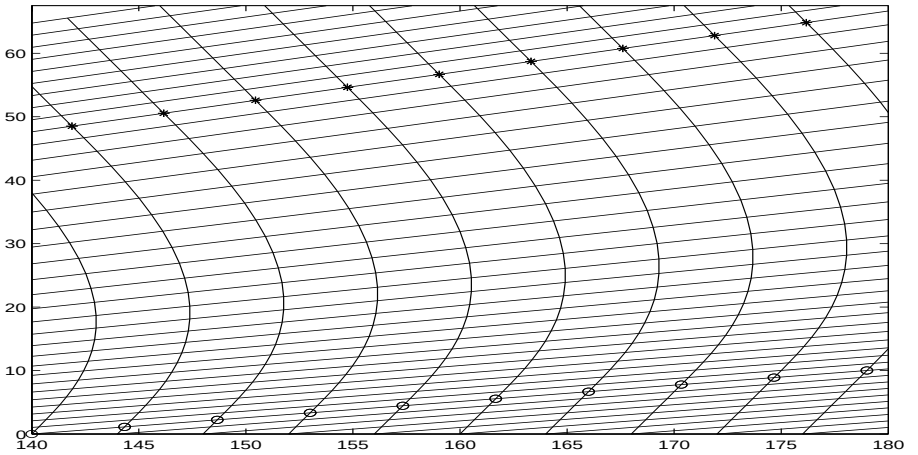


FIG. 5.

happens only after in the channel the target state has been reached, so the shock does not influence our control process.

THEOREM 7.1. *Let a subcritical or critical initial state* (R_1^+, R_1^-) *(with* $3R_1^+ + R_1^- > 0$, $R_1^+ + 3R_1^- \leq 0$*) and a supercritical or critical target state* (R_2^+, R_1^-)*, where the second Riemann invariant has the same value as for the initial state and* $R_2^+ > -R_1^-/3$*, be given. If the number* $|R_2^+ - R_1^+|$ *is sufficiently small, there exists a boundary control that steers the system in finite time from the initial state to the target state with a continuously differentiable solution. During the process, the nature of the boundary conditions changes at each boundary at most once. For every given positive constant the system can be steered in such a way from the initial to the target state that the absolute values of the derivatives of the state remain below this bound.*

Proof. Let $R_2^- = R_1^-$. The assumptions imply the inequality $R_2^+ \geq -3R_1^- \geq R_1^+$. If $R_2^+ = R_1^+$, the initial state equals the terminal state and there is nothing to prove. So in what follows, we can assume that $R_2^+ > R_1^+$. Let $\delta > 0$ be given. Define the function φ_1 by $\varphi_1(x) = R_2^+$ if $x \in (-\infty, -\delta]$, $\varphi_1(x) = R_1^+$ if $x \in [0, \infty)$, and extend φ_1

to a decreasing continuously differentiable function on the whole real axis such that

$$\varphi_1' \le (3/\delta)|R_2^+ - R_1^+|. \tag{7.1}$$

Moreover, choose φ_1 such that it is strictly decreasing on the interval $[-\delta, 0]$. Define φ_2 by $\varphi_2(x) = R_1^-$ for all $x \in (-\infty, \infty)$.

Now we consider the Cauchy problem for the de St. Venant system (2.1) with the initial data given by the function φ. Define the time T by the equation $T = (L+\delta)/[(3/4)R_2^+ + (1/4)R_1^-]$.

We can choose a number $c_0 > 0$ such that the following inequality is valid: $\max_{x\in(-\infty,\infty)}\{|\varphi_1(x)|,\ |\varphi_2(x)|\} \le c_0$. Then we can choose a number $k \ge 1$ such that for all (R^+, R^-) with $|(R^+, R^-)| \le 2c_0$ we have $\max\{|3R^+ + R^-|,\ |R^- + 3R^+|\}/4 \le k$. Define $a = -kT$ and $b = L + kT$. Then Theorem 3.IV from Cirina [4] implies the existence of a number $\beta > 0$ such that if

$$\max_{x\in\mathbb{R}}\{|\varphi_1'(x)|,\ |\varphi_2'(x)|\} \le \beta, \tag{7.2}$$

the de St. Venant system with initial state φ on $[a, b]$ has a unique continuously differentiable solution on the convex hull of the two sets $[a, b] \times \{0\}$ and $[0, L] \times [0, T]$. Let τ denote this region. Choose $\alpha = \beta\delta/3$. If $|R_2^+ - R_1^+| \le \alpha$, condition (7.1) implies (7.2). Therefore by Theorem 3.IV from [4], a C^1 solution with initial state φ on $[a, b]$ exists on the set τ. Let (S^+, S^-) denote this solution.

As in the proof of Theorem 5.1, the assertion that we can make the absolute values of the derivatives $\partial_x R^+$, $\partial_x R^-$, $\partial_t R^+$, $\partial_t R^-$ arbitrarily small by choosing α sufficiently small follows from the a priori bound in Theorem 3.IV from [4].

The slope of the characteristic curve that starts with slope $(3/4)R_2^+ + (1/4)R_2^-$ at the point $(x, t) = (-\delta, 0)$ is constant, since the Riemann invariant R^+ is constant on this curve and $R^- = R_1^-$. Hence after the time T, this curve has reached the x-value L, that is, the end L of the channel. In the area above this characteristic curve the Riemann invariants have the values (R_2^+, R_2^-); hence the state equals the desired target state in this area. This implies that the state in the points in $[0, L] \times \{T\}$ is the desired target state, i.e., after the time T the channel state has reached the target.

The initial values of the Riemann invariants are given by the decreasing function φ_1 and the constant function φ_2; so $R^-(x, t) = R_1^-$ is constant for the whole solution, and the values of the other Riemann invariant R^+ are constant on the corresponding characteristic curves. Therefore on the line (L, t), $t \ge 0$, the eigenvalue $(1/4)S^+ + (3/4)S^-$ is increasing with time. Since $(1/4)R_1^+ + (3/4)R_1^- \le 0$, $(1/4)R_2^+ + (3/4)R_2^- \ge 0$, and $R_2^+ > R_1^+$, there exists an interval $[t_L^1, t_L^2]$ consisting exactly of the points where the eigenvalue $(1/4)S^+ + (3/4)S^-$ is zero. If the initial state is critical, that is, if initially the eigenvalue is zero, that is, $(1/4)R_1^+ + (3/4)R_1^- = 0$, we have $t_L^1 = 0$. If the target state is critical, that is, if for the target state the eigenvalue is zero, that is, $(1/4)R_2^+ + (3/4)R_2^- = 0$, we have $t_L^2 = T$. If both the initial and the terminal state are not critical, the interval $[t_L^1, t_L^2]$ reduces to a single point t_L, since φ_1 is strictly decreasing on the interesting interval $[-\delta, 0]$; so in this case there is exactly one time t_L where this eigenvalue is zero at the right channel boundary.

Analogously, we see that also on the line $(0, t)$, $t \ge 0$, the eigenvalue $(1/4)S^+ + (3/4)S^-$ is increasing with time and vanishes exactly on an interval $[t_0^1, t_0^2]$. If the target state is critical, $t_0^2 = T$. If the initial state is critical, $t_0^1 = t_0^2 = 0$, and if both the initial and the terminal state are not critical, the interval $[t_0^1, t_0^2]$ reduces to a single point t_0, that is, in this case there is exactly one time t_0 where this eigenvalue is zero at the left channel boundary.

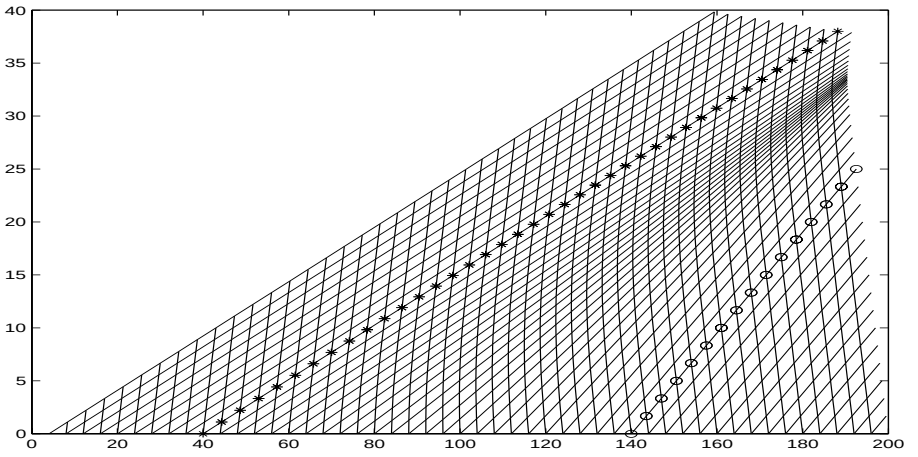


FIG. 6.

For $t \in [0, \infty)$, define $g_1(t) = S^+(0,t)$, $g_2(t) = S^-(0,t)$, and $g_3(t) = S^-(L,t)$. The boundary conditions for the problem of flow control for the channel are then defined in the following way:

At the left channel–boundary 0, $R^+(0,t) = g_1(t)$ and if $R^+(0,t) + 3R_1^- > 0$, $R^-(0,t) = g_2(t)$.

At the right channel–boundary L, if $R^+(L,t) + 3R_1^- < 0$, $R^-(L,t) = g_3(t)$.

Since the continuously differentiable function (S^+, S^-) satisfies the conditions defined above, it is clear that a classical solution of the initial-boundary problem on $[0, L]$, $t \geq 0$, with the above boundary conditions exists. Moreover, since the boundary conditions are such that the value of R^- remains always R_1^- and the value R^+ is prescribed at the left boundary for all times $t \geq 0$ to be the same as for S^+, the solution must in fact be equal to the restriction of (S^+, S^-) to $[0, L] \times [0, \infty)$. So we see that the system can be controlled from the initial state to the target state in finite time with a continuously differentiable solution and with at most one switching time at each end. □

Remark 2. If the initial state is subcritical (not critical) and the target state is supercritical (not critical), the switching points t_0 and t_L are connected by the line $M = \{(x,t) : (1/4)R^+(x,t) + (3/4)R^-(x,t) = 0\}$. This line is in fact a characteristic curve since $R^- = R_1^-$ is constant for the solution; hence (see also Remark 1)

$$M = \{(x,t) : R^+(x,t) = -3R_1^-\} = \{(x,t) : (3/4)R^+(x,t) + (1/4)R^-(x,t) = -2R_1^-\}.$$

Example 4. Consider the subcritical initial state with $R_1^+ = 3.3$, $R_1^- = -1.5$ and the supercritical target state with $R_2^+ = 5.7$, $R_1^- = -1.5$.

As in Example 2, define the functions φ_1 and φ_2 by cubic polynomials on the interval $[-\delta, 0]$.

Figure 6 shows the characteristic curves for the solution of the Cauchy problem with the initial data given by φ. The point zero is shifted to 140 and the point $-\delta$ is shifted to 40. The channel goes from $x = 140$ to $x = 180$. Above the characteristic curve marked by asterisks, the state is equal to the target state (R_2^+, R_2^-). Below the

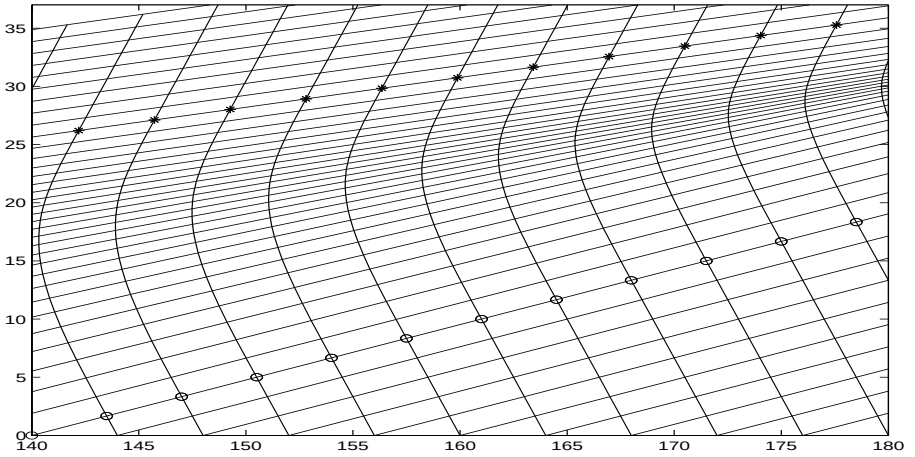


FIG. 7.

line marked by o's, the state equals the initial state (R_1^+, R_1^-). The control is obtained by cutting the rectangle $[0, L] \times [0, T]$ from the x-t plane; the boundary values yield the control functions. In Figure 7, the rectangle is shifted to $[140, 180] \times [0, 37]$. At time $T = 37$, the desired target state has been reached.

8. Proof of the general controllability result. For the proof of Theorem 3.1, we point out that starting with a critical state, we can reach any subcritical state as a consequence of Theorem 6.1 and Theorem 4.1. Starting from a subcritical state, we can reach any supercritical state with positive eigenvalues; this follows from Theorem 4.1, Theorem 7.1, and Theorem 5.3. From a supercritical state with positive eigenvalues, we can reach any subcritical state; this can be proved using Theorem 5.3, Theorem 6.1, and Theorem 4.1. Analogously, we see that starting from a subcritical state, we can reach any supercritical state with negative eigenvalues, and starting from a state of this type, we can go to any subcritical state.

Finally a combination of Theorem 4.1 and Theorem 7.1 implies that from a subcritical state we can control the system to any critical state. In this case Theorem 7.1 is applied in the following way: The critical target state (R_2^+, R_2^-) is prescribed and the initial state (R_1^+, R_2^-) in the statement of Theorem 7.1 is then chosen in such a way that it is subcritical: $R_1^+ = R_2^+ - \varepsilon$ with $\varepsilon > 0$ chosen sufficiently small. Since Theorem 4.1 implies that from any subcritical initial state we can reach this state (R_1^+, R_1^-), the assertion follows.

9. Conclusion. We have shown that for the de St. Venant system that describes the flow in a horizontal channel without friction, boundary control between arbitrary stationary states is possible in such a way that during the control process the system has a classical solution. This result raises many interesting questions: Can it be extended to the case of a system with a nonzero source term, for example, a steep channel with friction? Are phase transitions of the type that occur here, where eigenvalues of the system matrix change their sign during the control process, possible for general quasi-linear hyperbolic systems? These questions are left for future research.

Acknowledgments. The author wants to thank E.J.P. Georg Schmidt and G. Leugering for fruitful discussions in Darmstadt.

REFERENCES

[1] E. Arnold, H. Linke, and W. Siebert, *Ein modell–praediktives regelungsverfahren zur optimierten wasserbewirtschaftung des mittellandkanals und des elbe–seitenkanals*, Automatisierungstechnik, 47 (1999), pp. 399–407.

[2] C. Bardos, A. Y. Leroux, and J. C. Nedelec, *First order quasilinear equations with boundary conditions*, Comm. Partial Differential Equations, 4 (1979), pp. 1017–1034.

[3] M. Cirinà, *Boundary controllability of nonlinear hyperbolic systems*, SIAM J. Control, 7 (1969), pp. 198–212.

[4] M. Cirina, *Nonlinear hyperbolic problems with solutions on preassigned sets*, Michigan Math. J., 17 (1970), pp. 193–209.

[5] J. M. Coron, *Local controllability of a 1-D tank containing a fluid modeled by the shallow water equations*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 513–554.

[6] J. M. Coron, B. d'Andrea Novel, and G. Bastin, *A Lyapunov approach to control irrigation canals modeled by Saint-Venant equations*, in Advance in Control, Springer-Verlag, Berlin, 1999.

[7] M. Gugat and G. Leugering, *Global boundary controllability of the de St. Venant equations between steady states*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 1–11.

[8] M. Gugat, G. Leugering, K. Schittkowski, and E. J. P. G. Schmidt, *Modelling, stabilization, and control of flow in networks of open channels*, in Online Optimization of Large Scale Systems, Martin Groetschel et al., eds., Springer, Berlin, 2001, pp. 251–270.

[9] M. Gugat, G. Leugering, and E. J. P. G. Schmidt, *Global controllability between steady supercritical flows in channel networks*, Math. Methods Appl. Sci., to appear.

[10] G. Leugering and E. J. P. Georg Schmidt, *On the modelling and stabilization of flows in networks of open canals*, SIAM J. Control Optim., 41 (2002), pp. 164–180.

[11] P. O. Lindberg and A. Wolf, *Optimization of the short term operation of a cascade of hydro power stations*, in Optimal Control: Theory, Algorithms, and Applications, W. H. Hager and P. M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 326–345.

[12] E. J. P. Georg Schmidt, *On a non-linear wave equation and the control of an elastic string from one equilibrium to another*, J. Math. Anal. Appl., to appear.

[13] Li Ta-tsien, *Global Classical Solutions for Quasilinear Hyperbolic Systems*, Masson, Paris, 1994.

[14] T.-T. Li, B. Rao, and Y. Jin, *Semi-global C^1 solution and exact boundary controllability for reducible quasilinear hyperbolic systems*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 399–408.

[15] T.-T. Li, B. Rao, and Y. Jin, *Solution C^1 semi–globale et contrôlabilité exacte frontière de systèmes hyperboliques quasi linéaires réductibles*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 205–210.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 1071–1086

DEGREES OF EFFICIENCY AND DEGREES OF MINIMALITY*

ALBERTO ZAFFARONI†

Abstract. In this work we characterize different types of solutions of a vector optimization problem by means of a scalarization procedure. Usually different scalarizing functions are used in order to obtain the various solutions of the vector problem. Here we consider different kinds of solutions of the *same* scalarized problem. Our results allow us to establish a parallelism between the solutions of the scalarized problem and the various efficient frontiers: stronger solution concepts of the scalar problem correspond to more restrictive notions of efficiency. Besides the usual notions of weakly efficient and efficient points, which are characterized as global and strict global solutions of the scalarized problem, we also consider some restricted notions of efficiency, such as strict and proper efficiency, which are characterized as Tikhonov well-posed minima and sharp minima for the scalarized problem.

Key words. vector optimization, scalarization, proper efficiency, strict efficiency, sharp minima, well-posed minima

AMS subject classifications. 90C29, 90C31, 90C48, 49K40

DOI. 10.1137/S0363012902411532

1. Introduction.

Both theory and practice of vector optimization have always been closely related to scalarization procedures. The most widely used is probably the linear scalarization; for Pareto optimization problems, in which the outcome space is ordered componentwise, it consists of the consideration of a weighted sum of components of the objective function with nonnegative coefficients.

In the seminal paper by Kuhn and Tucker [21], for the first time the concept of proper efficiency was introduced, precisely in order to prove that the multipliers of all components of the objective function in the necessary optimality conditions are (strictly) positive. To reach this result one has to require, besides a constraint qualification, a further requirement on the efficient point to avoid anomalous features.

In this fashion properly efficient points can be characterized (under convexity assumptions) as the solutions of a linearly scalarized problem in which all coefficients are positive. From a geometric point of view this entails that an open halfplane exists, whose normal vector is strictly positive and is disjoint from the image set.

In the last fifty years the notion of proper efficiency has been described in a number of ways (see, e.g., [13, 27, 6] for references and comparisons); the various definitions emphasize different aspects (boundedness of trade-offs between objectives, disjointness between the ordering cone and some conical approximation of the image set, stability properties with respect to the ordering structure), but they can be seen as an extension of the primitive idea in that they can be geometrically described in terms of separation between the ordering cone and the image set by means of an open convex cone or an open convex set.

It is thus evident that the nature of proper efficiency, whose origin has a purely local nature (in the classical paper by Kuhn and Tucker its definition is given in terms of the derivatives of the objective and constraint functions), also entails a global character.

*Received by the editors July 15, 2002; accepted for publication (in revised form) January 8, 2003; published electronically July 8, 2003.
http://www.siam.org/journals/sicon/42-3/41153.html
†Dipartimento di Scienze Economiche e Matematico-Statistiche, Università di Lecce, Centro Ecotekne, 73100 Lecce, Italy (azaffaroni@economia.unile.it).

The deep connection with scalarization has always been stressed, and various ad hoc scalarization techniques have been devised to relate, in a nonconvex problem, proper efficiency according to the various definitions to an optimal scalar solution.

For instance, in the scalarization approach proposed by Jahn [17, 18], which consists in the minimization of the distance (according to an appropriate parametric norm) from some ideal point, two different families of norms are used to characterize efficient and properly efficient points, the former being related to the ordering cone and the latter to another cone which contains the ordering cone in its interior; more generally, to obtain more restrictive notions of efficiency as minimal solution of a scalarized problem, stronger monotonicity features must be imposed on the scalarizing function (see, for instance, [27, 6, 10, 28]).

Proceeding in this vein one finds that efficient and properly efficient points can be seen as minimal solutions of different scalar problems.

A different type of restriction on efficiency has been posed in order to control the asymptotic behavior of unbounded admissible sequences, thus obtaining the notion of strict efficiency [2].

The aim of this work is to characterize the various sets of solutions of a vector optimization problem by means of a unique scalarizing function. A solution belonging to a more restrictive set of solutions is found as an optimal scalar solution according to a more restrictive definition of minimality: we will consider strict (i.e., unique) minima, sharp minima, or others. In particular, we will refer to some growth conditions known in scalar optimization and to the notion of Tikhonov well-posedness.

In section 2 we will introduce six different types of solutions of a vector optimization problem and show their relationships. Section 3 is devoted to the scalarizing function; we point out that our scalarizing function (the Δ function, introduced by Hiriart-Urruty [14, 15] for obtaining nonsmooth optimality conditions) is a very simple tool which immediately emphasizes the geometry of the ordering relation and has a very simple form in specific problems, thus allowing for simple calculations. In section 4 the main results of the paper are given: the six types of efficient solutions presented above appear as solutions of the scalarized problem with increasingly more restrictive minimality properties. In section 5 we consider again the concepts introduced in section 2 with the aim of comparing them to other, maybe better known, notions of efficiency. Some equivalence results are proved there in infinite and finite dimensional spaces, which complement the results in [13].

2. Degrees of efficiency. Let Y be a normed space, $S \subset Y$ be the set of admissible points, and $K \subset Y$ a closed, convex, pointed cone inducing on Y the partial order relation given by $y_1 \geq y_2$ if and only if $y_1 - y_2 \in K$. We say that the set $\Theta \subset Y$ is a base for K if Θ is convex with $0 \notin \operatorname{cl}\Theta$ and $K \setminus \{0\} = \operatorname{cone}\Theta = \{y \in Y : y = \lambda\theta, \lambda > 0, \theta \in \Theta\}$, i.e., K is the cone generated by Θ. Here and in what follows, we denote with $\operatorname{cl} A$, $\operatorname{int} A$, ∂A, and A^c the closure, the interior, the boundary, and, respectively, the complement of the set $A \subseteq Y$. Moreover, we will denote with B (respectively, $\hat{B}$) the closed (respectively, open) unit ball in Y and with $d_A(y) = \inf\{\|a - y\|, a \in A\}$ the distance function to the set $A \subseteq Y$.

DEFINITION 2.1. *A point $y_0 \in S$ is said to be* efficient *(with respect to K) and denoted $y_0 \in E(S, K)$ (or $y_0 \in E(S)$ if no confusion arises) if*

$$(S - y_0) \cap -K = \{0\}.$$

DEFINITION 2.2. *In the case where* $\operatorname{int} K \neq \emptyset$, *a point* $y_0 \in S$ *is said to be* weakly efficient *(denoted* $y_0 \in WE(S)$*) if*

$$(S - y_0) \cap -\operatorname{int} K = \emptyset.$$

Various other conditions have been used to single out particular classes of efficient points with special features. The concept of strict efficiency was introduced in [2] in order to obtain upper semicontinuity of the section mapping $G(y) = S \cap (y - K)$ at an efficient point.

DEFINITION 2.3. *A point* $y_0 \in S$ *is called* strictly efficient *(denoted* $y_0 \in StE(S)$*) if for every* $\varepsilon > 0$ *there exists* $\delta > 0$ *such that*

$$(S - y_0) \cap (\delta B - K) \subseteq \varepsilon B.$$

An efficient point $y_0 \in S$ is strictly efficient if the points $y \in Y$, which are dominated by y_0, i.e., $y \leq y_0$, and are bounded away from y_0, cannot be arbitrarily close to the feasible region S.

Proper efficiency instead implies a control on the tangentially admissible directions close to the efficient point. It has been defined in a great number of ways; in [13] an attempt was made to classify the various definitions along three main classes whose elements are proved to be equivalent in Euclidean spaces. Further notions are analyzed in the paper by Zalinescu [27]. It readily becomes evident that most definitions of proper efficiency entail a restriction on the admissible region which has a global character. In what follows we will introduce a few of them; since the ones we will mention are often not the better known, we give in section 5 some comparisons with others, without any scope of completeness. The interested reader can refer to [6, 27, 13] for more details.

DEFINITION 2.4 (see [6]). *The point* $y_0 \in S$ *is called* superefficient *(*$y_0 \in SE(S)$*) if there exists* $M > 0$ *such that*

$$\operatorname{cl}\operatorname{cone}(S - y_0) \cap (B - K) \subseteq MB. \tag{2.1}$$

Remark 2.5. The inclusion (2.1) can be equivalently expressed (see [6]) as

$$\operatorname{cone}(S - y_0) \cap (B - K) \subseteq MB;$$

in what follows we will often use this equivalent formulation.

DEFINITION 2.6. *The point* $y_0 \in S$ *is* locally superefficient *(*$y_0 \in LSE(S)$*) if it is efficient and there exist* $\eta > 0$ *and* $M > 0$ *such that*

$$\operatorname{cl}\operatorname{cone}[(S - y_0) \cap \eta B] \cap (B - K) \subseteq MB.$$

It is obvious that $y_0 \in LSE(S)$ if and only if there exists $\eta > 0$ such that $y_0 \in SE(S \cap (y_0 + \eta B))$.

We will also refer to another notion of efficiency which is somehow intermediate between the previous ones.

DEFINITION 2.7. *The point* $y_0 \in S$ *is called* tightly properly efficient *(*$y_0 \in TPE(S)$*) if there exists an open convex set* $C \subset Y$ *with* $0 \in \partial C$ *satisfying* $(S - y_0) \cap C = \emptyset$ *and there exists* $\delta > 0$ *such that*

$$C^c \cap (\delta B - K) \subseteq B. \tag{2.2}$$

Our first aim is to clarify the relationships among the above concepts.

To understand the relation between superefficiency and tight efficiency, we will need the concept of a dilating cone. If Θ is a base for the cone K, then by definition $0 \notin \operatorname{cl}\Theta$ and $\|\theta\| > \delta > 0$ for some $\delta > 0$ and all $\theta \in \Theta$. For every $\varepsilon < \delta$ the cone $K_\varepsilon = \operatorname{cone}(\Theta + \varepsilon\hat{B})$, where $\hat{B}$ is the open unit ball in Y, is an open convex cone containing $K\backslash\{0\}$. Note that the assumption that a cone K has a base is equivalent to the existence of strictly positive continuous linear functionals, i.e., the set

$$K^{+i} = \{\ell \in Y' : \ell(k) > 0 \quad \forall k \in K \setminus \{0\}\}$$

is nonempty, where Y' is the (topological) dual space of Y. Moreover, if a closed cone K admits a base, then the base can be taken to be closed.

THEOREM 2.8. *If the cone K has a bounded base, then $SE(S) \subseteq TPE(S)$, i.e., every superefficient point is also tightly efficient.*

Proof. To simplify the presentation and without loss of generality, we will prove the result under the assumption that $y_0 = 0$. As the first step of the proof, we need to show that, if 0 is superefficient, there exists an open convex dilating cone K_ε such that

$$\operatorname{cl}\operatorname{cone} S \cap -K_\varepsilon = \emptyset.$$

This is exactly the content of Proposition 3.3 in [6].

Second, we prove that there exists $N > 0$ such that

$$(B - K) \cap (-K_\varepsilon)^c \subseteq NB. \tag{2.3}$$

Indeed, set $W = (-K_\varepsilon)^c$, and suppose relation (2.3) is false, i.e., there exists a sequence w_n, $n \in \mathbb{N}$, such that $w_n \in W$, $\|w_n\| > n$, and $w_n \in B - K$. The latter means that there exists a sequence $k_n \in K$ such that $\|w_n + k_n\| \leq 1$. Thus there exists a sequence $b_n \in B$ such that $w_n + k_n = b_n$, which can be rewritten as

$$w_n = b_n - k_n = b_n - \lambda_n\theta_n = \lambda_n(b_n/\lambda_n - \theta_n),$$

where $\theta_n \in \Theta$ and $\lambda_n > 0$. But this is a contradiction to $w_n \notin (-K_\varepsilon)^c$, since, by the boundedness of Θ, we have that $\lambda_n \to +\infty$ as $\|w_n\| \to +\infty$ and $1/\lambda_n$ becomes smaller than any fixed ε. Thus eventually $w_n \in \operatorname{cone}(\varepsilon B - \Theta)$.

Now the definition of tight efficiency is verified with $C = -K_\varepsilon$, by taking $\delta = 1/N$. ∎

Cones with a bounded base have been extensively used in the literature related to ordered spaces and vector optimization under a variety of different characterizations which were later proved to be equivalent. For instance, in [26] it is proved that a cone K has a bounded base if and only if it is supernormal (or nuclear) (see [16] for the definition) and if and only if there exists $\phi \in K^{+i}$ and a scalar α such that $\phi(k) \geq \alpha\|k\|$ for all $k \in K$ (Bishop–Phelps cone), and in [3] it is proved that a cone K has bounded base if and only if it allows plastering (see [20]).

To show that a tightly efficient point is both locally superefficient and strictly efficient, we need a preliminary result.

LEMMA 2.9. *The point y_0 is tightly efficient in S if and only if there exists an open convex set $C \subset Y$, with $0 \in \partial C$ satisfying $(S - y_0) \cap C = \{0\}$, and for every $\varepsilon > 0$ there exists $\delta' > 0$ such that*

$$C^c \cap (\delta' B - K) \subseteq \varepsilon B.$$

Proof. Possibly after a translation of S, assume that $y_0 = 0$ is a tightly properly efficient point for S and fix $\varepsilon > 0$. If $\varepsilon \geq 1$, the thesis follows with $\delta' = \delta$, as in Definition 2.7. So let $\varepsilon \in (0,1)$ and choose δ such that

$$(\delta B - K) \cap C^c \subseteq B.$$

This implies that

$$\varepsilon(\delta B - K) \cap \varepsilon(C^c) \subseteq \varepsilon B.$$

Noting that $\varepsilon(\delta B - K) = (\varepsilon\delta B - K)$, the "only if" part of the thesis follows by the observation that for an open convex set C with $0 \in \operatorname{cl} C$ and any $\varepsilon \in (0,1)$ it holds that $\varepsilon C \subseteq C$ and $C^c \subseteq (\varepsilon C)^c = \varepsilon(C^c)$. The reverse inclusion is obvious. ∎

THEOREM 2.10. *If the point y_0 is tightly efficient in S, then it is both strictly efficient and locally superefficient.*

Proof. Strict efficiency follows by simply comparing the definitions and using Lemma 2.9. Moreover, if $(\delta B - K) \cap C^c \subseteq \varepsilon B$, then it holds that $-K\backslash\{0\} \subseteq C$. Indeed, if $k \in K\backslash\{0\}$ and $-k \notin C$, then $-\alpha k \notin C$ for all $\alpha > 1$ (remember that $0 \in \operatorname{cl} C$) and $-\alpha k \in -K \subset \delta B - K$, and thus we have a contradiction.

To prove that $0 \in LSE(S)$ we first notice that

$$S \cap C = \emptyset \Rightarrow S \subset C^c \Rightarrow [1,\infty) \cdot S \subset C^c.$$

Hence $\operatorname{cone}(S \cap B) \subset B \cup [1,\infty) \cdot S \subset B \cup C^c$. Since $(\delta B - K) \cap C^c \subset B$ (for some $\delta > 0$), we have

$$(B - K) \cap C^c \subset (\delta B - K) \cap C^c \subset B$$

when $\delta \geq 1$ and

$$(B - K) \cap C^c \subset (B - K) \cap \delta^{-1} C^c \subset \delta^{-1} B$$

when $\delta < 1$. Thus, taking $M = \max\{1, \delta^{-1}\}$, we get

$$\operatorname{cone}(S \cap B) \cap (B - K) \subset (B \cap (B - K)) \cup (C^c \cap (B - K)) \subset MB. \qquad ∎$$

The opposite relation is true under the assumption that the cone K has a bounded base.

THEOREM 2.11. *If the cone K has a bounded base and the point y_0 is both strictly efficient and locally superefficient in S, then it is also tightly efficient.*

Proof. Take η such that the point y_0 is superefficient for the admissible set $S \cap (y_0 + \eta B)$ and, by strict efficiency, choose δ' such that

$$(S - y_0) \cap (\delta' B - K) \subseteq \eta B. \tag{2.4}$$

Then follow the proof of Theorem 2.8 to show that there exists $\varepsilon > 0$ such that

$$(S - y_0) \cap \eta B \cap (-K_\varepsilon) = \emptyset \tag{2.5}$$

(see again [6, Prop. 3.3]) and that there exists $N > 0$ such that

$$(B - K) \cap (-K_\varepsilon)^c \subseteq NB,$$

which is equivalent to

$$[(1/N)B - K] \cap (-K_\varepsilon)^c \subseteq B.$$

Now choose $\delta = \min(\delta', 1/N)$ and set $C = (\delta\hat{B} - K) \cap (-K_\varepsilon)$, where $\hat{B} = \operatorname{int} B$ denotes the open unit ball. Then C is open and convex, with $0 \in \partial C$. It follows from (2.4) that

$$(\delta B - K) \cap (S - y_0) \subseteq \eta B,$$

which, together with (2.5), yields

$$(\delta B - K) \cap (S - y_0) \cap -K_\varepsilon = \emptyset,$$

that is,

$$(S - y_0) \cap C = \emptyset.$$

Moreover,

$$\left[\frac{\delta}{2}B - K\right] \cap C^c = \left[\frac{\delta}{2}B - K\right] \cap [(\delta\hat{B} - K)^c \cup (-K_\varepsilon)^c] = \left[\frac{\delta}{2}B - K\right] \cap (-K_\varepsilon)^c \subseteq B,$$

which is our thesis. □

To summarize we can state that, if the cone K has a bounded base, then the following inclusions hold:

$$SE(S) \subseteq TPE(S) = LSE(S) \cap StE(S) \subseteq LSE(S) \cup StE(S) \subseteq E(S) \subseteq WE(S),$$

and simple examples in $\mathbb{R}^2$ show that all inclusions are strict. The six notions we introduced can thus be seen as six different degrees of efficiency for a vector optimization problem.

3. Scalarization. The theory and the methods of scalarization have always been of the utmost importance for solving a vector optimization problem. The linear scalarization is historically the first method proposed and the most widely known and used; it consists of the minimization, over the set S, of the function $\lambda \cdot y$, with $\lambda \in K^+$. Besides this, in order to treat nonconvex problems, the method of compromise solutions and its generalizations are of great relevance. It consists of the minimization of the distance from some reference objective, often not belonging to the admissible region, but dominating all available alternatives. This was originally done in the Paretian setting in which the ordering cone is simply the nonnegative orthant, with the use of the supremum norm of $Y = \mathbb{R}^p$. The main idea has successively been extended to the setting of general ordered vector space (see, e.g., [17, 18, 19] and references therein) by defining the norm as the Minkowski functional of the order interval $[-a, a] \equiv (-a + K) \cap (a - K)$, i.e.,

$$\|y\|_a = \inf\{\lambda > 0 : \lambda^{-1}y \in [-a, a]\},$$

where a is some point in the interior of the ordering cone K and the order interval $[-a, a]$ is a closed, convex set with nonempty interior.

For a fixed reference point $\ell \in Y$ such that $S \subset \ell + K$, Jahn characterizes weakly efficient (respectively, efficient) points as the nearest (respectively, unique nearest)

points to ℓ with respect to the distance induced by $\|\cdot\|_a$. To characterize properly efficient points, Jahn considers the nearest points to ℓ according to a different norm, defined by means of the dilating cone K_ε.

Closely related to this approach are some variants in which the scalarizing function is defined by means of a sort of Minkowski functional of sets related to the ordering cone, as, for instance, in [27, 10, 28]. In this case the scalarizing functional $\phi : X \to \overline{\mathbb{R}}$ is given by

$$\phi(x) = \inf\{\lambda \in \mathbb{R} : \ x \in \lambda e - A\},$$

where $e \in \operatorname{int} A$. If the set A coincides with the ordering cone K, then characterizations of efficient and weakly efficient points are obtained. However, to obtain properly efficient points as a minimal scalar solution, the scalarizing function must be defined with respect to some convex set or cone A which contains $K \setminus \{0\}$ in its interior, and hence every properly efficient point is the solution of a different scalarized problem.

The main results of this paper give a complete characterization of the different types of efficient points for a vector optimization problem by means of different degrees of minimality of a unique scalarized problem. This is obtained by means of a special scalarizing function.

DEFINITION 3.1. *For a set $A \subseteq Y$ let the function $\Delta_A : Y \to \mathbb{R} \cup \{\pm\infty\}$ be defined as*

$$\Delta_A(y) = d_A(y) - d_{Y\setminus A}(y),$$

with $d_\emptyset(y) = +\infty$.

The function Δ was introduced in [14, 15] to analyze the geometry of nonsmooth optimization problems and obtain necessary optimality conditions. It has later been used in [8, 1, 22, 23]. Its main properties are gathered together in the following proposition.

PROPOSITION 3.2. *If the set A is nonempty and $A \neq Y$, then*

(1) *Δ_A is real valued;*
(2) *Δ_A is 1-Lipschitzian;*
(3) *$\Delta_A(y) < 0$ for every $y \in \operatorname{int} A$, $\Delta_A(y) = 0$ for every $y \in \partial A$, and $\Delta_A(y) > 0$ for every $y \in \operatorname{int} A^c$;*
(4) *if A is closed, then it holds that $A = \{y : \ \Delta_A(y) \leq 0\}$;*
(5) *if A is convex, then Δ_A is convex;*
(6) *if A is a cone, then Δ_A is positively homogeneous;*
(7) *if A is a closed convex cone, then Δ_A is nonincreasing with respect to the ordering relation induced on Y, i.e., the following is true: if $y_1, y_2 \in Y$, then*

$$y_1 - y_2 \in A \qquad \Longrightarrow \qquad \Delta_A(y_1) \leq \Delta_A(y_2);$$

if A has a nonempty interior, then

$$y_1 - y_2 \in \operatorname{int} A \qquad \Longrightarrow \qquad \Delta_A(y_1) < \Delta_A(y_2).$$

Proof. Statements (1)–(4) and (6) are immediate. To prove (5) one can pass through a characterization (given in [15]) of Δ_A as an infimal convolution:

$$\Delta_A(y) = (\mu_A \nabla \|\cdot\|)(y) := \inf\{\mu_A(x) + \|y - x\| : \ x \in Y\}$$

, where the function $\mu_A(y) = +\infty$ for $y \notin A$ and $\mu_A(y) = -d_{Y\setminus A}(y)$ for $y \in A$ can be proved to be convex when A is convex. This result is stated without proof in [14]

and is equivalent to the concavity of the function d_{A^c} on A. To show that the latter holds when A is convex, one can notice that for any $x, y \in A$, the closed balls B_x and B_y centered in x and in y with radii $d_{A^c}(x)$ and $d_{A^c}(y)$, respectively, are contained in $\operatorname{cl} A$ (which is itself convex), as well as the set $H = \operatorname{conv}\{B_x \cup B_y\}$. Moreover, for every $\alpha \in [0,1]$, the ball centered in $\alpha x + (1-\alpha)y$ with radius $\alpha d_{A^c}(x) + (1-\alpha) d_{A^c}(y)$ is contained in H, and hence $d_{A^c}(\alpha x + (1-\alpha)y) \geq \alpha d_{A^c}(x) + (1-\alpha) d_{A^c}(y)$. The result then follows, since the convolution of a convex function with the norm is itself convex. To prove 7 use nonpositivity of Δ_A on A and subadditivity to write $0 \geq \Delta_A(y_1 - y_2) \geq \Delta_A(y_1) - \Delta_A(y_2)$; the second implication is proved analogously. □

For our purposes, let $A = -K$. Then the function $\Delta_{-K}(y)$ is sublinear and nondecreasing with respect to the ordering induced by K.

It is important to note that the use of the scalarizing function Δ_{-K} implies no assumption (explicit or implicit) of boundedness of the admissible region S. On the other hand such assumptions are required for most other scalarizations. For instance S has to be contained in a halfspace if the linear scalarization is used, or it has to be lower bounded, in the sense that there exists an element $\ell \in Y$ such that $s \geq \ell$ for every $s \in S$, if the distance from an ideal point is used.

This remark has some relevance in that the main differences among the different types of efficient points considered in our analysis disappear under boundedness assumptions.

We give now some examples of how the function Δ_{-K} looks for different choices of the space Y and its norm and the ordering cone K.

Examples.

1. Let $Y = \mathbb{R}^n$ with the Euclidean norm $\|\cdot\|_2$ and $K = \mathbb{R}^n_+$. Then it holds that $d_{-K}(y) = \|y^+\|$, where $y_i^+ = \max(y_i, 0)$, $i = 1, \ldots, n$, and

$$d_{Y\setminus -K}(y) = \begin{cases} 0 & \text{if} \quad y_i \geq 0 \text{ for some } i, \\ -\max_i y_i & \text{if} \quad y_i < 0 \ \forall i. \end{cases}$$

   Thus it holds that

$$\Delta_{-K}(y) = \begin{cases} \|y^+\| & \text{if} \quad y \notin -K, \\ \max y_i & \text{if} \quad y \in -K. \end{cases}$$

2. The function Δ_{-K} takes a more familiar form if we consider $Y = \mathbb{R}^n$ with the norm $\|y\|_\infty = \max |y_i|$; in this case we have

$$d_{-K}(y) = \begin{cases} \max_i y_i & \text{if} \quad y \notin -K, \\ 0 & \text{if} \quad y \in -K \end{cases}$$

   and

$$d_{Y\setminus -K}(y) = \begin{cases} 0 & \text{if} \quad y \notin -K, \\ -\max y_i & \text{if} \quad y \in -K, \end{cases}$$

   and thus, for all $y \in Y$,

$$\Delta_{-K}(y) = \max_i y_i.$$

3. The same reasoning can be applied to the case where $Y = \mathcal{C}(T)$, the space of continuous functions defined over the compact set T, with the supremum norm and the ordering cone of nonnegative functions $K = \{y \in Y : y(t) \geq$

$0\ \forall t \in T\}$. In this case, for those $y \in \mathcal{C}(T)$ for which there exists some $t \in T$ with $y(t) \geq 0$, it holds that $\Delta_{-K}(y) = d_{-K}(y) = \max_{t\in T} y(t)$, and for those $y \in \mathcal{C}(T)$ such that $y(t) < 0$ for all $t \in T$, then it holds that $\Delta_{-K}(y) = -d_{Y\setminus -K}(y) = \max\{y(t), t \in T\}$. Therefore

$$\Delta_{-K}(y) = \max_{t\in T} y(t)$$

for all $y \in Y$.

4. Analogously for the space $Y = L^\infty(I)$ of essentially bounded functions on the interval $I \subseteq \mathbb{R}$, with the usual norm and the cone K of nonnegative functions, it holds that

$$\Delta_{-K}(y) = \sup_{\omega\in I} y(\omega),$$

where sup means essential supremum.

5. For the spaces L^p, with $1 \leq p < \infty$ (with the usual norm $\|\cdot\|_p$), we have that the nonnegative orthant K has an empty interior. In this case we have $\Delta_{-K}(y) = d_{-K}(y) = \|y^+\|$, where $y^+ = \sup(y, 0)$ has the usual meaning of lattice theory.

4. Characterizations. We give in this section the main results of the paper. The various types of efficient solutions introduced in section 2 will be characterized by different degrees of minimality of the scalar solutions of the parametrized problem:

$$(P_p) \qquad \min_{y\in S} \Delta_{-K}(y - p),$$

with $p \in Y$.

Theorem 4.1 characterizes efficient solutions as the strict minimal points for the scalarized problem. This result is well known for other types of scalarizations and was proved in [22] with the scalarizing function Δ_{-K}. It also appears in [23] in an axiomatic setting. We prove it for completeness.

THEOREM 4.1. *Let $y_0 \in S$ be an admissible point. Then $y_0 \in E(S)$ if and only if there exists $\hat{y} \in Y$ such that y_0 is a unique (strict) global solution of $(P_{\hat{y}})$.*

Proof. If y_0 is efficient, then it is a strict minimum for (P_{y_0}). Indeed, if $y_0 \in E(S)$, then $y - y_0 \notin -K \setminus \{0\}$ for every $y \in S$, and we have that $\Delta_{-K}(y - y_0) = d_{-K}(y - y_0)$ is positive whenever $y \neq y_0$ and is null at $y = y_0$. To prove the converse, suppose that $\Delta_{-K}(y_0 - \hat{y}) < \Delta_{-K}(y - \hat{y})$ for all $y \in S$ with $y \neq y_0$. If there exists in S some point $y_1 \neq y_0$ such that $y_1 \leq y_0$, then it would hold that $y_1 - \hat{y} \leq y_0 - \hat{y}$ and, by Proposition 3.2(7) we get $\Delta_{-K}(y_1 - \hat{y}) \leq \Delta_{-K}(y_0 - \hat{y})$, which is absurd. □

THEOREM 4.2. *Let $y_0 \in S$ be an admissible point and suppose that $\operatorname{int} K \neq \emptyset$; then y_0 is weakly efficient in S if and only if there exists $\hat{y} \in Y$ such that y_0 is a global solution of $(P_{\hat{y}})$.*

Proof. The proof is analogous to the one of Theorem 4.1 and hence is omitted. □

We observe that, if the cone K has empty interior, then $\Delta_{-K} = d_{-K}$ and, for every $\hat{y} \in S$, $d_{-K}(y - \hat{y}) = d_{-K}(\hat{y} - \hat{y}) = 0$ for all $y \leq \hat{y}$, with $y \in S$. Hence if the point $\hat{y}$ is not itself efficient, then the scalarized problem will have solutions at all $y \in S$ with $y \leq \hat{y}$ and uniqueness will fail. In this case, Theorem 4.1 can be reformulated as follows.

THEOREM 4.3. *Let $y_0 \in S$ be an admissible point; then $y_0 \in E(S)$ if and only if y_0 is a unique (strict) global solution of (P_{y_0}).*

Proof. The proof is trivial. □

In what follows we will always use problem (P_{y_0}) to characterize efficiency properties of the point y_0. It is understood that, if $\operatorname{int} K \neq \emptyset$, then a greater generality can be achieved (as in Theorem 4.1) considering a different value $\hat{y}$ for the parameter.

If the point y_0 is strictly efficient, then we can prove a further property for the scalarizing function $\Delta_{-K}(\cdot - y_0)$.

THEOREM 4.4. *Let $y_0 \in S$ be an admissible point. Then y_0 is strictly efficient in S if and only if there exists a nondecreasing function $\gamma : \mathbb{R}^+ \to \mathbb{R}^+$ with $\gamma(0) = 0$ and $\gamma(t) > 0$ for all $t > 0$, such that $\Delta_{-K}(y - y_0) \geq \gamma(\|y - y_0\|)$ for all $y \in S$.*

Proof. Strict efficiency of y_0 can be rephrased as follows: for every $\varepsilon > 0$ there exists $\delta > 0$ such that $d_{-K}(y - y_0) > \delta$ for every $y \in S$ with $\|y - y_0\| > \varepsilon$. So suppose that the point y_0 is strictly efficient in S and consider the functions

$$\gamma_0(\varepsilon) = \inf\{d_{-K}(y - y_0) \,|\, y \in S, \|y - y_0\| \geq \varepsilon\}$$

and

$$\gamma(\varepsilon) = \min(\gamma_0(\varepsilon), 1).$$

It is easily seen that γ is nondecreasing, null at the origin, and positive elsewhere; moreover, for $y \in S$ it holds that

$$\Delta_{-K}(y - y_0) = d_{-K}(y - y_0) \geq \gamma(\|y - y_0\|).$$

If, on the other hand, there exists a nondecreasing function γ with the above properties and such that $\Delta_{-K}(y - y_0) \geq \gamma(\|y - y_0\|)$ for all $y \in S$, then it holds that $\Delta_{-K}(y - y_0) > 0$ for all $y \in S$ with $y \neq y_0$, yielding $y_0 \in E(S)$ and $\Delta_{-K}(y - y_0) = d_{-K}(y - y_0)$ for all $y \in S$. To show that $y_0 \in StE(S)$, with every $\varepsilon > 0$ we can associate $\delta = \inf\{d_{-K}(y - y_0) \,|\, \|y - y_0\| > \varepsilon\}$, and the result is proved. □

The definition of strict efficiency can be reformulated to offer a version of Tikhonov well-posedness (in the image) for a vector problem. The problem of minimizing an extended real valued function f defined over some metric space X is said to be Tikhonov well-posed when it admits a unique minimum x_0 and for every sequence x_n such that $f(x_n)$ converges to the infimal value $f(x_0)$, it holds that $x_n \to x_0$. We refer to the monograph by Dontchev and Zolezzi [11] for more details about well-posedness in optimization. A function γ with the properties mentioned in Theorem 4.4 is called a forcing function in [11] and a growth gage in [24].

Though it is well known by specialists in well-posedness and rather trivial, it should be recalled that the forcing property of a nondecreasing function $\gamma : \mathbb{R}^+ \to \mathbb{R}^+$ with $\gamma(0) = 0$ can be equivalently expressed as $\gamma(t) > 0$ for all $t > 0$ or $\gamma(t_n) \to 0 \Rightarrow t_n \to 0$.

In our formulation the objective function is $\Delta_{-K}(\cdot - y_0)$, defined over Y and extended to $+\infty$ outside the set S. The minimization problem (P_{y_0}) is Tikhonov well-posed if $\Delta_{-K}(y - y_0) > 0$ for all $y \in S$ with $y \neq y_0$ and

$$(\ y_n \in S, \quad d_{-K}(y_n - y_0) \to 0\ ) \qquad \Longrightarrow \qquad y_n \to y_0;$$

that is, if, when a sequence $y_n \in S$ is minimizing, i.e., it belongs to S and approaches the region dominated by y_0, then it must converge to y_0.

Theorem 4.5 makes clearer the interpretation of strict efficiency as a sort of well-posedness for the scalarized problem (P_{y_0}). Notice that the equivalence between

Tikhonov well-posedness and the existence of a forcing function is a known result; see, e.g., [29].

THEOREM 4.5. *The admissible point $y_0 \in S$ is strictly efficient for S if and only if y_0 is a solution of (P_{y_0}) and the problem (P_{y_0}) is Tikhonov well-posed.*

Proof. If y_0 is strictly efficient, then it is the unique solution for the scalarized problem (P_{y_0}) and there exists a forcing function γ such that $\Delta_{-K}(y - y_0) \geq \gamma(\|y - y_0\|)$. Hence all minimizing sequences, i.e., sequences $y_n \in S$ such that $\Delta_{-K}(y_n - y_0) \to 0$, must converge to y_0, since $\gamma(t_n) \to 0$ implies $t_n \to 0$. Conversely, if the problem (P_{y_0}) is well-posed, then y_0 is efficient and, for every $y \in S$, it holds that $d_{-K}(y - y_0) = \Delta_{-K}(y - y_0)$. Thus we can consider the function $\gamma(\varepsilon) = \inf\{d_{-K}(y - y_0) \mid \|y - y_0\| \geq \varepsilon\}$; it holds by construction that $d_{-K}(y - y_0) \geq \gamma(\|y - y_0\|)$, and it is easy to see that γ is nondecreasing on $[0, +\infty)$ with $\gamma(\varepsilon) > 0$ for $\varepsilon > 0$; hence y_0 is strictly efficient. ∎

As we have seen in Theorem 4.4, strict efficiency of the point y_0 guarantees the existence of a forcing function for $\Delta_{-K}(\cdot - y_0)$. But there is no control on the slope of such a function close to zero. This is where proper efficiency comes into the picture. Indeed, we will show that some properties of the forcing function γ can be strengthened when y_0 is properly efficient to obtain a linear growth for Δ_{-K}. This will hold near y_0 if y_0 is locally superefficient and will hold globally for $y_0 \in SE(S)$.

THEOREM 4.6. *Let $y_0 \in S$ be an admissible point. Then the following hold:*

(a) *y_0 is superefficient in S if and only if there exists $L > 0$ such that $\Delta_{-K}(y - y_0) \geq L\|y - y_0\|$ for all $y \in S$.*

(b) *y_0 is locally superefficient in S if and only if there exist $\eta > 0$ and $L > 0$ such that $\Delta_{-K}(y - y_0) \geq L\|y - y_0\|$ for all $y \in S \cap (y_0 + \eta B)$.*

Proof. (a) We first need to show that, for an efficient point $y_0 \in S$, the inclusion

$$\operatorname{cone}(S - y_0) \cap (B - K) \subseteq MB \tag{4.1}$$

can be expressed by saying that for each $y \in S$ and each $k \in K$ it holds that

$$\|y - y_0\| \leq M\|y + k - y_0\|. \tag{4.2}$$

Indeed, if (4.1) does not hold, then there exist a positive number λ and some $\bar{y} \in S$, $\bar{k} \in K$, and $\bar{b} \in B$ such that

$$\lambda(\bar{y} - y_0) = \bar{b} - \bar{k} \tag{4.3}$$

and

$$\lambda\|\bar{y} - y_0\| > M.$$

From (4.3) we derive

$$\|\lambda(\bar{y} - y_0) + \bar{k}\| \leq 1$$

and

$$\left\|\bar{y} - y_0 + \frac{\bar{k}}{\lambda}\right\| \leq \frac{1}{\lambda},$$

which, together with

$$\|\bar{y} - y_0\| > \frac{M}{\lambda},$$

give a contradiction to (4.2).

If conversely (4.2) does not hold, then there exist $\bar{y} \in S$ and $\bar{k} \in K$ such that

$$\|\bar{y} - y_0\| > M\|\bar{y} - y_0 + \bar{k}\|.$$

Then there exists $\bar{b} \in B$ such that

$$\bar{y} - y_0 + \bar{k} = \bar{b}\|\bar{y} - y_0 + \bar{k}\|,$$

and this yields

$$s = \frac{\bar{y} - y_0}{\|\bar{y} - y_0 + \bar{k}\|} = \bar{b} - \frac{\bar{k}}{\|\bar{y} - y_0 + \bar{k}\|} \in \operatorname{cone}(S - y_0) \cap B - K$$

and

$$\|s\| = \frac{\|\bar{y} - y_0\|}{\|\bar{y} - y_0 + \bar{k}\|} > M,$$

which means that (4.1) does not hold.

To conclude the proof it is enough to see that, given the point $y_0 \in S$, there exists $M > 0$ such that (4.2) holds if and only if $\|y + k - y_0\| \geq L\|y - y_0\|$ for $L = 1/M$, and this is equivalent to

$$d_{-K}(y - y_0) = \inf_{k \in K} \|y + k - y_0\| \geq L\|y - y_0\| \qquad \forall y \in S. \tag{4.4}$$

Notice at last that under each of the following assumptions separately, that $y_0 \in SE(S)$ or that Δ_{-K} is nonnegative, then d_{-K} coincides with Δ_{-K}, so that (4.4) proves the result.

(b) The proof of (b) is the same as in part (a) for $\|y - y_0\| \leq \eta$. □

We recall that in mathematical programming a point x_0 is called a sharp minimum for the function $f : X \to [-\infty, +\infty]$ relative to the set $A \subseteq X$ if there exists $\alpha > 0$ such that it holds that

$$f(x) \geq f(x_0) + \alpha\|x - x_0\|$$

for all $x \in A$. This has important consequences in convergence analysis of many iterative procedures (see, e.g., [25, 7] for details and references). Thus, if a point x_0 is a sharp minimum for the function f, then f admits a forcing function which is linear, $\gamma(t) = \alpha t$.

Obviously for a local sharp minimum there exists a forcing function with positive slope close to the minimum point.

On the other hand, note that the requirement that y_0 be locally superefficient does not imply that $\Delta_{-K}(\cdot - y_0)$ admits a forcing function, i.e., the largest nondecreasing function γ such that $\gamma(\|y - y_0\|)$ minorizes $\Delta_{-K}(y - y_0)$ might be identically zero. This is made clear by the following example.

Example 4.7. Let the function $G : \mathbb{R}^2 \to \mathbb{R}$ be defined as $G(y_1, y_2) = y_2 + y_1 e^{y_1}$ for $y_1 < y_2$ and $G(y_1, y_2) = y_1 + y_2 e^{y_2}$ for $y_1 \geq y_2$; take $S = \{(y_1, y_2) : G(y_1, y_2) \leq 0\}$ and $K = \mathbb{R}^2_+$. The origin is the only efficient point for S and it is also locally superefficient, but it is not strictly efficient. Indeed S is asymptotically close to $-K$ as shown by the sequence $y^n = (y^n_1, y^n_2) = (-n, ne^{-n}) \in S$, with $\|y^n\| \to +\infty$.

It follows from Theorems 2.10 and 2.11 that a tightly properly efficient solution $y_0 \in S$ can be characterized in terms of the scalarized problem (P_{y_0}) by requiring at

the same time the properties of the forcing function γ which hold for a well-posed minimum (γ is nondecreasing and positive outside the origin) and for a local sharp minimum (γ has a positive slope at the origin).

COROLLARY 4.8. *If the admissible point $y_0 \in S$ is tightly properly efficient, then for the problem (P_{y_0}) there exists a nondecreasing growth function with positive slope at the origin. If the cone K has a bounded base, then the converse is true.*

5. Further results about strict and proper efficiency. This concluding section is devoted to a deeper analysis of the notions of (restricted) efficiency defined in section 2. Indeed, the concept of strict efficiency has only recently been introduced and can be given an equivalent description in finite dimensional spaces, which sheds new light on the geometry of the admissible region. The concept of proper efficiency has a much longer history in vector optimization, and the underlying idea has been analytically described in a great number of ways. In [13] an attempt was made to classify the known definitions in three main classes, each collecting definitions which coincide in finite dimensional spaces. We will see that the notions of superefficiency, local superefficiency, and tight efficiency can be seen as representative examples of the three above-mentioned classes.

The reason why we restrict to finite dimensional spaces for some of the results of the present section is that we will need to assume that the ordering cone K has a (weakly) compact base Θ, and this is indeed true, in any Euclidean space, for any closed convex pointed cone K, while the same assumption proves to be very restrictive in infinite dimensional spaces (see [9]) and fails to hold for the nonnegative orthant in most common spaces.

It is immediately verified that, if K is pointed, $y_0 \in S$ is efficient exactly when $(S + K - y_0) \cap -K = \{0\}$ holds. In the case where S in unbounded, however, the set $S + K$ need not be closed even if both S and K are. Thus the condition $\mathrm{cl}\,(S + K - y_0) \cap -K = \{0\}$ is a stronger requirement on y_0 than only efficiency. We will see that the previous condition is related to strict efficiency.

THEOREM 5.1. *If Y is any normed space and $y_0 \in S$ is strictly efficient, then $\mathrm{cl}\,(S + K - y_0) \cap -K = \{0\}$. If Y is finite dimensional, the converse is true.*

Proof. The definition of strict efficiency can be rephrased as follows: for every $\varepsilon > 0$ there exists $\delta > 0$ such that $d_{-K}(y - y_0) > \delta$ for every $y \in S$ with $\|y - y_0\| > \varepsilon$. Hence, if there would exist sequences $y_n \in S$, $k_n \in K$, and some $k \in K\backslash\{0\}$ such that $y_n + k_n - y_0 \to -k$, then $y_n + k + k_n - y_0 \to 0$ and hence $d_{-K}(y_n - y_0) \to 0$; moreover, $y_n - y_0$ is outside some small ball around the origin since $y_n - y_0 = k + k_n$, $k \neq 0$, and K is pointed. This shows that y_0 is not strictly efficient.

On the other hand, if $y_0 \in S$ is not strictly efficient, then there exist $\varepsilon > 0$ and sequences $y_n \in S$ and $k_n \in K$ such that $\|y_n - y_0\| \geq \varepsilon$ and $\|y_n + k_n - y_0\| \to 0$. Write $k_n = \lambda_n \theta_n$ with $\lambda_n > 0$ and $\theta_n \in \Theta$, and take some $\alpha < \lambda_n$ for all $n \in \mathbb{N}$ (such a number α exists since $k_n \notin (\varepsilon/2)B$ and Θ is compact) to define $k'_n = (\lambda_n - \alpha)\theta_n$. Thus we obtain

$$y_n + k'_n - y_0 = y_n + k_n - y_0 - \alpha\theta_n \to -\alpha\theta \neq 0,$$

at least for some subsequence $\{\theta_{n_k}\} \subseteq \{\theta_n\}$. Hence $\mathrm{cl}\,(S + K - y_0) \cap -K\backslash\{0\} \neq \emptyset$. □

We come now to local superefficiency; we will show that a locally superefficient point can be described in terms of the definition of local proper efficiency given by Borwein in [5], based on the separation between the ordering cone and a local conical

approximation of the feasible region. For a set $A \subseteq Y$ and $x_0 \in \operatorname{cl} A$, we call the tangent cone to A at x_0 the set

$$T(A, x_0) = \{v \in X : \exists \beta_n > 0, \exists x_n \in A,\ x_n \to x_0 \text{ with } v = \lim_n \beta_n (x_n - x_0)\}.$$

THEOREM 5.2. *If Y is any normed space and the point y_0 is locally superefficient, then it is efficient and $T(S, y_0) \cap -K = \{0\}$. If Y is finite dimensional, the opposite relation is true.*

Proof. It has been proved in [6] that if $y_0 \in SE(S)$, then $\operatorname{cl}\operatorname{cone}(S - y_0) \cap -K = \{0\}$ (the last relation is another definition of proper efficiency, due to Borwein [5]), and hence $y_0 \in SE(S \cap (y_0 + \eta B))$ implies $\operatorname{cl}\operatorname{cone}[(S - y_0) \cap \eta B] \cap -K = \{0\}$; the first inclusion follows from the equality $T(S, y_0) = \cap_{\eta > 0} \operatorname{cl}\operatorname{cone}[(S - y_0) \cap \eta B]$. To prove the converse, we should show that there exist $M > 0$ and $\eta > 0$ such that

$$\operatorname{cone}[(S - y_0) \cap \eta B] \cap (B - K) \subseteq MB.$$

Suppose by contradiction that there exist sequences $\alpha_n > 0$ and $s_n \in S$ with $\|s_n - y_0\| \to 0$, $\|\alpha_n(s_n - y_0)\| \to +\infty$, and $d_{-K}(\alpha_n(s_n - y_0)) \leq 1$. The latter implies that there exists a sequence $k_n \in K$ such that $\|\alpha_n(s_n - y_0) + k_n\| \leq 1 + 1/n$, which can be rewritten as $\alpha_n(s_n - y_0) + k_n = (1 + 1/n)b_n$, with $b_n \in B$. Moreover, it holds that $K = \operatorname{cone}\Theta$, where Θ is compact, and then $k_n = \lambda_n \theta_n$ with $\lambda_n > 0$ and $\theta_n \in \Theta$. It also holds that $\lambda_n \to +\infty$, because $\|k_n\| \to +\infty$ and Θ is bounded. This yields

$$\alpha_n(s_n - y_0) = (1 + 1/n)b_n - \lambda_n \theta_n$$

and

$$\frac{\alpha_n}{\lambda_n}(s_n - y_0) = \frac{n+1}{n\lambda_n} b_n - \theta_n. \tag{5.1}$$

The right-hand side of (5.1) converges (up to subsequences) to $-\theta \in -\Theta \subset -K$, and the left-hand side converges to an element of $T(S, y_0)$. Since $0 \notin \Theta$ we have a contradiction. □

THEOREM 5.3. *Consider the following statements:*

(a) *the point y_0 is tightly properly efficient in S;*
(b) *there exists an open convex set $C \subset Y$ such that $-K \backslash \{0\} \subseteq C$ and $(S - y_0) \cap C = \emptyset$;*
(c) $T(S + K, y_0) \cap -K = \{0\}$.

If Y is any normed space, then it holds that (a) ⇒ (b) ⇒ (c). *If Y is finite dimensional, we have also that* (c) ⇒ (a) *and all statements are equivalent.*

Proof. To prove that (a) implies (b) it is enough to show that, since there exists an open convex set C with $0 \in \partial C$, and there exists δ such that $C^c \cap (\delta B - K) \subseteq B$, then it holds that $-K \backslash \{0\} \subseteq C$. If indeed $k \in K \backslash \{0\}$ and $-k \in C^c$, then $-\lambda k \in C^c$ for all $\lambda \geq 1$ and a contradiction arises.

For the proof that (b) implies (c) one can refer to [27]. To prove the last relation we will see that (c) implies both strict efficiency and local superefficiency, considering Theorem 2.11. The latter is trivial since, by Theorem 5.2, the tangent cone is isotone with respect to the set inclusion, and hence it satisfies the inclusion $T(S, y_0) \subseteq T(S + K, y_0)$. To finish suppose that $y_0 \notin StE(S)$; then there exists a sequence $y_n \in S$ such that $y_n - y_0 \notin \varepsilon B$ for some $\varepsilon > 0$ and $d_{-K}(y_n - y_0) \to 0$, which means that there exists a sequence $k_n \in K$ with $y_n + k_n - y_0 \to 0$ and $k_n \notin (\varepsilon/2)B$. Since we can

always write $k_n = \lambda_n \theta_n$ with $\theta_n \in \Theta$, it follows that λ_n does not converge to zero, i.e., there exists a subsequence (we again call it λ_n) with $\lambda_n > \beta$ for some $\beta > 0$.

Now take $\alpha_n = \|y_n + k_n - y_0\|^{1/2}$ (since λ_n is bounded away from zero and α_n vanishes, it eventually holds that $\alpha_n < \lambda_n$) and set $k'_n = (\lambda_n - \alpha_n)\theta_n$ to obtain $y_n + k'_n - y_0 = y_n + k_n - y_0 - \alpha_n \theta_n \to 0$ and $\alpha_n^{-1}(y_n + k'_n - y_0) \to -\theta \neq 0$. □

We observe that statement (c) in Theorem 5.3 is the definition of proper efficiency given by Borwein in [4] and that statement (b) is another definition of proper efficiency attributed to Gerstewitz in [27]. The equivalence between (b) and (c) was already proved in finite dimensional spaces in [12]. The results proved in this section complement the ones given in [13], to which we refer for more details on proper efficiency.

Acknowledgments. I wish to thank E. Miglierina and E. Molho for helpful discussion during the preparation of the paper and an anonymous referee whose suggestions have improved the presentation of some results.

REFERENCES

[1] T. Amahroq and A. Taa, *On Lagrange-Kuhn-Tucker multipliers for multiobjective optimization problems*, Optimization, 41 (1997), pp. 159–172.
[2] E. Bednarczuk and W. Song, *PC points and their application to vector optimization*, Pliska Stud. Math. Bulgar., 12 (1998), pp. 21–30.
[3] E. Bednarczuk, *A note on lower semicontinuity of minimal points*, Nonlinear Anal., 50 (2002), pp. 285–297.
[4] J. Borwein, *Proper efficient points for maximization with respect to cones*, SIAM J. Control Optim., 15 (1977), pp. 57–63.
[5] J. M. Borwein, *The geometry of Pareto efficiency over cones*, Math. Operationsforsch. Statist. Ser. Optim., 11 (1980), pp. 235–248.
[6] J. M. Borwein and D. Zhuang, *Superefficiency in vector optimization*, Trans. Amer. Math. Soc., 338 (1993), pp. 105–122.
[7] J. V. Burke and M. C. Ferris, *Weak sharp minima in mathematical programming*, SIAM J. Control Optim., 31 (1993), pp. 1340–1359.
[8] M. Ciligot-Travain, *On Lagrange-Kuhn-Tucker multipliers for Pareto optimization problems*, Numer. Funct. Anal. Optim., 15 (1994), pp. 689–693.
[9] J. P. Dauer and R. J. Gallagher, *Positive proper efficient points and related cone results in vector optimization theory*, SIAM J. Control Optim., 28 (1990), pp. 158–172.
[10] J. P. Dauer and O. A. Saleh, *A characterization of properly minimal points as solution of sublinear optimization problems*, J. Math. Anal. Appl., 178 (1993), pp. 227–246.
[11] A. L. Dontchev and T. Zolezzi, *Well-Posed Optimization Problems*, Lecture Notes in Math. 1543, Springer-Verlag, Berlin, 1993.
[12] W. B. Gearhart, *Characterization of properly efficient solutions by generalized scalarization methods*, J. Optim. Theory Appl., 41 (1983), pp. 491–502.
[13] A. Guerraggio, E. Molho, and A. Zaffaroni, *On the notion of proper efficiency in vector optimization*, J. Optim. Theory Appl., 82 (1994), pp. 1–19.
[14] J.-B. Hiriart-Urruty, *New concepts in nondifferentiable programming*, Bull. Soc. Math. France Mém., 60 (1979), pp. 57–85.
[15] J.-B. Hiriart-Urruty, *Tangent cones, generalized gradients and mathematical programming in Banach spaces*, Math. Oper. Res., 4 (1979), pp. 79–97.
[16] G. Isac and V. Postolica, *The Best Approximation and Optimization in Locally Convex Spaces*, Peter Lang, Frankfurt am Main, Germany, 1993.
[17] J. Jahn, *Scalarization in vector optimization*, Math. Programming, 29 (1984), pp. 203–218.
[18] J. Jahn, *A characterization of properly minimal elements of a set*, SIAM J. Control Optim., 23 (1985), pp. 649–656.
[19] P. Q. Khanh, *Optimality conditions via norm scalarization in vector optimization*, SIAM J. Control Optim., 31 (1993), pp. 646–658.
[20] M. A. Krasnoselskii, *Positive Solutions of Operator Equations*, Noordhoof, Groningen, The Netherlands, 1964.

[21] H. W. Kuhn and A. W. Tucker, *Nonlinear Programming*, in Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, CA, 1951, pp. 481–492.
[22] E. Miglierina, *Characterizations of solutions of multiobjective optimization problems*, Rend. Circ. Mat. Palermo (2), 50 (2001), pp. 153–164.
[23] E. Miglierina and E. Molho, *Scalarization and its stability in vector optimization*, J. Optim. Theory Appl., 114 (2002), pp. 657–670.
[24] J.-P. Penot, *Conditioning convex and nonconvex problems*, Optim. Theory Appl., 90 (1996), pp. 535–554.
[25] B. T. Polyak, *Introduction to Optimization*, Optimization Software, New York, 1987.
[26] X. D. H. Truong, *Existence and density results for proper efficiency in cone compact sets*, J. Optim. Theory Appl., 111 (2001), pp. 173–194.
[27] C. Zalinescu, *On two definitions of proper efficiency*, in Optimization in Mathematical Physics, B. Brosowski and E. Martensen, eds., Peter Lang, Frankfurt am Main, Germany, 1987, pp. 77–86
[28] X. Y. Zheng, *Scalarization of Henig proper efficient points in a normed space*, J. Optim. Theory Appl., 105 (2000), pp. 233–247.
[29] T. Zolezzi, *On equiwellset minimum problems*, Appl. Math. Optim., 4 (1978), pp. 209–223.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 1087–1117

OPTIMAL SHAPE DESIGN OF INDUCTOR COILS FOR SURFACE HARDENING*

DIETMAR HÖMBERG† AND JAN SOKOŁOWSKI‡

Abstract. We study a mathematical model for induction hardening of steel. It consists of a vector potential formulation of Maxwell's equations coupled with a heat equation and an evolution equation for the volume fraction of the high temperature phase in steel called *austenite*. An important task for practical applications of induction hardening is to find the optimal coupling distance between inductor and workpiece. To this end we control the volume fraction of austenite with respect to perturbations of the coupling distance. The coil is modeled as a tube and is defined by a regular curve. We formulate the shape optimization problem over the set of admissible curves and prove the existence of an optimal curve. We apply the material derivative method for the shape sensitivity analysis of the state system. Finally, the shape gradient is specified for an optimal curve and the first order necessary optimality conditions are established.

Key words. shape optimization, surface hardening, heat equation, Maxwell's equations, optimality conditions

AMS subject classifications. Primary 49Q10, 49Q12; Secondary 35J05, 35J50, 35B37

DOI. 10.1137/S0363012900375822

1. Introduction.

Electromagnetic induction provides a method of heating electrically conducting materials. The basic components of an induction heating system are depicted in Figure 1.1. An alternating current flows through the induction coil (in what follows called the inductor). It generates an alternating magnetic field which in turn induces eddy currents in the workpiece. These dissipate energy, bring about heating, and lead to the growth of the high temperature phase austenite in the workpiece made of steel.

Since the magnitude of the eddy currents decreases with growing distance from the workpiece surface because of the frequency dependent skin-effect, induction heating is a suitable heat source for surface hardening if the current frequency has been chosen big enough. After heating, the workpiece is quenched by spray-water cooling and another phase transition leads to the desired hardening effect in the boundary layers of the workpiece.

An important task during the planning of an induction heat treatment is to find the optimal coupling distance between inductor and workpiece in order to obtain a desired heating pattern. This is illustrated in Figure 1.2. In all the examples shown, the goal is to produce a uniform hardening depth. In (a) a conical workpiece shall be heated by an inductor of the shape of a cylindrical spiral. To compensate for the bigger distance between inductor and workpiece in the upper part, the turn spacing there is narrower compared to the lower part.

In (b), because of the workpiece's geometry, heat will concentrate in the lower corners of the workpiece cross-section if the coupling distance is everywhere the same.

*Received by the editors July 24, 2000; accepted for publication (in revised form) January 30, 2003; published electronically July 8, 2003. The authors gratefully acknowledge the financial support of DAAD (Germany) and the Ministry of Foreign Affairs (France) within the scope of the French-German project PROCOPE 98158.

http://www.siam.org/journals/sicon/42-3/37582.html

†Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstraße 39, D-10117 Berlin, Germany (hoemberg@wias-berlin.de).

‡Institut Elie Cartan, BP 239, 54506 Vandoeuvre lès Nancy, France (sokolows@iecn.u-nancy.fr).

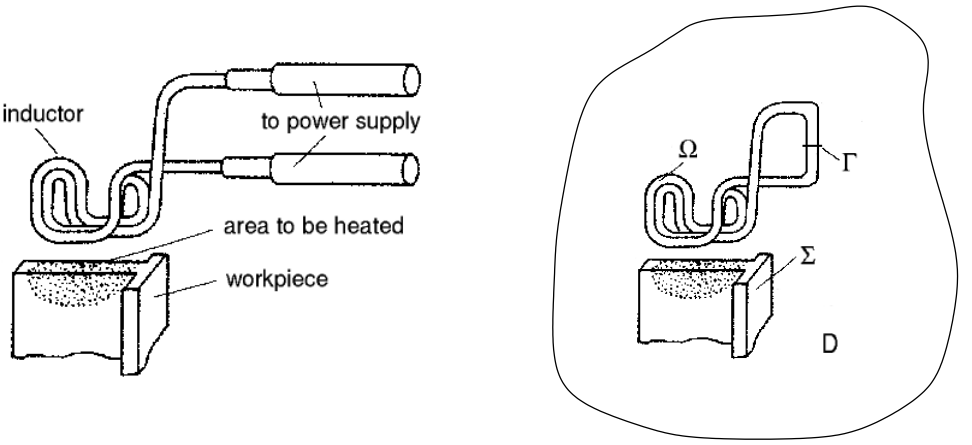


FIG. 1.1. *Induction heating: real process (left) and notation of domains in idealized setting.*

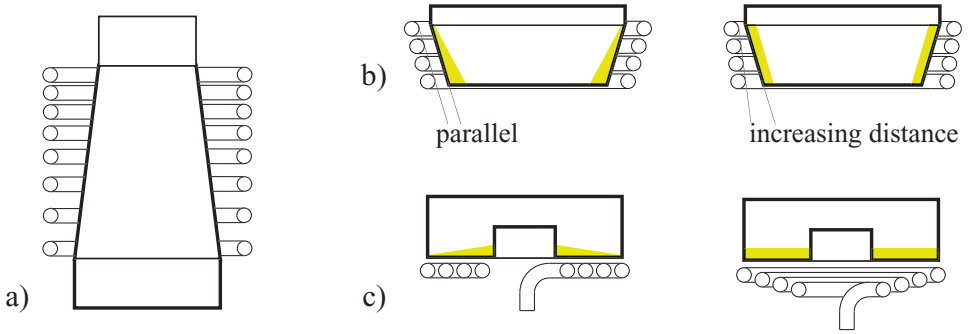


FIG. 1.2. *Adjustment of induction heating patterns by varying the turn spacing* (a) *or the coupling distance* (b), (c).

The remedy is to increase the coupling distance in the lower part of the workpiece leading to a uniform penetration depth.

Example (c) depicts the typical situation of a hole in an otherwise plane workpiece surface. The inductor on the left-hand side with a uniform coupling distance leads to an uneven hardening pattern and possibly even to a melting of hole edges. A better result can be achieved when the coupling distance between inductor and workpiece is increased locally around the hole.

There are already a lot of papers on modeling, analysis, and simulation of induction heating, e.g., [3], [4], [7], [10], [12], [13], [16], [17], [18], [19], [24], [25], [26]. In [2], an optimal control problem for a two-dimensional (2D) induction heating setting has been considered. Mathematical models for phase transitions in steel have been considered in, e.g., [15], [16], [17], [18], and [23].

In this paper, for the first time a design problem for the three-dimensional (3D) induction heating process including phase transitions is investigated. The main novelty of our approach is that we describe the inductor coil as a tube generated from a closed regular curve. To obtain information about the shape gradient, we first use the speed method. Then we show that we can relate the speed vector-field to

perturbations of the curve, which defines the coil. Probably this new procedure of characterizing the optimal configuration of tubes will admit further applications, for instance, in optimal design problems related to the flow of liquids through pipelines.

In the next section we derive a mathematical model for induction heating and prove its well-posedness. Moreover, we provide some numerical results showing the effect of varying coupling distance between workpiece and inductor. In section 3 we formulate the shape design problem for the inductor coil modeled as a tube with circular cross-section. Here, the main issue is to define a reasonable set of admissible domains excluding intersections of parts of the inductor. The existence of an optimal design is proved in section 4 using regularity results of sets with positive *reach* as introduced by Federer [11]. In section 5 we derive a necessary optimality condition and show how one can relate the shape gradient to perturbations of the generating curve. The necessary estimates for the sensitivity analysis are provided in section 6.

2. The mathematical model.

2.1. The state equations. Since we cannot model the hardening machine itself, we restrict ourselves to the following idealized geometric setting (cf. Figure 1.1 (right)). Let $D \subset \mathbb{R}^3$ be the hold-all domain, which contains the inductor Ω and the workpiece Σ. We call $G = \Omega \cup \Sigma$ the set of conductors and define the space-time cylinder $Q = \Sigma \times (0, T)$.

Since we do not consider the hardening machine in our model, we assume that the inductor Ω is a closed tube. Inside we fix a section Γ and model the current density which is generated by the hardening machine by an interface condition on Γ.

The following mathematical model is a simplified version of the electro-thermo-mechanical model of induction hardening derived in [18]. For further details about modeling, we refer the reader to this paper.

In eddy current problems we can neglect displacement currents, and hence we consider the following set of Maxwell equations:

$$\text{curl } H = J, \tag{2.1a}$$

$$\text{curl } E = -B_t, \tag{2.1b}$$

$$\text{div } B = 0. \tag{2.1c}$$

Here, E is the electric field, B the magnetic induction, H the magnetic field, and J the current density. In addition, we use the linear constitutive relations

$$J = \sigma E \quad \text{in } D, \tag{2.2a}$$

$$B = \mu H \quad \text{in } D, \tag{2.2b}$$

with the magnetic permeability μ and the electric conductivity σ. We assume zero current density outside conductors, i.e.,

$$\sigma(x) = \begin{cases} \sigma_0 > 0 & \text{in } \bar{G}, \\ 0 & \text{in } D \setminus \bar{G}. \end{cases}$$

The magnetic permeability takes different values in the workpiece made of steel and in the surrounding air. The coil is usually made of copper, which has approximately the same permeability as air. Hence we assume

$$\mu(x) = \begin{cases} \mu_1 & \text{in } D \setminus \bar{\Sigma}, \\ \mu_2 & \text{in } \bar{\Sigma}. \end{cases} \tag{2.3}$$

Using (2.1a)–(2.1c) one now introduces the magnetic vector potential A and the scalar potential ϕ such that

$$B = \text{ curl } A, \tag{2.4}$$

$$E = -A_t - \text{ grad } \phi. \tag{2.5}$$

Then, Maxwell's equations (2.1a)–(2.1c) can be rewritten in the following way:

$$\sigma A_t + \text{ curl } \Big(\frac{1}{\mu} \text{ curl } A\Big) + \sigma \text{ grad } \phi = 0 \quad \text{ in } D. \tag{2.6}$$

In view of Ohm's law (2.2a) and of (2.5) the current density can be viewed as the sum of an induced part $(-\sigma A_t)$ and an impressed part, stemming from the scalar potential $(-\sigma \text{ grad } \phi)$. Assuming the continuity equation to hold for the impressed part, the scalar potential ϕ is determined by

$$\text{div } (\sigma \text{ grad } \phi) = 0 \quad \text{ in } G. \tag{2.7a}$$

On the boundaries of the set of conductors, we assume a homogenous Neumann condition, i.e.,

$$\sigma_0 \frac{\partial \phi}{\partial \nu} = 0 \text{ in } \partial\Sigma \cup \partial\Omega. \tag{2.7b}$$

In the section Γ we supply current via an interface condition, i.e.,

$$\Big[\sigma_0 \frac{\partial \phi}{\partial \tilde{\nu}}\Big] = j_s \quad \text{ on } \Gamma. \tag{2.7c}$$

Here, j_s is the external source current density, $[f(x)]$ denotes the jump of a function $f(x)$ across the interface Γ, and $\tilde{\nu}$ is a unit normal vector on Γ.

The system (2.7a)–(2.7c) can be solved separately in each conductor. In the workpiece Σ, we obtain a homogenous, linear Neumann problem. Hence its solution is constant in Σ. Since only the gradient enters in (2.6), we can restrict the domain of ϕ to the coil Ω.

To solve the interface problem (2.7a)–(2.7c) in the coil Ω, we introduce the quotient space $H^1(\Omega)/\mathbb{R}$ with norm

$$\|\bar{u}\|_{H^1(\Omega)/\mathbb{R}} = \inf_{u \in \bar{u}} \|u\|_{H^1(\Sigma)}.$$

According to [14, Theorem 1.9], the functional

$$\bar{u} \mapsto \Big(\int_\Sigma |\nabla u|^2 \, dx \Big)^{1/2} \tag{2.8}$$

is an equivalent norm on $H^1(\Omega)/\mathbb{R}$. In what follows we will not distinguish between $\bar{u}$ and u.

Now we turn to the Maxwell equation (2.6). Assuming that the tangential component of A vanishes on ∂D, i.e.,

$$n \times A = 0, \tag{2.9}$$

we introduce the Hilbert space

$$\mathbf{X} = \{v \in \mathbf{L}^2(D)\,;\ \text{curl } v \in \mathbf{L}^2(D),\ \text{div } v \in \mathrm{L}^2(D) \text{ and } n \times v|_{\partial D} = 0\}.$$

Here and in what follows, $\mathbf{L}$ denotes the vector-valued counterpart $\mathbf{L} = [L]^3$ for any real-valued Sobolev space L. For ∂D of class $C^{1,1}$, $\mathbf{X}$ is a closed subspace of $\mathbf{H}^1(D)$, equipped with the norm

$$\|v\|^2_{\mathbf{X}} = \|\text{curl } v\|^2_{\mathbf{L}^2(D)} + \|\text{div } v\|^2_{L^2(D)}.$$

A good measure for the hardness penetration depth in the workpiece is the formation of austenite during heating, which can be described by the following initial value problem derived by Leblond and Devaux [23] (for details, we refer to [12]):

$$z(0) = 0, \tag{2.10a}$$

$$z_t(t) = \frac{1}{\mathcal{T}(\theta)}\Big[z_{eq}(\theta) - z\Big]^+. \tag{2.10b}$$

Here, z is the volume fraction of austenite and θ the temperature, $z_{eq}(\theta) \in [0,1]$ is an equilibrium fraction of austenite, and $\mathcal{T}(\theta)$ a time constant. The model is completed by a semilinear energy balance equation. For convenience, we recall the complete model of induction hardening of steel:

(P) Find $(A, \phi, \theta, z) \in L^\infty(0,T;\mathbf{X}) \times H^1(0,T;H^1(\Omega)/\mathbb{R}) \times W^{2,1}_3(Q)$
$\times W^{1,\infty}(0,T;L^\infty(\Sigma))$ such that

$$\sigma_0 \int_\Omega \nabla\phi \cdot \nabla u \; dx + \int_\Gamma j_s \varphi \, dx = 0 \text{ for all } \varphi \in H^1(\Omega)/\mathbb{R}, \tag{2.11a}$$

$$A(0) = A_0 \ \text{ in } D, \tag{2.11b}$$

$$\sigma_0 \int_G A_t \cdot v \; dx + \int_D \frac{1}{\mu} \text{ curl } A \cdot \text{ curl } v \; dx + \int_D \frac{1}{\mu} \text{ div } A \text{ div } v \, dx$$
$$+ \sigma_0 \int_\Omega \nabla\phi \cdot v \; dx = 0 \ \text{ for all } v \in \mathbf{X}, \text{ a.e. in } (0,T), \tag{2.11c}$$

$$\theta(0) = \theta_0 \quad \text{in } \Sigma, \tag{2.11d}$$

$$\frac{\partial\theta}{\partial\nu} = 0 \quad \text{in } \partial\Sigma \times (0,T), \tag{2.11e}$$

$$\rho c_p \theta_t - k\Delta\theta = -\rho L z_t + \sigma_0 |A_t|^2 \quad \text{in } Q, \tag{2.11f}$$

$$z(0) = 0 \quad \text{in } \Sigma, \tag{2.11g}$$

$$z_t = \frac{1}{\mathcal{T}(\theta)}[z_{eq}(\theta) - z]^+ \quad \text{in } Q, \tag{2.11h}$$

with $W^{2,1}_p(Q) = W^{1,p}(0,T;L^p(\Sigma)) \cap L^p(0,T;W^{2,p}(\Sigma))$.

The first term on the right-hand side of (2.11f) measures the latent heat inside the workpiece Σ, which is consumed during the formation of austenite. The second one describes the Joule heating $\sigma_0|E|^2$; cf. (2.5). Note that $\nabla\phi \equiv 0$ in Σ. ρ, c_p, k, L are density, specific heat at constant pressure, heat conductivity, and latent heat.

Remark 2.1. Usually, (2.6) is complemented by the Coulomb gauge $\operatorname{div} A = 0$ to ensure uniqueness (cf., e.g., [18]). To simplify the model, we have chosen here to include a divergence part in the bilinear form in (2.11c) as a penalty term, as it is often done in electrical engineering (cf., e.g., [21]). Otherwise, the shape sensitivity analysis requires the appropriate transformation of the vector potential A in the fixed domain setting and becomes technically involved.

We make the following assumptions:

(H1) $\bar\Omega \subset D$, $\bar\Sigma \subset D, \bar\Omega \cap \bar\Sigma = \emptyset$, and $\partial\Omega$, $\partial\Sigma$, ∂D are of class $C^{1,1}$.

(H2) σ_0, ρ, c_p, k, and L are positive constants.

(H3) $A_0 \in \mathbf{X} \cap \mathbf{H}^2(D)$, $\theta_0 \in W^{2,3}(\Sigma)$.

(H4) $\mu(x) = \mu_2\chi_\Sigma + \mu_1(1-\chi_\Sigma)$, with constants $0 < \mu_1 < \mu_2$.

(H5) $j_s \in H^1(0,T;H^{-1/2}(\Gamma))$, such that $\int_\Gamma j_s\,dx = 0$ and $\int_\Gamma j_{s,t}\,dx = 0$.

(H6) There exists $y_0 \in \mathbf{X}$, such that

$$\sigma_0 \int_G y_0 \cdot\ v\ dx + \int_D \frac{1}{\mu}\ \operatorname{curl} A_0 \cdot\ \operatorname{curl} v\,dx + \int_D \frac{1}{\mu}\ \operatorname{div} A_0\ \operatorname{div} v\,dx + \sigma_0 \int_\Omega \nabla\phi(0)\cdot v\ dx = 0$$

for all $v \in \mathbf{X}$.

(H7) $0 < \mathcal{T}_* \le \mathcal{T}(x) \le \mathcal{T}^* < \infty$ for all $x \in \mathbb{R}$, $\|\mathcal{T}\|_{C^2(\mathbb{R})} \le M$.

(H8) $0 \le z_{eq}(x) \le 1$ for all $x \in \mathbb{R}$; $\|z_{eq}\|_{C^2(\mathbb{R})} \le M$.

(H9) $[x]^+ = \frac{1}{2}(x + |x|)\mathcal{H}$ with $\mathcal{H} \in C^{2,1}(\mathbb{R})$, a monotone approximation of the Heaviside function.

Remark 2.2. The first part of assumption (H5) is the usual compatibility condition to ensure solvability of an elliptic Neumann problem. The second part of (H5) and (H6) are due to the fact that we will differentiate the equations for scalar and vector potential with respect to time to obtain higher regularity.

2.2. A weak solution to the state equations. Problem (P) is only sequentially coupled and can be solved by solving consecutively the subproblems (2.11a), (2.11b), (2.11c), and (2.11d)–(2.11h).

For the first one, we have the following lemma.

LEMMA 2.3. *Assume* (H1), (H2), *and* (H5); *then* (2.11a) *has a unique solution* $\phi \in H^1(0,T;H^1(\Omega)/\mathbb{R})$ *such that*

$$\|\nabla\phi\|_{H^1(0,T;\mathbf{L}^2(\Omega))} \le C\|j_s\|_{H^1(0,T;H^{-1/2}(\Gamma))}, \tag{2.12}$$

with a constant $C > 0$.

Proof. The proof follows from the Lax–Milgram lemma and the fact that we may differentiate (2.11a) with respect to time because of (H5). ∎

For the vector potential equation (2.11b), (2.11c) we have the following lemma.

LEMMA 2.4. *Assume* (H1)–(H6); *then* (2.11b), (2.11c) *has a unique solution* $A \in L^\infty(0,T;\mathbf{X})$, *satisfying the estimate*

$$\|A\|_{L^\infty(0,T;\mathbf{X})} + \|A_t\|_{L^\infty(0,T;\mathbf{L}^6(G))} \le C_1 + C_2\|j_s\|_{H^1(0,T;H^{-1/2}(\Gamma))}, \tag{2.13}$$

where the constant C_1 *depends on* A_0 *and* y_0 *(cf.* (H6)*).*

Proof. To prove the existence of a unique weak solution one can use, e.g., Rothe's method of implicit time discretization as described in the monograph [20]. The first part of the a priori estimate follows from inserting $v = A_t$ into (2.11c) and integrating

in time. To obtain the second part one can formally differentiate (2.11c) with respect to t. Then, we substitute $y = A_t$ and solve the system

$$y(0) = y_0 \quad \text{in } D,$$
$$\sigma_0 \int_G y_t \cdot v \, dx + \int_D \frac{1}{\mu} \operatorname{curl} y \cdot \operatorname{curl} v \, dx + \int_D \frac{1}{\mu} \operatorname{div} y \operatorname{div} v \, dx + \sigma_0 \int_\Omega \nabla \phi_t \cdot v \, dx = 0 \text{ for all } v \in \mathbf{X}, \text{ a.e. in } (0,T).$$

Testing with $v = y_t$ and integrating in time we obtain an estimate for y in $L^\infty(0,T;\mathbf{X})$. Owing to the compatibility condition (H6) we can recover that $y = A_t$ a.e. in G. Hence we can use the embedding $\mathbf{H}^1(G) \subset \mathbf{L}^6(G)$ and obtain the second part of (2.13). ∎

LEMMA 2.5. *Assume* (H7)–(H9)*; then the following are true:*

(1) *Let* $\theta \in L^1(Q)$*; then* (2.11g), (2.11h) *has a unique solution satisfying*

$$0 \le z(x,t) < 1 \quad \text{a.e. in } Q, \tag{2.14}$$

and

$$\|z\|_{W^{1,\infty}(0,T;L^\infty(\Sigma))} \le C, \tag{2.15}$$

with a constant $C > 0$ *independent of* θ.

(2) *Let* $\theta_k \to \theta$ *strongly in* $L^1(\Sigma)$. *Then*

$$z_k \longrightarrow z \quad \text{strongly in } W^{1,p}(0,T;L^p(\Sigma)) \text{ for } p \in [1,\infty),$$

where z_k *and* z *are the solution to* (2.11g), (2.11h) *corresponding to* θ_k *and* θ*, respectively.*

(3) *Let* $\theta_1, \theta_2 \in L^p(Q)$, $p \in [1,\infty)$, *and* z_1, z_2 *be the corresponding solutions to* (2.11g), (2.11h)*; then there exists a constant* $C > 0$ *such that*

$$\|z_1 - z_2\|_{W^{1,p}(0,T;L^p(\Sigma))} \le C \|\theta_1 - \theta_2\|_{W^{1,p}(0,T;L^p(\Sigma))}.$$

Proof. The existence of a unique local solution to (2.11g), (2.11h) is a direct consequence of the theorem of Carathéodory; see, e.g., [31, p. 1044]. Using (H7)–(H9) and the theory of differential inequalities (cf. [15, Lemma 2.1], we obtain (2.14), whereas (2.15) is a direct consequence of (H7)–(H9).

Assertion (2) follows from Lebesgue's lemma.

To prove (3), let $\theta^i \in L^p(Q)$, $i = 1, 2$, and define $\bar\theta = \theta_1 - \theta_2$; then $\bar z = z^1 - z^2$ solves

$$\bar z_t = f(\theta^1, z^1) - f(\theta^2, z^2), \tag{2.16}$$

where $f(\theta, z)$ denotes the right-hand side of (2.11h). In view of (H7)–(H9), f is Lipschitz continuous. Hence we can test (2.16) with $\bar z^{p-1}$ and apply Young's inequality to obtain

$$\begin{aligned} \frac{1}{p} \int_\Sigma \|\bar z(t)\|^p \, dx &\le c_1 \int_0^t \int_\Sigma |\bar z|^p \, dx \, ds + c_2 \int_0^t \int_\Sigma |\bar\theta| |\bar z|^{p-1} \, dx \, ds \\ &\le \Big(c_1 + c_2 \frac{p-1}{p} \Big) \int_0^t \int_\Sigma |\bar z|^p \, dx \, ds + \frac{c_2}{p} \int_0^t \int_\Sigma |\bar\theta|^p \, dx \, ds. \end{aligned}$$

Now we can apply Gronwall's lemma and use (2.16) once again to conclude the proof. ∎

Before considering the heat equation (2.11f), we recall the following results from the linear theory of parabolic equations.

LEMMA 2.6 (see [22, Theorem 9.1]). *Let $g \in L^p(Q)$ and $u_0 \in W^{1,p}(\Sigma)$ for some $p \in (1,\infty)$. Then there exists a constant $C > 0$ such that the unique solution to*

$$\begin{aligned} u_t - \Delta u &= g && \text{in } Q,\\ \frac{\partial u}{\partial \nu} &= 0 && \text{in } \partial\Sigma \times (0,T),\\ u(0) &= u_0 && \text{in } \Sigma \end{aligned}$$

satisfies the estimate

$$\|u\|_{W_p^{2,1}(Q)} \le C\Big(\|u_0\|_{W^{1,p}(\Sigma)} + \|g\|_{L^p(Q)}\Big).$$

For later use we also note the following embedding theorem [26, eq. (3.9)], written down for dim $\Sigma = 3$.

LEMMA 2.7. *Let $k = 0, 1$, $p \ge q$, $2 - k - 5(\frac{1}{q} - \frac{1}{p}) \ge 0$; then the embedding*

$$W_q^{2,1}(Q) \subset W_p^{k,0}(Q)$$

is continuous. The inclusion is compact if the last inequality is strict.

LEMMA 2.8. *Assume* (H1)–(H9)*; then* (2.11d)–(2.11h) *has a unique solution (θ, z), such that*

$$\|(\theta, z)\|_{W_3^{2,1}(Q)\times W^{1,\infty}(0,T;L^\infty(\Sigma))} \le C.$$

The constant C depends on A_t and θ_0.

Proof. The existence can be proved, e.g., using the Schauder fixed point theorem. The a priori estimate is a direct consequence of Lemma 2.5 and Lemma 2.6.

To prove uniqueness, we take the difference of two solutions $\bar\theta = \theta^1 - \theta^2$ which satisfies

$$\rho c_p \bar\theta_t - k\Delta\bar\theta = -\rho L(z_t^1 - z_t^2) \qquad \text{in } Q,$$

$$\frac{\partial\bar\theta}{\partial\nu} = 0 \text{ in } \partial\Sigma \times (0,T), \qquad \bar\theta(0) = 0 \text{ in } \Sigma.$$

Using Lemma 2.5(3), Lemma 2.6, and Hölder's inequality, we can infer

$$\|\bar\theta\|^3_{W_3^{2,1}(Q_t)} \le c_1 \int_0^t \int_\Sigma \left| \int_0^s \bar\theta_\xi \, d\xi \right|^3 dx\, ds \le T^2 c_1 \int_0^t \int_0^s \int_\Sigma |\bar\theta_\xi|^3 \, dx d\xi \, ds \le c_2 \int_0^t \|\bar\theta\|^3_{W_3^{2,1}(Q_s)} ds,$$

where $Q_t = \Sigma \times (0,t)$. Note that $\bar\theta_\xi$ is short for $\frac{\partial\bar\theta(x,\xi)}{\partial\xi}$. Now the assertion follows from Gronwall's lemma. ∎

Summarizing the results of Lemmas 2.3–2.5 and Lemma 2.8 we obtain the following theorem.

THEOREM 2.9. *Assume* (H1)–(H9)*; then problem* (P) *has a unique solution.*

FIG. 2.1. *Cylinder-symmetric workpiece with uniform inductor distance and the corresponding distribution of austenite.*

FIG. 2.2. *Cylinder-symmetric workpiece with varying inductor distance and the corresponding distribution of austenite.*

2.3. Numerical simulations for the state equations. The goal of this section is to demonstrate by numerical simulations that our model of induction hardening shows the correct behavior with respect to variations of the coupling distance between inductor and workpiece (cf. Figure 1.2), thereby justifying again the formulation of our shape optimization problem for induction hardening in the following section.

We use the commercial software package ANSYS to compute 3D simulations of the heating part of an induction hardening process, i.e., of equations (2.11a)–(2.11h). However, in order to avoid the problem of different time scales for the rapidly oscillating vector and scalar potentials on the one hand and for the slower heat diffusion on the other hand, we use harmonic approximation of (2.11c) for the numerical solution. This means we replace $A(x,t)$ with $A(x)e^{i\omega t}$ and make the same ansatz for ϕ. Moreover, we replace the Joule heat term $\sigma_0|A_t|^2$ in the energy balance (2.11f) by its average value for one period, i.e., by $\frac{\sigma_0\omega^2}{2}|A|^2$. The phase transition model (2.11h) has been added to ANSYS with the help of FORTRAN user routines. The physical data have been chosen for the steel 42 CrMo 4.

Owing to limitations in the number of nodes for this 3D problem, the results do not look as smooth and symmetric as they should. However, the principal behavior, which has already been discussed in connection with Figure 1.2, is clearly visible. In Figure 2.1 we have a cylinder-symmetric workpiece, and the inductor lies at a uniform distance from the workpiece. This geometric configuration produces a strip of austenite along the lateral cylinder boundary. Then in Figure 2.2, we have nearly the same situation except that the distance between inductor and workpiece now varies. The result is that austenite is only produced in the part of the cylinder that is close to the inductor.

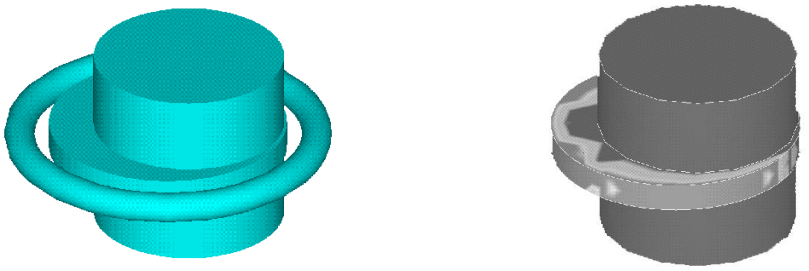


FIG. 2.3. *Nonsymmetric workpiece with uniform inductor distance and the corresponding distribution of austenite.*

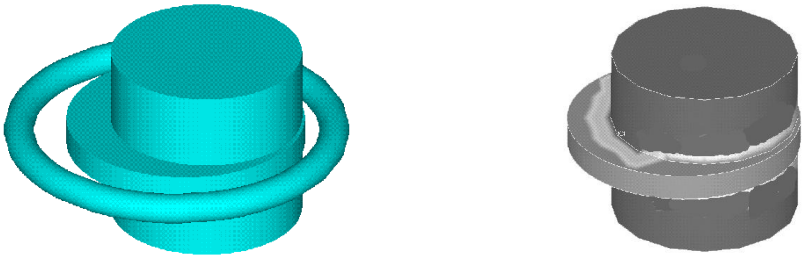


FIG. 2.4. *Nonsymmetric workpiece with varying inductor distance and the corresponding distribution of austenite.*

While Figures 2.1 and 2.2 are of rather academic nature, the next two figures depict more realistic situations. They show a nonsymmetric workpiece, where the center of mass is not in the middle, for instance, one might think of a camshaft. In Figure 2.3 the inductor is at a uniform distance from the workpiece. As a result, heat diffuses in the part with the bigger mass, and austenite is only produced in the parts with less mass.

Thus the strategy to achieve a uniform penetration of austenite should be to change the distance between inductor and workpiece, such that it is smaller on the side of the workpiece where the bigger part of the mass is concentrated. This has been done in Figure 2.4, and the result indeed is an approximately uniform distribution of austenite.

For the further implementation of a numerical procedure for the optimal design problem, the automatic geometry generation creates severe difficulties. For instance, it is easy to define a space curve as a B-spline using keypoints. Then one can create the tube by dragging a circle along the curve. These operations are standard in most CAD tools. However, to create disjoint domains necessary for grid generation, we have to apply Boolean operations. It turns out that these operations are not stable with respect to small perturbations of the curve defining the tube. This problem prevents the development of a solution strategy for the design problem using the same software as for the state equations. Hence an important step towards the solution of the design problem is the development of a stable CAD tool. However, this is beyond the scope of this paper and is the subject of current research.

3. The shape design problem. To decide whether the coupling distance between inductor and workpiece has been properly chosen, we measure the volume

fraction of austenite at the end-time T and compare it to a desired volume fraction $\bar{z}$; i.e., we consider the following cost functional of tracking type:

$$\mathcal{J}(\Omega) = \int_\Sigma \Big(z(x,T) - \bar{z}\Big)^2 \, dx. \tag{3.1}$$

Note that the cost functional depends on Ω only implicitly, through the solution to the Maxwell equation (2.11c).

Inductor coils are manufactured from copper tubes with approximately quadratic or circular cross-section. For convenience, we will only consider tubes with circular cross-section. These tubes can easily be generated from curves $\gamma : [0, 2\pi] \longrightarrow \mathbb{R}^3$, for which we assume

$$|\gamma'(s)| \neq 0 \text{ for all } s \in [0,2\pi],\ \gamma(0) = \gamma(2\pi), \text{ and } \frac{\gamma'(0)}{|\gamma'(0)|} = \frac{\gamma'(2\pi)}{|\gamma'(2\pi)|} = e_1. \tag{3.2}$$

Hence γ is a regular, closed curve, which starts and ends in the origin and has a tangent vector parallel to e_1 (the unit-vector in the x_1-direction) in the origin. In doing so we tacitly assume that the section Γ of the tube, where the source current is supplied, lies in a plane orthogonal to e_1 centered around the origin.

Then we define the corresponding tube of radius R by

$$\Omega(\gamma) = \{x \in \mathbb{R}^3 \,|\, d(\boldsymbol{\Gamma}_\gamma, x) \leq R\}, \tag{3.3}$$

whereas its lateral boundary is given by

$$\partial\Omega(\gamma) = \{x \in \mathbb{R}^3 \,|\, d(\boldsymbol{\Gamma}_\gamma, x) = R\}. \tag{3.4}$$

Here $d(\boldsymbol{\Gamma}_\gamma, x) = \inf\{|x - \gamma(s)| \,:\, s \in [0,2\pi]\}$ is the distance between x and the trace of γ, defined by

$$\boldsymbol{\Gamma}_\gamma = \{\gamma(s) \,|\, s \in [0,2\pi]\}.$$

From a physical point of view it is indispensable to avoid self-intersections of the tube. They can easily happen, for instance, when the curvature becomes too big (i.e., $> \frac{1}{R}$). But even if we exclude this by imposing an explicit bound on the curvature, it could happen that remote parts of the curve intersect with each other (cf. Figure 3.1).

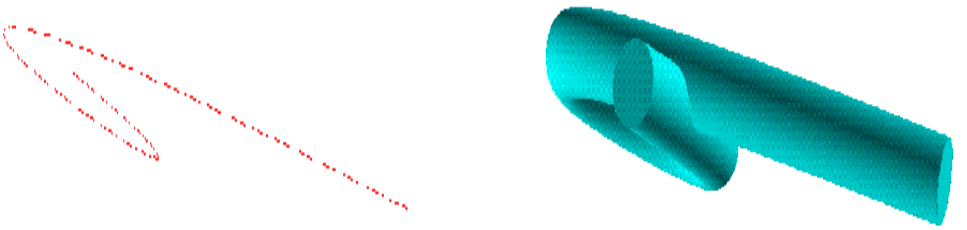


FIG. 3.1. *Curves without double point may lead to tubes with intersecting parts.*

Hence it is desirable to create a certain surrounding of the curve in which all points have a unique projection onto it. This can be achieved by adopting a concept from differential geometry introduced by Federer [11].

DEFINITION 3.1 (Federer [11]). *Let B be a closed subset of $\mathbb{R}^n$; then we call* Unp (B) *the set of all points $x \in \mathbb{R}^3$, for which there exists a unique projection onto B, and introduce the mapping ξ_B : Unp $(B) \longrightarrow B$, $x \mapsto b$, where b is uniquely given by $d(B,x) = |x-b|$.*

For $y \in B$ we define

$$\text{reach } (B,y) = \sup\Big\{ r \,\Big|\, \{x \in \mathbb{R}^3\,,\, |x-y| < r\} \subset \text{Unp } (B)\Big\}, \qquad \textit{and}$$
$$\text{reach } (B) = \inf\Big\{ \text{reach } (B,y) \,\Big|\, y \in B\Big\}.$$

In other words, *reach* of a subset $B \subset \mathbb{R}^n$ is the largest ε such that for all x in an ε-surrounding of B, there exists a unique projection onto B. If B is convex, then reach $(B) = \infty$. On the other hand, if B is concave with a re-entrant corner, e.g., an L-shaped domain, then reach $(B) = 0$.

For our purposes, we demand

$$\text{reach } (\boldsymbol{\Gamma}_\gamma) \geq R + \delta, \tag{3.5}$$

where $\delta > 0$ is a given positive parameter and R is the tube radius. Thus, we avoid situations as depicted in Figure 3.1 and also too narrow twists of the curve. Moreover, we gain a certain smoothness of $\partial\Omega(\gamma)$.

LEMMA 3.2. *Let $\boldsymbol{\Gamma}_\gamma$ be the trace of a curve γ satisfying* (3.2) *and* (3.5)*; then the boundary $\partial\Omega(\gamma)$ of the corresponding tube $\Omega(\gamma)$ is of class $C^{1,1}$.*

For the proof we need some differentiability properties of the distance function, which can be found in [11].

LEMMA 3.3. *For every closed and nonempty subset B of $\mathbb{R}^n$, there hold*

(1) $|d(B,x) - d(B,y)| \leq |x-y|$ *for all $x, y \in \mathbb{R}^n$;*

(2) $d(B,.)$ *is continuously differentiable on* Int(Unp $(B) \setminus B$) *and*

$$\text{grad } d(B,x) = \frac{1}{d(B,x)}(x - \xi_B(x));$$

(3) *if $0 < s < r <$ reach (B), then* grad $d(B,.)$ *is Lipschitzian on $\{x | s \leq d(B,x) \leq r\}$ and ξ_B is Lipschitzian on $\{x | d(B,x) \leq r\}$.*

Proof of Lemma 3.2. Let us define

$$F(x) = d(\boldsymbol{\Gamma}_\gamma, x) - R;$$

then $\partial\Omega(\gamma)$ is given by the zero level-set of F. In view of Lemma 3.3(3), F is $C^{1,1}$ in a surrounding of this level-set and grad $F(x) \neq 0$. Hence the assertion follows from a standard application of the implicit function theorem. ∎

The Hausdorff distance between two closed sets A, $B \subset \mathbb{R}^n$ can be defined as

$$d_H(A,B) = \sup_{x \in C} \Big| d(A,x) - d(B,x) \Big|$$

for any $C \subset \mathbb{R}^n$ such that $A \cup B \subset C$. The following lemma states that the family of bounded sets with reach bounded away from zero is closed with respect to the Hausdorff metric.

LEMMA 3.4. *Let $\varepsilon > 0$. If B_i, $i \in \mathbb{N}$, and B are closed subsets of $\mathbb{R}^n$ such that* reach $(B_i) \geq \varepsilon$ *for $i \in \mathbb{N}$ and $d(B_i,x) \longrightarrow d(B,x)$ uniformly for $x \in C$, whenever C is a compact subset of $\{x \in \mathbb{R}^n \,|\, d(B,x) < \varepsilon\}$, then*

$$\text{reach } (B) \geq \varepsilon$$

and

$$\xi_{B_i}(x) \xrightarrow[i\to\infty]{} \xi_B(x)$$

uniformly for $x \in C$.

Now we can introduce the set of admissible curves. Let δ be a positive constant; then we define

$$U_{ad} = \Big\{\gamma \in [W^{2,\infty}(0,2\pi)]^3 \,\Big|\, \gamma \text{ satisfies (3.2) and (3.5)}, \ \overline{\Omega}(\gamma) \subset (D \setminus \overline{B_\delta}(\Sigma)), \ \kappa_\gamma(s) \le \frac{1}{R} \text{ for a.e. } s \in [0,2\pi]\Big\}.$$

Note that for regular curves the curvature κ_γ is defined by

$$\kappa_\gamma(s) = \frac{|\gamma'(s) \times \gamma''(s)|}{|\gamma'|^3}.$$

Thus, admissible curves are regular and closed, with pointwise bounded curvature; the generated tubes do not touch the workpiece or ∂D and cannot have self-intersections because of the reach condition.

Remark 3.5. Note that there is no artificial boundedness assumption in the definition of U_{ad}. Choosing $\gamma \in [W^{2,\infty}(0,2\pi)]^3$ seems to be natural, because then the curve regularity corresponds to the $C^{1,1}$ regularity of the tube surface that could be obtained in Lemma 3.3 using the reach condition.

The curvature condition that we have imposed explicitly can also be derived from the reach condition. To this end one has to take a smooth approximation of the curve, parametrize the tube surface locally using the Frenet-frame, and compute a normal to the surface.

Now, we define the corresponding set of admissible domains by

$$\mathcal{U}_{ad} = \{\Omega(\gamma)\,,\ \gamma \in U_{ad}\}$$

and give a precise definition of our control problem, which reads as

(CP) minimize $J(\Omega)$, given by (3.1),
subject to
$\Omega \in \mathcal{U}_{ad}$
and the state equations (2.11a)–(2.11h).

Note that (CP) is a nonconvex optimization problem due to the nonconvexity of $\mathcal{U}_{ad}$ and of the cost functional.

4. The existence of an optimal domain Ω^*.

THEOREM 4.1. *Assume* (H1)–(H9)*; then* (CP) *admits a solution* $\Omega^* \in \mathcal{U}_{ad}$.

Proof. We take a minimizing sequence $\{\Omega_n\}$, such that $J(\Omega_n) \longrightarrow \inf J$ for $n \to \infty$. We have $\Omega_n = \Omega(\widetilde{\gamma}_n)$ and $\widetilde{\gamma}_n \in U_{ad}$. Owing to the embedding $W^{2,\infty}(0,2\pi) \subset C^{1,\alpha}[0,2\pi]$, which is compact for $\alpha < 1$, $\widetilde{\gamma}_n'$ is a continuous function and there exists a constant $\widetilde{c}_n > 0$ such that

$$|\widetilde{\gamma}_n'(s)| \ge \widetilde{c}_n \quad \text{for all } s \in [0,2\pi]. \tag{4.1}$$

Obviously, this estimate is not uniform in n, so we proceed as in [29] and introduce a reparametrization of $\widetilde{\gamma}_n$. Let $\varphi_n : [0,2\pi] \to [0,l_n]$ measure the length l_n of $\widetilde{\gamma}_n$, i.e.,

$$\varphi_n(s) = \int_0^s |\widetilde{\gamma}_n'| ds.$$

Owing to (4.1), φ_n is invertible and its inverse $\psi_n : [0, l_n] \to [0, 2\pi]$ satisfies

$$\psi_n'(\tau) = \frac{1}{|\widetilde{\gamma}_n'|},$$
$$\psi_n''(\tau) = -\frac{\widetilde{\gamma}_n' \cdot \widetilde{\gamma}_n''}{|\widetilde{\gamma}_n'|^4}$$

for $s = \psi_n(\tau)$ and a.e. $\tau \in [0, l_n]$.

Hence ψ_n is a C^1 diffeomorphism and the composition $\widetilde{\gamma}_n \circ \psi_n$ is in $W^{2,\infty}(0, l_n)$. Now we introduce the function

$$\Lambda_n : [0, 2\pi] \to [0, l_n], \quad \Lambda(s) = \frac{l_n}{2\pi} s$$

and define

$$\gamma_n = \widetilde{\gamma}_n \circ \psi_n \circ \Lambda_n. \tag{4.2}$$

It follows that $\gamma_n \in [W^{2,\infty}(0, 2\pi)]^3$,

$$|\gamma_n'(s)| = \frac{l_n}{2\pi} \text{ for all } s \in [0, 2\pi], \tag{4.3}$$

and $\gamma_n(0) = \gamma_n(l) = 0$, $\frac{\gamma_n'(0)}{|\gamma_n'(0)|} = \frac{\gamma_n'(2\pi)}{|\gamma_n'(2\pi)|} = e_1$. Moreover, since γ_n is just a reparametrization of $\widetilde{\gamma}_n$, both curves have the same trace. Hence $\Omega(\widetilde{\gamma}_n) = \Omega(\gamma_n)$, the reach condition is satisfied, and due to the parametric invariance of curvature we can conclude $\gamma_n \in U_{ad}$.

In order to obtain uniform estimates for γ_n we first remark that

$$2\pi(R + \delta) \le l_n \le \frac{|D|}{\pi R^2}, \tag{4.4}$$

where $|D|$ is the Lebesgue-measure of the domain D. The first part of the inequality follows directly from the reach condition. Suppose it would not hold. Since the curve γ_n is closed, we could take a ball with radius $\widetilde{R} < R + \delta$ and move it towards the curve until it has at least two different points of contact with the curve but no points where the curve intersects the ball. Hence the center of the ball has no unique projection on the curve; this implies reach $(\boldsymbol{\Gamma}_n) \le \widetilde{R}$, a contradiction to reach $(\boldsymbol{\Gamma}_n) \ge R + \delta$.

Due to the classical theorem of Pappus (see, e.g., [1, Corollary 6.9.10]) the tube volume is given by

$$|\Omega(\gamma_n)| = \pi R^2 l_n.$$

Since the tube cannot have self-intersections because of the reach condition, its volume is bounded by the volume of D. This is the second part of (4.4).

Now, we exploit the curvature functional. Utilizing (4.3), we see that

$$0 = \frac{\partial}{\partial s} |\gamma_n'(s)|^2 = 2\gamma_n' \cdot \gamma_n''.$$

Hence γ_n' is perpendicular to γ_n'', and we have for a.e. $s \in [0, 2\pi]$

$$\kappa_{\gamma_n}(s) = \frac{|\gamma_n'(s) \times \gamma_n''(s)|}{|\gamma_n'(s)|^3} = \frac{4\pi^2}{l_n^2} |\gamma_n''(s)|.$$

Taking into account also (4.3) and (4.4), there exists a constant c_1 independent of n such that

$$\|\gamma_n\|_{W^{2,\infty}(0,2\pi)} \le c_1,$$

and we can extract a subsequence such that

$$\gamma_{n'} \longrightarrow \gamma^* \quad \begin{array}{ll} \text{weakly* in} & W^{2,\infty}(0,2\pi), \\ \text{strongly in} & C^{1,\alpha}[0,2\pi] \text{ for } 0 \le \alpha < 1. \end{array} \tag{4.5}$$

Obviously, γ^* satisfies $\gamma^*(0) = \gamma^*(l) = 0$ and $\frac{\gamma^{*\prime}(0)}{|\gamma^{*\prime}(0)|} = \frac{\gamma^{*\prime}(2\pi)}{|\gamma^{*\prime}(2\pi)|} = e_1$. Invoking (4.3) and (4.4) we also find that γ^* is regular.

Next, we want to prove

$$\text{reach } (\boldsymbol{\Gamma}^*) \ge R + \delta, \tag{4.6}$$

where $\boldsymbol{\Gamma}^* = \boldsymbol{\Gamma}_{\gamma^*}$ is the trace of γ^* in $\mathbb{R}^n$. Utilizing Lemma 3.4, it suffices to show

$$d(\boldsymbol{\Gamma}_{n'}, x) \longrightarrow d(\boldsymbol{\Gamma}^*, x) \qquad \text{uniformly in } x \in C, \tag{4.7}$$

where

$$C \subset\subset \{x \in \mathbb{R}^3 \,:\, d(\boldsymbol{\Gamma}^*, x) < R + \delta\} \tag{4.8}$$

and $\boldsymbol{\Gamma}_n = \boldsymbol{\Gamma}_{\gamma_n}$. For $x \in C$, we define

$$f_{n'}(x) = d(\boldsymbol{\Gamma}_{n'}, x) \quad \text{ and } \quad f(x) = d(\boldsymbol{\Gamma}, x).$$

Since $C \subset \text{Unp } (\boldsymbol{\Gamma}^*)$, there exists exactly one $s \in [0, 2\pi]$ such that $d(\boldsymbol{\Gamma}^*, x) = |\gamma^*(s) - x|$. We obtain

$$\begin{aligned} |f_{n'}(x) - f(x)| &\le \big| |\gamma_{n'}(s) - x| - |\gamma^*(s) - x| \big| \\ &\le |\gamma_{n'}(s) - \gamma^*(s)|, \end{aligned}$$

and, invoking (4.5), we have

$$f_{n'}(x) \longrightarrow f(x) \quad \text{for } x \in C. \tag{4.9}$$

Moreover, since $\{f_{n'}\}$ is equicontinuous (cf. Lemma 3.3(1)), we can use the Arcela–Ascoli theorem to conclude that there exists a subsequence satisfying

$$f_{n''} \longrightarrow f \text{ uniformly in } C.$$

Since the limit does not depend on the subsequence, the whole sequence $\{f_{n'}\}$ converges to f.

Thus we have proved (4.7) and can apply Lemma 3.4 to get reach $(\boldsymbol{\Gamma}^*) \ge R + \delta$.

According to Lemma 3.2, all the domains $\Omega(\gamma)$ with reach $(\boldsymbol{\Gamma}) \ge R + \delta$ are of class $C^{1,1}$; in particular, they satisfy the uniform cone condition (cf., e.g., [30]). Hence we can conclude the following from Chenais [6]:

(1) All domains $\Omega \in \mathcal{U}_{ad}$ have the uniform extension property; i.e., for every $f \in H^m(\Omega)$ there exists an extension $\tilde{f} \in H^m(D)$ such that

$$\|\tilde{f}\|_{H^m(D)} \le K \|f\|_{H^m(\Omega)}, \tag{4.10}$$

where the constant K depends on M in our case.

(2) There exists a subsequence $\{\Omega_{n'}\}$ such that the corresponding characteristic functions satisfy

$$\chi_{\Omega_{n'}} \longrightarrow \chi_{\Omega^*} \quad \text{strongly in } L^p(D),\ p \in [1,\infty), \tag{4.11}$$

where χ_Ω is the characteristic function of Ω.

Note that (4.11) also implies $\overline{\Omega}(\gamma^*) \subset (D \setminus \overline{B_\delta}(\Sigma))$, which in turn implies $\gamma^* \in U_{ad}$ and $\Omega^* \in \mathcal{U}_{ad}$.

To finish the proof, we have to pass to the limit in the state equations. This will be done in the following lemmas using (4.10), (4.11). As a by-product we will obtain the strong convergence in $L^2(\Sigma)$ for $z_n(.,T)$; thus we can also pass to the limit in J, whence it follows that Ω^* is a solution to (CP). ∎

We begin with the equation for the scalar potential (2.11a). Denoting

$$\widetilde{\nabla\phi_{n'}}(x) = \begin{cases} \nabla\phi_{n'}(x), & x \in \Omega_n, \\ 0, & x \in D \setminus \Omega_n, \end{cases}$$

we have

$$\sigma_0 \int_D \chi_n \widetilde{\nabla\phi_{n'}} \cdot \nabla u \, dx + \int_\Gamma j_g u \, dx = 0 \qquad \text{for all } u \in H^1(D)/\mathbb{R} \tag{4.12}$$

and obtain the following lemma.

LEMMA 4.2. *There exists a subsequence satisfying*

$$\widetilde{\nabla\phi_{n'}} \longrightarrow \widetilde{\nabla\phi} \qquad \textit{strongly in } H^1(0,T;\mathbf{L}^2(D)). \tag{4.13}$$

Proof. Since $\{\widetilde{\nabla\phi_{n'}}\}$ is bounded in $L^2(0,T;\mathbf{L}^2(D))$, (4.13) holds weakly in $L^2(0,T;\mathbf{L}^2(D))$. Moreover, taking $\nabla u = \widetilde{\nabla\phi_{n'}}$ in (4.12) we have

$$\int_D |\widetilde{\nabla\phi_{n'}}|^2 \, dx = -\frac{1}{\sigma_0} \int_\Gamma j_g \phi_n \, dx \longrightarrow -\frac{1}{\sigma_0} \int_\Gamma j_g \phi \, dx = \int_D |\widetilde{\nabla\phi}|^2 \, dx$$

and, thus, strong convergence in $L^2(0,T;\mathbf{L}^2(D))$. Differentiating (4.12) formally with respect to t and reasoning as above completes the proof. ∎

Now, we consider the equation for the magnetic vector potential (2.11c). Note that the permeability μ is independent of n (cf. (H4)).

Defining $G_{n'} = \Omega_{n'} \cup \Sigma$, we rewrite (2.11c) as

$$\sigma_0 \int_{G_{n'}} A_{n',t} \cdot v \, dx + \int_D \frac{1}{\mu} \operatorname{curl} A_{n'} \cdot \operatorname{curl} v \, dx + \int_D \frac{1}{\mu} \operatorname{div} A_{n'} \operatorname{div} v \, dx + \sigma_0 \int_D \widetilde{\nabla\phi_{n'}} \cdot v \, dx = 0.$$

Proceeding as in the proof of Lemma 2.4, we can obtain a priori estimates similar to (2.13), where the bounds are independent of n'. In addition, a close inspection of the proof of Lemma 2.4 shows that we have the additional estimate

$$\|A_t^{n'}\|_{L^\infty(0,T;\mathbf{H}^1(\Sigma)) \cap H^1(0,T;L^2(\Sigma))} \le c_1$$

with a constant c_1 independent of n. Since the embedding $\mathbf{H}^1(\Sigma) \subset \mathbf{L}^p(\Sigma)$ is compact for $p < 6$, we can apply [27, Corollary 4] to conclude that the embedding $L^\infty(0,T;\mathbf{H}^1(\Sigma)) \cap H^1(0,T;L^2(\Sigma)) \subset C(0,T;\mathbf{L}^p(\Sigma))$ is compact for $p < 6$.

Thus, we obtain the following lemma.

LEMMA 4.3. *There exists a subsequence* $\{A^{n'}\}$ *satisfying*

$$\begin{aligned} A_{n'} &\longrightarrow A \ \textit{weakly-star in}\ L^\infty(0,T;\mathbf{X}),\\ \chi_{G_{n'}} A_{n',t} &\longrightarrow \chi_G A_t \ \ \textit{weakly in}\ L^2(0,T;L^2(D)),\\ A_{n',t_{|\Sigma}} &\longrightarrow A_{t_{|\Sigma}} \ \ \textit{strongly in}\ C(0,T;L^4(\Sigma)). \end{aligned}$$

The equations for temperature and phase transition (2.11d)–(2.11h) depend only implicitly on the shape of Ω_n, namely through A_n. From Lemma 2.8 we obtain

$$\|z_{n'}\|_{W^{1,\infty}(0,T;L^\infty(\Sigma))} + \|\theta_{n'}\|_{W_3^{2,1}(Q)} \le c_2,$$

with a constant c_2 independent of n'. In view of the compact embedding $W_3^{2,1}(Q) \subset L^p(0,T;W^{1,p}(\Sigma))$ for $p < 15/2$ (cf. Lemma 2.7), we obtain the following lemma.

LEMMA 4.4. *There exist subsequences* $\{\theta_{n'}\}$, $\{z_{n'}\}$ *satisfying*

$$\begin{aligned} \theta_{n'} &\longrightarrow \theta \qquad \textit{weakly in}\ W_3^{2,1}(Q),\\ &\qquad\quad \textit{strongly in}\ L^p(0,T;W^{1,p}(\Sigma))\ \textit{for}\ p < 15/2,\\ z_{n'} &\longrightarrow z \qquad \textit{weakly* in}\ W^{1,\infty}(0,T;L^\infty(\Sigma))\\ &\qquad\quad \textit{strongly in}\ W^{1,p}(0,T;L^p(\Sigma))\ \textit{for all}\ p \in [1,\infty). \end{aligned}$$

Utilizing Lebesgue's convergence theorem, we can pass to the limit in the state equations (2.11a)–(2.11h) and in the cost functional (3.1), which concludes the proof of Theorem 4.1.

5. Necessary optimality conditions.

5.1. Shape sensitivity analysis. To investigate the sensitivity of solutions to the state system (2.11a)–(2.11h) with respect to perturbations of the shape of the inductor Ω, we use the standard method (cf. [28, sect. 2.9]). To this end we forget for a moment that Ω has been generated from a curve γ.

We introduce a vector-field V satisfying

$$V \in C(-\tau_1,\tau_1;C_0^2(D,\mathbb{R}^3)), \qquad \operatorname{supp} V \subset (B_{\delta_1}(\Omega) \setminus B_{\delta_2}(\Gamma)), \tag{5.1}$$

with positive constants τ_1 and $\delta_{1,2}$.

Hence the velocity field is chosen in such a way that the inductor can be perturbed, except for a small region around the interface Γ, where current is supplied and in reality the inductor is fixed to the hardening machine. Moreover, we tacitly assume that δ_1 has been chosen small enough to assure $\bar{\Sigma} \cap \operatorname{supp} V = \emptyset$.

Now we construct a family of mappings

$$\mathcal{T}_\tau(V) : \mathbb{R}^3 \ni X \longrightarrow x_\tau \in \mathbb{R}^3,$$

where x_τ satisfies the initial value problem

$$\begin{aligned} \frac{dx_\tau}{d\tau} &= V(\tau, x_\tau),\\ x_0 &= X. \end{aligned}$$

Then we define a family of perturbations of a given initial configuration Ω by

$$\Omega_\tau = \mathcal{T}_\tau(V)(\Omega).$$

All equations defined in Ω_τ can be transported to the fixed domain Ω, using the transformation $T_\tau^{-1} : \Omega_\tau \to \Omega$. Note that, by construction, we have $\Omega_0 = \Omega$ and $\overline{\Omega_\tau} \cap \bar{\Sigma} = \emptyset$ for all $\tau \in (-\tau_1, \tau_1)$ if τ_1 has been chosen small enough. Moreover, the interface Γ, where the source current is supplied, remains invariant under the perturbations of Ω, and we have $\mathcal{T}_\tau(V)(D) = D$.

Remark 5.1. In what follows we indicate functions on Ω_τ with subscript τ and functions transported to the fixed domain Ω with superscript τ, i.e., $f^\tau = f_\tau \circ T_\tau$.

DEFINITION 5.2. *All the state variables depend on the shape of Ω_τ, either explicitly as A^τ and ϕ^τ or implicitly as θ^τ and z^τ. For all these quantities, we call*

$$\dot{f} = \lim_{\tau \to 0} \frac{f^\tau - f}{\tau}$$

the strong material derivative of f, whenever the limit exists in the strong sense.

In section 6 (Theorem 6.8) we will prove that the strong material derivatives exist. As a corollary to Theorem 6.8 we obtain the following estimate.

COROLLARY 5.3. *There exists a constant $C > 0$ such that*

$$\begin{aligned} &\|\nabla\dot{\phi}\|_{H^1(0,T;\mathbf{L}^2(\Omega))} + \|\dot{A}\|_{L^\infty(0,T;X)} + \|\dot{A}_t\|_{L^{10/3}(0,T;L^{10/3}(G))} \\ &\qquad + \|\dot{\theta}\|_{W^{2,1}_{5/3}(Q)} + \|\dot{z}\|_{W^{1,5}(0,T;L^5(\Sigma))} \le C\|V(0)\|_{C^1(D)}. \end{aligned} \tag{5.2}$$

Remark 5.4. We would like to emphasize that the same result on the shape differentiability can be derived using the perturbation of identity technique. In that case one has to consider

$$\Omega_\zeta = T_\zeta(\Omega), \quad \text{where } T_\zeta(x) = (I + \zeta\Theta)(x)$$

with the field Θ and the shape parameter ζ, which replaces the parameter τ.

5.2. The structure theorem. From Corollary 5.3 we can infer that the functional $J(\Omega)$ defined in (3.1) is differentiable, in the sense that there exists the limit

$$dJ(\Omega; V) = \lim_{\tau \to 0} \frac{1}{\tau}(J(\Omega_\tau) - J(\Omega)), \tag{5.3}$$

where $\Omega_\tau = T_\tau(V)(\Omega)$. Moreover, we can conclude the following corollary.

COROLLARY 5.5. *The mapping*

$$V \mapsto dJ(\Omega; V) : C_0^1(D; \mathbb{R}^3) \to \mathbb{R}$$

is linear and continuous.

Thus, we can apply the structure theorem (cf. [28]) and obtain the following lemma.

LEMMA 5.6. *Let $\partial\Omega$ be of class C^2; then there exists a distribution $g_{\partial\Omega}$ with support in $\partial\Omega$, the shape gradient, such that $g_{\partial\Omega} \in \mathcal{D}_1'(\partial\Omega)$, and for all $V \in C_0^1(D; \mathbb{R}^3)$ there holds*

$$dJ(\Omega; V) = \langle g_{\partial\Omega}, V \cdot \nu \rangle_{\mathcal{D}_1'(\partial\Omega) \times \mathcal{D}_1(\partial\Omega)},$$

where ν is the outer unit normal vector on the boundary of the tube Ω.

Now we relate the perturbations of $\partial\Omega$ by means of $T_\tau(V)$ with perturbations of the curve γ in the form $\gamma_\varepsilon = \gamma + \varepsilon\omega$, where $\omega : [0, 2\pi] \to \mathbb{R}^3$ is a given curve. To

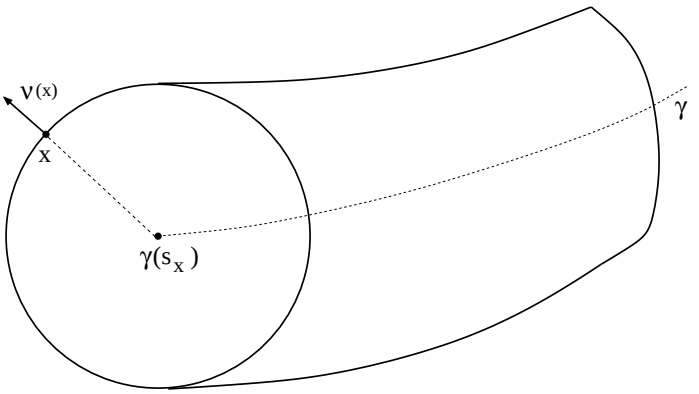


FIG. 5.1. *Computing a normal to the tube.*

ensure that the interface Γ, where the source current is supplied, remains unperturbed, we assume

(H10) $\omega(s) = \omega(2\pi - s) = 0 \quad$ for all $s \in [0, \delta]$ for some $\delta > 0$.

Note that although ω is not a regular curve, γ_ε is regular for ϵ small enough and for sufficiently smooth ω.

Now let $x(\varepsilon) \in \partial\Omega(\gamma_\varepsilon)$ be a point on the surface of the perturbed tube. We define a function $g(\varepsilon)$ by

$$\begin{aligned} g(\varepsilon) &= d\big(\boldsymbol{\Gamma}_{\gamma_\varepsilon}, x(\varepsilon)\big)^2 = \inf_{s\in[0,2\pi]} |x(\varepsilon) - \gamma_\varepsilon(s)|^2 \\ &= |x(\varepsilon) - \gamma(s_{x(\varepsilon)}) - \varepsilon\omega(s_{x(\varepsilon)})|^2. \end{aligned}$$

With the notation of Lemma 3.2, $\gamma(s_{x(\varepsilon)}) = \xi_{\Gamma_{\gamma_\varepsilon}}(x(\varepsilon))$ is the projection of $x(\varepsilon) \in \partial\Omega(\gamma_\varepsilon)$ on the curve γ_ε. Differentiating g at $\varepsilon = 0$ (cf., e.g., [9, Theorem 9.2.1]) and assuming tacitly that $x(\varepsilon)$ is differentiable at $\varepsilon = 0$, we obtain

$$\begin{aligned} 0 = g'(0) &= 2\Big(x(0) - \gamma(s_{x(0)})\Big) \cdot \Big(\dot{x}(0) - \omega(s_{x(0)})\Big) \\ &= 2R\nu(x) \cdot \Big(\dot{x} - \omega(s_x)\Big), \end{aligned}$$

where $\nu(x)$ is the outer unit normal vector to $\partial\Omega(\gamma)$ and $\gamma(s_x) = \xi_{\Gamma_\gamma}(x)$ is the projection of $x \in \partial\Omega(\gamma)$ on the curve γ (cf. Figure 5.1). Hence we obtain for the normal component of the speed vector-field

$$V_\nu(\omega)(x) = \frac{1}{R}\Big(x - \gamma(s_x)\Big) \cdot \omega(s_x). \tag{5.4}$$

Now, let $\eta \in C_0^\infty(B_\delta(\partial\Omega))$ such that $\eta(x) = 1$ if $x \in \partial\Omega$. Then we define the autonomous speed vector-field by

$$V(\omega)(x) = \frac{1}{R^2}\eta(x)\Big((x - \gamma(s_x)) \cdot \omega(s_x)\Big)\Big(x - \gamma(s_x)\Big). \tag{5.5}$$

Obviously, the normal component of the field V on $\partial\Omega$ coincides with $V_\nu(\omega)$ for $x \in \partial\Omega$. However, to obtain C^1 regularity for V, we have to increase the regularity of γ and ω. In addition to (3.2), we assume

(H11) $\gamma, \omega \in C^2[0,2\pi]$ and $\gamma''(0) = \gamma''(2\pi)$.

Since a curve has no interior points, the distance function with respect to a curve coincides with the signed distance function introduced by Delfour and Zolésio. According to [9, Theorem 5.4.3] and (H11), the distance function $d(\boldsymbol{\Gamma}_\gamma, .)$ is of class C^2 in an $(R+\delta)$-neighborhood of γ. Thus $\partial\Omega$ is of class C^2 and from Lemma 3.3(2) we can infer that the projection $\xi_{\boldsymbol{\Gamma}_\gamma} = \gamma(s_x)$, which shows up in (5.5), is a C^1 function.

Altogether we can conclude that the speed vector-field defined in (5.5) is admissible for Lemma 5.6.

Now let γ^* be an optimal solution to (CP) and $\gamma_\varepsilon = \gamma^* + \varepsilon\omega$ be an admissible perturbation of the curve γ^*. Then

$$J(\Omega(\gamma_\varepsilon)) \geq J(\Omega(\gamma^*)), \tag{5.6}$$

and in view of Corollaries 5.3 and 5.5 we have

$$J(\Omega(\gamma_\varepsilon)) = J(\Omega(\gamma^*)) + \varepsilon\, dJ(\Omega(\gamma^*); V(\omega)) + o(\varepsilon).$$

Hence from (5.6) it follows that

$$dJ(\Omega(\gamma^*); V(\omega)) \geq 0.$$

To summarize, we have the following necessary optimality condition.

THEOREM 5.7. *Assume* (H1)–(H11)*, and let the optimal curve* γ^* *satisfy* $\gamma^* \in \widetilde{U}_{ad} = U_{ad} \cap [C^2(0,2\pi)]^3$*; then there exists a shape gradient* $g_{\partial\Omega^*}$ *with support in* $\partial\Omega^*$*, such that*

$$\langle g_{\partial\Omega^*}, V_\nu(\omega)\rangle \mathcal{D}_1'(\partial\Omega^*) \times \mathcal{D}_1(\partial\Omega^*) \;\geq\; 0$$

for all $\omega \in T_{\widetilde{U}_{ad}}(\gamma^*)$*, where* $V_\nu(\omega)$ *is given by* (5.4) *and* $T_{\widetilde{U}_{ad}}(\gamma^*)$ *denotes the tangent set to* $\widetilde{U}_{ad}$ *at* γ^**.*

Remark 5.8. For numerical purposes it is possible to identify the shape gradient $g_{\partial\Omega^*}$ by using the Lagrange formalism [5]. However, this approach usually requires an additional regularity analysis of the state system to assure the differentiability of the corresponding Lagrangian and therefore is omitted in this paper.

Remark 5.9. It is not difficult to show, using similar arguments as in section 6, that the mapping $\gamma \mapsto dJ(\Omega(\gamma); V)$ is continuous with respect to weak-star convergence in $W^{2,\infty}$ on a sufficiently small open neighborhood of the set $U_{ad} \subset W^{2,\infty}(0,2\pi)$. Such a continuity implies the Frechet differentiability of the mapping $\gamma \mapsto J(\Omega(\gamma))$. Let us point out that the way we identify the gradient of the mapping $\gamma \mapsto J(\Omega(\gamma))$—taking into account the structure theorem—is important for numerical applications. By direct differentiation with respect to γ an equivalent form of the derivative can be obtained; however, the tangential component of the resulting vector-field on $\partial\Omega$ can be possibly taken into account, which may imply additional error when using the discretization of the continuous gradient for numerical solution of the optimization problem.

6. Strong material derivatives. In this section we provide all the estimates that finally lead to the proof of Corollary 5.3.

6.1. Transformation to the fixed domain. The following lemma describes the transport of div and grad to the fixed domain. The proof can be found in [28, sect. 2]. Note that the Jacobian of T_τ is denoted by $\mathcal{D}T_\tau$. Moreover, for any matrix B, the transposed one is denoted by ${}^*B$.

LEMMA 6.1. *Let* $B_1(\tau) = {}^*\mathcal{D}T_\tau^{-1}$*; then we have*
(1)

$$(\text{grad } \varphi) \circ T_\tau = \Big(B_1(\tau)\nabla\Big)\Big(\varphi \circ T_\tau\Big) \quad \textit{for all } \varphi \in H^1(D),$$

(2)

$$(\text{div } \psi) \circ T_\tau = \Big(B_1(\tau)\nabla\Big) \cdot \Big(\psi \circ T_\tau\Big) \quad \textit{for all } \psi \in \mathbf{H}^1(D).$$

(3)

$$(\text{curl } \psi) \circ T_\tau = \Big(B_1(\tau)\nabla\Big) \times \Big(\psi \circ T_\tau\Big) \quad \textit{for all } \psi \in \mathbf{H}^1(D).$$

Using Lemma 6.1, we obtain for (2.11a), with $\varphi \in H^1(\Omega_\tau)/\mathbb{R}$,

$$\begin{aligned} -\int_\Gamma j_g \varphi \, dx &= \sigma_0 \int_{\Omega_\tau} \nabla\phi_\tau \cdot \nabla\varphi \, dx \\ &= \sigma_0 \int_\Omega \det(\mathcal{D}T_\tau)\Big(\nabla\phi_\tau \cdot \nabla\varphi\Big) \circ T_\tau \, dx \\ &= \sigma_0 \int_\Omega B_2(\tau)\nabla\phi^\tau \cdot \nabla(\varphi \circ T_\tau) \, dx \end{aligned}$$

with

$$\beta(\tau) = \det(\mathcal{D}T_\tau) \text{ and } B_2(\tau) = \beta(\tau)\, {}^*B_1(\tau)B_1(\tau).$$

Hence (2.11a) is replaced with

$$-\int_\Gamma j_g \varphi \, dx = \alpha_0(\tau, \phi^\tau, \varphi) \qquad \text{for all } \varphi \in H^1(\Omega)/\mathbb{R}, \tag{6.1}$$

and

$$\alpha_0(\tau, \phi^\tau, \varphi) := \sigma_0 \int_\Omega B_2(\tau)\nabla\phi^\tau \cdot \nabla\varphi \, dx.$$

Now we turn to the Maxwell equation (2.11c). For the first term, we obtain

$$\begin{aligned} \sigma_0 \int_{G_\tau} \frac{\partial A_\tau}{\partial t} v \, dx \; &= \; \sigma_0 \int_G \beta(\tau) A_t^\tau \cdot (v \circ T_\tau) \, dx \\ &=: \alpha_1(\tau, A_t^\tau, v \circ T_\tau). \end{aligned}$$

For the next term, we utilize Lemma 6.1(2) and (3) to obtain

$$\begin{aligned} &\int_D \frac{1}{\mu} \text{ curl } A_\tau \cdot \text{ curl } v \, dx + \int_D \frac{1}{\mu} \text{ div } A_\tau \text{ div } v \, dx \\ &= \int_D \frac{\beta(\tau)}{\mu}\Big(\text{ curl } A_\tau \cdot \text{ curl } v\Big) \circ T_\tau \, dx + \int_D \frac{\beta(\tau)}{\mu}\Big(\text{ div } A_\tau \text{ div } v\Big) \circ T_\tau \, dx \\ &= \int_D \frac{\beta(\tau)}{\mu} \{(B_1(\tau)\nabla) \times A^\tau\} \cdot \{(B_1(\tau)\nabla) \times (v \circ T_\tau)\} \, dx \\ &\qquad + \int_D \frac{\beta(\tau)}{\mu} \{(B_1(\tau)\nabla) \cdot A^\tau\} \{(B_1(\tau)\nabla) \cdot (v \circ T_\tau)\} \, dx \\ &=: \alpha_2(\tau, A^\tau, v \circ T_\tau). \end{aligned}$$

For the last term in (2.11c), we have

$$\begin{aligned}\sigma_0 \int_{\Omega_\tau} \nabla\phi_\tau \cdot v \, dx &= \sigma_0 \int_\Omega \beta(\tau)\Big(\nabla\phi_\tau \cdot v\Big) \circ T_\tau \, dx \\ &= \sigma_0 \int_\Omega B_3(\tau)\nabla\phi^\tau \cdot (v \circ T_\tau) \, dx \\ &=: F(\tau, \phi^\tau, v \circ T_\tau),\end{aligned}$$

with $B_3(\tau) = \beta(\tau)B_1(\tau)$.

Altogether, we have replaced (2.11c) with

$$\alpha_1(\tau, A_t^\tau, v) + \alpha_2(\tau, A^\tau, v) + F(\tau, \phi^\tau, v) = 0 \qquad \text{for all } v \in X, \tag{6.2a}$$

$$A_0^\tau = A_0 \circ T_\tau. \tag{6.2b}$$

Remark 6.2. Another possibility to transport the divergence operator to the fixed domain is to use the formula (cf. [28])

$$(\operatorname{div} \psi) \circ T_\tau = \frac{1}{\beta(\tau)} \operatorname{div} \Big(\beta(\tau)\mathcal{D}T_\tau^{-1}(\psi \circ T_\tau)\Big) \quad \text{for all } \psi \in \mathbf{H}^1(D).$$

It shows that functions that are divergence free on Ω_τ generally lose this property when transported to the fixed domain. Hence the application of the Coulomb gauge $\operatorname{div} A = 0$ (cf. Remark 2.1) can be managed by the introduction of an auxiliary unknown function

$$\eta^\tau = \beta(\tau)\mathcal{D}T_\tau^{-1}A^\tau,$$

which in view of the above formula would give

$$\operatorname{div} A_\tau = 0 \quad \text{in } \Omega_\tau \quad \Longleftrightarrow \quad \operatorname{div} \eta^\tau = 0 \quad \text{in } \Omega.$$

However, we avoid such a transformation for the sake of simplicity.

6.2. Stability estimates.

LEMMA 6.3. *B_1, B_2, B_3, β are differentiable. For $|\tau| \le \tau_1$ and τ_1 small enough, we have*

$$\begin{aligned}\beta(\tau) &= 1 + \tau\beta'(0) + o(\tau), \\ B_i(\tau) &= I + \tau B_i'(0) + o(\tau), \qquad i = 1, \dots, 3.\end{aligned}$$

The derivatives at $\tau = 0$ are given by

$$\begin{aligned}\beta'(0) &= \operatorname{div} V(0), \\ B_1'(0) &= -{}^*\mathcal{D}V(0), \\ B_2'(0) &= \operatorname{div} V(0)I - 2\varepsilon(V(0)), \\ B_3'(0) &= \operatorname{div} V(0)I - {}^*\mathcal{D}V(0).\end{aligned}$$

Here, $\varepsilon(V(0))$ is the symmetrized part of $\mathcal{D}V(0)$, i.e., $\varepsilon(V(0)) = \frac{1}{2}(\mathcal{D}V(0) + {}^*\mathcal{D}V(0))$.

For the proof, we refer again to [28, sect. 2.13].

A particular consequence of Lemma 6.3 is the following corollary.

COROLLARY 6.4. *Let $|\tau| \le \tau_1$ and τ_1 be small enough. Then there exist real-valued functions g_i satisfying $g_i(\tau) = o(\tau)$, $i = 0, \dots, 3$, and bilinear forms $\tilde\alpha_i(\tau, \,.\,, \,.\,)$, $i = 0, 1, 2$, and $\tilde F(\tau, \,.\,, \,.\,)$ such that the following are valid:*

(1) *For all* $\varphi_1, \varphi_2 \in H^1(\Omega)/\mathbb{R}$, *we have*

$$\begin{aligned}
\alpha_0(\tau, \varphi_1, \varphi_2) &= \alpha_0(0, \varphi_1, \varphi_2) + \tau\alpha_{0,\tau}(0, \varphi_1, \varphi_2) + \tilde{\alpha}_0(\tau, \varphi_1, \varphi_2),\\
\alpha_{0,\tau}(0, \varphi_1, \varphi_2) &= \sigma_0 \int_\Omega B_2'(0)\nabla\varphi_1\nabla\varphi_2\, dx,\\
\left|\tilde{\alpha}_0(\tau, \varphi_1, \varphi_2)\right| &\le g_0(\tau)\|\nabla\varphi_1\|_{\mathbf{L}^2(\Omega)}\|\nabla\varphi_2\|_{\mathbf{L}^2(\Omega)}.
\end{aligned}$$

(2) *For all* $v_1, v_2 \in \mathbf{L}^2(D)$, *we have*

$$\begin{aligned}
\alpha_1(\tau, v_1, v_2) &= \alpha_1(0, v_1, v_2) + \tau\alpha_{1,\tau}(0, v_1, v_2) + \tilde{\alpha}_1(\tau, v_1, v_2),\\
\alpha_{1,\tau}(0, v_1, v_2) &= \sigma_0 \int_G \beta'(0) v_1 \cdot v_2\, dx,\\
\left|\tilde{\alpha}_1(\tau, v_1, v_2)\right| &\le g_1(\tau)\|v_1\|_{\mathbf{L}^2(G)}\|v_2\|_{\mathbf{L}^2(G)}.
\end{aligned}$$

(3) *For all* $v_1, v_2 \in \mathbf{X}$, *we have*

$$\begin{aligned}
\alpha_2(\tau, v_1, v_2) &= \alpha_2(0, v_1, v_2) + \tau\alpha_{2,\tau}(0, v_1, v_2) + \tilde{\alpha}_2(\tau, v_1, v_2),\\
\alpha_{2,\tau}(0, v_1, v_2) &= \int_D \frac{\beta'(0)}{\mu}\Big(\operatorname{curl} v_1 \cdot \operatorname{curl} v_2 + \operatorname{div} v_1 \operatorname{div} v_2\Big)\, dx\\
&\quad + \int_D \frac{1}{\mu}[(B_1'(0)\nabla) \times v_1] \cdot \operatorname{curl} v_2\, dx\\
&\quad + \int_D \frac{1}{\mu}\operatorname{curl} v_1 \cdot [(B_1'(0)\nabla) \times v_2]\, dx\\
&\quad + \int_D \frac{1}{\mu}[(B_1'(0)\nabla) \cdot v_1] \operatorname{div} v_2\, dx\\
&\quad + \int_D \frac{1}{\mu} \operatorname{div} v_1 \cdot [(B_1'(0)\nabla) \cdot v_2]\, dx,\\
|\tilde{\alpha}_2(\tau, v_1, v_2)| &\le g_2(\tau)\|v_1\|_{\mathbf{X}}\|v_2\|_{\mathbf{X}}.
\end{aligned}$$

(4) *For all* $\varphi \in H^1(\Omega)/\mathbb{R}$ *and* $v \in \mathbf{X}$, *we have*

$$\begin{aligned}
F(\tau, \varphi, v) &= F(0, \varphi, v) + \tau F_{,\tau}(0, \varphi, v) + \tilde{F}(\tau, \varphi, v),\\
F_{,\tau}(0, \varphi, v) &= \sigma_0 \int_\Omega B_3'(0)\nabla\varphi \cdot v\, dx,\\
\left|\tilde{F}(\tau, \varphi, v)\right| &\le g_4(\tau)\|\nabla\varphi\|_{\mathbf{L}^2(\Omega)}\|v\|_{\mathbf{X}}.
\end{aligned}$$

Using Corollary 6.4, we can prove the following stability result.

LEMMA 6.5. *Assume* (H1)–(H9) *and* (5.1)*; then there exists a constant* $C > 0$ *such that*

(1) $\|\nabla\phi^\tau - \nabla\phi\|_{H^1(0,T;\mathbf{L}^2(\Omega))} \le C \cdot |\tau|$,
(2) $\|A^\tau - A\|_{L^2(0,T;\mathbf{X})} + \|A_t^\tau - A_t\|_{L^{10/3}(0,T;\mathbf{L}^{10/3}(G))} \le C \cdot |\tau|$,
(3) $\|\theta^\tau - \theta\|_{W^{2,1}_{5/3}(Q)} \le C \cdot |\tau|$,
(4) $\|z^\tau - z\|_{W^{1,5}(0,T;L^5(\Sigma))} \le C \cdot |\tau|$.

Remark 6.6. (z^τ, θ^τ) is the solution to (2.11d)–(2.11h), where A_t in (2.11f) has been replaced with A_t^τ. In view of (5.1), we have $A_t^\tau = A_{\tau,t}$ on Σ.

For the proof, we need the following interpolation result.

LEMMA 6.7. *Let $u \in L^\infty(0,T;L^2(\Sigma)) \cap L^2(0,T;H^1(\Sigma))$; then there holds*

$$\int_0^T \|u(t)\|^{10/3}_{L^{10/3}(\Sigma)}\,dt \le \Big(\int_0^T \|u(t)\|^2_{L^6(\Sigma)}\,dt\Big)\,\|u\|^{4/3}_{L^\infty(0,T;L^2(\Sigma))}.$$

Proof. Owing to Riesz's convexity theorem (cf. [31, (A113)]), we have

$$\|u\|_{L^r(\Sigma)} \le \|u\|^{1-\Theta}_{L^{q_1}(\Sigma)}\|u\|^{\Theta}_{L^{q_2}(\Sigma)}$$

for all $u \in L^{q_1}(\Sigma) \cap L^{q_2}(\Sigma)$ with $1 \le q_1, q_2 < \infty$, $0 < \Theta < 1$, and $\frac{1}{r} = \frac{1-\Theta}{q_1} + \frac{\Theta}{q_2}$. Invoking the continuous embedding $H^1(\Sigma) \subset L^6(\Sigma)$, the assertion follows by defining $q_1 = 6$, $q_2 = 2$, $\Theta = \frac{2}{5}$, and $r = \frac{10}{3}$. ∎

Proof of Lemma 6.5. According to Lemma 6.3, we can write

$$\beta(\tau) = 1 + \tau\beta'(\xi_0), \quad B_i(\tau) = I + \tau B_i'(\xi_i), \quad i = 1,2,3, \tag{6.3}$$

for τ small enough and $\xi_i \in [0,\tau]$, $i = 0,\dots,3$. Note that $\beta(\tau) \ge c_{\tau_1} > 0$ for $|\tau| \le \tau_1$, if the latter has been chosen small enough, and that the B_i's are positive definite for $|\tau| \le \tau_1$.

Using (H2) and (H5), this immediately gives

$$\|\nabla\phi^\tau\|_{H^1(0,T;\mathbf{L}^2(\Omega))} \le c_1, \tag{6.4}$$

independent of τ. Moreover, we can use (6.1) and (6.3) to write

$$\begin{aligned} 0 &= \alpha_0(\tau,\phi^\tau,\varphi) - \alpha_0(0,\phi,\varphi) \\ &= \alpha_0(0,\phi^\tau-\phi,\varphi) + \tau\int_\Omega B_2'(\xi)\nabla\phi^\tau\cdot\nabla\varphi\,dx. \end{aligned}$$

Inserting $\varphi = \phi^\tau - \phi$ and using Young's inequality, we obtain

$$\|\nabla\phi^\tau - \nabla\phi\|_{L^2(0,T;\mathbf{L}^2(\Omega))} \le c_2|\tau|.$$

Since the same estimate holds true for $\phi_t^\tau - \phi_t$, assertion (1) is proved.

We insert $v = A_t^\tau$ into (6.2a), use (6.3), and integrate in time to obtain for the first term

$$\sigma_0\int_0^t\int_\Sigma \beta(\tau)A_s^\tau\cdot A_s^\tau\,dx\,ds \ge c_{\tau_1}\sigma_0\int_0^t\int_\Sigma |A_s^\tau|^2\,dx\,ds.$$

The second term gives

$$\begin{aligned} \int_0^t \alpha_2(\tau,A^\tau,A_s^\tau)\,ds &= \int_0^t\int_D \frac{\beta(\tau)}{\mu}\,\{(B_1(\tau)\cdot\nabla)\times A^\tau\}\cdot\{(B_1(\tau)\cdot\nabla)\times A_s^\tau\}\,dx\,ds \\ &\quad + \int_0^t\int_D \frac{\beta(\tau)}{\mu}\,\{(B_1(\tau)\cdot\nabla)\cdot A^\tau\}\,\{(B_1(\tau)\cdot\nabla)\cdot A_s^\tau\}\,dx\,ds \\ &= \frac{1}{2}\int_0^t\int_D \frac{\beta(\tau)}{\mu}\frac{\partial}{\partial s}\big|\,(B_1(\tau)\cdot\nabla)\times A^\tau\big|^2\,dx\,ds \\ &\quad + \frac{1}{2}\int_0^t\int_D \frac{\beta(\tau)}{\mu}\frac{\partial}{\partial s}\big|\,(B_1(\tau)\cdot\nabla)\cdot A^\tau\big|^2\,dx\,ds \\ &\ge \frac{1}{2\mu_2}\int_D \big|\,\mathrm{curl}\;A^\tau(t)\big|^2dx + \frac{1}{2\mu_2}\int_D \big|\,\mathrm{div}\;A^\tau(t)\big|^2\,dx + \tau\widetilde{g}(A^\tau(t)) - c_3, \end{aligned}$$

with a function $\widetilde{g}$ satisfying $\widetilde{g}(A^\tau(t)) \le c_4 \|A^\tau(t)\|^2_{\mathbf{X}}$. For the last term in (6.2a), we apply Young's inequality and obtain

$$\int_0^t F(\tau, \phi^\tau, A^\tau_s)\, ds \le \frac{c_{\tau_1}}{2}\sigma_0 \int_0^t \int_\Omega |A^\tau_s|^2\, dx\, ds + c_5 \int_0^t \int_\Omega |\nabla \phi^\tau|^2\, dx\, ds.$$

Invoking (6.4) and choosing τ small enough, we finally obtain

$$\|A^\tau\|_{L^\infty(0,T;\mathbf{X})} + \|A^\tau_t\|_{L^2(0,T;L^2(G))} \le c_6. \tag{6.5}$$

Now we differentiate (6.2a) formally with respect to time and insert $v = A^\tau_{tt}$. Defining

$$A^\tau_{0,t} = y \circ T_\tau \tag{6.6}$$

(cf. (H6) and (6.2b)), analogously to the derivation of the previous estimate, we get

$$\|A^\tau_t\|_{L^\infty(0,T;\mathbf{X})} + \|A^\tau_{tt}\|_{L^2(0,T;L^2(G))} \le c_7. \tag{6.7}$$

Next, we take the difference of (6.2a) for A^τ and A and obtain

$$\begin{aligned} 0 &= \alpha_1(\tau, A^\tau_t, v) + \alpha_2(\tau, A^\tau, v) + F(\tau, \phi^\tau, v) - \alpha_1(0, A_t, v) - \alpha_2(0, A, v) - F(0, \phi, v) \\ &= \alpha_1(0, A^\tau_t - A_t, v) + \alpha_2(0, A^\tau - A, v) + F(0, \phi^\tau - \phi, v) \\ &\qquad + G_0(\tau, \phi^\tau, v) + G_1(\tau, A^\tau_t, v) + G_2(\tau, A^\tau, v), \end{aligned} \tag{6.8}$$

with $G_0(\tau, \phi^\tau, v) = F(\tau, \phi^\tau, v) - F(0, \phi^\tau, v)$, $G_1(\tau, A^\tau_t, v) = \alpha_1(\tau, A^\tau_t, v) - \alpha_1(0, A^\tau_t, v)$, and $G_2(\tau, A^\tau, v) = \alpha_2(\tau, A^\tau, v) - \alpha_2(0, A^\tau, v)$ satisfying (cf. (6.3))

$$\begin{aligned} |G_0(\tau, \varphi, v)| &\le c_8 |\tau|\, \|\nabla \varphi\|_{\mathbf{L}^2(\Omega)} \|v\|_{\mathbf{X}}, \\ |G_1(\tau, v_1, v_2)| &\le c_9 |\tau|\, \|v_1\|_{\mathbf{L}^2(G)} \|v_2\|_{\mathbf{L}^2(G)}, \\ |G_2(\tau, v_1, v_2)| &\le c_{10} |\tau|\, \|v_1\|_{\mathbf{X}} \|v_2\|_{\mathbf{X}}. \end{aligned}$$

Inserting $v = A^\tau - A$ into (6.8) and integrating in time leads to

$$\begin{aligned} &\frac{\sigma_0}{2} \int_G |A^\tau(t) - A(t)|^2\, dx + \int_0^t \int_D \frac{1}{\mu} |\operatorname{curl}\,(A^\tau - A)|^2\, dx\, dt \\ &\le \sigma_0 \int_0^t \int_D |\nabla(\phi^\tau - \phi) \cdot (A^\tau - A)|\, dx\, dt + \frac{\sigma_0}{2} \int_G |A^\tau_0 - A_0|^2\, dx \\ &\quad + |\tau| c_8 \int_0^t \|\nabla \phi^\tau\|_{\mathbf{L}^2(\Omega)} \|A^\tau - A\|_{\mathbf{X}} + |\tau| c_9 \int_0^t \|A^\tau_t\|_{\mathbf{L}^2(G)} \cdot \|A^\tau - A\|_{\mathbf{L}^2(G)} \\ &\quad + |\tau| c_{10} \int_0^t \|A^\tau\|_{\mathbf{X}} \cdot \|A^\tau - A\|_{\mathbf{X}}. \end{aligned}$$

Applying the inequalities of Young and Gronwall and using (6.2b), we obtain

$$\|A^\tau - A\|_{L^\infty(0,T;L^2(G))} + \|A^\tau - A\|_{L^2(0,T;\mathbf{X})} \le c_{11} |\tau|. \tag{6.9}$$

Moreover, using (6.8) once again as well as (6.9), we obtain

$$\int_0^t \alpha_1(0, A^\tau_s - A_s, v)\, ds \le c_{12} |\tau| \|v\|_{L^2(0,t;\mathbf{X})}. \tag{6.10}$$

As before, we now differentiate (6.8) formally with respect to time, insert $v = A_t^\tau - A_t$, and make the same computations as before, but use (6.6) instead of (6.2b). Thus we obtain

$$\|A_t^\tau - A_t\|_{L^\infty(0,T;L^2(G))} + \|A_t^\tau - A_t\|_{L^2(0,T;\mathbf{X})} \le c_{13}|\tau|$$

and, similar to (6.10),

$$\int_0^t \alpha_1(0, A_{ss}^\tau - A_{ss}, v)\, ds \le c_{14}|\tau|\|v\|_{L^2(0,t;\mathbf{X})}. \tag{6.11}$$

To conclude the proof of assertion (2), we apply Lemma 6.7 with $u = A_t^\tau - A_t$, i.e.,

$$\|A_t^\tau - A_t\|_{L^{10/3}(0,T;\mathbf{L}^{10/3}(\Sigma))} \le c_{15} \cdot |\tau|. \tag{6.12}$$

To prove assertion (3), we define $\bar\theta = \theta^\tau - \theta$ and $\bar z = z^\tau - z$ (cf. Remark 6.6). Then $\bar\theta$ solves

$$\rho c_p \bar\theta_t - k\Delta\bar\theta = -\rho L \bar z_t + \sigma_0(A_t^\tau - A_t)\cdot(A_t^\tau + A_t) \text{ in } Q$$

$$\frac{\partial\bar\theta}{\partial\nu} = 0 \text{ in } \Sigma\times(0,T), \quad \bar\theta(0) = 0 \text{ in } \Sigma.$$

In view of Lemma 2.6, we can apply Hölder's inequality, Lemma 2.5(3), and (6.12) to infer

$$\begin{aligned}\|\bar\theta\|^{5/3}_{W^{2,1}_{5/3}(Q_t)} &\le c_{16}\int_0^t\int_\Sigma |\bar z_s|^{5/3}\,dx\,ds \\ &\quad + c_{17}\Big(\int_0^t\int_\Sigma |A_s^\tau - A_s|^{10/3}\,dx\,ds\Big)^{1/2}\Big(\int_0^t\int_\Sigma |A_s^\tau + A_s|^{10/3}\,dx\,ds\Big)^{1/2} \\ &\le c_{18}\int_0^t \|\bar\theta\|^{5/3}_{W^{2,1}_{5/3}(Q_s)} + c_{19}|\tau|^{5/3}.\end{aligned}$$

Then assertion (3) follows from Gronwall's lemma whereas assertion (4) is a direct consequence of (3), Lemma 2.5(3), and the continuous embedding $W^{2,1}_{5/3}(Q) \subset L^5(Q)$ (cf. Lemma 2.7). □

6.3. Material derivatives. Our main result in this section is the following theorem.

THEOREM 6.8. *Assume* (H1)–(H9) *and* (5.1)*; then the following are valid:*

(1) *The strong material derivative* $\nabla\dot\phi$ *exists in* $H^1(0,T;\mathbf{L}^2(\Omega))$,
$\dot A$ *exists in* $L^\infty(0,T;X)$ *and* $W^{1,10/3}(0,T;L^{10/3}(G))$,
$\dot z$ *exists in* $W^{1,5/2}(0,T;L^{5/2}(\Sigma))$,
$\dot\theta$ *exists in* $W^{2,1}_{5/3}(Q)$.

(2) $(\dot\phi, \dot A, \dot z, \dot\theta)$ *satisfy the linearized state equations*

$$\alpha_0(0,\dot\phi,\varphi) + \alpha_{0,\tau}(0,\phi,\varphi) = 0 \quad \textit{for all } \varphi \in H^1(\Omega)/\mathbb{R}, \tag{6.13a}$$

$$\begin{aligned}\alpha_1(0,\dot A_t, v) + \alpha_2(0,\dot A, v) + F(0,\dot\phi, v) + F_{,\tau}(0,\phi,v)\\ + \alpha_{1,\tau}(0,A_t,v) + \alpha_{2,\tau}(0,A,v) = 0 \quad \textit{for all } v \in \mathbf{X},\end{aligned} \tag{6.13b}$$

$$\dot A_0 - \mathcal{D}A_0 V(0) = 0 \quad \textit{in } D, \tag{6.13c}$$

$$\dot z_t - \frac{\partial f}{\partial\theta}\dot\theta - \frac{\partial f}{\partial z}\dot z = 0 \quad \textit{in } Q, \tag{6.13d}$$

$$\dot z(0) = 0 \quad \textit{in } \Sigma, \tag{6.13e}$$

$$\rho c_p \dot\theta_t - k\Delta\dot\theta + \rho L\dot z_t - 2\sigma_0 A_t\cdot\dot A_t = 0 \quad \textit{in } Q, \tag{6.13f}$$

$$\frac{\partial\dot\theta}{\partial\nu} = 0 \quad \textit{in } \partial\Sigma\times(0,T), \tag{6.13g}$$

$$\dot\theta(0) = 0 \quad \textit{in } \Sigma, \tag{6.13h}$$

where f is the right-hand side of (2.11h).

Proof. Similar to the proof of Theorem 2.9, one can show that (6.13a)–(6.13h) has a unique solution $(\dot\phi, \dot A, \dot z, \dot\theta)$; hence we omit this part of the proof. To prove assertion (3), we first test (6.13a) with $\dot\phi$. According to Corollary 6.4 and Lemma 6.3, we obtain

$$\sigma_0\int_0^t\int_\Omega |\nabla\dot\phi|^2\,dx\,ds \le \|B_2'(0)\|_{C^1(D)}\Big(\int_0^t\int_\Omega |\nabla\dot\phi|^2\,dx\,ds\Big)^{1/2}\Big(\int_0^t\int_\Omega |\nabla\phi|^2\,dx\,ds\Big)^{1/2}.$$

Using Young's inequality and (2.12), we obtain the estimate for $\nabla\dot\phi$. Then we again differentiate formally with respect to t and obtain the estimate for $\nabla\dot\phi_t$.

Next, we test (6.13b) with $\dot A$ and obtain

$$\begin{aligned}&\frac{\sigma_0}{2}\int_G |\dot A(t)|^2\,dx + \int_0^t\int_D \frac{1}{\mu}|\operatorname{curl}\dot A|^2\,dx\,ds + \int_0^t\int_D\frac{1}{\mu}|\operatorname{div}\dot A|^2\,dx\,ds\\ &\le \kappa_1\int_0^t \|\nabla\dot\phi\|_{\mathbf{L}^2(\Omega)}\|\dot A\|_{\mathbf{L}^2(\Omega)}\,ds + \kappa_1\|B_3'(0)\|_{C^1(D)}\int_0^t\|\nabla\phi\|_{\mathbf{L}^2(\Omega)}\|\dot A\|_{\mathbf{L}^2(\Omega)}\,ds\\ &\quad + \|\beta'(0)\|_{C^1(D)}\int_\Sigma \|A_s\|_{\mathbf{L}^2(G)}\|\dot A\|_{\mathbf{L}^2(G)}\,ds\\ &\quad + c_1\|B_1'(0)\|_{C^1(D)}\int_0^t\|A\|_{\mathbf{X}}\|\dot A\|_{\mathbf{X}}\,ds + \frac{\sigma_0}{2}\int_G|\dot A(0)|^2\,dx.\end{aligned}$$

Now we apply Young's inequality, Gronwall's lemma, and (6.13c) to infer

$$\|\dot A\|_{L^\infty(0,T;\mathbf{L}^2(G))} + \|\dot A\|_{L^2(0,T;\mathbf{X})} \le c_1\|V(0)\|_{C^1(D)}.$$

Using the corresponding initial condition for $\dot A_t$ (cf. (6.6)), we differentiate (6.13b) formally with respect to time and insert $v = \dot A$ to obtain

$$\|\dot A_t\|_{L^\infty(0,T;\mathbf{L}^2(G))} + \|\dot A_t\|_{L^2(0,T;\mathbf{X})} \le c_2\|V(0)\|_{C^1(D)}.$$

A further application of Lemma 6.7 then yields

$$\|\dot A_t\|_{L^{10/3}(0,T;L^{10/3}(G))} \le c_3\|V(0)\|_{C^1(D)}. \tag{6.14}$$

Next, we remark that similar to the derivation of Lemma 2.5(3), we can infer that

$$\|\dot{z}\|^p_{W^{1,p}(0,T;L^p(\Sigma))} \le c_4 \|\dot{\theta}\|_{L^p(Q)}.$$

Then in the light of (6.14), the last part of inequality (5.2) in Corollary 5.3 follows as in the proof of Lemma 6.5(3) and (4).

It remains to show that the solutions to (6.13a)–(6.13h) are the strong material derivatives. To this end, let

$$\psi^\tau = \frac{1}{\tau}(\phi^\tau - \phi) - \dot{\phi}; \tag{6.15}$$

then, according to Corollary 6.4, (6.1), and (6.13a), ψ^τ satisfies

$$\begin{aligned}\alpha_0(0,\psi^\tau,\varphi) &= -\frac{1}{\tau}\Big(\alpha_0(\tau,\phi^\tau,\varphi) - \alpha_0(0,\phi^\tau,\varphi)\Big) - \alpha_0(0,\dot{\phi},\varphi)\\ &= \alpha_{0,\tau}(0,\phi-\phi^\tau,\varphi) - \frac{1}{\tau}\tilde{\alpha}_0(\tau,\phi^\tau,\varphi).\end{aligned}$$

Integrating in time, inserting $\varphi = \psi^\tau$, and using Corollary 6.4 once again, we obtain

$$\|\nabla\psi^\tau\|_{L^2(0,T;\mathbf{L}^2(\Omega))} \xrightarrow[\tau\to 0]{} 0. \tag{6.16}$$

Since the same computations hold for $\nabla\phi_t$, the first part of assertion (1) is proved.

Next, defining

$$p^\tau = \frac{1}{\tau}(A^\tau - A) - \dot{A},$$

and using (6.13b) and Corollary 6.4, we see that p^τ satisfies

$$\begin{aligned}\alpha_1(0,p_t^\tau,v) + \alpha_2(0,p^\tau,v) &= -\frac{1}{\tau}\Big(F(\tau,\phi^\tau,v) - F(0,\phi,w)\Big)\\ &\quad -\frac{1}{\tau}\Big(\alpha_1(\tau,A_t^\tau,v) - \alpha_1(0,A_t^\tau,v)\Big) - \frac{1}{\tau}\Big(\alpha_2(\tau,A^\tau,v) - \alpha_2(0,A^\tau,v)\Big)\\ &\quad +F(0,\dot{\phi},v) + F_{,\tau}(0,\phi,v) + \alpha_{1,\tau}(0,A_t,v) + \alpha_{2,\tau}(0,A,v)\\ &= -F(0,\psi^\tau,v) - F_{,\tau}(0,\phi^\tau-\phi,v) + \frac{1}{\tau}\tilde{F}(\tau,\phi^\tau,v)\\ &\quad -\alpha_{1,\tau}(0,A_t^\tau - A_t,v) - \alpha_{2,\tau}(0,A^\tau - A,v)\\ &\quad -\frac{1}{\tau}\tilde{\alpha}_1(\tau,A_t^\tau,v) - \frac{1}{\tau}\tilde{\alpha}_2(\tau,A^\tau,v).\end{aligned} \tag{6.17}$$

We take $v = p^\tau$, integrate in time, and use Hölder's inequality to obtain

$$\begin{aligned}&\frac{\sigma_0}{2}\int_G |p^\tau|^2\,dx - \frac{\sigma_0}{2}\int_G |p_0^\tau|^2\,dx + \int_0^t\int_D \frac{1}{\mu}|\operatorname{curl} p^\tau|^2\,dxds + \int_0^t\int_D \frac{1}{\mu}(\operatorname{div} p^\tau)^2\,dxds\\ &\le \sigma_0\int_0^t \|\nabla\psi^\tau\|_{\mathbf{L}^2(\Omega)}\|p^\tau\|_{\mathbf{L}^2(\Omega)}\,ds + c_5\int_0^t \|\nabla\phi^\tau - \nabla\phi\|_{\mathbf{L}^2(\Omega)}\|p^\tau\|_{\mathbf{L}^2(\Omega)}\,ds\\ &\quad +\frac{1}{\tau}g_3(\tau)\int_0^t \|\nabla\phi^\tau\|_{\mathbf{L}^2(\Omega)}\|p^\tau\|_{\mathbf{L}^2(\Omega)}\,ds + \int_0^t\int_G \beta'(0)(A_s^\tau - A_s)\cdot p^\tau\,dx\,ds\\ &\quad +c_6\int_0^t \|A^\tau - A\|_{\mathbf{X}}\|p^\tau\|_{\mathbf{X}}\,ds + \frac{1}{\tau}g_1(\tau)\int_0^t \|A_s^\tau\|_{\mathbf{L}^2(G)}\|p^\tau\|_{\mathbf{L}^2(G)}\\ &\quad +\frac{1}{\tau}g_2(\tau)\int_0^t \|A^\tau\|_{\mathbf{X}}\|p^\tau\|_{\mathbf{X}}\,ds.\end{aligned}$$

Using (6.2b), the second term in (6.17) gives

$$\int_G |p_0^\tau|^2\,dx = \int_G \Big|\frac{1}{\tau}(A_0 \circ T_\tau - A_0) - \dot{A}_0\Big|^2\,dx.$$

According to [28, sect. 2.14], $\tau \mapsto A_0 \circ T_\tau$ is differentiable with

$$\frac{d}{d\tau}(A_0 \circ T_\tau)\Big|_{\tau=0} = DA_0V(0);$$

hence

$$A_0 \circ T_\tau = A_0 + \tau \mathcal{D}A_0V(0) + o(\tau).$$

Hence we obtain

$$\int_G |p_0^\tau|^2\,dx \xrightarrow[\tau\to 0]{} 0.$$

Regarding (5.1) and Lemma 6.3, $\beta'(0)p^\tau \in \mathbf{X}$ a.e. in $(0,T)$. Thus, we apply (6.10) to infer

$$\begin{aligned}\int_0^t \int_G \beta'(0)(A_s^\tau - A_s)\cdot p^\tau\,dx\,ds &= \frac{1}{\kappa_1}\int_0^t \alpha_1(0, A_s^\tau - A_s, \beta'(0)p^\tau)\,ds\\ &\le c_7|\tau|\|p^\tau\|_{L^2(0,t;\mathbf{X})}.\end{aligned}$$

Then we apply Young's inequality, Corollary 6.4, (6.16), and Gronwall's lemma to conclude

$$\|p^\tau\|^2_{L^\infty(0,T;\mathbf{L}^2(G))} + \|p^\tau\|^2_{L^2(0,T;\mathbf{X})} \xrightarrow[\tau\to 0]{} 0.$$

Now we differentiate (6.17) formally with respect to time, repeat the same considerations as before (but use (6.6) as initial value instead of (6.2b)), and obtain

$$\|p_t^\tau\|^2_{L^\infty(0,T;\mathbf{L}^2(G))} + \|p_t^\tau\|^2_{L^2(0,T;\mathbf{X})} \xrightarrow[\tau\to 0]{} 0.$$

A further application of Lemma 6.7 finally yields

$$\|p_t^\tau\|_{L^{10/3}(0,T;\mathbf{L}^{10/3}(G))} \xrightarrow[\tau\to 0]{} 0. \tag{6.18}$$

To prove the differentiability of θ^τ and z^τ, we define

$$\begin{aligned}q^\tau &= \frac{1}{\tau}(\theta^\tau - \theta) - \dot{\theta},\\ r^\tau &= \frac{1}{\tau}(z^\tau - z) - \dot{z};\end{aligned}$$

then (q^τ, r^τ) solve

$$\rho c_p q_t^\tau - k\Delta q^\tau = -\rho L r_t^\tau + \sigma_0\tau|\dot{A}_t|^2 + \sigma_0 p_t^\tau\cdot\Big(2A_t + 2\tau\dot{A}_t + \tau p_t^\tau\Big), \tag{6.19a}$$

$$\begin{aligned}r_t^\tau &= \frac{1}{\tau}\Big(f(\theta^\tau, z^\tau) + f(\theta, z)\Big) - \frac{\partial f}{\partial\theta}(\theta, z)\dot{\theta} - \frac{\partial f}{\partial z}(\theta, z)\dot{z}\\ &=: G(\tau),\end{aligned} \tag{6.19b}$$

$$\frac{\partial q^\tau}{\partial\nu} = 0, \qquad q^\tau = 0, \qquad r^\tau(0) = 0. \tag{6.19c}$$

Owing to (H7)–(H9), we can apply Taylor's formula to develop $G(\tau)$ and obtain (with a constant $\xi \in [0,1]$)

$$\begin{aligned}
|G(\tau)| &= \Big|\frac{1}{\tau}\Big(f(\theta+\tau(q^\tau+\dot\theta), z+\tau(r^\tau+\dot z)) - f(\theta,z)\Big) - \frac{\partial f}{\partial\theta}(\theta,z)\dot\theta - \frac{\partial f}{\partial z}(\theta,z)\dot z\Big| \\
&= \Big|(q^\tau+\dot\theta)\frac{\partial f}{\partial\theta}(\theta+\xi\tau(q^\tau+\dot\theta), z+\xi\tau(r^\tau+\dot z)) \\
&\qquad + (r^\tau+\dot z)\frac{\partial f}{\partial z}(\theta+\xi\tau(q^\tau+\dot\theta), z+\xi\tau(r^\tau+\dot z)) - \frac{\partial f}{\partial\theta}(\theta,z)\dot\theta - \frac{\partial f}{\partial z}(\theta,z)\dot z\Big| \\
&\le c_8|q^\tau| + c_9|r^\tau| + |\dot\theta|\Big|\frac{\partial f}{\partial\theta}(\theta+\xi(\theta^\tau-\theta), z+\xi(z^\tau-z)) - \frac{\partial f}{\partial\theta}(\theta,z)\Big| \\
&\qquad + |\dot z|\Big|\frac{\partial f}{\partial z}(\theta+\xi(\theta^\tau-\theta), z+\xi(z^\tau-z)) - \frac{\partial f}{\partial z}(\theta,z)\Big| \\
&\le c_8|q^\tau| + c_9|r^\tau| + c_{10}|\dot\theta||\theta^\tau-\theta| + c_{11}|\dot\theta||z^\tau-z| + c_{12}|\dot z||\theta^\tau-\theta| + c_{13}|\dot z||z^\tau-z|.
\end{aligned}$$

Owing to (6.18) and (5.2), the last term of the right-hand side of (6.19a) will be in $L^{5/3}(0,T;L^{5/3}(\Sigma))$. Thus we try to get an estimate for $G(\tau)$ in the same space. To this end, we apply the inequalities of Hölder and Young and use (5.2) to obtain

$$\begin{aligned}
&\int_0^t\int_\Sigma |G(\tau)|^{5/3}\,dx\,ds \le c_{14}\int_0^t\int_\Sigma |q^\tau|^{5/3}\,dx\,ds + c_{15}\int_0^t\int_\Sigma |r^\tau|^{5/3}\,dx\,ds \\
&+ c_{16}\Big(\int_0^t\int_\Sigma |\theta^\tau-\theta|^{10/3}\,dx\,ds\Big)^{1/2} + c_{17}\Big(\int_0^t\int_\Sigma |z^\tau-z|^{10/3}\,dx\,ds\Big)^{1/2}. \qquad (6.20)
\end{aligned}$$

Next, we test (6.19b) with $(r^\tau)^{2/3}$, use the estimate above, and apply the inequalities of Young and Gronwall, as well as the stability estimates of Lemma 6.5, to obtain

$$\frac{3}{5}\int_\Sigma |r^\tau|^{5/3}\,dx \le c_{18}|\tau|^{5/3} + c_{19}\int_0^t\int_\Sigma |q^\tau|^{5/3}\,dx\,ds.$$

Using the last estimate and (6.20) we go back to (6.19b) and conclude

$$\int_0^t\int_\Sigma |r_s^\tau|^{5/3}\,dx\,ds \le c_{20}|\tau|^{5/3} + c_{21}\int_0^t\int_\Sigma |q^\tau|^{5/3}\,dx\,ds. \qquad (6.21)$$

Now we can proceed again as in the proof of Lemma 6.5(3); i.e., we apply Lemma 2.6 to (6.19a) and use (6.21) and (5.2) to obtain (recall that $g(\tau) = o(1)$ if and only if $g(\tau) \to 0$ for $\tau \to 0$)

$$\|q^\tau\|_{W^{2,1}_{5/3}(Q)} = o(1). \qquad (6.22)$$

Using the embedding $W^{2,1}_{5/3}(Q) \subset L^5(Q)$ (cf. Lemma 2.7), we can go back, estimate $G(\tau)$ again (this time in $L^{5/2}(Q)$), and obtain finally

$$\|r^\tau\|_{W^{1,5/2}(0,T;L^{5/2}(\Sigma))} \xrightarrow[\tau\to 0]{} 0. \qquad \square$$

REFERENCES

[1] M. Berger and B. Gostiaux, *Differential Geometry: Manifolds, Curves, and Surfaces*, Springer-Verlag, New York, 1988.

[2] O. Bodart and R. Touzani, *Optimal control of electromagnetic induction heating* 1. *A two-control parameter model*, Preprint, Laboratoire de Mathématiques Appliquées, Université Blaise Pascal (Clermont-Ferrand 2), France, 1996.

[3] A. Bossavit, *Free boundaries in induction heating*, Control Cybernet., 14 (1985), pp. 69–96.
[4] A. Bossavit and J.-F. Rodrigues, *On the electromagnetic "induction heating" problem in bounded domains*, Adv. Math. Sci. Appl., 4 (1994), pp. 79–92.
[5] J. Cea, *Conception optimale ou identification de formes, calcul rapide de la dèrivèe directionelle de la fonction coût*, M2AN Math. Model. Numer. Anal., 20 (1986), pp. 371–402.
[6] D. Chenais, *On the existence of a solution in a domain identification problem*, J. Math. Anal. Appl., 52 (1975), pp. 189–219.
[7] S. Clain, *Numerical modelling of induction heating for two-dimensional geometries*, Math. Models Methods Appl. Sci., 3 (1993), pp. 271–281.
[8] F.W. Curtis, *High Frequency Induction Heating*, McGraw–Hill, New York, 1950.
[9] M.C. Delfour and J.P. Zolésio, *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*, Adv. Des. Control 4, SIAM, Philadelphia, 2001.
[10] L.R. Egan and E.P. Furlani, *A computer simulation of an induction heating system*, IEEE Trans. Mag., 27 (1991), pp. 4343–4354.
[11] H. Federer, *Curvature measures*, Trans. Amer. Math. Soc., 93 (1959), pp. 418–491.
[12] J. Fuhrmann and D. Hömberg, *Numerical simulation of surface heat treatments*, Internat. J. Numer. Methods Heat Fluid Flow, 9 (1999), pp. 705–724.
[13] J. Fuhrmann, D. Hömberg, and M. Uhle, *Numerical simulation of induction hardening of steel*, COMPEL, 18 (1999), pp. 482–493.
[14] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer, Berlin, 1986.
[15] D. Hömberg, *A mathematical model for the phase transitions in eutectoid carbon steel*, IMA J. Appl. Math., 54 (1995), pp. 31–57.
[16] D. Hömberg, *A numerical simulation of the Jominy end-quench test*, Acta Mater., 44 (1996), pp. 4375–4385.
[17] D. Hömberg, *Irreversible phase transitions in steel*, Math. Methods Appl. Sci., 20 (1997), pp. 59–77.
[18] D. Hömberg, *A mathematical model for induction hardening including mechanical effects*, Nonlinear Anal. Real World Appl., to appear.
[19] R.H.W. Hoppe and R. Kornhuber, *Multi-grid solution of two coupled Stefan equations arising in induction heating of large steel slabs*, Internat. J. Numer. Methods Engrg., 30 (1990), pp. 779–801.
[20] J. Kačur, *Method of Rothe in Evolution Equations*, Teubner-Texte Math. 80, Teubner, Leipzig, Germany, 1985.
[21] A. Kost, *Numerische Methoden in der Berechnung elektromagnetischer Felder*, Springer, Berlin, 1994.
[22] O.A. Ladyzenskaja, V.A. Solonnikov, and N.N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Amer. Math. Soc. Transl. 23, AMS, Providence, RI, 1968.
[23] J.-B. Leblond and J. Devaux, *A new kinetic model for anisothermal metallurgical transformations in steels including effect of austenite grain size*, Acta Met., 32 (1984), pp. 137–146.
[24] J. Rappaz and M. Swierkosz, *Mathematical modelling and numerical simulation of induction heating processes*, Appl. Math. Comput. Sci., 6 (1996), pp. 207–221.
[25] J. Rappaz and M. Swierkosz, *A boundary element method for solving an external vector potential problem. Application to electromagnetic heating*, J. Comput. Phys., 11 (1997), pp. 145–150.
[26] J.-F. Rodrigues, *A nonlinear parabolic system arising in thermomechanics and in thermomagnetism*, Math. Models Methods Appl. Sci., 2 (1992), pp. 271–281.
[27] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.
[28] J. Sokołowski and J.-P. Zolésio, *Introduction to Shape Optimization*, Springer Ser. Comput. Math. 16, Springer, Berlin, 1992.
[29] H. von der Mosel, *Elastic knots in Euclidean 3-space*, Ann. Inst. Henri Poincaré Anal. Non Linéaire, 16 (1999), pp. 137–166.
[30] J. Wloka, *Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1987.
[31] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, Vol. II, Springer, New York, 1990.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 3, pp. 1118–1142

MINIMIZATION OF RISK AND LINEAR QUADRATIC OPTIMAL CONTROL THEORY*

MICHAEL KOHLMANN† AND SHANJIAN TANG‡

Abstract. This article is concerned with the optimal control problem for the linear stochastic system $X_t = x + \int_0^t (A_s X_s + B_s u_s + f_s)\,ds + \int_0^t \sum_{i=1}^d [C_i(s)X_s + D_i(s)u_s + g_i(s)]\,dw_i(s)$ with the convex risk functional $J(u) = EM(X_T) + E\int_0^T G(t, X_t, u_t)\,dt$. In order to guarantee the existence of an optimal control without any (weak) compactness assumption on the admissible control set, we assume that the risk function M is coercive and that $\sum_{i=1}^d D_i^* D_i$ is uniformly positive, rather than to assume like in the control literature that the running risk function G is coercive with respect to the control variable. In this new setting, the running risk function G may be independent of the control variable, and therefore the so-called singular linear-quadratic (LQ) stochastic control problem is included. A rigorous theory is developed for the general stochastic LQ problem with random coefficients, and the bounded mean oscillation–martingale theory is used to account for the concerned integrability. It plays a crucial role in the following exposition: (a) to connect the stochastic LQ problem to two associated backward stochastic differential equations (BSDEs)—one is an $n \times n$ symmetric matrix-valued nonlinear Riccati BSDE and the other is an n-dimensional linear BSDE with unbounded coefficients; (b) to show that the latter BSDE has an adapted solution pair of the suitably necessary regularity. This seems to be the first application in a stochastic LQ theory of the BMO-martingale theory, which roots in harmonic analysis. Furthermore, with the help of an a priori estimate on the risk functional, existence and uniqueness of the solutions of backward stochastic Riccati differential equations (BSRDEs) in the singular case is reduced to the regular case via a perturbation method, and then a new existence and uniqueness result on BSRDEs is obtained for the singular case.

Key words. minimization of risk, linear-quadratic stochastic control, nonlinear backward stochastic Riccati equation, BMO-martingale

AMS subject classifications. 49N10, 91B28, 93C05

DOI. 10.1137/S0363012900372465

1. Introduction.

Let $(\Omega, \mathcal{F}_T, P, \{\mathcal{F}_t, 0 \le t \le T\})$ be a fixed complete probability space on which is defined a standard $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted d-dimensional Brownian motion $\{w(t) \equiv (w_1(t), \dots, w_d(t))^*, 0 \le t \le T\}$. Assume that $\{\mathcal{F}_t, 0 \le t \le T\}$ is the completion, by the totality $\mathcal{N}$ of all null sets of $\mathcal{F}_T$, of the natural filtration $\{\mathcal{F}_t^w, 0 \le t \le T\}$ generated by w.

Consider the optimal control problem of the linear stochastic system

$$\text{(1.1)} \qquad \begin{cases} dX_s &= [AX_s + Bu(s) + f_s]\,ds \\ &\quad + \displaystyle\sum_{i=1}^d [C_i X_s + D_i u(s) + g_i(s)]\,dw_i(s), \quad 0 < s \le T, \\ X_0 &= x, \qquad u(t) \in R^m, \end{cases}$$

under the cost functional

*Received by the editors May 22, 2000; accepted for publication (in revised form) February 15, 2003; published electronically July 8, 2003. This work was supported by the Center of Finance and Econometrics, University of Konstanz.

http://www.siam.org/journals/sicon/42-3/37246.html

†Department of Mathematics and Statistics, University of Konstanz, D-78457, Konstanz, Germany (Michael.Kohlmann@uni-konstanz.de).

‡Department of Mathematics, Fudan University, Shanghai 200433, China (sjtangk@online.sh.cn). This author was supported by a research fellowship from the Alexander von Humboldt Foundation and by the National Natural Science Foundation of China under grant 79790130.

$$J(u;0,x) = EM(X_T) + E\int_0^T G(s, X_s, u_s)\,ds. \tag{1.2}$$

Here $M(x)$ is $\mathcal{F}_T$-measurable for each $x \in R^n$ and is uniformly convex in x, and $G(t,x,u)$ is $\{\mathcal{F}_t, 0 \le t \le T\}$-measurable and convex in (x,u). Assume that all the coefficients A, B, C_i, D_i are $\{\mathcal{F}_t, 0 \le t \le T\}$-progressively measurable bounded matrix-valued processes, defined on $\Omega \times [0,T]$, of suitable dimensions. $u(\cdot)$ is the control, which is required to take values in a previously given nonempty closed convex subset U of the m-dimensional Euclidean space R^m and to be adapted to $\{\mathcal{F}_t, 0 \le t \le T\}$.

We are concerned with the existence of optimal controls without any direct (weak) compactness assumption on the admissible class of controls . In the control literature, to formulate such an existence theory, apart from the convexity assumption on M and G, it is usually assumed that G is *coercive* in the *control* variable u, that is, there are some constants $\delta_1 > 0, \delta_2 > 0$, and $p > 1$ such that

$$G(t,x,u) \ge \delta_1 |u|^p - \delta_2 \quad \forall\ (t,x,v) \in [0,T] \times R^n \times R^m.$$

The latter coercivity implies that the cost functional $J(\cdot;0,x)$ is coercive, and therefore an optimal control is a priori known to lie within a weakly compact subset of the admissible class of controls. Then with the obvious functional analytical arguments (see Ekeland and Temam [13, Corollary 2.1 and Proposition 2.12]), the existence of an optimal control is obtained. We refer the reader to Lions [24] for optimal control of systems governed by PDEs and to Bismut [4, 6] and to Yong and Zhou [35] for optimal control of stochastic systems.

In section 2, we shall formulate a different existence result. Rather than assume that G is *coercive* in the *control* variable u, we assume instead that M is *coercive* in the *state* variable x. Furthermore, we assume that $\sum_{i=1}^d D_i^* D_i$ is uniformly positive. Under the two assumptions, the cost functional $J(\cdot;0,x)$ is proved to be still coercive. The proof relies on an a priori estimate on the concerned stochastic system.

In the above new setting, the running risk functional may be independent of the control variable. Therefore, the so-called singular stochastic linear-quadratic (LQ) problem is included. In particular, the mean-variance hedging problem is covered as a special one-dimensional case. The two problems are important in stochastic control theory and mathematical finance, respectively.

Section 3 is devoted to the study of the general stochastic LQ problem. The first part aims at characterizing the solution of the LQ problem in terms of two associated backward stochastic differential equations (BSDEs). In this consideration, the routine arguments in the case of constant coefficients meet with, in our general case, new integrability difficulties due to the appearance of possibly unbounded coefficients. These difficulties reflect the essential feature of the general LQ problem with random coefficients. To have the necessary rigor, the BMO-martingale theory has to be introduced to account for the required integrability. Here, BMO is the abbreviation for "bounded mean oscillation." The second part is devoted to the reduction of backward stochastic Riccati differential equations (BSRDEs) in the singular case to the regular case via a perturbation method. As a by-product of this reduction, the generalization of the existence and uniqueness result on BSRDEs in the regular case of Bismut [6] and Peng [27] is obtained in the singular case.

Finally, we conclude the paper in section 4 by giving some comments.

In summary, the contributions of the paper include the following:

1. A new existence result is formulated, in which the singular stochastic LQ problem and the mean-variance hedging problem are included.

2. New difficulties are identified in the general nonhomogeneous stochastic LQ problem, and the BMO-martingale theory is introduced to overcome these difficulties.
3. The existence and uniqueness result for BSRDEs in the singular case is reduced to the regular case. A new existence and uniqueness result for BSRDEs in some singular case is established.

These results will find applications in the so-called mean-variance hedging (see, e.g., [12]), which is extensively studied in the mathematical finance literature.

We conclude this section by introducing some notations.

M^* is the transpose of the vector or matrix M. $|M| := \sqrt{\Sigma_{ij} m_{ij}^2}$ for any vector or matrix $M = (m_{ij})$. $\langle M_1, M_2 \rangle$ is the inner product of the two vectors M_1 and M_2. R^n is the n-dimensional Euclidean space. S^n is the Euclidean space of all $n \times n$ symmetric matrices. S_+^n is the set of all $n \times n$ nonnegative definite matrices. $C(t, T; H)$ is the H-valued continuous functions on $[t, T]$, endowed with the maximum norm for a given Hilbert space H.

Let H be a given Euclidean space and $p \geq 1$. $\mathcal{L}_{\mathcal{F}}^p(t, T; H)$ is the space of H-valued $\{\mathcal{F}_s, t \leq s \leq T\}$-adapted L^p-integrable stochastic processes f on $[t, T]$, endowed with the norm $(E \int_t^T |f_s|^p \, ds)^{1/p}$. $\mathcal{L}_{\mathcal{F}}^\infty(t, T; H)$ is the space of H-valued $\{\mathcal{F}_s, t \leq s \leq T\}$-adapted essentially bounded stochastic processes f on $[t, T]$, endowed with the norm $\operatorname{ess\,sup}_{s,\omega} |f_s|$. $L^p(\Omega, \mathcal{F}_T, P; H)$ is the space of H-valued L^p-integrable random variables on $(\Omega, \mathcal{F}_T, P)$. $R := R^1$; $\mathcal{L}_{\mathcal{F}}^p(t, T) := \mathcal{L}_{\mathcal{F}}^p(t, T; R)$; $C(t, T) := C(t, T; R)$. $L^\infty(\Omega, \mathcal{F}_T, P; C([0, T]; H))$ is the space of $C([0, T]; H)$-valued, essentially maximum-norm-bounded random variables f on the fixed probability space $(\Omega, \mathcal{F}_T, P)$, endowed with the norm $\operatorname{ess}_{\omega \in \Omega} \sup_{0 \leq t \leq T} |f(t, \omega)|$. $L^\infty(\Omega, \mathcal{F}_T, P; C([0, T]; S_+^n))$ consists of those elements of $L^\infty(\Omega, \mathcal{F}_T, P; C([0, T]; S^n))$ which take values in S_+^n. $\mathcal{M}_{\mathcal{F}}^p(t, T; H)$ is the set of all $\{\mathcal{F}_s, t \leq s \leq T\}$-adapted processes ψ which take values in H and are such that $\|\psi\|_{\mathcal{M}^p}^p := E |\int_t^T |\psi_s|^2 \, ds|^{p/2} < +\infty$. $\mathcal{S}_{\mathcal{F}}^p(t, T; H)$ is the set of all the continuous $\{\mathcal{F}_s, t \leq s \leq T\}$-adapted processes ψ which take values in H and are such that $\|\psi\|_{\mathcal{S}_{\mathcal{F}}^p}^p := E[\sup_{t \leq s \leq T} |\psi_s|^p] < +\infty$. Note also that $\mathcal{L}_{\mathcal{F}}^p(t, T; H), \mathcal{L}_{\mathcal{F}}^\infty(t, T; H)$, $L^p(\Omega, \mathcal{F}_T, P; H), \mathcal{M}_{\mathcal{F}}^p(t, T; H)$, and $\mathcal{S}_{\mathcal{F}}^p(t, T; H)$ are occasionally abbreviated as $\mathcal{L}_{\mathcal{F}}^p$, $\mathcal{L}_{\mathcal{F}}^\infty, L^p, \mathcal{M}_{\mathcal{F}}^p$, and $\mathcal{S}_{\mathcal{F}}^p$, respectively, when the context is clear. This often happens especially when they appear in the subscripts.

Throughout this paper, by a square integrable process f we mean that it satisfies $E \int_0^T |f_s|^2 \, ds < \infty$.

2. The risk minimization problem.

2.1. The linear stochastic control system. Let $(\Omega, \mathcal{F}_T, P, \{\mathcal{F}_t, 0 \leq t \leq T\})$ be a fixed complete probability space on which is defined a standard $\{\mathcal{F}_t, 0 \leq t \leq T\}$-adapted d-dimensional Brownian motion $\{w(t) := (w_1(t), \ldots, w_d(t))^*, 0 \leq t \leq T\}$. Assume that $\{\mathcal{F}_t, 0 \leq t \leq T\}$ is the completion, by the totality $\mathcal{N}$ of all null sets of $\mathcal{F}_T$, of the natural filtration $\{\mathcal{F}_t^w, 0 \leq t \leq T\}$ generated by w.

Consider the optimal control problem of the linear stochastic system parameterized by the initial data $(x, t) \in R^n \times [0, T]$:

$$
\left\{
\begin{array}{rcl}
dX_s &=& [AX_s + Bu(s) + f_s] \, ds \\
&& + \sum_{i=1}^d [C_i X_s + D_i u(s) + g_i(s)] \, dw_i(s), \quad t < s \leq T, \\
X_t &=& x.
\end{array}
\right. \tag{2.1}
$$

We introduce the following two assumptions on the control system (2.1).

(A1) Let A, B, C_i, D_i be $\{\mathcal{F}_t,\ 0 \le t \le T\}$-progressively measurable bounded matrix-valued processes, defined on $\Omega \times [0,T]$, of dimensions $n \times n$, $n \times m$, $n \times n$, $n \times m$, respectively. The $d+1$ processes $g_i, i = 1, \dots, d$, and f belong to $\mathcal{L}^2_{\mathcal{F}}(0,T;R^n)$.

(A2) There is a real deterministic positive number δ such that $\sum_{i=1}^d D_i^* D_i \ge \delta I_{m\times m}$.

For given initial data (t,x) and control u such that $\int_t^T |u|^2\,ds < \infty$ a.s., the above control system (2.1) has a unique strong solution X (see Gal'chuk [15]). It will be denoted by $X^{t,x;u}$ whenever it is necessary to indicate its dependence on the triplet (x,t,u). We have the following lemma.

LEMMA 2.1. *Let assumption* (A1) *be satisfied. Then, for* $1 \le p < \infty$, *there is a deterministic positive constant* β *such that for any* $u \in \mathcal{M}^p_{\mathcal{F}}(t,T;R^m)$, $f \in \mathcal{L}_{\mathcal{F}}(t,T;R^n)$, *and* $(g_1,\dots,g_d) \in (\mathcal{M}^p_{\mathcal{F}}(t,T;R^n))^d$, *we have*

$$E^{\mathcal{F}_t} \max_{t\le s\le T} |X_s|^p \le \beta\left(|x|^p + E^{\mathcal{F}_t}\left| \int_t^T |f_s|\,ds \right|^p + E^{\mathcal{F}_t}\left| \int_t^T (|g(s)|^2 + |u_s|^2)\,ds \right|^{p/2} \right).$$

For the proof the reader is referred to Protter [31, Chapter V, Lemma 2, p. 196].

We have the following a priori estimate for system (2.1).

LEMMA 2.2. *Let assumptions* (A1) *and* (A2) *be satisfied. If* $u \in \mathcal{M}^p(0,T;R^m)$ *for some* $p \in [1,\infty)$, *then there is a deterministic constant* $\beta_{p,\delta,T} > 0$ *which depends on* (p,δ,T) *and the uniform bound of* $A, B, C := (C_1,\dots,C_d)$ *such that*

$$\begin{aligned} &E^{\mathcal{F}_t}\left| \int_t^T |u_s|^2\,ds \right|^{p/2} + E^{\mathcal{F}_t} \sup_{t\le s\le T} |X_s|^p \\ &\le \beta_{p,\delta,T} E^{\mathcal{F}_t}\left(|X_T|^p + \int_t^T |f_s|^p\,ds + \left| \int_t^T |g(s)|^2\,ds \right|^{p/2} \right). \end{aligned} \tag{2.2}$$

The proof is an adaptation of El Karoui, Peng, and Quenez [14, pp. 54–57], incorporating assumption (A2). It is left to the diligent reader as an exercise.

Lemma 2.2 is an a priori estimate for BSDEs. We remark here that BSDEs were initially introduced by Bismut [3] for the linear case and were later developed by Pardoux and Peng [25] for the general Lipschitz nonlinear case.

Throughout this paper, we write $\beta_{\delta,T}$ for $\beta_{2,\delta,T}$.

2.2. Admissible class of controls. Assume that U is a nonempty closed convex subset of R^m. Denote by U^t_{ad} the set of all the $\{\mathcal{F}_s, t \le s \le T\}$-adapted processes $\{u_s, t \le s \le T\}$ with values in U such that $\int_t^T |u(s)|^2\,ds < \infty$ a.s. Define

$$U^{t,p}_{ad} := \left\{ u \in U^t_{ad} : E\left(\int_t^T |u_s|^2\,ds \right)^{p/2} < \infty \right\}; \tag{2.3}$$

it is convex and closed. For given initial data (t,x), the control system (2.1) has a unique strong solution $X^{t,x;u}$. The following presents a multidimensional generalization of the closedness of attainable contingent claims in Pham, Rheinländer, and Schweizer [30].

PROPOSITION 2.3. *Let assumptions* (A1) *and* (A2) *be satisfied. Then, for every* $x \in R^n$, *the attainable set* $\mathcal{R}_p(0,x;T)$ *at time* T *when starting from point* x *at time*

0, *of system* (2.1), *defined by*

$$\mathcal{R}_p(0,x;T) := \{X_T^{0,x;u} : u \in U_{ad}^{0,p}\} \tag{2.4}$$

is convex and closed in $\mathcal{L}_{\mathcal{F}}^p(0,T;R^m)$.

Proof. Let $\{\xi_k\}_{k=1}^{\infty}$ be a sequence of elements of $\mathcal{R}_p(0,x;T)$ converging to ξ strongly in $\mathcal{L}_{\mathcal{F}}^p(0,T;R^n)$. We shall show that there is $u \in U_{ad}^{0,p}$ such that $\xi = X_T^{0,x;u}$.

By the definition of $\mathcal{R}_p(0,x;T)$, there are $u_k \in U_{ad}^{0,p}$ such that

$$\xi_k = X_T^{0,x;u_k}, \quad k = 1,2\ldots.$$

Since the pair $(\delta X, \delta u) := (X^{0,x;u_k} - X^{0,x;u_l}, u_k - u_l)$ satisfies the SDE

$$d\delta X_t = (A_t \delta X_t + B_t \delta u_t)\,dt + \sum_{i=1}^{d} [C_i(t)\delta X_t + D_i(t)\delta u_t]\,dw_i(t) \tag{2.5}$$

with the terminal condition $\delta X_T = \xi_k - \xi_l$, we derive from Lemma 2.2

$$\|u_k - u_l\|^p_{\mathcal{M}_{\mathcal{F}}^p(0,T;R^m)} \le \beta_{p,\delta,T} E|\xi_k - \xi_l|^p, \quad k,l = 1,2,\ldots. \tag{2.6}$$

This last inequality implies that $\{u_k\}_{k=1}^{\infty}$ converges strongly in $\mathcal{M}^p(0,T;R^m)$. Since $\mathcal{M}_{\mathcal{F}}^p(0,T;R^m)$ is complete, there is necessarily a limit $u \in \mathcal{M}_{\mathcal{F}}^p(0,T;R^m)$ such that $u_k \to u$ strongly in $\mathcal{M}_{\mathcal{F}}^p(0,T;R^m)$. While $X^{0,x;u_k} - X^{0,x;u} = X^{0,x;u_k-u}$, we deduce from Lemma 2.2 that $\xi_k = X_T^{0,x;u_k} \to X_T^{0,x;u}$ strongly in $\mathcal{L}_{\mathcal{F}}^p(0,T;R^n)$. Therefore $\xi = X_T^{0,x;u}$. Since the set $U_{ad}^{0,p}$ is strongly closed in $\mathcal{M}_{\mathcal{F}}^p(0,T;R^m)$, we also have $u \in U_{ad}^{0,p}$. The proof is then complete. □

2.3. The risk functional. We make the following assumptions.

(A3)$_p$ Let $M : R^n \times \Omega \to R$ and $G : [0,T] \times R^n \times R^m \times \Omega \to R$ be convex in (x,u), and for each $(x,u) \in R^n \times R^m$, they are $\mathcal{F}_T$-measurable and $\{\mathcal{F}_t,\, 0 \le t \le T\}$-progressively measurable, respectively. Moreover, there are deterministic numbers $p > 1$ and $\delta_1 > 0$ such that for all $(t,x) \in [0,T] \times R^n$,

$$M(x) \le \delta_1(1 + |x|^p), \tag{2.7}$$

$$G(t,x,u) \le \delta_1(1 + |x|^p + |u|^{p\wedge 2}), \tag{2.8}$$

and

$$\begin{aligned} &|M(x_1) - M(x_2)| + |G(t,x_1,u_1) - G(t,x_2,u_2)| \\ &\quad \le \delta_1(1 + |x_1|^{p-1} + |x_2|^{p-1})|x_1 - x_2| \\ &\qquad + \delta_1(1 + |u_1|^{p\wedge 2-1} + |u_2|^{p\wedge 2-1})|u_1 - u_2|. \end{aligned} \tag{2.9}$$

(A4)$_p$ The coercivity condition with index $p > 1$ holds for the terminal cost function M: there are deterministic constants $\delta_2 > 0$ and $\delta_3 > 0$ such that

$$M(x) \ge -\delta_2 + \delta_3|x|^p. \tag{2.10}$$

There is a deterministic constant $\delta_4 > 0$ such that $G(t,x,u) \ge -\delta_4$ for all (t,x,u).

Note that $M(x) := |x|^p$ and $G(t,x,u) := |x|^p$ satisfy assumptions (A3)$_p$ and (A4)$_p$.

Consider the risk functional parameterized by the initial data $(t,x)\in[0,T]\times R^n$:

$$J(u;t,x):=E^{\mathcal{F}_t}M(X_T^{t,x;u})+E^{\mathcal{F}_t}\int_t^T G(s,X_s^{t,x;u},u_s)\,ds. \tag{2.11}$$

It has the following useful properties.

THEOREM 2.4. *Let assumptions* (A1), (A2), $(A3)_p$, *and* $(A4)_p$ *be satisfied for some* $p>1$. *Then, the risk functional* $J(\cdot;0,x)$ *defined on* $U_{ad}^{0,p}$ *is bounded from below, finite-valued, convex, continuous, and coercive, i.e.,*

$$J(u;0,x)\to+\infty \quad \text{if } ||u||_{\mathcal{M}_{\mathcal{F}}^p}\to+\infty. \tag{2.12}$$

Moreover, it has the following coercivity, growth, and continuity:

$$\begin{aligned}
&\beta_1(|x|^p+||u||_{\mathcal{M}_{\mathcal{F}}^p}^p)-\|f\|_{\mathcal{L}_{\mathcal{F}}^p}^p-\|g\|_{\mathcal{M}_{\mathcal{F}}^p}^p-\beta_2\le J(u;0,x)\\
&\quad\le\beta_3(1+|x|^p+||u||_{\mathcal{M}_{\mathcal{F}}^p}^{p\wedge 2}+\|f\|_{\mathcal{L}_{\mathcal{F}}^p}^p+\|g\|_{\mathcal{M}_{\mathcal{F}}^p}^p),\\
&|J(u_1;0,x_1)-J(u_2;0,x_2)|\\
&\quad\le\beta_4(1+|x_1|^{p-1}+|x_2|^{p-1}+||u_1||_{\mathcal{M}_{\mathcal{F}}^p}^{p-1}+||u_2||_{\mathcal{M}_{\mathcal{F}}^p}^{p-1}+\|f\|_{\mathcal{L}_{\mathcal{F}}^p}^{p-1}+\|g\|_{\mathcal{M}_{\mathcal{F}}^p}^{p-1})\\
&\quad\times(|x_1-x_2|+||u_1-u_2||_{\mathcal{M}_{\mathcal{F}}^p})\\
&\quad+\beta_4(1+\|u_1\|_{\mathcal{M}_{\mathcal{F}}^p}^{(p\wedge 2-1)}+\|u_2\|_{\mathcal{M}_{\mathcal{F}}^p}^{(p\wedge 2-1)})\|u_1-u_2\|_{\mathcal{M}_{\mathcal{F}}^p}.
\end{aligned} \tag{2.13}$$

Here β_1,β_2,β_3, *and* β_4 *are positive constants.*

Proof. The convexity of $J(\cdot;0,x)$ is obvious. The continuity and coercivity of $J(\cdot;0,x)$ can be derived from inequalities (2.13).

Inequalities (2.13) are consequences of the assumptions and an application of Hölder's inequality, Lemma 2.1, or Lemma 2.2. From the coercivity assumption on M, the lower bound assumption on G, and Lemma 2.2, we derive the first inequality of (2.13):

$$\begin{aligned}
J(u;0,x) &\ge \delta_3E|X_T|^p-\delta_2-\delta_4T\\
&\ge \delta_3\beta_{p,\delta,T}^{-1}(|x|^p+||u||_{\mathcal{M}_{\mathcal{F}}^p}^p)-\|f\|_{\mathcal{L}_{\mathcal{F}}^p}^p-\|g\|_{\mathcal{M}_{\mathcal{F}}^p}^p-\delta_2-\delta_4T.
\end{aligned} \tag{2.14}$$

From the continuity assumption on M and G, we have

$$\begin{aligned}
&|J(u_1;0,x_1)-J(u_2;0,x_2)|\\
&\quad\le\delta_1E[(1+|X_1(T)|^{p-1}+|X_2(T)|^{p-1})|X_1(T)-X_2(T)|]\\
&\qquad+\delta_1E\int_0^T[(1+|X_1|^{p-1}+|X_2|^{p-1})|X_1-X_2|\\
&\qquad\qquad+(1+|u_1|^{p\wedge 2-1}+|u_2|^{p\wedge 2-1})|u_1-u_2|]\,ds,
\end{aligned} \tag{2.15}$$

where $X_1(t)=X_t^{0,x_1;u_1}$ and $X_2(t)=X_t^{0,x_2;u_2}$. In view of the Schwarz inequality, we have (note that $p^{-1}+q^{-1}=1$)

$$\begin{aligned}
&|J(u_1;0,x_1)-J(u_2;0,x_2)|\\
&\quad\le\delta_1\|1+|X_1(T)|^{p-1}+|X_2(T)|^{p-1}\|_{L^q}\|X_1(T)-X_2(T)\|_{L^p}\\
&\qquad+\delta_1\|1+|X_1|^{p-1}+|X_2|^{p-1}\|_{\mathcal{L}_{\mathcal{F}}^q}\|X_1-X_2\|_{\mathcal{L}_{\mathcal{F}}^p}\\
&\qquad+\delta_1E\left\{\left(\int_0^T(1+|u_1|^{p\wedge 2-1}+|u_2|^{p\wedge 2-1})\,ds\right)^{1/2}\left(\int_0^T|u_1-u_2|\,ds\right)^{1/2}\right\}
\end{aligned}$$

$$
\begin{aligned}
&\le \delta_1(1+\||X_1(T)|^{p-1}\|_{L^q}+\||X_2(T)|^{p-1}\|_{L^q})\|X_1(T)-X_2(T)\|_{L^p}\\
&\quad+\beta_4(1+\||X_1|^{p-1}\|_{\mathcal{L}^q_{\mathcal{F}}}+\||X_2|^{p-1}\|_{\mathcal{L}^q_{\mathcal{F}}})\|X_1-X_2\|_{\mathcal{L}^p_{\mathcal{F}}}\\
&\quad+\beta_4 E\left[1+\left(\int_0^T|u_1|^{2(p\wedge 2-1)}\,ds\right)^{1/2}+\left(\int_0^T|u_2|^{2(p\wedge 2-1)}\,ds\right)^{1/2}\right]\\
&\quad\times\left(\int_0^T|u_1-u_2|^2\,ds\right)^{1/2}\\
&\le \delta_1(1+\|X_1(T)\|_{L^p}^{p-1}+\|X_2(T)\|_{L^p}^{p-1})\|X_1(T)-X_2(T)\|_{L^p}\\
&\quad+\beta_4(1+\|X_1\|_{\mathcal{L}^p_{\mathcal{F}}}^{p-1}+\|X_2\|_{\mathcal{L}^p_{\mathcal{F}}}^{p-1})\|X_1-X_2\|_{\mathcal{L}^p_{\mathcal{F}}}\\
&\quad+\beta_4 E\left[1+\left(\int_0^T|u_1|^2\,ds\right)^{(p\wedge 2-1)/2}+\left(\int_0^T|u_2|^2\,ds\right)^{(p\wedge 2-1)/2}\right]\\
&\quad\times\left(\int_0^T|u_1-u_2|^2\,ds\right)^{1/2}.
\end{aligned}
\tag{2.16}
$$

Since for $i=1,2$

$$
\begin{aligned}
&E\left\{\left(\int_0^T|u_i|^2\,ds\right)^{(p\wedge 2-1)/2}\left(\int_0^T|u_1-u_2|^2\,ds\right)^{1/2}\right\}\\
&\le\left(E\left(\int_0^T|u_i|^2\,ds\right)^{q(p\wedge 2-1)/2}\right)^{1/q}\left(E\left(\int_0^T|u_1-u_2|^2\,ds\right)^{p/2}\right)^{1/p}\\
&\le\|u_i\|^{(p\wedge 2-1)}_{\mathcal{M}^{q(p\wedge 2-1)}_{\mathcal{F}}}\|u_1-u_2\|_{\mathcal{M}^p_{\mathcal{F}}}\\
&\le\|u_i\|^{(p\wedge 2-1)}_{\mathcal{M}^p_{\mathcal{F}}}\|u_1-u_2\|_{\mathcal{M}^p_{\mathcal{F}}}\quad\text{(noting that } q(p\wedge 2-1)\le p),
\end{aligned}
\tag{2.17}
$$

we have

$$
\begin{aligned}
&|J(u_1;0,x_1)-J(u_2;0,x_2)|\\
&\quad\le\delta_1(1+\|X_1(T)\|_{L^p}^{p-1}+\|X_2(T)\|_{L^p}^{p-1})\|X_1(T)-X_2(T)\|_{L^p}\\
&\qquad+\beta_4(1+\|X_1\|_{\mathcal{S}^p_{\mathcal{F}}}^{p-1}+\|X_2\|_{\mathcal{S}^p_{\mathcal{F}}}^{p-1})\|X_1-X_2\|_{\mathcal{L}^p_{\mathcal{F}}}\\
&\qquad+\beta_4(1+\|u_1\|^{(p\wedge 2-1)}_{\mathcal{M}^p_{\mathcal{F}}}+\|u_2\|^{(p\wedge 2-1)}_{\mathcal{M}^p_{\mathcal{F}}})\|u_1-u_2\|_{\mathcal{M}^p_{\mathcal{F}}}.
\end{aligned}
\tag{2.18}
$$

Using Lemma 2.1, we get the last inequality of (2.13).

In a similar way, the second inequality of (2.13) can be proved. The proof is complete. □

2.4. Minimization of the risk. For given initial data $(0,x)$, consider the risk minimization problem:

$$
\min_{u\in U^0_{ad}} J(u;0,x). \tag{2.19}
$$

We have immediately the following theorem.

THEOREM 2.5. *Let assumptions* (A1), (A2), *and* $(\mathrm{A4})_p$ *for some* $p>1$ *be satisfied. Then, problem* (2.19) *is equivalent to the following:*

$$
\min_{u\in U^{0,p}_{ad}} J(u;0,x). \tag{2.20}
$$

THEOREM 2.6. *Let assumptions* (A1), (A2), $(A3)_p$, *and* $(A4)_p$ *be satisfied for some* $p > 1$. *Then, the risk minimization problem* (2.19) *has an optimal control* $u \in U_{ad}^{0,p}$ *for each* $x \in R^n$. *Furthermore, if* M *is strictly convex, then the problem has a unique optimal control* $u \in U_{ad}^{0,p}$ *for each* $x \in R^n$.

Proof. From Theorem 2.4 and Proposition 2.12 of Ekeland and Temam [13], the convex, coercive, and continuous (of course lower-semicontinuous) functional $J(\cdot; 0, x)$ defined on the strongly closed subset $U_{ad}^{0,p} \subset \mathcal{M}_{\mathcal{F}}^p(0, T; R^m)$ (note that $\mathcal{M}_{\mathcal{F}}^p(0, T; R^m)$ is a reflexive Banach space) attains its infimum.

It is easy to show that the strict convexity of M implies the uniqueness of the terminal optimal control state and thus the optimal control (in view of Lemma 2.2). The proof is complete. □

3. The quadratic case: Singular stochastic LQ problem.

3.1. Introduction. This section is devoted to the study of the singular stochastic LQ problem, as a particular case of the stochastic control problem discussed in the preceding section. That is, we consider the case in which the risk functions M and G are both quadratic functions, i.e.,

$$M(x) := \langle M(x - \xi), x - \xi \rangle, \quad G(t, x, u) = \langle N_t u, u \rangle + \langle Q_t(x - q_t), x - q_t \rangle. \tag{3.1}$$

We introduce the following assumptions.

(A5) The matrix-valued processes N, Q, M are nonnegative, uniformly bounded in (t, ω), and $\{\mathcal{F}_t, 0 \le t \le T\}$-progressively measurable, of dimensions $m \times m$, $n \times n$, $n \times n$, respectively. The admissible control class is $U_{ad} := \mathcal{L}_{\mathcal{F}}^2(0, T; R^m)$.

(A6) The control weighting process N is uniformly positive: $N \ge \delta I_{m \times m}$ for some positive constant δ.

(A6)′ The terminal state weighting process M is uniformly positive: $M \ge \delta I_{n \times n}$ for some positive constant δ.

(A7) The random variable $\xi \in L^2(\Omega, \mathcal{F}_T, P; R^n)$, and the process $q \in \mathcal{L}_{\mathcal{F}}^2(0, T; R^n)$.

For subsequent convenience of reference, we give the following definitions.

DEFINITION 3.1. *The underlying optimal control problem is called regular if assumption* (A6) *is required. It will be called singular if assumption* (A6)′ *is required.*

From Lemma 2.2, we get the following lemma.

LEMMA 3.2. *Let assumptions* (A1), (A2), (A5), (A6)′, *and* (A7) *be satisfied. Then, for every* $u \in \mathcal{L}_{\mathcal{F}}^2(0, T; R^m)$, *we have*

$$\begin{aligned} J(u; t, x) \quad \ge \quad & \frac{\delta}{2}[\beta_{\delta,T}^{-1}|x|^2 + \beta_{\delta,T}^{-1} E^{\mathcal{F}_t} \int_t^T |u_s|^2 \, ds - E^{\mathcal{F}_t} \int_t^T (|f_s|^2 + |g(s)|^2) \, ds] \\ & + (\delta - \frac{2}{\delta}) E|\xi|^2. \end{aligned} \tag{3.2}$$

Define the value function $V(t, x)$ of the above LQ problem as follows:

$$V(t, x) := \operatorname{ess} \inf_{u \in \mathcal{L}_{\mathcal{F}}^2(t, T; R^m)} J(u; t, x). \tag{3.3}$$

See Karatzas and Shreve [16] for the definition of infimum of a set of stochastic processes. From the definition, it follows that for each fixed $x \in R^n$, $V(t, x)$ is an $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted real-valued stochastic process.

The rest of this section is organized as follows. The second subsection is devoted to existence and equivalent conditions of optimal controls, and a useful estimate, which states a dependence property of the optimal control and the pair of associated adjoint

processes on the 5-tuple (x, f, g, ξ, q). The third contains some useful properties of BMO-martingales. In the fourth, the LQ problem is explicitly solved in terms of the solutions of two associated BSDEs. Finally, in the fifth, the two BSDEs are studied.

3.2. Optimal controls: Existence, equivalent conditions, and dependence. A general existence and uniqueness of optimal controls was given in the regular case by Bismut [4, 6]. Concerning the singular case, Theorem 2.6 immediately implies the following theorem.

THEOREM 3.3. *Let assumptions* (A1), (A2), (A5), (A6)′, *and* (A7) *be satisfied. Then, there is a unique optimal control* $\widehat{u} \in \mathcal{L}^2_{\mathcal{F}}(0,T;R^m)$.

Proceeding identically as Bismut [4, 6], we can prove the following theorem.

THEOREM 3.4. *Let assumptions* (A1), (A2), (A5), (A6)′, *and* (A7) *be satisfied. Then* $\widehat{u} \in \mathcal{L}^2_{\mathcal{F}}(0,T;R^m)$ *is optimal if and only if there is a pair of processes* $(y, z := (z_1, \cdots, z_d))$ *which solves the BSDE (called the adjoint equation)*

$$(3.4)\quad \begin{cases} dy_t &= -\left[A_t^* y_t + \sum_{i=1}^d C_i(t)^* z_i(t) + Q_t(\widehat{X}_t - q_t)\right] dt + \sum_{i=1}^d z_i(t) dw_i(t), \\ y_T &= M(\widehat{X}_T - \xi) \end{cases}$$

with $\widehat{X} := X^{0,x;\widehat{u}}$ *and*

$$(3.5)\quad y \in \mathcal{L}^2_{\mathcal{F}}(0,T;R^n) \cap L^2(\Omega, \mathcal{F}_T, P; C([0,T];R^n)),\ z \in (\mathcal{L}^2_{\mathcal{F}}(0,T;R^n))^d,$$

such that

$$(3.6)\quad y_s^* B_s + \sum_{i=1}^d z_i(s)^* D_i(s) + N_s \widehat{u}_s = 0 \quad a.s.a.e.$$

Also see Peng [29] and Tang and Li [34] for the necessity.

We have the following useful estimate.

THEOREM 3.5. *Let assumptions* (A1), (A2), (A5), *and* (A6)′ *be satisfied. Let* $\widehat{u}$ *be an optimal control and* (y, z) *be defined by* (3.4). *Then, there is a positive constant* β *such that*

$$(3.7)\quad \begin{aligned} &E \sup_{0\le t\le T} (|\widehat{X}_t|^2 + |y_t|^2) + E\int_0^T (|z(t)|^2 + |\widehat{u}_t|^2)\, dt \\ &\quad \le \beta \left[|x|^2 + E|\xi|^2 + E\int_0^T (|f_t|^2 + |g(t)|^2 + |q_t|^2)\, dt \right]. \end{aligned}$$

Proof. According to the classical a priori estimate for BSDE (3.4) (see El Karoui, Peng, and Quenez [14]), we have

$$(3.8)\quad \begin{aligned} &E \sup_{0\le t\le T} |y_t|^2 + E\int_0^T |z(t)|^2\, dt \\ &\quad \le \beta \left[E(|\widehat{X}_T|^2 + |\xi|^2) + E\int_0^T (|f_t|^2 + |g(t)|^2 + |\widehat{X}_t|^2 + |q_t|^2)\, dt \right]. \end{aligned}$$

Using Itô's formula to compute $\langle y_t, \widehat{X}_t\rangle$, in view of (3.6), we obtain

$$(3.9)\quad \begin{aligned} &E\langle M(\widehat{X}_T - \xi), \widehat{X}_T\rangle + E\int_0^T [\langle Q_t(\widehat{X}_t - q_t), \widehat{X}_t\rangle + \langle N_t \widehat{u}_t, \widehat{u}_t\rangle]\, dt \\ &\quad = E\int_0^T \left[\langle y_t, f_t\rangle + \sum_{i=1}^d \langle z_i(t), g_i(t)\rangle \right] dt + \langle y_0, x\rangle. \end{aligned}$$

Therefore, in view of assumptions (A6)′ and (A5),

$$\begin{aligned}\delta E|\widehat{X}_T|^2 &\le E\langle M\widehat{X}_T,\widehat{X}_T\rangle\\ &\le E\langle M\xi,\widehat{X}_T\rangle + E\int_0^T \langle Q_t q_t,\widehat{X}_t\rangle\,dt\\ &\quad + E\int_0^T \left[\langle y_t,f_t\rangle + \sum_{i=1}^d \langle z_i(t),g_i(t)\rangle\right] dt + \langle y_0,x\rangle.\end{aligned} \tag{3.10}$$

Using Cauchy's inequality, in view of (3.8), we have

$$\begin{aligned}\delta E|\widehat{X}_T|^2 &\le \frac{1}{2\varepsilon}E|M\xi|^2 + \frac{\varepsilon}{2}E|\widehat{X}_T|^2 + E\int_0^T \left(\frac{1}{2\varepsilon}|Q_t q_t|^2 + \frac{\varepsilon}{2}|\widehat{X}_t|^2\right) dt\\ &\quad + E\int_0^T \left[\frac{\varepsilon}{2}|y_t|^2 + \frac{1}{2\varepsilon}|f_t|^2 + \frac{\varepsilon}{2}|z(t)|^2 + \frac{1}{2\varepsilon}|g(t)|^2\right] dt + \frac{\varepsilon}{2}|y_0|^2 + \frac{1}{2\varepsilon}|x|^2\\ &\le \beta_\varepsilon\left[|x|^2 + E|\xi|^2 + E\int_0^T (|f_t|^2 + |g(t)|^2 + |q_t|^2)\,dt\right]\\ &\quad + \frac{\varepsilon}{2}(\beta T + \beta + 1)\left[E|\widehat{X}_T|^2 + E\int_0^T |\widehat{X}_t|^2\,dt\right]\end{aligned} \tag{3.11}$$

for arbitrarily small positive number ε and a positive real number β_ε. In view of Lemma 2.2, we conclude from the last inequality that for sufficiently small $\varepsilon > 0$ and another positive real number β'_ε

$$E|\widehat{X}_T|^2 \le \beta'_\varepsilon\left[|x|^2 + E|\xi|^2 + E\int_0^T (|f_t|^2 + |g(t)|^2 + |q_t|^2)\,dt\right]. \tag{3.12}$$

Again noting Lemma 2.2 and the estimate (3.8), we get the desired estimate (3.7). □

3.3. The BMO-martingale. In this subsection, we prove some properties for a BMO-martingale, which will be used in the next two subsections.

DEFINITION 3.6. *A $\{\mathcal{F}_t, 0 \le t \le T\}$-martingale $\{\mathcal{M}_t, 0 \le t \le T\}$ is called a BMO-martingale if there is a deterministic positive constant β such that for every stopping time $\tau \le T$,*

$$E^{\mathcal{F}_\tau}|\mathcal{M}_T - \mathcal{M}_\tau|^2 \le \beta \quad a.s. \tag{3.13}$$

A BMO-martingale has the following nice properties, which will be used in the next two subsections.

LEMMA 3.7. *Assume that $\int_0^\cdot \langle Y, dw\rangle \in L^2(\Omega,\mathcal{F}_T,P)$ is a BMO-martingale. Then, we have the following:*

(i) *If $X_1(t) = X_1(0) + \int_0^t x_1(s)\,ds$, where x_1 is $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted such that $\int_0^\cdot |x_1|\,ds \in L^2(\Omega,\mathcal{F}_T,P)$, then $YX_1 \in \mathcal{L}^2_{\mathcal{F}}(0,T;R^d)$. Moreover, there is a positive constant β_Y depending on Y while independent of X_1 such that*

$$E\int_0^T |YX_1|^2\,ds \le \beta_Y\left(E|X_1(0)|^2 + E|\int_0^T |x_1|\,ds|^2\right). \tag{3.14}$$

(ii) *If* $X_2(t) = X_2(0) + \int_0^t \langle x_2(s), dw(s)\rangle$, *where* $x_2 \in \mathcal{L}^2_{\mathcal{F}}(0,T;R^d)$, *then* $X_2 Y \in \mathcal{L}^2_{\mathcal{F}}(0,T;R^d)$. *Moreover, there is a positive constant* β_Y *depending on* Y *while independent of* X_2 *such that*

$$E\int_0^T |YX_2|^2\,ds \le \beta_Y\left(E|X_2(0)|^2 + E\int_0^T |x_2|^2\,ds\right). \tag{3.15}$$

(iii) *If* $X(t) := X(0) + \int_0^t x_1(s)\,ds + \int_0^t \langle x_2(s), dw(s)\rangle$ *such that* x_1 *is* $\{\mathcal{F}_t, 0 \le t \le T\}$*-adapted satisfying* $\int_0^{\cdot} |x_1|\,ds \in L^2(\Omega,\mathcal{F}_T,P)$ *and* $x_2 \in \mathcal{L}^2_{\mathcal{F}}(0,T;R^d)$, *then* $YX \in \mathcal{L}^2_{\mathcal{F}}(0,T;R^d)$. *Moreover, there is a positive constant* β_Y *depending on* Y *while independent of* X *such that*

$$E\int_0^T |YX|^2\,ds \le \beta_Y\left(E|X(0)|^2 + E\int_0^T |x_2|^2\,ds + E\left|\int_0^T |x_1|\,ds\right|^2\right). \tag{3.16}$$

Proof. Assertion (iii) is a combination of assertions (i) and (ii). It follows directly from Bañuelos and Bennett [1, Theorem 1.1(i)] that $\int_0^T X_2(t)\langle Y_t, dw(t)\rangle$ is square-integrable and assertion (ii) then follows. It remains to prove assertion (i).

Define for $k = 1, 2, \ldots,$

$$\tau_k = \inf\left\{s \in [0,T] : \int_0^s |x_1(r)|\,dr \ge k\right\}$$

with $\tau_k = \infty$ if the underlying set is empty. Then, $\tau_k, k = 1, 2, \ldots,$ are stopping times. Denote by χ an indicator function. We have $YX_1\chi_{[0,\tau_k]} \in \mathcal{L}^2_{\mathcal{F}}(0,T;R^d)$ for $k = 1, 2, \ldots.$ In view of Fatou's lemma, it is sufficient to prove that the quantity $E\int_0^T |YX_1|^2\chi_{[0,\tau_k]}\,ds = E\int_0^{T_k} |YX_1|^2\,ds$ is uniformly bounded with respect to k with $T_k := T \wedge \tau_k$.

For all $a \in \mathcal{L}^2_{\mathcal{F}}(0,T;R^d)$, since $\int_0^{\cdot} |\langle Y, a\rangle|\,ds$ is the quadratic variation of the BMO-martingale $\int_0^{\cdot}\langle Y, dw\rangle$ and the square-integrable martingale $\int_0^{\cdot} \operatorname{sign}(\langle Y, a\rangle)\langle a, dw\rangle$, we have from Bañuelos and Bennett [1, Theorem 1.1(iii)] that $\int_0^T |\langle Y, a\rangle|\,ds \in L^2(\Omega,\mathcal{F}_T,P)$ and, moreover,

$$\left\|\int_0^T |\langle Y, a\rangle|\,ds\right\|_{L^2(\Omega,\mathcal{F}_T,P)} \le \beta\|a\|_{\mathcal{L}^2_{\mathcal{F}}(0,T;R^d)}. \tag{3.17}$$

Therefore,

$$\begin{aligned}
E\int_0^{T_k} \langle X_1 Y, a\rangle\,ds &= E\left\{\int_0^{T_k} X_1(t)\,d\int_0^t \langle Y_s, a_s\rangle\,ds\right\}\\
&= E\left\{X_1(T_k)\int_0^{T_k}\langle Y_s, a_s\rangle\,ds\right\} + E\left\{\int_0^{T_k} x_1(t)\int_0^t \langle Y_s, a_s\rangle\,ds\,dt\right\}\\
&\le E\left\{\left(|X_1(0)| + \int_0^T |x_1|\,ds\right)\int_0^T |\langle Y_s, a_s\rangle|\,ds\right\}\\
&\quad + E\left\{\int_0^T |\langle Y_s, a_s\rangle|\,ds\int_0^T |x_1|\,dt\right\}.
\end{aligned} \tag{3.18}$$

Using the Cauchy–Schwarz inequality, we have

$$
\begin{aligned}
&E\int_0^{T_k}\langle X_1 Y, a\rangle\,ds\\
&\quad\le \left\|\left(|X_1(0)|+\int_0^T |x_1|\,ds\right)\right\|_{L^2(\Omega,\mathcal{F}_T,P)} \left\|\int_0^T |\langle Y_s,a_s\rangle|\,ds\right\|_{L^2(\Omega,\mathcal{F}_T,P)}\\
&\qquad+\left\|\int_0^T |\langle Y_s,a_s\rangle|\,ds\right\|_{L^2(\Omega,\mathcal{F}_T,P)} \left\|\int_0^T |x_1|\,dt\right\|_{L^2(\Omega,\mathcal{F}_T,P)}\\
&\quad\le 2\left\|\left(|X_1(0)|+\int_0^T |x_1|\,ds\right)\right\|_{L^2(\Omega,\mathcal{F}_T,P)} \beta\|a\|_{\mathcal{L}^2_{\mathcal{F}}(0,T;R^d)}.
\end{aligned}
\tag{3.19}
$$

This implies that

$$
\left(E\int_0^{T_k}|YX_1|^2\,ds\right)^{1/2}\le 2\beta\left\|\left(|X_1(0)|+\int_0^T|x_1|\,ds\right)\right\|_{L^2(\Omega,\mathcal{F}_T,P)}. \qquad \square \tag{3.20}
$$

We refer the reader to Kazamaki [17] for further information on BMO-martingales.

3.4. Connection with two BSDEs. Define $\Gamma : [0,T]\times \mathcal{S}^n_+ \times R^{n\times d}\to R^{m\times n}$ by

$$
\Gamma(\cdot,S,L)=-\left(N+\sum_{i=1}^d D_i^*SD_i\right)^{-1}\left(B^*S+\sum_{i=1}^d D_i^*SC_i+\sum_{i=1}^d D_i^*L_i\right) \tag{3.21}
$$

and

$$
\widehat{A}:=A+B\Gamma(\cdot,K,L),\quad \widehat{C}_i:=C_i+D_i\Gamma(\cdot,K,L),\quad i=1,\dots,d. \tag{3.22}
$$

Observe that the function Γ also depends upon the coefficients B, C, D, and N. Occasionally in the following, we shall write $\Gamma(t,S,L;N)$ to indicate the dependence on N.

The optimal control and the value function of the LQ problem will be explicitly connected to the two BSDEs:

$$
\begin{cases}
dK_t=-\Bigg[A^*K_t+K_tA+\displaystyle\sum_{i=1}^d C_i^*K_tC_i+Q\\
\qquad\qquad+\displaystyle\sum_{i=1}^d (C_i^*L_i(t)+L_i(t)C_i)\\
\qquad\qquad-\Gamma(t,K_t,L(t))\left(N+\displaystyle\sum_{i=1}^d D_i^*K_tD_i\right)\Gamma(t,K_t,L(t))\Bigg)^*\Bigg]\,dt\\
\qquad\qquad+\displaystyle\sum_{i=1}^d L_i(t)\,dw_i(t),\qquad 0\le t<T,\\
K_T=M
\end{cases}
\tag{3.23}
$$

and

$$\begin{cases} d\psi_t = -\left[\widehat{A}^*\psi_t + \sum_{i=1}^d \widehat{C}_i^*(\phi_i - Kg_i) \right. \\ \qquad\qquad \left. - Kf - \sum_{i=1}^d L_i g_i + Qq\right] dt + \sum_{i=1}^d \phi_i\, dw_i, \\ \psi_T = M\xi, \end{cases} \tag{3.24}$$

where (K, L) is an $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted solution of BSRDE (3.23). More precisely, we have the following theorem.

THEOREM 3.8. (i) *Let assumptions* (A1), (A2), (A5), (A6)′, *and* (A7) *be satisfied.* (ii) *Let* (K, L) *be an* $\{\mathcal{F}_t, 0 \le t \le T\}$*-adapted solution of BSRDE* (3.23) *with* $K \in \mathcal{L}^\infty_{\mathcal{F}}(0, T; S^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0, T]; S^n_+))$, $L \in (\mathcal{L}^2_{\mathcal{F}}(0, T; S^n))^d$, *and* K *being uniformly positive.* (iii) *Let* (ψ, ϕ) *be an* $\{\mathcal{F}_t, 0 \le t \le T\}$*-adapted solution of BSDE* (3.24) *such that* $E \sup_{0\le t\le T} |\psi_t|^2 < \infty$ *and* $\phi \in (\mathcal{L}^2_{\mathcal{F}}(0, T; R^n))^d$.

Then the optimal control $\widehat{u}$ *exists uniquely with the feedback law*

$$\widehat{u} = \Gamma(\cdot; K, L)\widehat{X} + u^0, \tag{3.25}$$

where $\widehat{X} := X^{t,x;\widehat{u}}$ *and the functions* Γ *and* u^0 *are defined, respectively, by* (3.21) *and*

$$u^0 := \left(N + \sum_{i=1}^d D_i^* K D_i\right)^{-1} \left[B^*\psi + \sum_{i=1}^d D_i^*(\phi_i - Kg_i)\right], \quad t \le s \le T. \tag{3.26}$$

The value function $V(t, x), (t, x) \in [0, T] \times R^n$ *is given by*

$$V(t, x) = \langle K_t x, x\rangle - 2\langle \psi_t, x\rangle + V^0(t), \quad (t, x) \in [0, T] \times R^n, \tag{3.27}$$

where $V^0(t)$ *is defined by*

$$\begin{aligned} V^0(t) := {} & E^{\mathcal{F}_t}\langle M\xi, \xi\rangle + E^{\mathcal{F}_t}\int_t^T \langle Qq, q\rangle\, ds - 2E^{\mathcal{F}_t}\int_t^T \langle \psi, f\rangle\, ds \\ & + E^{\mathcal{F}_t}\int_t^T \sum_{i=1}^d [\langle Kg_i, g_i\rangle - 2\langle \phi_i, g_i\rangle]\, ds \\ & - E^{\mathcal{F}_t}\int_t^T \left\langle \left(N + \sum_{i=1}^d D_i^* K D_i\right) u^0, u^0 \right\rangle ds. \end{aligned} \tag{3.28}$$

Note that in view of the assumptions on the integrability of (K, L) and (ψ, ϕ), all the integrals appearing in the formulas are well defined.

Our proof of Theorem 3.8 requires the following BMO-property for BSRDE (3.23).

THEOREM 3.9. *Let assumptions* (A1) *and* (A5) *be satisfied. If* (K, L) *is an* $\{\mathcal{F}_t, 0 \le t \le T\}$*-adapted solution of BSRDE* (3.23) *such that* $K \in \mathcal{L}^\infty_{\mathcal{F}}(0, T; S^n_+)$, $L := (L_1, \dots, L_d) \in (\mathcal{L}^2_{\mathcal{F}}(0, T; S^n))^d$, *and* $(N + \sum_{i=1}^d D_i^* K D_i)$ *is uniformly positive, then* $\int_0^\cdot \sum_{i=1}^d L_i(s)\, dw_i(s)$ *is a BMO-martingale.*

The proof can be reduced to proving the following fact: there is a deterministic positive real number β such that for any stopping time τ

$$E^{\mathcal{F}_\tau}\int_\tau^T |L_s|^2\, ds \le \beta. \tag{3.29}$$

It is an a priori estimate for the second component L of any $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted solution (K, L) of BSRDE (3.23), but it involves stopping times. The reader is referred to Kohlmann and Tang [21] for further details of the proof.

In general, the martingale $\mathcal{M} =: \int_0^{\cdot} \sum_{i=1}^d L_i(s)\, dw_i(s)$ and L are not necessarily bounded. However, Theorem 3.9 states that any moment of the absolute increment $|\mathcal{M}_T - \mathcal{M}_\tau|$ has a bounded expectation conditioned on $\mathcal{F}_\tau$ uniformly with respect to any stopping time τ. Roughly speaking, Theorem 3.9 implies that the integrand L of the stochastic integral $\mathcal{M}$ is bounded in a sense of some mean values.

The following lemma will also be used in our proof of Theorem 3.8.

LEMMA 3.10. *Let all the assumptions of Theorem* 3.8 *be satisfied. Let u be an $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted m-dimensional process such that $\int_0^T |u_s|^2\, ds < \infty$ a.s. and $X := X^{t,x;u}$. Define for $k = 1, 2, \ldots,$*

$$\tau_k = \inf\left\{ s \in [t,T] : \int_t^s |u_r|^2\, dr \ge k, \int_t^s |L(r)|^2\, dr \ge k, |X_s| \ge k, |\psi_s| \ge k \right\}$$

and $\tau_k = \infty$ if the underlying set is empty. Also define

$$T_k := T \wedge \tau_k, \quad k = 1, 2, \ldots.$$

Then,

$$\begin{aligned}
&E^{\mathcal{F}_t}[\langle K_{T_k} X_{T_k}, X_{T_k}\rangle - 2\langle \psi_{T_k}, X_{T_k}\rangle + \langle M\xi, \xi\rangle] \\
&\quad + E^{\mathcal{F}_t}\left[\int_t^{T_k} \langle Q(X-q), X-q\rangle\, ds + \int_t^{T_k} \langle Nu, u\rangle\, ds\right] \\
&= \langle K_t x, x\rangle - 2\langle \psi_t, x\rangle + E^{\mathcal{F}_t}\left[\langle M\xi, \xi\rangle + \int_t^{T_k} \langle Qq, q\rangle\, ds\right] \\
&\quad - 2E^{\mathcal{F}_t}\int_t^{T_k} \langle \psi, f\rangle\, ds + E^{\mathcal{F}_t}\int_t^{T_k} \sum_{i=1}^d [\langle K g_i, g_i\rangle - 2\langle \phi_i, g_i\rangle]\, ds \\
&\quad + E^{\mathcal{F}_t}\int_t^{T_k} \left\langle \left(N + \sum_{i=1}^d D_i^* K D_i\right)(\widetilde{u} - u^0), \widetilde{u} - u^0 \right\rangle ds \\
&\quad - E^{\mathcal{F}_t}\int_t^{T_k} \left\langle \left(N + \sum_{i=1}^d D_i^* K D_i\right) u^0, u^0 \right\rangle ds.
\end{aligned} \tag{3.30}$$

The proof is a straightforward adaptation of Yong and Zhou [35, pp. 315–317] in the case of constant coefficients to our case.

Now let us prove Theorem 3.8.

Proof. In what follows, we shall show that the control $\widehat{u}$ given by (3.25) is square-integrable.

First, it is well defined. To see this, putting it into system (2.1), we get

$$\begin{cases} d\widehat{X} = (\widehat{A}\widehat{X} + Bu^0 + f)\, ds + \sum_{i=1}^d (\widehat{C}_i \widehat{X} + D_i u^0 + g_i)\, dw_i, \ t < s \le T, \\ \widehat{X}_t = x. \end{cases} \tag{3.31}$$

Note that u^0, f, g are all square-integrable. The system (3.31) has a unique strong solution (see Gal'chuk [15]), simply denoted by $\widehat{X}$. Therefore, $\widehat{u}$ is well defined. However, it is not yet clear that it is square-integrable.

We now show that $\widehat{u}$ is square-integrable. Proceeding identically as in Lemma 3.10, we define the stopping times T_k, which are associated with $\widehat{u}$, and obtain the following:

$$\begin{aligned}
&E^{\mathcal{F}_t}[\langle K_{T_k}\widehat{X}_{T_k},\widehat{X}_{T_k}\rangle - 2\langle\psi_{T_k},\widehat{X}_{T_k}\rangle + \langle M\xi,\xi\rangle] \\
&+ E^{\mathcal{F}_t}\left[\int_t^{T_k}\langle Q(\widehat{X}-q),\widehat{X}-q\rangle\,ds + \int_t^{T_k}\langle N\widehat{u},\widehat{u}\rangle\,ds\right] \\
&= \langle K_t x,x\rangle - 2\langle\psi_t,x\rangle + E^{\mathcal{F}_t}\left[\langle M\xi,\xi\rangle + \int_t^{T_k}\langle Qq,q\rangle\,ds\right] \\
&- 2E^{\mathcal{F}_t}\int_t^{T_k}\langle\psi,f\rangle\,ds + E^{\mathcal{F}_t}\int_t^{T_k}\sum_{i=1}^{d}[\langle Kg_i,g_i\rangle - 2\langle\phi_i,g_i\rangle]\,ds \\
&- E^{\mathcal{F}_t}\int_t^{T_k}\left\langle\left(N+\sum_{i=1}^{d}D_i^*KD_i\right)u^0,u^0\right\rangle\,ds.
\end{aligned} \tag{3.32}$$

This equality is the basis of the subsequent analysis.

Note that K is uniformly positive, i.e., there is a deterministic positive real number γ such that $K \geq \gamma I_{n\times n}$ a.s.a.e. Therefore, we have

$$\begin{aligned}
&\langle K_{T_k}\widehat{X}_{T_k},\widehat{X}_{T_k}\rangle - 2\langle\psi_{T_k},\widehat{X}_{T_k}\rangle \\
&\geq \gamma|\widehat{X}_{T_k}|^2 - \left(\frac{\gamma}{2}|\widehat{X}_{T_k}|^2 + \frac{2}{\gamma}|\psi_{T_k}|^2\right) \\
&= \frac{\gamma}{2}|\widehat{X}_{T_k}|^2 - \frac{2}{\gamma}|\psi_{T_k}|^2.
\end{aligned} \tag{3.33}$$

Noting the equality (3.32) and the fact that $Q \geq 0$ and $N \geq 0$, we get

$$\begin{aligned}
&E^{\mathcal{F}_t}\left[\frac{\gamma}{2}|\widehat{X}_{T_k}|^2 - \frac{2}{\gamma}|\psi_{T_k}|^2\right] \\
&\leq \langle K_t x,x\rangle - 2\langle\psi_t,x\rangle + E^{\mathcal{F}_t}\int_t^{T_k}\langle Qq,q\rangle\,ds \\
&-2E^{\mathcal{F}_t}\int_t^{T_k}\langle\psi,f\rangle\,ds + E^{\mathcal{F}_t}\int_t^{T_k}\sum_{i=1}^{d}[\langle Kg_i,g_i\rangle - 2\langle\phi_i,g_i\rangle]\,ds.
\end{aligned} \tag{3.34}$$

Applying Lemma 2.2 for $p=2$, we have

$$\begin{aligned}
&E^{\mathcal{F}_t}\int_t^{T_k}|\widehat{u}_s|^2\,ds + E^{\mathcal{F}_t}\sup_{t\leq s\leq T_k}|\widehat{X}_s|^2 \\
&\leq \beta_{\delta,T}E^{\mathcal{F}_t}\left\{|\widehat{X}_{T_k}|^2 + \int_t^{T_k}|f_s|^2\,ds + \int_t^{T_k}|g(s)|^2\,ds\right\} \\
&\leq \frac{2\beta_{\delta,T}}{\gamma}E^{\mathcal{F}_t}\left\{\frac{2}{\gamma}|\psi_{T_k}|^2 + \langle K_t x,x\rangle - 2\langle\psi_t,x\rangle + \int_t^{T_k}\langle Qq,q\rangle\,ds\right\} \\
&-\frac{2\beta_{\delta,T}}{\gamma}E^{\mathcal{F}_t}\left\{2\int_t^{T_k}\langle\psi,f\rangle\,ds - \int_t^{T_k}\sum_{i=1}^{d}[\langle Kg_i,g_i\rangle - 2\langle\phi_i,g_i\rangle]\,ds\right\} \\
&+\beta_{\delta,T}E^{\mathcal{F}_t}\left\{\int_t^{T_k}|f_s|^2\,ds + \int_t^{T_k}|g(s)|^2\,ds\right\}.
\end{aligned} \tag{3.35}$$

Noting that

$$\lim_{k\to\infty}\tau_k=+\infty,\quad \lim_{k\to\infty}T_k=T\quad \text{a.s.},\tag{3.36}$$

and the fact that ψ is uniformly integrable, passing to the limit and applying Fatou's lemma, we have

$$\begin{aligned}
&E^{\mathcal{F}_t}\int_t^T|\widehat{u}_s|^2\,ds+E^{\mathcal{F}_t}\sup_{t\le s\le T}|\widehat{X}_s|^2\\
&\le\frac{2\beta_{\delta,T}}{\gamma}E^{\mathcal{F}_t}\left\{\frac{2}{\gamma}|\psi_T|^2+\langle K_tx,x\rangle-2\langle\psi_t,x\rangle+\int_t^T\langle Qq,q\rangle\,ds\right\}\\
&\quad-\frac{2\beta_{\delta,T}}{\gamma}E^{\mathcal{F}_t}\left\{2\int_t^T\langle\psi,f\rangle\,ds-\int_t^T\sum_{i=1}^d[\langle Kg_i,g_i\rangle-2\langle\phi_i,g_i\rangle]\,ds\right\}\\
&\quad+\beta_{\delta,T}E^{\mathcal{F}_t}\left\{\int_t^T|f_s|^2\,ds+\int_t^T|g(s)|^2\,ds\right\}.
\end{aligned}\tag{3.37}$$

In particular, take $t=0$, and we see that $\widehat{u}\in\mathcal{L}^2_{\mathcal{F}}(0,T;R^m)$.

Finally let us show that $\widehat{u}$ given by (3.25) is optimal and that the formula (3.27) holds. For every fixed $u\in\mathcal{L}^2_{\mathcal{F}}(0,T;R^m)$, since X is a semimartingale and $\int_0^{\cdot}\sum_{i=1}^d L_i\,dw_i$ is a BMO-martingale, then we deduce from Lemma 3.7 that $\widetilde{u}\in\mathcal{L}^2_{\mathcal{F}}(0,T;R^m)$. This is crucial in the following limiting arguments.

Note that ψ is uniformly square-integrable. X is uniformly square-integrable for every fixed $u\in\mathcal{L}^2_{\mathcal{F}}(0,T;R^m)$. Also note that $f,g:=(g_1,\dots,g_d)$, q are square-integrable, that K is uniformly bounded, $\phi:=(\phi_1,\dots,\phi_d)$, and u^0 are square-integrable. Then passing to the limit in (3.30), it follows from Lebesgue's dominated convergence theorem that for every $u\in\mathcal{L}^2_{\mathcal{F}}(0,T;R^m)$,

$$\begin{aligned}
J(u;t,x)&:=E^{\mathcal{F}_t}[\langle K_TX_T,X_T\rangle-2\langle\psi_T,X_T\rangle+\langle M\xi,\xi\rangle]\\
&\quad+E^{\mathcal{F}_t}\left[\int_t^T\langle Q(X-q),X-q\rangle\,ds+\int_t^T\langle Nu,u\rangle\,ds\right]\\
&=\langle K_tx,x\rangle-2\langle\psi_t,x\rangle+E^{\mathcal{F}_t}\left[\langle M\xi,\xi\rangle+\int_t^T\langle Qq,q\rangle\,ds\right]\\
&\quad-2E^{\mathcal{F}_t}\int_t^T\langle\psi,f\rangle\,ds+E^{\mathcal{F}_t}\int_t^T\sum_{i=1}^d[\langle Kg_i,g_i\rangle-2\langle\phi_i,g_i\rangle]\,ds\\
&\quad+E^{\mathcal{F}_t}\int_t^T\left\langle\left(N+\sum_{i=1}^d D_i^*KD_i\right)(\widetilde{u}-u^0),\widetilde{u}-u^0\right\rangle\,ds\\
&\quad-E^{\mathcal{F}_t}\int_t^T\left\langle\left(N+\sum_{i=1}^d D_i^*KD_i\right)u^0,u^0\right\rangle\,ds.
\end{aligned}\tag{3.38}$$

Since $\widehat{u}\in\mathcal{L}^2_{\mathcal{F}}(0,T;R^m)$, we derive from the last equality that $\widehat{u}$ is optimal and the value function is given by (3.27). □

3.5. Solution of the two BSDEs.

3.5.1. The linear unbounded BSDE (3.24). Note that $\widehat{A}$ and $\widehat{C}$ depend on L in general, and thus they might not be uniformly bounded. In this case, we have no theorem to guarantee the existence and the uniqueness of an $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted solution, though BSDE (3.24) is linear.

We shall prove the following theorem.

THEOREM 3.11. *Let assumptions* (A1), (A2), (A5), (A6)′, *and* (A7) *be satisfied. Let* (K, L) *be an* $\{\mathcal{F}_t, 0 \le t \le T\}$*-adapted solution of BSRDE* (3.23) *such that* $K \in \mathcal{L}^\infty_{\mathcal{F}}(0, T; S^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0, T]; S^n_+))$, $L \in (\mathcal{L}^2_{\mathcal{F}}(0, T; S^n_+))^d$, *and* K *is uniformly positive. Then, BSDE* (3.24) *has a unique* $\{\mathcal{F}_t, 0 \le t \le T\}$*-adapted solution such that* $E \sup_{0\le t\le T} |\psi_t|^2 < \infty$ *and* $\phi \in (\mathcal{L}^2_{\mathcal{F}}(0, T; R^n))^d$. *Moreover, there is a deterministic positive constant* β *such that*

$$
\begin{aligned}
&E \sup_{0\le t\le T} |\psi_t|^2 + E \int_0^T |\phi(t)|^2 \, dt \\
&\quad \le \beta \left[E|\xi|^2 + E \int_0^T (|f_t|^2 + |g(t)|^2 + |q_t|^2) \, dt \right].
\end{aligned} \tag{3.39}
$$

Proof. From Theorem 2.6, the assumptions imply that there is a unique optimal control $\widehat{u} \in \mathcal{L}^2_{\mathcal{F}}(0, T; R^m)$. Let $x = 0$. From Theorem 3.4, it follows that there is a 4-tuple $(\widehat{X}, y, z, \widehat{u})$ such that (3.4), (3.5), and (3.6) are all satisfied. Using Itô's formula, we check out that the pair (ψ, ϕ) (with $\phi := (\phi_1, \dots, \phi_d)$) defined by

$$
\psi := K\widehat{X} - y, \ \phi_i := K(C_i\widehat{X} + D_i\widehat{u} + g_i) + L_i\widehat{X} - z_i \tag{3.40}
$$

solves BSDE (3.24).

Since K is uniformly positive, $N + \sum_{i=1}^d D_i^* K D_i$ is uniformly positive. From Theorem 3.9, it follows that $\int_0^\cdot \sum_{i=1}^d L_i(s) \, dw_i(s)$ is a BMO-martingale. From Lemma 3.7, it follows that $L_i\widehat{X} \in \mathcal{L}^2_{\mathcal{F}}(0, T; R^n)$ for $i = 1, \dots, d$. Therefore, $\phi \in (\mathcal{L}^2_{\mathcal{F}}(0, T; R^n))^d$.

Since K is uniformly bounded and $x = 0$, from (3.40) and Lemma 3.7(iii), we deduce

$$
\begin{aligned}
E \sup_{0\le t\le T} |\psi_t|^2 &\le \beta E \sup_{0\le t\le T} (|\widehat{X}_t|^2 + |y_t|^2), \\
E \int_0^T \sum_{i=1}^d |L_i\widehat{X}_t|^2 \, dt &\le \beta E \left| \int_0^T |A\widehat{X} + B\widehat{u} + f| \, dt \right|^2 \\
&\quad + \beta E \int_0^T \sum_{i=1}^d |C_i\widehat{X} + D_i\widehat{u} + g_i|^2 \, dt \\
&\le \beta E \int_0^T (|\widehat{X}| + |\widehat{u}|^2 + |f|^2 + |g|^2) \, dt, \\
E \int_0^T |\phi(t)|^2 \, dt &\le \beta E \int_0^T \sum_{i=1}^d |C_i\widehat{X} + D_i\widehat{u} + g_i - z_i|^2 \, dt \\
&\quad + \beta E \int_0^T \sum_{i=1}^d |L_i\widehat{X}_t|^2 \, dt \\
&\le \beta E \int_0^T (|\widehat{X}| + |\widehat{u}|^2 + |z|^2 + |f|^2 + |g|^2) \, dt.
\end{aligned} \tag{3.41}
$$

In view of Theorem 3.5 and the fact that $x = 0$, the desired estimate (3.39) is obtained.

Now it remains to show the uniqueness of adapted solutions for BSDE (3.24). Let $(\widetilde{\psi}, \widetilde{\phi})$ be any $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted solution such that $E \sup_{0\le t\le T} |\widetilde{\psi}_t|^2 < \infty$ and $\widetilde{\phi} \in (\mathcal{L}^2_{\mathcal{F}}(0,T;R^n))^d$. Define

$$(3.42)\ \widetilde{y} := K\widehat{X} - \widetilde{\psi}, \quad \widetilde{z}_i := K(C_i\widehat{X} + D_i\widehat{u} + g_i) + L_i\widehat{X} - \widetilde{\phi}_i,\ \widetilde{z} := (\widetilde{z}_1, \dots, \widetilde{z}_d),$$

where $\widetilde{\phi}_i$ is the ith component of $\widetilde{\phi}$. Using Itô's formula, we can verify that $(\widetilde{y}, \widetilde{z})$ solves BSDE (3.4) in the sense of (3.5). According to the uniqueness of solutions of BSDE (3.4), we have $y = \widetilde{y}$ and $z = \widetilde{z}$. Therefore, in view of (3.40) and (3.42), we have $\psi = \widetilde{\psi}$ and $\phi = \widetilde{\phi}$. The desired uniqueness is proved. □

We remark that the square-integrability of ϕ is necessary for formulas (3.27) and (3.28) to make sense. (Note that $g \in (\mathcal{L}^2_{\mathcal{F}}(0,T;R^n))^d$.) In the above proof, it is far from being obvious, and it is derived from the BMO-property of $\int_0^{\cdot} \sum_{i=1}^d L_i\, dw_i(s)$.

3.5.2. BSRDE (3.23). We first give some historical comments.

When the coefficients A, B, C, D, M, Q, N are all deterministic, the second unknown variable L in BSRDE (3.23) can be taken to be zero. In this case, it reduces to an ordinary Riccati differential matrix equation, and the existence and uniqueness of its solution was proved by Wonham [36, 37] using Bellman's principle quasi linearization (see Bellman [2]) and a monotone argument. Also see Bismut [5] for a different approach.

The study concerning the case of random coefficients is dated back at least to Bismut [4, 6]. In the two papers, he derived the general form of BSRDE (3.23) starting from the stochastic maximum principle for the stochastic LQ problem. However, he could only treat a simple case in which the second unknown variable L does not appear in the nonlinear term of the drift. His method is to construct a contraction mapping. So, his method is quite different from Wonham's approach to the case of deterministic coefficients.

Later, Peng [27, 28] rediscovered the general form of BSRDE (3.23) starting from the dynamic programming principle and proved in the spirit of Wonham the existence and uniqueness result on BSRDE in the simple case.

The above-mentioned works are concerned only with the regular case, i.e., they make assumption (A6). For the singular case, i.e., under assumptions (A2) and (A6)′ instead of (A6), the study in the literature is concentrated on the case of deterministic coefficients: see Chen, Li, and Zhou [7], Kohlmann and Zhou [22], and Chen and Zhou [11] for a variety of results. Note also that Chen and Yong [8, 9, 10] contain a result on local solutions of BSRDEs in the case of random coefficients.

In the rest of this subsection, we shall first describe a general scheme to reduce BSRDE (3.23) in the singular case to the regular case, so that we can prove the existence and uniqueness result in the singular case from that in the regular case. Then, in view of Bismut and Peng's existence and uniqueness result on BSRDE (3.23) in the regular case, we derive from the above reduction scheme the counterpart in the singular case.

Throughout the rest of this subsection, we assume that $f = 0, g_i = 0, \xi = 0, q = 0$. We first describe the reduction scheme.

Consider a sequence of $m \times m$ uniformly positive, decreasing, symmetric matrix processes $\{N_k(t) : 0 \le t \le T\}_{k=1}^{\infty}$ of $\mathcal{L}^\infty(0,T;S^m)$ such that N_k converges to N strongly in $\mathcal{L}^\infty(0,T;S^m)$ as $k \to \infty$. Then consider the following perturbation of

BSRDE (3.23):

$$
\begin{cases}
dK^k = -\Bigg[A^*K^k + K^kA + \sum_{i=1}^d C_i^*K^kC_i + Q + \sum_{i=1}^d (C_i^*L_i^k + L_i^kC_i) \\
\qquad\qquad -\Gamma(t,K^k,L^k;N_k)\left(N_k + \sum_{i=1}^d D_i^*K^kD_i\right)\Gamma(t,K^k,L^k;N_k)^*\Bigg]\,dt \\
\qquad\quad + \sum_{i=1}^d L_i^k\,dw_i(t), \qquad 0 \le t < T, \\
K_T^k = M.
\end{cases} \tag{3.43}
$$

It is associated with the LQ problem, denoted by $\mathcal{P}_k$:

$$
\min_{u\in\mathcal{L}^2_{\mathcal{F}}(t,T;R^m)} J_k(u;t,x) \tag{3.44}
$$

with

$$
J_k(u;t,x) = E^{\mathcal{F}_t}\langle M\xi,\xi\rangle + E^{\mathcal{F}_t}\int_t^T (\langle Q_sX_s,X_s\rangle + \langle N_k(s)u_s,u_s\rangle)\,ds. \tag{3.45}
$$

The value function of problem $\mathcal{P}_k$ will be denoted by $V_k(t,x), (t,x)\in[0,T]\times R^n$.

We introduce the following assumption.

(A8) *If assumptions* (A1), (A2), (A5), *and* (A6)′ *are satisfied, then for each* $k = 1,2,\ldots$, *BSRDE* (3.43) *has a unique* $\{\mathcal{F}_t, 0\le t\le T\}$*-adapted solution* (K^k,L^k) *such that* $K^k\in\mathcal{L}^\infty_{\mathcal{F}}(0,T;S^n_+)\cap L^\infty(\Omega,\mathcal{F}_T,P;C([0,T];S^n_+))$, K^k *is uniformly positive in* (t,ω), *and* $L^k\in(\mathcal{L}^2_{\mathcal{F}}(0,T;S^n))^d$.

We have the following theorem.

THEOREM 3.12. *Let assumptions* (A1), (A2), (A5), (A6)′, *and* (A8) *be satisfied. Then, BSRDE* (3.23) *has a unique* $\{\mathcal{F}_t, 0\le t\le T\}$*-adapted solution* (K,L) *such that*

$$
K\in\mathcal{L}^\infty_{\mathcal{F}}(0,T;S^n_+)\cap L^\infty(\Omega,\mathcal{F}_T,P;C([0,T];S^n_+)) \quad\text{and}\quad L\in(\mathcal{L}^2_{\mathcal{F}}(0,T;S^n))^d.
$$

Moreover, $K(t,\omega)$ *is uniformly positive with respect to* (t,ω).

Proof. We develop a perturbation method.

Step 1. Note that if f,g,q,ξ are all zero objects, BSDE (3.24) associated with problem $\mathcal{P}_k$ has a unique zero solution. Assumption (A8) implies that BSRDE (3.43) has a unique $\{\mathcal{F}_t, 0\le t\le T\}$-adapted solution (K^k,L^k) such that $K^k\in\mathcal{L}^\infty_{\mathcal{F}}(0,T;S^n_+)$ and $L^k\in(\mathcal{L}^2_{\mathcal{F}}(0,T;S^n))^d$. Then, we conclude from Theorem 3.8 and the remarks in section 4 that problem $\mathcal{P}_k$ has a unique optimal control with the closed form:

$$
\widehat{u}^k := \Gamma(\cdot,K^k,L^k;N_k)\widehat{X}^k. \tag{3.46}
$$

The value function is

$$
V_k(t,x) := \langle K_t^kx,x\rangle. \tag{3.47}
$$

Moreover, we can show that $K_t^k(\omega)$ is uniformly positive in (t,ω,k), i.e., almost surely,

$$
K_t^k \ge \delta\beta_{\delta,T}^{-1}I_{n\times n} \quad \forall\, t\in[0,T]
$$

for a positive deterministic real number δ which is independent of k,t, and ω. Since $V_k(t,x) \le J_k(0;t,x) \le |x|^2\exp(\beta_1(T-t))$ for some positive constant β_1 which is

independent of k, K_t^k is also bounded from the above uniformly with respect to (t,ω,k): $K_t^k \le \exp(\beta_1 T) I_{n\times n}$.

In the next three steps, we show that the sequence $\{(K^k, L^k)\}_{k=1}^{\infty}$ converges to a limit (K,L) strongly in $\{\mathcal{L}_{\mathcal{F}}^{\infty}(0,T;S_+^n)\cap L^{\infty}(\Omega,\mathcal{F}_T,P;C([0,T];S_+^n))\}\times(\mathcal{L}_{\mathcal{F}}^{2}(0,T;S^n))^d$ with $K_t \ge \delta\beta_{\delta,T}^{-1} I_{n\times n}$.

Step 2. We shall prove in this step that for fixed $x\in R^n$, $V_k(t,x)$ converges to $V(t,x)$ strongly both in $\mathcal{L}_{\mathcal{F}}^{\infty}(0,T;R)$ and in $L^{\infty}(\Omega,\mathcal{F}_T,P;C([0,T];R))$.

Let $\widehat{u}$ be the optimal control of the original LQ problem, i.e., $V(t,x)=J(\widehat{u};t,x)$. Then,

$$\begin{aligned} V(t,x) &\le V_k(t,x) \le J_k(\widehat{u};t,x) \\ &= J(\widehat{u};t,x) + E^{\mathcal{F}_t}\int_t^T \langle (N_k-N)\widehat{u},\widehat{u}\rangle\, ds \\ &= V(t,x) + E^{\mathcal{F}_t}\int_t^T \langle (N_k-N)\widehat{u},\widehat{u}\rangle\, ds. \end{aligned} \tag{3.48}$$

Therefore,

$$\begin{aligned} |V_k(t,x)-V(t,x)| &\le E^{\mathcal{F}_t}\int_t^T \langle (N_k-N)\widehat{u},\widehat{u}\rangle\, ds \\ &\le \|N_k-N\|_{\mathcal{L}_{\mathcal{F}}^{\infty}} E^{\mathcal{F}_t}\int_t^T |\widehat{u}|^2\, ds. \end{aligned} \tag{3.49}$$

On the other hand, there is a constant $\beta_1>0$ such that

$$J(0;t,x) \le |x|^2 \exp(\beta_1(T-t)). \tag{3.50}$$

Noting the positivity of M and Lemma 3.2, we have

$$J(\widehat{u};t,x) \ge \delta E^{\mathcal{F}_t}|X_T^{t,x;\widehat{u}}|^2 \ge \delta\beta_{\delta,T}^{-1} E^{\mathcal{F}_t}\int_t^T |\widehat{u}_s|^2\, ds. \tag{3.51}$$

Since

$$J(\widehat{u};t,x) = V(t,x) \le J(0;t,x),$$

we have

$$\delta\beta_{\delta,T}^{-1} E^{\mathcal{F}_t}\int_t^T |\widehat{u}(s)|^2\, ds \le |x|^2 \exp(\beta_1(T-t)). \tag{3.52}$$

Concluding the above, we have

$$|V_k(t,x)-V(t,x)| \le \delta^{-1}\beta_{\delta,T}|x|^2 \exp(\beta_1(T-t))\|N_k-N\|_{\mathcal{L}_{\mathcal{F}}^{\infty}}.$$

This implies the desired result.

Step 3. From the preceding two steps, we conclude that the value function V has a quadratic expression. More precisely, there is an $\{\mathcal{F}_t, 0\le t\le T\}$-adapted stochastic process $K\in\mathcal{L}_{\mathcal{F}}^{\infty}(0,T;S_+^n)\cap L^{\infty}(\Omega,\mathcal{F}_T,P;C([0,T];S_+^n))$ such that

$$V(t,x) = \langle K_t x, x\rangle \quad \forall (t,x)\in[0,T]\times R^n \text{ a.s.} \tag{3.53}$$

Moreover, K^k converges to K strongly in the two spaces

$$\mathcal{L}^\infty_{\mathcal{F}}(0,T;S^n_+) \quad \text{and} \quad L^\infty(\Omega,\mathcal{F}_T,P;C([0,T];S^n_+)),$$

and K is uniformly positive with respect to $(t,\omega)\in[0,T]\times\Omega : K_t \geq \delta\beta^{-1}_{\delta,T}I_{n\times n}$ (in view of Lemma 3.2).

Step 4. We shall prove that $\{L^k\}_{k=1}^\infty$ is a Cauchy sequence in $(\mathcal{L}^2_{\mathcal{F}}(0,T;S^n))^d$.

First, we show that $\{L^k\}_{k=1}^\infty$ is bounded in $\mathcal{L}^2_{\mathcal{F}}(0,T;R^d)$. In view of BSRDE (3.43), use Itô's formula to compute $|K^k_t|^2$. Let $H(t,K,L)$ be the nonlinear term of the drift in (3.23). We observe that since $K^k\geq 0$, $H(\cdot,K^k,L^k)\leq 0$, we have

$$2\text{ tr }[K^kH(\cdot,K^k,L^k)] = 2\text{ tr }[(K^k)^{1/2}H(\cdot,K^k,L^k)(K^k)^{1/2}]\leq 0.$$

Therefore, this term can be left out in our subsequent arguments. Now it is standard to show that $\{L^k\}_{k=1}^\infty$ is bounded in $\mathcal{L}^2_{\mathcal{F}}(0,T;R^d)$ (see Pardoux and Peng [26] and El Karoui, Peng, and Quenez [14], for example).

Now we show that $\{L^k\}_{k=1}^\infty$ is a Cauchy sequence in $\mathcal{L}^2_{\mathcal{F}}(0,T;R^d)$. For this purpose, use Itô's formula to compute $|K^k_t-K^l_t|^2$. We get the following:

$$\begin{aligned} &E|K^k_t-K^l_t|^2 + E\int_t^T |L^k(s)-L^l(s)|^2\,ds \\ &\quad = 2E\int_t^T \text{tr }\{(K^k_s-K^l_s)[\Delta_s(K^k_s-K^l_s,L^k(s)-L^l(s)) \\ &\qquad\qquad +H(s,K^k_s,L^k(s))-H(s,K^l_s,L^l(s))]\}\,ds, \end{aligned} \tag{3.54}$$

where the random matrix $\Delta_s(K,L)$ is defined for each triplet $(s,K,L)\in[0,T]\times S^n\times(S^n)^d$ by

$$\Delta_s(K,L) := A^*_sK + KA_s + \sum_{i=1}^d [C_i(s)^*L_i + C_i(s)^*KC_i(s) + L_iC_i(s)].$$

In view of assumption (A2) and the fact that K^k is uniformly bounded and uniformly positive (see Step 1), $(N_k+\sum_{i=1}^d D^*_iK^kD_i)^{-1}$ and $(N_l+\sum_{i=1}^d D^*_iK^lD_i)^{-1}$ are bounded uniformly in (k,l). Therefore the second integral in (3.54) is less than the product of $\|K^k-K^l\|_{\mathcal{L}^\infty_{\mathcal{F}}(0,T;S^n_+)}$ and

$$2E\int_0^T \left[(2|A|+|C|^2)|K^k-K^l| + 2|C||L^k-L^l| + |H(s,K^k,L^k)| + |H(s,K^l,L^l|\right]\,ds.$$

It is clear that the last quantity is less than a positive constant times the term $(1+\|K^k\|^2_{\mathcal{L}^\infty_{\mathcal{F}}}+\|K^l\|^2_{\mathcal{L}^\infty_{\mathcal{F}}}+\|L^k\|^2_{\mathcal{L}^2_{\mathcal{F}}}+\|L^l\|^2_{\mathcal{L}^2_{\mathcal{F}}})$, and it is thus bounded uniformly in (k,l) (since $\{(K^k,L^k)\}_{k=1}^\infty$ is bounded). Since

$$\lim_{k,l\to\infty}\|K_k-K_l\|_{\mathcal{L}^\infty_{\mathcal{F}}(0,T;S^n_+)}=0,$$

we conclude that the second integral in (3.54) converges to zero as k and l tend to ∞, and then we have the desired result.

Step 5. Let L be the limit in $(\mathcal{L}^2_{\mathcal{F}}(0,T;S^n))^d$ of the Cauchy sequence $\{L^k\}_{k=1}^\infty$. By Step 3, $K\in\mathcal{L}^\infty_{\mathcal{F}}(0,T;S^n_+)\cap L^\infty(\Omega,\mathcal{F}_T,P;C([0,T];S^n_+))$ is uniformly positive. Therefore, it is meaningful to take the limit in BSRDEs (3.43) by letting $k\to\infty$. As a result, (K,L) is shown to be an $\{\mathcal{F}_t, 0\leq t\leq T\}$-adapted solution to BSRDE (3.23).

Step 6. Finally, we show the uniqueness. It is a consequence of the representation (3.27). In fact, assume that (K, L) and $(\widetilde{K}, \widetilde{L})$ are two $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted solutions of BSRDE (3.23) with

$$K, \widetilde{K} \in \mathcal{L}^{\infty}_{\mathcal{F}}(0, T; S^n_+) \cap L^{\infty}(\Omega, \mathcal{F}_T, P; C([0, T]; S^n_+)) \quad \text{and} \quad L, \widetilde{L} \in (\mathcal{L}^2_{\mathcal{F}}(0, T; S^n))^d.$$

Then, we have

$$\begin{cases} d\delta K_s = -[\Delta_s(\delta K_s, \delta L(s)) + H(s, K_s, L(s)) - H(s, \widetilde{K}_s, \widetilde{L}(s))]\, ds \\ \qquad\qquad + \sum_{i=1}^{d} \langle \delta L_i(s), dw_i(s) \rangle, \\ \delta K_T = 0 \end{cases} \tag{3.55}$$

with $\delta K := K - \widetilde{K}$ and $\delta L := L - \widetilde{L}$. Moreover, in view of Theorem 3.8, the two symmetric $n \times n$ matrices K and $\widetilde{K}$ are uniformly positive. Therefore, in view of assumption (A2), both $(N + \sum_{i=1}^{d} D_i^* K D_i)^{-1}$ and $(N + \sum_{i=1}^{d} D_i^* \widetilde{K} D_i)^{-1}$ are bounded though the uniform positivity of the matrix N is not guaranteed (the nonnegativity is sufficient!). The two *difficult* terms $H(t, K_t, L(t))$ and $H(t, \widetilde{K}_t, \widetilde{L}(t))$—which appear in the drift of the last equation, distinct from $\Delta(\delta K, \delta L)$ due to the nonlinearity—are in $\mathcal{L}_{\mathcal{F}}(0, T; R^{n \times n})$. This integrability, together with the fact that $\delta K \in \mathcal{L}^{\infty}_{\mathcal{F}}(0, T; S^n)$, justify the following Itô's formula in its expected form:

$$\begin{aligned} &E|\delta K_0|^2 + E \int_0^T |\delta L|^2\, ds \\ &\quad = 2E \int_0^T \delta K_s [\Delta_s(\delta K_s, \delta L(s)) + H(s, K_s, L(s)) - H(s, \widetilde{K}_s, \widetilde{L}(s))]\, ds. \end{aligned} \tag{3.56}$$

Noting that K and $\widetilde{K}$ have the same representation (3.27), we have $V(t, x) = \langle Kx, x \rangle = \langle \widetilde{K} x, x \rangle$ for all $x \in R^n$, which implies $\delta K = 0$. Putting this equality into (3.56), we have $E \int_0^T |\delta L|^2\, ds = 0$. This implies that $L = \widetilde{L}$. □

In what follows, we address a case in which assumption (A8) is true automatically.

To avoid the appearance of L in the nonlinear term of the drift, Bismut [6] and Peng [27] imposed the following assumption.

(A9) Let the nonnegative integer $d_0 \le d$. Denote by $\{\mathcal{G}_t, 0 \le t \le T\}$ the P-augmented natural filtration generated by the $(d - d_0)$-dimensional Brownian motion $(w_{d_0+1}, \ldots, w_d)$. Let the matrix-valued processes A, B, C, D, M, Q, N be $\{\mathcal{G}_t, 0 \le t \le T\}$-progressively measurable. The matrix-valued random variable M is $\mathcal{G}_T$-measurable. $D_{d_0+1} = \cdots = D_d = 0$.

Under assumption (A9), BSRDE (3.23) reduces to

$$\begin{cases} dK_t = -\Bigg[A^* K_t + K_t A + \sum_{i=1}^{d} C_i^* K_t C_i + Q \\ \qquad + \sum_{i=d_0+1}^{d} (C_i^* L_i(t) + L_i(t) C_i) - \left(K_t B + \sum_{i=1}^{d} C_i^* K_t D_i \right) \\ \qquad \times \left(N + \sum_{i=1}^{d_0} D_i^* K_t D_i \right)^{-1} \left(K_t B + \sum_{i=1}^{d} C_i^* K_t D_i \right)^* \Bigg]\, dt \\ \qquad + \sum_{i=d_0+1}^{d} L_i(t)\, dw_i(t), \qquad 0 \le t < T, \\ K_T = M. \end{cases} \tag{3.57}$$

The above Riccati equation is easier, for the second unknown variable L appears in the drift in a linear way.

Bismut [6] and Peng [27] showed that assumption (A8) is true if assumption (A9) is satisfied. According to Theorem 3.12, we obtain immediately the following theorem.

THEOREM 3.13. *Let assumptions* (A1), (A2), (A5), (A6)′, *and* (A9) *be satisfied. Then, BSRDE* (3.57) *has a unique* $\{\mathcal{G}_t, 0 \le t \le T\}$*-adapted solution* $(K; L_{d_0+1}, \dots, L_d)$ *such that* $K \in \mathcal{L}^{\infty}_{\mathcal{G}}(0,T;S^n_+) \cap L^{\infty}(\Omega, \mathcal{G}_T, P; C([0,T]; S^n_+))$ *and* $(L_{d_0+1}, \dots, L_d) \in (\mathcal{L}^2_{\mathcal{G}}(0,T;S^n))^{d-d_0}$. *Moreover,* $K(t,\omega)$ *is uniformly positive with respect to* (t,ω) *almost surely,* $K_t \ge \delta\beta^{-1}_{\delta,T} I_{n\times n}$, *for all* $t \in [0,T]$.

4. Concluding comments. In this paper, a new framework for the stochastic LQ problem is developed where the cost may be independent of the control variable and the coefficients are allowed to be random. In this setting, the two assumptions (A2) and (A6)′ play a crucial role. However, in Theorems 3.8 and 3.11, if both assumptions are replaced with assumption (A6) and the assumption that K is uniformly positive is removed, the assertions in both theorems are still true. To see this comment, it is sufficient to mention the following points.

In the regular case, $N + \sum_{i=1}^d D_i^* K D_i$ is still uniformly positive and $\int_0^{\cdot} \sum_{i=1}^d L_i\, dw_i$ is still a BMO-martingale (in view of Theorem 3.9).

In the regular case, to derive from the formula (3.32) the fact that $\widehat{u} \in \mathcal{L}^2_{\mathcal{F}}(0,T;R^m)$, it suffices to note the following points. We have for any deterministic constant $\alpha_1 > 0$,

$$
\begin{aligned}
&E^{\mathcal{F}_t}[\langle K_{T_k}\widehat{X}_{T_k}, \widehat{X}_{T_k}\rangle - 2\langle \psi_{T_k}, \widehat{X}_{T_k}\rangle] \\
&+ E^{\mathcal{F}_t}\left[\int_t^{T_k} \langle Q(\widehat{X}-q), \widehat{X}-q\rangle\, ds + \int_t^{T_k} \langle N\widehat{u}, \widehat{u}\rangle\, ds\right] \\
&\ge E^{\mathcal{F}_t}\left[-2\langle \psi_{T_k}, \widehat{X}_{T_k}\rangle + \int_t^{T_k} \langle N\widehat{u}, \widehat{u}\rangle\, ds\right] \\
&\ge E^{\mathcal{F}_t}\left[-\alpha_1^{-1}|\psi_{T_k}|^2 - \alpha_1 |\widehat{X}_{T_k}|^2 + \delta \int_t^{T_k} |\widehat{u}|^2\, ds\right].
\end{aligned} \tag{4.1}
$$

In view of Lemma 2.1, we have

$$
\begin{aligned}
&E^{\mathcal{F}_t}[\langle K_{T_k}\widehat{X}_{T_k}, \widehat{X}_{T_k}\rangle - 2\langle \psi_{T_k}, \widehat{X}_{T_k}\rangle] \\
&+E^{\mathcal{F}_t}\left[\int_t^{T_k} \langle Q(\widehat{X}-q), \widehat{X}-q\rangle\, ds + \int_t^{T_k} \langle N\widehat{u}, \widehat{u}\rangle\, ds\right] \\
\ge &E^{\mathcal{F}_t}\left[(\delta - \alpha_1\beta)\int_t^{T_k} |\widehat{u}|^2\, ds - \alpha_1^{-1}|\psi_{T_k}|^2 - \alpha_1\beta \int_t^{T_k} (T|f|^2 + |g|^2)\, ds - \alpha_1\beta |x|^2\right].
\end{aligned} \tag{4.2}
$$

We now choose a sufficiently small $\delta_1 > 0$, to say $\delta_1 = \frac{\delta}{2\beta}$, such that $\delta - \delta_1\beta > 0$. Then, combining (3.32) and (4.2), we conclude through the same limiting analysis as in the singular case that $\widehat{u} \in \mathcal{L}^2_{\mathcal{F}}(0,T;R^m)$.

The proof of Theorem 3.11 involves Theorem 3.5. However, in Theorem 3.5, the two assumptions (A2) and (A6)′ can be replaced with assumption (A6). To see this, it is enough to note the following. The formula (3.9) is still true. It implies the following:

$$\begin{aligned}\delta E\int_0^T |\widehat{u}_t|^2\,dt &\le E\int_0^T \langle N_t\widehat{u}_t,\widehat{u}_t\rangle\,dt\\ &\le E\int_0^T \left[\langle y_t,f_t\rangle+\sum_{i=1}^d\langle z_i(t),g_i(t)\rangle\right]dt+\langle y_0,x\rangle.\end{aligned}\tag{4.3}$$

In view of Lemma 2.1, we see that

$$\begin{aligned}E\sup_{0\le t\le T}|\widehat{X}_t|^2 &\le \frac{\beta}{\delta}\left(E\int_0^T\left[\langle y_t,f_t\rangle+\sum_{i=1}^d\langle z_i(t),g_i(t)\rangle\right]dt+\langle y_0,x\rangle\right)\\ &\quad+\beta\left(|x|^2+E\int_0^T(|f_t|^2+|g(t)|^2)\,dt\right).\end{aligned}\tag{4.4}$$

Then, combining the last inequality with (3.8), we have through a similar analysis as in the singular case the desired estimate (3.7).

Finally, we remark that the solution of BSRDE (3.23) in general is still left open in this paper. However, some subsequent works give new existence and uniqueness results on BSRDE (3.23). See, for instance, Kohlmann and Tang [18, 19, 20] and Lim and Zhou [23].

Note added in proof. The general existence and uniqueness for the solution of BSRDE (3.23) has been proved recently. The reader is referred to Tang [32, 33].

Acknowledgments. Both authors would like to thank the referees (especially the second referee) and the Associate Editor for their careful reading, critical comments, and helpful suggestions, which make the present version more readable.

REFERENCES

[1] R. Bañuelos and A. Bennett, *Paraproducts and commutators of martingale transforms*, Proc. Amer. Math. Soc., 103 (1988), pp. 1226–1234.

[2] R. Bellman, *Functional equations in the theory of dynamic programming, positivity and quasilinearity*, Proc. Natl. Acad. Sci. USA, 41 (1955), pp. 743–746.

[3] J.-M. Bismut, *Conjugate convex functions in optimal stochastic control*, J. Math. Anal. Appl., 44 (1973), pp. 384–404.

[4] J.-M. Bismut, *Linear quadratic optimal stochastic control with random coefficients*, SIAM J. Control Optim., 14 (1976), pp. 419–444.

[5] J.-M. Bismut, *On optimal control of linear stochastic equations with a linear-quadratic criterion*, SIAM J. Control Optim., 15 (1977), pp. 1–4.

[6] J.-M. Bismut, *Controle des systems lineares quadratiques: applications de l'integrale stochastique*, in Séminaire de Probabilités XII, Lecture Notes in Math. 649, C. Dellacherie, P. A. Meyer et M. Weil, eds., Springer-Verlag, Berlin, 1978, pp. 180–264.

[7] S. Chen, X. Li, and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs*, SIAM J. Control Optim., 36 (1998), pp. 1685–1702.

[8] S. Chen and J. Yong, *Stochastic linear-quadratic optimal control problems*, Appl. Math. Optim., 43 (2001), pp. 21–45.

[9] S. Chen and J. Yong, *Stochastic linear quadratic optimal control problems with random coefficients*, Chinese Ann. Math. Ser. B, 21 (2000), pp. 323–338.

[10] S. Chen and J. Yong, *Solvability of a stochastic linear quadratic optimal control problem*, in Applied Probability, AMS/IP Stud. Adv. Math. 26, R. Chan et al., eds., AMS, Providence, RI, 2002, pp. 35–43.

[11] S. Chen and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weights costs.* II, SIAM J. Control Optim., 39 (2000), pp. 1065–1081.

[12] D. Duffie and H. R. Richardson, *Mean-variance hedging in continuous time*, Ann. Appl. Probab., 1 (1991), pp. 1–15.

[13] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North–Holland, Amsterdam, Elsevier, New York, 1976.

[14] N. El Karoui, S. Peng, and M. C. Quenez, *Backward stochastic differential equations in finance*, Math. Finance, 7 (1997), pp. 1–71.
[15] L. I. Gal'chuk, *Existence and uniqueness of a solution for stochastic equations with respect to semimartingales*, Theory Probab. Appl., 23 (1978), pp. 751–763.
[16] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, Berlin, Heidelberg, 1999.
[17] N. Kazamaki, *Continuous Exponential Martingales and BMO*, Springer-Verlag, Berlin, Heidelberg, 1994.
[18] M. Kohlmann and S. Tang, *Global adapted solution of one-dimensional backward stochastic Riccati equations, with application to the mean-varaince hedging*, Stochastic Process. Appl., 97 (2002), pp. 255–288.
[19] M. Kohlmann and S. Tang, *Multidimensional backward stochastic Riccati equations and applications*, SIAM J. Control Optim., 41 (2003), pp. 1696–1721.
[20] M. Kohlmann and S. Tang, *New developments in backward stochastic Riccati equations and their applications*, in Mathematical Finance, M. Kohlmann and S. Tang, eds., Birkhauser, Basel, 2001, pp. 194–214.
[21] M. Kohlmann and S. Tang, *A BMO-Property for Backward Stochastic Riccati Differential Equations*, preprint.
[22] M. Kohlmann and X. Y. Zhou, *Relationship between backward stochastic differential equations and stochastic controls: A linear-quadratic approach*, SIAM J. Control Optim., 38 (2000), pp. 1392–1407.
[23] A. Lim and X. Zhou, *Mean-Variance Portfolio Selection with Random Parameters*, preprint, Chinese University of Hong Kong, Hong Kong, China, 2000.
[24] J. L. Lions, *Contrôle Optimal des Systèmes Gouvernés par des Équations aux Derivés Partielles*, Dunod, Paris, 1968.
[25] E. Pardoux and S. Peng, *Adapted solution of backward stochastic equation*, Systems Control Lett., 14 (1990), pp. 55–61.
[26] E. Pardoux and S. Peng, *Backward stochastic differential equations and quasi-linear parabolic partial differential equations*, in Stochastic Partial Differential Equations and Their Applications, Lecture Notes in Control and Inform. Sci. 176, B. L. Rozovskii and R. S. Sowers, eds., Springer-Verlag, Berlin, Heidelberg, New York, 1992, pp. 200–217.
[27] S. Peng, *Stochastic Hamilton–Jacobi–Bellman equations*, SIAM J. Control Optim., 30 (1992), pp. 284–304.
[28] S. Peng, *Open problems on backward stochastic differential equations*, in Control of Distributed Parameter and Stochastic Systems, S. Chen, et al., eds., Kluwer Academic Publishers, Boston, MA, 1999, pp. 265–273.
[29] S. Peng, *A general stochastic maximum principle for optimal control problems*, SIAM J. Control Optim., 28 (1990), pp. 966–979.
[30] H. Pham, T. Rheinländer, and M. Schweizer, *Mean-variance hedging for continuous processes: New proofs and examples*, Finance Stoch., 2 (1998), pp. 173–198.
[31] P. Protter, *Stochastic Integration and Differential Equations. A New Approach*, Springer-Verlag, Berlin, 1990.
[32] S. Tang, *General linear quadratic optimal stochastic control problems with random coefficients: Linear stochastic Hamilton systems and backward stochastic Riccati equations*, SIAM J. Control Optim., 42 (2003), pp. 53–75.
[33] S. Tang, *Financial mean-variance problems and stochastic LQ problems: Linear stochastic Hamilton systems and backward stochastic Riccati equations*, in Recent Developments in Mathematical Finance, J. Yong, ed., World Scientific, River Edge, NJ, 2002, pp. 190–203.
[34] S. Tang and X. Li, *Necessary conditions for optimal control of stochastic systems with random jumps*, SIAM J. Control Optim., 32 (1994), pp. 1447–1475.
[35] J. Yong and X. Zhou, *Stochastic Controls. Hamiltonian Systems and HJB equations*, Springer-Verlag, New York, 1999.
[36] W. M. Wonham, *On a matrix Riccati equation of stochastic control*, SIAM J. Control, 6 (1968), pp. 681–697.
[37] W. M. Wonham, *Random differential equations in control theory*, in Probabilistic Methods in Applied Mathematics, Vol. 2, A. T. Bharucha-Reid, ed., Academic Press, New York, London, 1970, pp. 131–212.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1143–1166

ON ACTOR-CRITIC ALGORITHMS*

VIJAY R. KONDA† AND JOHN N. TSITSIKLIS†

Abstract. In this article, we propose and analyze a class of actor-critic algorithms. These are two-time-scale algorithms in which the critic uses temporal difference learning with a linearly parameterized approximation architecture, and the actor is updated in an approximate gradient direction, based on information provided by the critic. We show that the features for the critic should ideally span a subspace prescribed by the choice of parameterization of the actor. We study actor-critic algorithms for Markov decision processes with Polish state and action spaces. We state and prove two results regarding their convergence.

Key words. reinforcement learning, Markov decision processes, actor-critic algorithms, stochastic approximation

AMS subject classifications. 93E35, 68T05, 62L20

DOI. 10.1137/S0363012901385691

1. Introduction.

Many problems in finance, communication networks, operations research, and other fields can be formulated as dynamic programming problems. However, the dimension of the state space in these formulations is often too large for the problem to be tractable. Moreover, the underlying dynamics are seldom known and are often difficult to identify. Reinforcement learning and neuro-dynamic programming [5, 19] methods try to overcome these difficulties by combining simulation-based learning and compact representations of policies and value functions. The vast majority of these methods falls into one of the following two categories:

(a) Actor-only methods work with a parameterized family of policies. The gradient of the performance, with respect to the actor parameters, is directly estimated by simulation, and the parameters are updated in a direction of improvement [8, 10, 16, 23]. A possible drawback of such methods is that the gradient estimators may have a large variance. Furthermore, as the policy changes, a new gradient is estimated independently of past estimates. Hence, there is no "learning" in the sense of accumulation and consolidation of older information.

(b) Critic-only methods rely exclusively on value function approximation and aim at learning an approximate solution to the Bellman equation, which will then hopefully prescribe a near-optimal policy. Such methods are indirect in the sense that they do not try to optimize directly over a policy space. A method of this type may succeed in constructing a "good" approximation of the value function yet lack reliable guarantees in terms of near-optimality of the resulting policy.

Actor-critic methods [2] aim at combining the strong points of actor-only and critic-only methods. The critic uses an approximation architecture and simulation to learn a value function, which is then used to update the actor's policy parameters in a

*Received by the editors February 28, 2001; accepted for publication (in revised form) November 24, 2002; published electronically August 6, 2003. This research was partially supported by the NSF under contract ECS-9873451 and by the AFOSR under contract F49620-99-1-0320. A preliminary version of this paper was presented at the 1999 Neural Information Processing Systems conference [13].

http://www.siam.org/journals/sicon/42-4/38569.html

†Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 (konda@alum.mit.edu, jnt@mit.edu).

direction of performance improvement. Such methods, as long as they are gradient-based, may have desirable convergence properties, in contrast to critic-only methods for which convergence is guaranteed in rather limited settings. They also hold the promise of delivering faster convergence (due to variance reduction) than actor-only methods. On the other hand, theoretical understanding of actor-critic methods has been limited to the case of lookup table representations of policies and value functions [12].

In this paper, we propose some actor-critic algorithms in which the critic uses linearly parameterized approximations of the value function, and we provide a convergence proof. The algorithms are based on the following important observation: since the number of parameters that the actor has to update is relatively small (compared to the number of states), the critic need not attempt to compute or approximate the exact value function, which is a high-dimensional object. In fact, we show that the critic should ideally compute a certain "projection" of the value function onto a low-dimensional subspace spanned by a set of "basis functions," which are *completely determined* by the parameterization of the actor. This key insight was also derived in simultaneous and independent work [20] that also included a discussion of certain actor-critic algorithms.

The outline of the paper is as follows. In section 2, we state a formula for the gradient of the average cost in a Markov decision process with finite state and action space. We provide a new interpretation of this formula, and use it in section 3 to derive our algorithms. In section 4, we consider Markov decision processes and the gradient of the average cost in much greater generality and describe the algorithms in this more general setting. In sections 5 and 6, we provide an analysis of the asymptotic behavior of the critic and actor, respectively. The appendix contains a general result concerning the tracking ability of linear stochastic iterations, which is used in section 5.

2. Markov decision processes and parameterized families of randomized stationary policies. Consider a Markov decision process with finite state space $\mathbb{X}$ and finite action space $\mathbb{U}$. Let $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ be a given one-stage cost function. Let $p(y|x,u)$ denote the probability that the next state is y, given that the current state is x and the current action is u. A *randomized stationary policy* (RSP) is a mapping μ that assigns to each state x a probability distribution over the action space $\mathbb{U}$. We consider a set of RSPs $\{\mu_\theta; \theta \in \mathbb{R}^n\}$, parameterized in terms of a vector θ. For each pair $(x,u) \in \mathbb{X} \times \mathbb{U}$, $\mu_\theta(u|x)$ denotes the probability of taking action u when the state x is encountered, under the policy corresponding to θ. Hereafter, we will not distinguish between the parameter of an RSP and the RSP itself. Therefore, whenever we refer to an "RSP θ," we mean the RSP corresponding to parameter vector θ. Note that, under any RSP, the sequence of states $\{X_k\}$ and the sequence of state-action pairs $\{X_k, U_k\}$ of the Markov decision process form Markov chains with state spaces $\mathbb{X}$ and $\mathbb{X} \times \mathbb{U}$, respectively. We make the following assumption about the family of policies.

Assumption 2.1 (finite case).

(a) For every $x \in \mathbb{X}$, $u \in \mathbb{U}$, and $\theta \in \mathbb{R}^n$, we have $\mu_\theta(u|x) > 0$.

(b) For every $(x,u) \in \mathbb{X} \times \mathbb{U}$, the mapping $\theta \mapsto \mu_\theta(u|x)$ is twice differentiable. Furthermore, the $\mathbb{R}^n$-valued function $\theta \to \nabla \ln \mu_\theta(u|x)$ is bounded and has a bounded first derivative, for any fixed x and u.*

*Throughout the paper, ∇ will stand for the gradient with respect to the vector θ.

(c) For every $\theta \in \mathbb{R}^n$, the Markov chains $\{X_k\}$ and $\{X_k, U_k\}$ are irreducible and aperiodic, with stationary probabilities $\pi_\theta(x)$ and $\eta_\theta(x,u) = \pi_\theta(x)\mu_\theta(u|x)$, respectively, under the RSP θ.

(d) There is a positive integer N, state $x^* \in \mathbb{X}$, and $\epsilon_0 > 0$ such that, for all $\theta_1, \ldots, \theta_N \in \mathbb{R}^n$,

$$\sum_{k=1}^{N} [P(\theta_1) \cdots P(\theta_k)]_{xx^*} \geq \epsilon_0 \qquad \forall x \in \mathbb{X},$$

where $P(\theta)$ denotes the transition probability matrix for the Markov chain $\{X_k\}$ under the RSP θ. (We use here the notation $[P]_{xx^*}$ to denote the (x, x^*) entry of a matrix P.)

The first three parts of the above assumption are natural and easy to verify. The fourth part assumes that the probability of reaching x^*, in a number of transitions that is independent of θ, is uniformly bounded away from zero. This assumption is satisfied if part (c) of the assumption holds, and the policy probabilities $\mu_\theta(u|x)$ are all bounded away from zero uniformly in θ (see [11]).

Consider the average cost function $\bar{\alpha} : \mathbb{R}^n \to \mathbb{R}$, defined by

$$\bar{\alpha}(\theta) = \sum_{x \in \mathbb{X}, u \in \mathbb{U}} c(x,u)\eta_\theta(x,u).$$

A natural approach to minimize $\bar{\alpha}(\theta)$ over RSPs θ is to start with a policy θ_0 and improve it using gradient descent. To do this, we will rely on a formula for $\nabla\bar{\alpha}(\theta)$ to be presented shortly.

For each $\theta \in \mathbb{R}^n$, let $V_\theta : \mathbb{X} \to \mathbb{R}$ be a "differential cost function," i.e., a solution of the Poisson equation:

$$\bar{\alpha}(\theta) + V_\theta(x) = \sum_u \mu_\theta(u|x) \left[c(x,u) + \sum_y p(y|x,u)V_\theta(y) \right].$$

Intuitively, $V_\theta(x)$ can be viewed as the "disadvantage" of state x: it is the expected future excess cost—on top of the average cost—incurred if we start at state x. It plays a role similar to that played by the more familiar value function that arises in total or discounted cost Markov decision problems. Finally, for every $\theta \in \mathbb{R}^n$, we define the Q-value function $Q_\theta : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ by

$$Q_\theta(x,u) = c(x,u) - \bar{\alpha}(\theta) + \sum_y p(y|x,u)V_\theta(y).$$

We recall the following result, as stated in [16]. (Such a result has been established in various forms in [7, 8, 10] and elsewhere.)

THEOREM 2.2. *We have*

$$\nabla\bar{\alpha}(\theta) = \sum_{x,u} \eta_\theta(x,u)Q_\theta(x,u)\psi_\theta(x,u), \tag{2.1}$$

where

$$\psi_\theta(x,u) = \nabla \ln \mu_\theta(u|x).$$

In [16], the quantity $Q_\theta(x,u)$ in the above formula is interpreted as the expected excess cost incurred over a certain renewal period of the Markov chain $\{X_n, U_n\}$, under the RSP μ_θ, and is then estimated by means of simulation, leading to actor-only algorithms. Here, we provide an alternative interpretation of the formula in Theorem 2.2, as an inner product, and arrive at a different set of algorithms.

For any $\theta \in \mathbb{R}^n$, we define the inner product $\langle \cdot, \cdot \rangle_\theta$ of two real-valued functions Q_1, Q_2 on $\mathbb{X} \times \mathbb{U}$, viewed as vectors in $\mathbb{R}^{|\mathbb{X}||\mathbb{U}|}$, by

$$\langle Q_1, Q_2 \rangle_\theta = \sum_{x,u} \eta_\theta(x,u) Q_1(x,u) Q_2(x,u).$$

(We will be using the above notation for vector- or matrix-valued functions as well.) With this notation, we can rewrite the formula (2.1) as

$$\frac{\partial}{\partial \theta_i} \bar{\alpha}(\theta) = \langle Q_\theta, \psi_\theta^i \rangle_\theta, \qquad i = 1, \ldots, n,$$

where ψ_θ^i stands for the ith component of ψ_θ. Let $\| \cdot \|_\theta$ denote the norm induced by this inner product on $\mathbb{R}^{|\mathbb{X}||\mathbb{U}|}$. For each $\theta \in \mathbb{R}^n$, let Ψ_θ denote the span of the vectors $\{\psi_\theta^i;\ 1 \le i \le n\}$ in $\mathbb{R}^{|\mathbb{X}||\mathbb{U}|}$.

An important observation is that although the gradient of $\bar{\alpha}$ depends on the function Q_θ, which is a vector in a possibly very high-dimensional space $\mathbb{R}^{|\mathbb{X}||\mathbb{U}|}$, the dependence is only through its inner products with vectors in Ψ_θ. Thus, instead of "learning" the function Q_θ, it suffices to learn its projection on the low-dimensional subspace Ψ_θ.

Indeed, let $\Pi_\theta : \mathbb{R}^{|\mathbb{X}||\mathbb{U}|} \mapsto \Psi_\theta$ be the projection operator defined by

$$\Pi_\theta Q = \arg \min_{\hat{Q} \in \Psi_\theta} \|Q - \hat{Q}\|_\theta.$$

Since

$$\langle Q_\theta, \psi_\theta^i \rangle_\theta = \langle \Pi_\theta Q_\theta, \psi_\theta^i \rangle_\theta, \qquad i = 1, \ldots, n, \tag{2.2}$$

it is enough to know the projection of Q_θ onto Ψ_θ to compute $\nabla \bar{\alpha}$.

3. Actor-critic algorithms. We view actor-critic algorithms as stochastic gradient algorithms on the parameter space of the actor. When the actor parameter vector is θ, the job of the critic is to compute an approximation of the projection $\Pi_\theta Q_\theta$, which is then used by the actor to update its policy in an approximate gradient direction. The analysis in [21, 22] shows that this is precisely what temporal difference (TD) learning algorithms try to do, i.e., to compute the projection of an exact value function onto a subspace spanned by feature vectors. This allows us to implement the critic by using a TD algorithm. (Note, however, that other types of critics are possible, e.g., based on batch solution of least squares problems, as long as they aim at computing the same projection.)

We note some minor differences with the common usage of TD. In our context, we need the projection of q-functions rather than value functions. But this is easily achieved by replacing the Markov chain $\{x_t\}$ in [21, 22] with the Markov chain $\{X_k, U_k\}$. A further difference is that [21, 22] assume that the decision policy and the feature vectors are fixed. In our algorithms, the decision policy as well as the features need to change as the actor updates its parameters. As suggested by the

results of [12, 6, 14], this need not pose any problems, as long as the actor parameters are updated on a slower time-scale.

We are now ready to describe two actor-critic algorithms, which differ only as far as the critic updates are concerned. In both variants, the critic is a TD algorithm with a linearly parameterized approximation architecture for the Q-value function, of the form

$$Q_\theta^r(x,u) = \sum_{j=1}^{m} r^j \phi_\theta^j(x,u),$$

where $r = (r^1, \ldots, r^m) \in \mathbb{R}^m$ denotes the parameter vector of the critic. The features ϕ_θ^j, $j = 1, \ldots, m$, used by the critic are dependent on the actor parameter vector θ and are chosen so that the following assumptions are satisfied.

Assumption 3.1 (critic features).

(a) For every $(x,u) \in \mathbb{X} \times \mathbb{U}$ the map $\theta \to \phi_\theta(x,u)$ is bounded and differentiable, with a bounded derivative.

(b) The span of the vectors ϕ_θ^j, $j = 1, \ldots, m$, in $\mathbb{R}^{|\mathbb{X}||\mathbb{U}|}$, denoted by Φ_θ, contains Ψ_θ.

Note that the formula (2.2) still holds if Π_θ is redefined as the projection onto Φ_θ, as long as Φ_θ contains Ψ_θ. The most straightforward choice would be to let the number m of critic parameters be equal to the number n of actor parameters, and $\phi_\theta^i = \psi_\theta^i$ for each i. Nevertheless, we allow the possibility that $m > n$ and that Φ_θ properly contains Ψ_θ, so that the critic can use more features than are actually necessary. This added flexibility may turn out to be useful in a number of ways:

(a) It is possible that for certain values of θ, the feature vectors ψ_θ^i are either close to zero or are almost linearly dependent. For these values of θ, the operator Π_θ becomes ill-conditioned, which can have a negative effect on the performance of the algorithms. This might be avoided by using a richer set of features ϕ_θ^i.

(b) For the second algorithm that we propose, which involves a TD(λ) critic with $\lambda < 1$, the critic can only compute an approximate—rather than exact—projection. The use of additional features can result in a reduction of the approximation error.

To avoid the above first possibility, we choose features for the critic so that our next assumption is satisfied. To understand that assumption, note that if the functions $\underline{1}$ and ϕ_θ^j, $j = 1, \ldots, m$, are linearly independent for each θ, then there exists a positive function $a(\theta)$ such that

$$||r'\hat{\phi}_\theta||_\theta^2 \geq a(\theta)|r|^2,$$

where $|r|$ is the Euclidean norm of r and $\hat{\phi}_\theta$ is the projection of ϕ_θ on the subspace orthogonal to the function $\underline{1}$. (Here and throughout the rest of the paper, $\underline{1}$ stands for a function which is identically equal to 1.) Our assumption below involves the stronger requirement that the function $a(\cdot)$ be uniformly bounded away from zero.

Assumption 3.2. There exists $a > 0$, such that for every $r \in \mathbb{R}^m$ and $\theta \in \mathbb{R}^n$

$$||r'\hat{\phi}_\theta||_\theta^2 \geq a|r|^2,$$

where

$$\hat{\phi}_\theta(x,u) = \phi_\theta(x,u) - \sum_{\bar{x},\bar{u}} \eta_\theta(\bar{x},\bar{u})\phi_\theta(\bar{x},\bar{u}).$$

Along with the parameter vector r, the critic stores some auxiliary parameters: a scalar estimate α of the average cost and an m-vector $\hat{Z}$ which represents Sutton's eligibility trace [5, 19]. The actor and critic updates take place in the course of a simulation of a single sample path of the Markov decision process. Let $r_k, \hat{Z}_k, \alpha_k$ be the parameters of the critic, and let θ_k be the parameter vector of the actor, at time k. Let $(\hat{X}_k, \hat{U}_k)$ be the state-action pair at that time. Let $\hat{X}_{k+1}$ be the new state, obtained after action $\hat{U}_k$ is applied. A new action $\hat{U}_{k+1}$ is generated according to the RSP corresponding to the actor parameter vector θ_k. The critic carries out an update similar to the average cost TD method of [22]:

$$\begin{aligned} \alpha_{k+1} &= \alpha_k + \gamma_k(c(\hat{X}_{k+1}, \hat{U}_{k+1}) - \alpha_k), \\ r_{k+1} &= r_k + \gamma_k d_k \hat{Z}_k, \end{aligned} \tag{3.1}$$

where the TD d_k is defined by

$$d_k = c(\hat{X}_k, \hat{U}_k) - \alpha_k + r_k' \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) - r_k' \phi_{\theta_k}(\hat{X}_k, \hat{U}_k),$$

and where γ_k is a positive step-size parameter. The two variants of the critic differ in their update of $\hat{Z}_k$, which is as follows.

TD(1) *critic.*

$$\begin{aligned} \hat{Z}_{k+1} &= \hat{Z}_k + \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) \qquad \text{if} \quad \hat{X}_{k+1} \neq x^* \\ &= \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) \qquad \text{otherwise,} \end{aligned}$$

where x^* is the special state introduced in Assumption 2.1.

TD(λ) *critic,* $0 < \lambda < 1$.

$$\hat{Z}_{k+1} = \lambda \hat{Z}_k + \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}).$$

Actor. Finally, the actor updates its parameter vector according to

$$\theta_{k+1} = \theta_k - \beta_k \Gamma(r_k) r_k' \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) \psi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}), \tag{3.2}$$

where $\Gamma(\cdot)$ is a scalar that controls the step-size β_k of the actor, taking into account the current estimate r_k of the critic.

Note that we have used $\hat{X}_k$, $\hat{U}_k$, and $\hat{Z}_k$ to denote the simulated processes in the above algorithm. Throughout the paper we will use hats to denote the simulated processes that are used to update the parameters in the algorithm, and X_k, U_k, and Z_k to denote processes in which a fixed RSP θ is used.

To understand the actor update, recall the formulas (2.1) and (2.2). According to these formulas, if the projection $\hat{Q}_\theta$ of Q_θ onto the subspace Φ_θ (which contains Ψ_θ) was known for the current value of $\theta \in \mathbb{R}^n$, then $\hat{Q}_{\theta_k}(\hat{X}_k, \hat{U}_k)\psi_{\theta_k}(\hat{X}_k, \hat{U}_k)$ would be a reasonable estimate of $\nabla \bar{\alpha}(\theta_k)$, because the steady-state expected value of the former is equal to the latter. However, $\hat{Q}_{\theta_k}(\hat{X}_k, \hat{U}_k)$ is not known, and it is natural to use in its place the critic's current estimate, which is $Q^{r_k}_{\theta_k}(\hat{X}_k, \hat{U}_k) = r_k' \phi_\theta(\hat{X}_k, \hat{U}_k)$. For the above scheme to converge, it is then important that the critic's estimate be accurate (at least asymptotically). This will indeed be established in section 5, under the following assumption on the step-sizes.

Assumption 3.3.

(a) The step-sizes β_k and γ_k are deterministic and nonincreasing and satisfy

$$\sum_k \beta_k = \sum_k \gamma_k = \infty,$$

$$\sum_k \beta_k^2 < \infty, \quad \sum_k \gamma_k^2 < \infty, \quad \text{and} \quad \sum_k \left(\frac{\beta_k}{\gamma_k}\right)^d < \infty$$

for some $d > 0$.

(b) The function $\Gamma(\cdot)$ is assumed to satisfy the following inequalities for some positive constants $C_1 < C_2$:

$$|r|\Gamma(r) \in [C_1, C_2] \quad \forall r \in \mathbb{R}^m,$$
$$|\Gamma(r) - \Gamma(\hat{r})| \leq \frac{C_2|r - \hat{r}|}{1 + |r| + |\hat{r}|} \quad \forall r, \hat{r} \in \mathbb{R}^n. \tag{3.3}$$

The following result on the convergence properties of the actor is established in section 6 in much greater generality.

THEOREM 3.4. *Under Assumptions* 2.1 *and* 3.1–3.3*, the following hold.*

(a) *In the actor-critic algorithm with a TD*(1) *critic,* $\liminf_k |\nabla\bar{\alpha}(\theta_k)| = 0$*, w.p.*1.

(b) *For each* $\epsilon > 0$*, there exists* λ *sufficiently close to* 1 *such that, in the actor-critic algorithm with a TD*(λ) *critic,* $\liminf_k |\nabla\bar{\alpha}(\theta_k)| < \epsilon$*, w.p.*1.

The algorithms introduced in this section are only two out of many possible variations. For instance, one can also consider "episodic" problems, in which one starts from a given initial state x^* and runs the process until a random termination time (at which time the process is reinitialized at x^*), with the objective of minimizing the expected total cost until termination. In this setting, the average cost estimate α_k is unnecessary and is removed from the critic update formula. If the critic parameter r_k were to be reinitialized each time that x^* is entered, one would obtain a method closely related to Williams's REINFORCE algorithm [23]. Such a method does not involve any value function learning, because the observations during one episode do not affect the critic parameter r during another episode. In contrast, in our approach, the observations from all past episodes affect the current critic parameter r, and in this sense, the critic is "learning." This can be advantageous because, as long as θ is changing slowly, the observations from recent episodes carry useful information on the Q-value function under the current policy.

The analysis of actor-critic methods for total and/or discounted cost problems is similar to (in fact, a little simpler than) that for the average cost case; see [20, 11].

4. Algorithms for Polish state and action spaces. In this section, we consider actor-critic algorithms for Markov decision processes with Polish (complete, separable, metric) state and action spaces. The algorithms are the same as for the case of finite state and action spaces and therefore will not be repeated in this section. However, we will restate our assumptions in the general setting, as the notation and the theory is quite technical. Throughout, we will use the abbreviation *w.p.*1 for the phrase *with probability* 1. We will denote norms on real Euclidean spaces with $|\cdot|$ and norms on Hilbert spaces by $||\cdot||$. For a probability measure ν and a ν-integrable function f, $\nu(f)$ will denote the expectation of f with respect to ν. Finally, for any Polish space $\mathbb{X}$, $\mathcal{B}(\mathbb{X})$ denotes its countably generated Borel σ-field.

4.1. Preliminaries. Consider a Markov decision process in which the state space $\mathbb{X}$ and the action space $\mathbb{U}$ are Polish spaces, and with a transition kernel $p(dy|x, u)$ which for every (x, u) defines a probability measure on $\mathbb{X}$. In the finite case, we had considered a parameterized family of randomized stationary policies (RSPs) described by a parameterized family of probability mass functions. Similarly, we now consider a family of parameterized RSPs specified by a parameterized family

of probability density functions. More specifically, let ν be a fixed measure on the action space $\mathbb{U}$. Let $\{\mu_\theta; \theta \in \mathbb{R}^n\}$ be a family of positive measurable functions on $\mathbb{X} \times \mathbb{U}$ such that for each $x \in \mathbb{X}$, $\mu_\theta(\cdot|x)$ is a probability density function with respect to $\nu(du)$, i.e.,

$$\int \mu_\theta(u|x)\nu(du) = 1 \quad \forall x, \theta.$$

This parameterized family of density functions can be viewed as a parameterized family of RSPs where, for each $\theta \in \mathbb{R}^n$, the probability distribution of an action at state x under RSP θ is given by $\mu_\theta(u|x)\nu(du)$.

Note that the state-action process $\{X_k, U_k\}$ of a Markov decision process controlled by any fixed RSP is a Markov chain. For each θ, let $\mathbf{P}_{\theta,x}$ denote the probability law of the state-action process $\{X_k, U_k\}$ in which the starting state X_0 is x. Let $\mathbf{E}_{\theta,x}$ denote expectation with respect to $\mathbf{P}_{\theta,x}$.

Assumption 4.1 (irreducibility and aperiodicity). For each $\theta \in \mathbb{R}^n$, the process $\{X_k\}$ controlled by RSP θ is irreducible and aperiodic.

For the details on the notion of irreducibility for general state space Markov chains, see [17]. Under Assumption 4.1, it follows from Theorem 5.2.2 of [17] that for each $\theta \in \mathbb{R}^n$, there exists a set of states $\mathbb{X}_0(\theta) \in \mathcal{B}(\mathbb{X})$, a positive integer $N(\theta)$, a constant $\delta_\theta > 0$, and a probability measure ϑ_θ on $\mathbb{X}$, such that $\vartheta_\theta(\mathbb{X}_0(\theta)) = 1$ and

$$\mathbf{P}_{\theta,x}(X_{N(\theta)} \in B) \geq \delta_\theta \vartheta_\theta(B) \qquad \forall \theta \in \mathbb{R}^n, \quad x \in \mathbb{X}_0(\theta), \quad B \in \mathcal{B}(\mathbb{X}).$$

We will now assume that such a condition holds uniformly in θ. This is one of the most restrictive of our assumptions. It corresponds to a "stochastic stability" condition, which holds uniformly over all policies.

Assumption 4.2 (uniform geometric ergodicity).

(a) There exists a positive integer N, a set $\mathbb{X}_0 \in \mathcal{B}(\mathbb{X})$, a constant $\delta > 0$, and a probability measure ϑ on $\mathbb{X}$, such that

$$\mathbf{P}_{\theta,x}(X_N \in B) \geq \delta \vartheta(B) \qquad \forall \theta \in \mathbb{R}^n, \quad x \in \mathbb{X}_0, \quad B \in \mathcal{B}(\mathbb{X}). \tag{4.1}$$

(b) There exists a function $L : \mathbb{X} \to [1, \infty)$ and constants $0 \leq \rho < 1$, $b > 0$, such that, for each $\theta \in \mathbb{R}^n$,

$$\mathbf{E}_{\theta,x}[L(X_1)] \leq \rho L(x) + b I_{\mathbb{X}_0}(x) \qquad \forall x \in \mathbb{X}, \tag{4.2}$$

where $I_{\mathbb{X}_0}(\cdot)$ is the indicator function of the set $\mathbb{X}_0$. We call a function L satisfying the above condition a stochastic Lyapunov function.

We note that in the finite case, Assumption 2.1(d) implies that Assumption 4.2 holds. Indeed, the first part of Assumption 4.2 is immediate, with $\mathbb{X}_0 = \{x^*\}$, $\delta_\theta = \epsilon_0$, and ϑ equal to a point mass at state x^*. To verify the second part, consider the first hitting time τ of the state x^*. For a sequence $\{\theta_k\}$ of values of the actor parameter, consider the time-varying Markov chain obtained by using policy θ_k at time k. For $s > 1$, consider the function

$$L(x) = \sup_{\{\theta_k\}} E\left[s^\tau \middle| X_0 = x \right].$$

Assumption 2.1(d) guarantees that $L(\cdot)$ is finite when s is sufficiently close to 1. Then it is a matter of simple algebraic calculations to see that $L(\cdot)$ satisfies (4.2).

Using geometric ergodicity results (Theorem 15.0.1) in [17], it can be shown that if Assumption 4.2 is satisfied, then for each $\theta \in \mathbb{R}^n$ the Markov chains $\{X_k\}$ and $\{X_k, U_k\}$ have steady-state distributions $\pi_\theta(dx)$ and

$$\eta_\theta(dx, du) = \pi_\theta(dx)\mu_\theta(u|x)\nu(du),$$

respectively. Moreover, the steady state is reached at a geometric rate (see Lemma 4.3 below). For any $\theta \in \mathbb{R}^n$, we will use $\langle \cdot, \cdot \rangle_\theta$ and $|| \cdot ||_\theta$ to denote the inner product and the norm, respectively, on $\mathcal{L}^2(\eta_\theta)$. Finally, for any $\theta \in \mathbb{R}^n$, we define the operator P_θ on $\mathcal{L}^2(\eta_\theta)$ by

$$\begin{aligned}(P_\theta Q)(x,u) &= \mathbf{E}_\theta[Q(X_1, U_1) \mid X_0 = x, U_0 = u] \\ &= \int Q(y, \bar{u})\mu_\theta(\bar{u}|y)p(dy|x,u)\nu(d\bar{u}) \quad \forall (x,u) \in \mathbb{X} \times \mathbb{U},\ Q \in \mathcal{L}^2(\eta_\theta).\end{aligned}$$

For the finite case, we introduced certain boundedness assumptions on the maps $\theta \mapsto \psi_\theta(x,u)$ and $\theta \mapsto \phi_\theta(x,u)$ and their derivatives. For the more general case considered here, these bounds may depend on the state-action pair (x,u). We wish to bound the rate of growth of such functions, as (x,u) changes, in terms of the stochastic Lyapunov function L. Toward this purpose, we introduce a class $\mathcal{D}$ of functions that satisfy the desired growth conditions.

We will say that a parameterized family of functions $f_\theta : \mathbb{X} \times \mathbb{U} \mapsto \mathbb{R}$ belongs to $\mathcal{D}$ if there exists a function $q : \mathbb{X} \times \mathbb{U} \mapsto \mathbb{R}$ and constants C, K_d $(d \geq 1)$, such that

$$f_\theta(x,u) \leq Cq(x,u) \qquad \forall\ x \in \mathbb{X},\ u \in \mathbb{U},\ \theta \in \mathbb{R}^n$$

and

$$\mathbf{E}_{\theta,x}\left[|q(x, U_0)|^d\right] \leq K_d L(x) \qquad \forall\ \theta, x, d \geq 1.$$

For easy reference, we collect here various useful properties of the class $\mathcal{D}$. The proof is elementary and is omitted.

LEMMA 4.3. *Consider a process $\{\hat{X}_k, \hat{U}_k\}$ driven by RSPs θ_k which change with time but in a nonanticipative manner (i.e., θ_k is completely determined by $(\hat{X}_l, \hat{U}_l)$, $l \leq k$). Assume that $\mathbf{E}[L(\hat{X}_0)] < \infty$.*

(a) *The sequence $\mathbf{E}[L(\hat{X}_k)]$, $k = 1, 2, \ldots$, is bounded.*

(b) *If the parametric class of functions f_θ belongs to $\mathcal{D}$, then for any $d \geq 1$ and any (possibly random) sequence $\{\tilde{\theta}_k\}$*

$$\sup_k \mathbf{E}\left[\left|f_{\tilde{\theta}_k}(\hat{X}_k, \hat{U}_k)\right|^d\right] < \infty.$$

(c) *In particular, the above boundedness property holds when θ_k and $\tilde{\theta}_k$ are held fixed at some θ, for all k, so that the process $\{\hat{X}_k, \hat{U}_k\}$ is time-homogeneous.*

(d) *If $f_\theta \in \mathcal{D}$, then the maps $(x,u) \to \mathbf{E}_{\theta,x}[f_\theta(x, U_0)]$ and $(x,u) \to (P_\theta f_\theta)(x,u)$ also belong to $\mathcal{D}$, and*

$$f_\theta \in \mathcal{L}^d(\eta_\theta) \qquad \forall \theta \in \mathbb{R}^n,\ d \geq 1.$$

(e) *For any function $f \in \mathcal{D}$, the steady-state expectation $\pi_\theta(f)$ is well-defined and a bounded function of θ, and there exists a constant $C > 0$ such that*

$$|\mathbf{E}_{\theta,x}[f(X_k, U_k)] - \pi_\theta(f)| \leq C\rho^k L(x) \qquad \forall x \in \mathbb{X},\ \theta \in \mathbb{R}^n. \tag{4.3}$$

(f) *If the parametric classes of functions f_θ and g_θ belong to $\mathcal{D}$, then*

$$f_\theta + g_\theta \in \mathcal{D}, \qquad f_\theta g_\theta \in \mathcal{D}.$$

The next two assumptions will be used to show that the average cost is a smooth function of the policy parameter θ. In the finite case, their validity is an automatic consequence of Assumption 2.1.

Assumption 4.4 (differentiability).

(a) For every $x \in \mathbb{X}$, $u \in \mathbb{U}$, and $\theta \in \mathbb{R}^n$, we have $\mu_\theta(u|x) > 0$.

(b) The mapping $\theta \mapsto \mu_\theta(u|x)$ is twice differentiable. Furthermore, $\psi_\theta(x,u) = \nabla \ln \mu_\theta(u|x)$ and its derivative belong to $\mathcal{D}$.

(c) For every θ_0, there exists $\epsilon > 0$ such that the class of functions

$$\{\nabla \mu_\theta(u|x)/\mu_{\bar\theta}(u|x),\ |\theta - \theta_0| \le \epsilon,\ |\bar\theta - \theta_0| \le \epsilon\}$$

(parameterized by θ and $\bar\theta$) belongs to $\mathcal{D}$.

Assumption 4.5. The cost function $c(\cdot,\cdot)$ belongs to $\mathcal{D}$.

Under the above assumptions we wish to prove that a gradient formula similar to (2.1) is again valid. By Assumption 4.5 and Lemma 4.3, $c \in \mathcal{L}^2(\eta_\theta)$ and therefore the average cost function can be written as

$$\bar\alpha(\theta) = \int c(x,u)\pi_\theta(dx)\mu_\theta(u|x)\nu(du) = \langle c, \underline{1}\rangle_\theta.$$

We say that $Q \in \mathcal{L}^2(\eta_\theta)$ is a solution of the Poisson equation with parameter θ if Q satisfies

$$Q = c - \bar\alpha(\theta)\underline{1} + P_\theta Q. \tag{4.4}$$

Using Proposition 17.4.1 from [17], one can easily show that a solution to the Poisson equation with parameter θ exists and is unique up to a constant. That is, if Q_1, Q_2 are two solutions, then $Q_1 - Q_2$ and $\underline{1}$ are collinear in $\mathcal{L}^2(\eta_\theta)$. One obvious family of solutions to the Poisson equation is

$$Q_\theta(x,u) = \sum_{k=0}^{\infty} \mathbf{E}_{\theta,x}\left[\,(c(X_k,U_k) - \bar\alpha(\theta))|\, U_0 = u\right].$$

(The convergence of the above series is a consequence of (4.3).)

There are other (e.g., regenerative) representations of solutions to the Poisson equation which are useful both for analysis and for derivation of algorithms. For example, Glynn and L'Ecuyer [9] use regenerative representations to show that the steady-state expectation of a function is differentiable under certain assumptions. We use similar arguments to prove that the average cost function $\bar\alpha(\cdot)$ is twice differentiable with bounded derivatives. Furthermore, it can be shown that there exist solutions $\hat{Q}_\theta(x,u)$ to the Poisson equation that are differentiable in θ. From a technical point of view, our assumptions are similar to those provided by Glynn and L'Ecuyer [9]. The major difference is that [9] concerns Markov chains $\{X_k\}$ that have the recursive representation

$$X_{k+1} = f(X_k, W_k),$$

where W_k are i.i.d., whereas we allow the distribution of W_k (which is U_k in our case) to depend on X_k. Furthermore, the formula for the gradient of steady-state

expectations that we derive here is quite different from that of [9] and makes explicit the role of the Poisson equation in gradient estimation. The following theorem holds for any solution $Q_\theta : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ of the Poisson equation with parameter θ. We provide only an outline of the proof and refer the reader to [15] for the details.

THEOREM 4.6. *Under Assumptions* 4.1, 4.2, 4.4, *and* 4.5,

$$\nabla \bar{\alpha}(\theta) = \langle \psi_\theta, Q_\theta \rangle_\theta.$$

Furthermore, $\nabla \bar{\alpha}(\theta)$ *has bounded derivatives.*

Proof. (Outline) Using regenerative representations and likelihood ratio methods, we can show that $\bar{\alpha}(\theta)$ is differentiable and that there exists a parameterized family $\{\hat{Q}_\theta(x,u)\}$ of solutions to the Poisson equation, belonging to $\mathcal{D}$, such that the map $\theta \to \hat{Q}_\theta(x,u)$ is differentiable for each (x,u), and such that the family of functions $\nabla \hat{Q}_\theta(x,u)$ belongs to $\mathcal{D}$ (see [15]). Then one can differentiate both sides of equation (4.4) with respect to θ to obtain

$$\nabla \bar{\alpha}(\theta) \underline{1} + \nabla \hat{Q}_\theta = P_\theta(\psi_\theta \hat{Q}_\theta) + P_\theta(\nabla \hat{Q}_\theta).$$

(This step involves an interchange of differentiation and integration justified by uniform integrability.) Taking inner product with $\underline{1}$ on both sides of the above equation and using that $\nabla \hat{Q}_\theta \in \mathcal{L}^2(\eta_\theta)$ and

$$\langle \underline{1}, P_\theta f \rangle_\theta = \langle \underline{1}, f \rangle_\theta \qquad \forall f \in \mathcal{L}^2(\eta_\theta),$$

we obtain $\nabla \bar{\alpha}(\theta) = \langle \hat{Q}_\theta, \psi_\theta \rangle_\theta = \langle Q_\theta, \psi_\theta \rangle_\theta$, where the second equality follows from the fact that $Q_\theta - \hat{Q}_\theta$ and $\underline{1}$ are necessarily collinear and the easily verified fact $\langle \underline{1}, \psi_\theta \rangle_\theta = 0$.

Since ψ_θ and $\hat{Q}_\theta$ are both differentiable with respect to θ, with the derivatives belonging to $\mathcal{D}$, the formula

$$\nabla \bar{\alpha}(\theta) = \langle \psi_\theta, Q_\theta \rangle_\theta = \langle \underline{1}, \psi_\theta Q_\theta \rangle$$

implies that $\nabla \bar{\alpha}(\theta)$ is also differentiable with bounded derivative. □

Before we move on to present the algorithms for Polish state and action spaces, we illustrate how the above assumptions can be verified in the context of a simple inventory control problem.

Example 4.7. Consider a facility with $X_k \in \mathbb{R}$ amount of stock at the beginning of the kth period, with negative stock representing the unsatisfied (or backlogged) demand. Let $D_k \geq 0$ denote the random demand during the kth period. The problem is to determine the amount of stock to be ordered at the beginning of the kth period, based on the current stock and the previous demands. If $U_k \geq 0$ represents the amount of stock ordered at the beginning of the kth period, then the cost incurred is assumed to be

$$c(X_k, U_k) = h \max(0, X_k) + b \max(0, -X_k) + pU_k,$$

where p is the price of the material per unit, b is the cost incurred per unit of backlogged demand, and h is the holding cost per unit of stock in the inventory. Moreover, the evolution of the stock X_k is given by

$$X_{k+1} = X_k + U_k - D_k, \quad k = 0, 1, \ldots.$$

If we assume that the demands $D_k, k = 0, 1 \ldots$, are nonnegative and i.i.d. with finite mean, then it is well known (e.g., see [4]) that there is an optimal policy μ^* of the form

$$\mu^*(x) = \max(S - x, 0)$$

for some $S > 0$ depending on the distribution of D_k. A good approximation for policies having the above form is the family of randomized policies in which S is chosen at random from the density

$$p_\theta(s) = \frac{1}{2T}\text{sech}^2\left(\frac{s - \bar{s}(\theta)}{T}\right),$$

where $\bar{s}(\theta) = e^\theta C/(1 + e^\theta)$. The constant C is picked based on our prior knowledge of an upper bound on the parameter S in an optimal policy. To define the family of density functions $\{\mu_\theta\}$ for the above family of policies, let $\nu(du)$ be the sum of the Dirac measure at 0 and the Lebesgue measure on $[0, \infty)$. Then the density functions are given by

$$\mu_\theta(0|x) = \frac{1}{2}\left(1 + \tanh\left(\frac{x - \bar{s}(\theta)}{T}\right)\right),$$
$$\mu_\theta(u|x) = \frac{1}{2T}\text{sech}^2\left(\frac{x + u - \bar{s}(\theta)}{T}\right), \qquad u > 0.$$

The dynamics of the stock in the inventory, when controlled by policy μ_θ, are described by

$$X_{k+1} = \max(X_k, S_k) - D_k, \qquad k = 0, 1 \ldots,$$

where the $\{S_k\}$ are i.i.d. with density p_θ and independent of the demands D_k and the stock X_k. It is easy to see that the Markov chain $\{X_k\}$ is irreducible. To prove that the Markov chain is aperiodic, it suffices to show that (4.1) holds with $N = 1$. Indeed, for $\mathbb{X}_0 = [-a, a]$, $x \in \mathbb{X}_0$, and a Borel set B consider

$$\begin{aligned}\mathbf{P}_{\theta,x}(X_1 \in B) &= \mathbf{P}_{\theta,x}(\max(x, S_0) - D_0 \in B),\\ &\geq \mathbf{P}_{\theta,x}(S_0 - D_0 \in B, S_0 \geq a),\\ &\geq \int_B \int_{a-t}^{\infty} \left(\inf_\theta p_\theta(t + y)\right) D(dy)dt,\end{aligned}$$

where $D(dy)$ is the probability distribution of D_0 and $\vartheta(dy)$ is the right-hand side appropriately normalized. This normalization is possible because the above integral is positive when $B = \mathbb{X}_0$.

To prove the Lyapunov condition (4.2), assume that D_k has exponentially decreasing tails. In other words, assume that there exists $\gamma > 0$ such that

$$\mathbf{E}[\exp(\gamma D_0)] < \infty.$$

We first argue intuitively that the function

$$L(x) = \exp(\bar{\gamma}|x|)$$

for some $\bar{\gamma}$ with $\min(\gamma, \frac{1}{T}) > \bar{\gamma} > 0$ is a good candidate Lyapunov function. To see this, note that the desired inequality (4.2) requires the Lyapunov function to decrease

by a common factor outside some set $\mathbb{X}_0$. Let us try the set $\mathbb{X}_0 = [-a, a]$ for a sufficiently larger than C. If the inventory starts with a stock larger than a, then no stock is ordered with very high probability (since S_0 is most likely less than C) and therefore the stock decreases by D_0, decreasing the Lyapunov function by a factor of $\mathbf{E}[\exp(-\bar{\gamma}D_0)] < 1$. If the inventory starts with a large backlogged demand, then most likely new stock will be ordered to satisfy all the backlogged demand decreasing the Lyapunov function to almost 1. This can be made precise as follows:

$$\begin{aligned}\mathbf{E}_{\theta,x}[L(X_1)] &= \mathbf{E}_{\theta,x}[\exp(\bar{\gamma}|\max(x, S_0) - D_0|)] \\ &= \exp(\bar{\gamma}x)\mathbf{P}_{\theta,x}(S_0 \le x)\mathbf{E}_{\theta,x}[\exp(-\bar{\gamma}D_0); D_0 \le x] \\ &\quad + \exp(-\bar{\gamma}x)\mathbf{P}_{\theta,x}(S_0 \le x)\mathbf{E}_{\theta,x}[\exp(\bar{\gamma}D_0); D_0 > x] \\ &\quad + \mathbf{E}_{\theta,x}[\exp(\bar{\gamma}|S_0 - D_0|); S_0 > x].\end{aligned}$$

Note that the third term is bounded uniformly in θ, x since $\bar{\gamma} < \min(\frac{1}{T}, \gamma)$. The first term is bounded when x is negative, and the second term is bounded when x is positive. Therefore the Lyapunov function decreases by a factor of $\mathbf{E}[\exp(-\bar{\gamma}D_0)] < 1$ when $x > a$ and decreases by a factor of $\mathbf{P}(S_0 \le -a)\mathbf{E}[\exp(\bar{\gamma}D_0)] < 1$ for a sufficiently large. The remaining assumptions are easy to verify.

4.2. Critic. In the finite case, the feature vectors were assumed to be bounded. This assumption is seldom satisfied for infinite state spaces. However, it is reasonable to impose some bounds on the growth of the feature vectors, as in the next assumption.

Assumption 4.8 (critic features).

(a) The family of functions $\phi_\theta(x, u)$ belongs to $\mathcal{D}$.

(b) For each (x, u), the map $\theta \mapsto \phi_\theta(x, u)$ is differentiable, and the family of functions $\nabla\phi_\theta(x, u)$ belongs to $\mathcal{D}$.

(c) There exists some $a > 0$, such that

$$||r'\hat{\phi}_\theta||_\theta^2 \ge a|r|^2 \qquad \forall\theta \in \mathbb{R}^n, r \in \mathbb{R}^m, \tag{4.5}$$

where $\hat{\phi}_\theta = \phi_\theta - \langle\phi_\theta, \underline{1}\rangle_\theta \underline{1}$.

(d) For each $\theta \in \mathbb{R}^n$, the subspace Φ_θ in $\mathcal{L}^2(\eta_\theta)$ spanned by the features ϕ_θ^i, $i = 1, \dots, m$, of the critic contains the subspace Ψ_θ spanned by the functions ψ_θ^j, $j = 1, \dots, n$, i.e.,

$$\Phi_\theta \supset \Psi_\theta \quad \forall\theta \in \mathbb{R}^n.$$

4.2.1. TD(1) critic. For the TD(1) critic, we will strengthen Assumption 4.2 by adding the following condition.

Assumption 4.9. The set $\mathbb{X}_0$ consists of a single state x^*, and

$$\mathbf{E}_{\theta,x^*}[\phi_\theta(x^*, U_0)] = 0 \quad \forall\theta \in \mathbb{R}^n.$$

The requirement that there is a single state that is hit with positive probability is quite strong but is satisfied in many practical situations involving queuing systems, as well as for systems that have been made regenerative using the splitting techniques of [1] and [18]. The assumption that the expected value of the features at x^* is zero is automatically satisfied in the special case where $\phi_\theta = \psi$. Furthermore, for features of the form $\phi_\theta(x)$ that do not depend on u, the assumption is easily satisfied by enforcing the condition $\phi_\theta(x^*) = 0$. It is argued in [11] that besides ψ_θ, there is little benefit in using additional features that depend on u. Therefore, the assumption imposed here is not a major restriction.

5. Convergence of the critic. In this section, we analyze the convergence of the critic in the algorithms described above, under the assumptions introduced in section 4, together with Assumption 3.3 on the step-sizes. If θ_k was held constant at some value θ, it would follow (similar to [22], which dealt with the finite case) that the critic parameters converge to some $\bar{r}(\theta)$. In our case, θ_k changes with k, but slowly, and this will allow us to show that $r_k - \bar{r}(\theta_k)$ converges to zero. To establish this, we will cast the update of the critic as a linear stochastic approximation driven by Markov noise, specifically in the form of (A.1) in Appendix A. We will show that the critic update satisfies all the hypotheses of Theorem A.7 of Appendix A, and the desired result (Theorem 5.7) will follow. The assumptions of the result in Appendix A are similar to the assumptions of a result (Theorem 2) used in [22]. Therefore, the proof we present here is similar to that in [22], modulo the technical difficulties due to more general state and action spaces. We start with some notation.

For each time k, let

$$\hat{Y}_{k+1} = (\hat{X}_k, \hat{U}_k, \hat{Z}_k),$$
$$R_k = \begin{pmatrix} L\alpha_k \\ r_k \end{pmatrix}$$

for some deterministic constant $L > 0$, whose purpose will be clear later. Let $\mathcal{F}_k$ be the σ-field generated by $\{Y_l, R_l, \theta_l,\ l \leq k\}$. For $y = (x, u, z)$, define

$$h_\theta(y) = \begin{pmatrix} Lc(x,u) \\ zc(x,u) \end{pmatrix},$$

$$G_\theta(y) = \begin{pmatrix} 1 & 0 \\ z/L & \tilde{G}_\theta(y) \end{pmatrix},$$

where

$$\tilde{G}_\theta(y) = z\left(\phi'_\theta(x,u) - (P_\theta \phi_\theta)'(x,u)\right).$$

It will be shown later that the steady-state expectation of $\tilde{G}_\theta(y)$ is positive definite. The constant L is introduced because when it is chosen small enough, we will be able to show that the steady-state expectation of $G_\theta(y)$ is also positive definite.

The update (3.1) for the critic can be written as

$$R_{k+1} = R_k + \gamma_k(h_{\theta_k}(\hat{Y}_{k+1}) - G_{\theta_k}(\hat{Y}_{k+1})R_k + \xi_k R_k),$$

which is a linear iteration with Markov-modulated coefficients and ξ_k is a martingale difference given by

$$\xi_k = \begin{bmatrix} 0 \\ \hat{Z}_k\left(\phi'_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) - (P_{\theta_k}\phi'_{\theta_k})(\hat{X}_k, \hat{U}_k)\right) \end{bmatrix}.$$

To apply Theorem A.7 to this update equation, we need to prove that it satisfies Assumptions A.1–A.6. We will verify these assumptions for the two cases $\lambda = 1$ and $\lambda < 1$ separately.

Assumption A.1 follows from our Assumption 3.3. Assumption A.2 is trivially satisfied. To verify Assumption A.4, we use the actor iteration (3.2) to identify H_{k+1} with $\Gamma(r_k) r'_k \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1})\psi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1})$. Because of Assumption 3.3(b), the

term $\Gamma(r_k)r_k$ is bounded. Furthermore, since ψ_θ and ϕ_θ belong to $\mathcal{D}$ (Assumptions 4.4 and 4.8), Lemma 4.3(b) implies that $\mathbf{E}[|H_k|^d]$ is bounded. This, together with Assumption 3.3(a), shows that Assumption A.4 is satisfied. In the next two subsections, we will concentrate on showing that Assumptions A.3, A.5, and A.6 are satisfied.

5.1. TD(1) critic. Define a process Z_k in terms of the process $\{X_k, U_k\}$ of section 4.1 (in which the policy is fixed) as follows:

$$Z_0 = \phi_\theta(X_0, U_0), \quad Z_{k+1} = I\{X_{k+1} \neq x^*\}Z_k + \phi_\theta(X_{k+1}, U_{k+1}),$$

where I is the indicator function. Note that the process $\{Z_k\}$ depends on the parameter θ. Whenever we use this process inside an expectation or a probability measure, we will assume that the parameter of this process is the same as the parameter of the probability or expectation. It is easy to see that $Y_{k+1} = (X_k, U_k, Z_k)$ is a Markov chain. Furthermore, the transition kernel of this process, when the policy parameter is θ, is the same as that of $\{\hat{Y}_k\}$ when the actor parameter is fixed at θ.

Let τ be the stopping time defined by

$$\tau = \min\{k > 0 \mid X_k = x^*\}.$$

For any $\theta \in \mathbb{R}^n$, define T_θ and Q_θ by

$$T_\theta(x,u) = \mathbf{E}_{\theta,x}[\tau \mid U_0 = u],$$
$$Q_\theta(x,u) = \mathbf{E}_{\theta,x}\left[\sum_{k=0}^{\tau-1}(c(X_k,U_k) - \bar{\alpha}(\theta)) \;\Big|\; U_0 = u\right].$$

LEMMA 5.1. *The families of functions T_θ and Q_θ both belong to $\mathcal{D}$.*

Proof. The fact that $T_\theta \in \mathcal{D}$ follows easily from the assumption that $\mathbb{X}_0 = x^*$ (Assumption 4.9) and the uniform ergodicity Assumption 4.2. Using Theorem 15.2.5 of [17], we obtain that $\mathbf{E}_{\theta,x}[Q_\theta(x,U_0)]^d \leq K'_d L(x)$ for some $K'_d > 0$, so that $\mathbf{E}_{\theta,x}[Q_\theta(x,U_0)]^d$ also belongs to $\mathcal{D}$. Since

$$Q_\theta(x,u) = c(x,u) - \bar{\alpha}(\theta) + \mathbf{E}_{\theta,x}[Q_\theta(X_1,U_1)|U_0 = u]$$

is a sum of elements of $\mathcal{D}$, it follows that Q_θ also belongs to $\mathcal{D}$. □

Using simple algebraic manipulations and Assumption 4.9, we obtain, for every $\theta \in \mathbb{R}^n$,

$$\mathbf{E}_{\theta,x^*}\left[\sum_{k=0}^{\tau-1}\Big(\big(c(X_k,U_k) - \bar{\alpha}(\theta)\big)Z_k - \langle Q_\theta, \phi_\theta\rangle_\theta\Big)\right] = 0,$$
$$\mathbf{E}_{\theta,x^*}\left[\sum_{k=0}^{\tau-1}\Big(Z_k\big(\phi'_\theta(X_k,U_k) - \phi'_\theta(X_{k+1},U_{k+1})\big) - \langle \phi_\theta, \phi'_\theta\rangle_\theta\Big)\right] = 0.$$

This implies that the steady-state expectations of $h_\theta(y)$ and $G_\theta(y)$ are given by

$$\bar{h}(\theta) = \begin{pmatrix} L\bar{\alpha}(\theta) \\ \bar{h}_1(\theta) + \bar{\alpha}(\theta)\bar{Z}(\theta) \end{pmatrix},$$
$$\bar{G}(\theta) = \begin{pmatrix} 1 & 0 \\ \bar{Z}(\theta)/L & \bar{G}_1(\theta) \end{pmatrix},$$

where

$$\bar{h}_1(\theta) = \langle Q_\theta, \phi_\theta \rangle_\theta, \qquad \bar{Z}(\theta) = \langle T_\theta, \phi_\theta \rangle_\theta, \qquad \bar{G}_1(\theta) = \langle \phi_\theta, \phi'_\theta \rangle_\theta.$$

For $y = (x, u, z)$, we define

$$\hat{h}_\theta(y) = \mathbf{E}_{\theta,\bar{x}} \left[\sum_{k=0}^{\tau-1} (h_\theta(Y_k) - \bar{h}(\theta)) \;\Big|\; Y_0 = y \right],$$

$$\hat{G}_\theta(y) = \mathbf{E}_{\theta,\bar{x}} \left[\sum_{k=0}^{\tau-1} (G_\theta(Y_k) - \bar{G}(\theta)) \;\Big|\; Y_0 = y \right],$$

and it can be easily verified that part (a) of Assumption A.3 is satisfied. Note that we have been working with families of functions that belong to $\mathcal{D}$, and which therefore have steady-state expectations that are bounded functions of θ (Lemma 4.3(e)). In particular, $\bar{G}(\cdot)$ and $\bar{h}(\cdot)$ are bounded, and part (b) of Assumption A.3 is satisfied.

To verify the other parts of Assumption A.3, we will need the following result.

LEMMA 5.2. *For every $d > 1$, $\sup_k \mathbf{E}[|\hat{Z}_k|^d] < \infty$.*

Proof. Let $\hat{W}_k$ denote the vector $(\hat{X}_k, \hat{U}_k, \hat{Z}_k, r_k, \alpha_k, \theta_k)$. Since the step-size sequences $\{\gamma_k\}$ and $\{\beta_k\}$ are deterministic, $\{\hat{W}_k\}$ forms a time-varying Markov chain. For each k, let $\mathbf{P}_{k,\hat{w}}$ denote the conditional law of the process $\{\hat{W}_n\}$ given that $\hat{W}_k = \hat{w}$. Define a sequence of stopping times for the process $\{\hat{W}_n\}$ by letting

$$\hat{\tau}_k = \min\{n > k : \hat{X}_n = x^*\}.$$

For $1 < t < 1/\rho$, define

$$V_k^{(d)}(\hat{w}) = \mathbf{E}_{k,\hat{w}} \left[\sum_{l=k}^{\tau_k - 1} t^{l-k} (1 + |\hat{Z}_l|^d) \right],$$

which can be verified to be finite, due to uniform geometric ergodicity and the assumption that ϕ_θ belongs to $\mathcal{D}$. It is easy to see that $V_k^{(d)}(\hat{W}_k) \geq |\hat{Z}_k|^d$. Therefore, it is sufficient to prove that $\mathbf{E}[V_k^{(d)}(\hat{W}_k)]$ is bounded.

We will now show that $V_k^{(d)}(\hat{w})$ acts as a Lyapunov function for the algorithm. Indeed,

$$\begin{aligned}
V_k^{(d)}(\hat{w}) &\geq \mathbf{E}_{k,\hat{w}} \left[\sum_{l=k+1}^{\tau_k - 1} t^{l-k} (1 + |\hat{Z}_l|^d) \right] \\
&= \mathbf{E}_{k,\hat{w}} \left[\sum_{l=k+1}^{\tau_k - 1} t^{l-k} (1 + |\hat{Z}_l|^d) I\{\hat{X}_{k+1} \neq x^*\} \right] \\
&= t \mathbf{E}_{k,\hat{w}} \left[V_{k+1}^{(d)}(\hat{W}_{k+1}) I\{\hat{X}_{k+1} \neq x^*\} \right] \\
&= t \mathbf{E}_{k,\hat{w}} \left[V_{k+1}^{(d)}(\hat{W}_{k+1}) \right] - t \mathbf{E}_{k,\hat{w}} \left[V_{k+1}^{(d)}(\hat{W}_{k+1}) I\{\hat{X}_{k+1} = x^*\} \right].
\end{aligned}$$

Using the geometric ergodicity condition (4.2), some algebraic manipulations, and the fact that ϕ_θ belongs to $\mathcal{D}$, we can verify that $\mathbf{E}_{k,\hat{w}}[V_{k+1}^{(d)}(\hat{W}_1) I\{\hat{X}_1 = x^*\}]$ is bounded by some constant C. We take expectations of both sides of the preceding inequality, with $\hat{w}$ distributed as the random variable $\hat{W}_k$, and use the property

$$\mathbf{E}\Big[\mathbf{E}_{k,\hat{W}_k} \big[V_{k+1}^{(d)}(\hat{W}_{k+1}) \big] \Big] = \mathbf{E}[V_{k+1}^{(d)}(\hat{W}_{k+1})]$$

to obtain

$$\mathbf{E}[V_k^{(d)}(\hat{W}_k)] \geq t\mathbf{E}[V_{k+1}^{(d)}(\hat{W}_{k+1})] - C.$$

Since $t > 1$, $\mathbf{E}[V_k^{(d)}(\hat{W}_k)]$ is bounded, and the result follows. □

To verify part (c) of Assumption A.3, note that $\hat{h}_\theta(\cdot), \hat{G}_\theta(\cdot), h_\theta(\cdot)$, and $G_\theta(\cdot)$ are affine in z, of the form

$$f_\theta^{(1)}(\cdot) + z f_\theta^{(2)}(\cdot),$$

for some functions $f_\theta^{(i)}$ that belong to $\mathcal{D}$. Therefore, Holder's inequality and Lemma 5.2 can be used to verify part (c) of Assumption A.3. As in the proof of Theorem 4.6, likelihood ratio methods can be used to verify Assumptions parts (d) and (e) of Assumption A.3; see [15] for details. Assumption A.5 follows from Holder's inequality, Lemma 5.2, and part (b) of Lemma 4.3.

Finally, the following lemma verifies Assumption A.6.

LEMMA 5.3. *There exist L and $\epsilon > 0$ such that for all $\theta \in \mathbb{R}^n$ and $R \in \mathbb{R}^{m+1}$,*

$$R'\bar{G}(\theta)R \geq \epsilon |R|^2.$$

Proof. Let $R = (\alpha, r)$, where $\alpha \in \mathbb{R}$ and $r \in \mathbb{R}^m$. Using the definition of $\bar{G}(\theta)$, and Assumption 4.8(c) for the first inequality, we have

$$\begin{aligned} R'\bar{G}(\theta)R &= ||r'\phi_\theta||_\theta^2 + |\alpha|^2 + r'\bar{Z}(\theta)\alpha/L \\ &\geq a|r|^2 + |\alpha|^2 - r'\bar{Z}(\theta)\alpha/L \\ &\geq \min(a,1)|R|^2 - |\bar{Z}(\theta)|(|r|^2 + |\alpha|^2)/2L \\ &= \left(\min(a,1) - \frac{\bar{Z}(\theta)}{2L}\right)|R|^2. \end{aligned}$$

We can now choose $L > \sup_\theta |\bar{Z}(\theta)|/\min(a,1)$, which is possible because $\bar{Z}(\theta)$ is bounded (it is the steady-state expectation of a function in $\mathcal{D}$). □

5.2. TD(λ) critic. To analyze the TD(λ) critic, with $0 < \lambda < 1$, we redefine the process Z_k as

$$Z_{k+1} = \lambda Z_k + \phi_\theta(X_{k+1}, U_{k+1}).$$

As in the case of TD(1), we consider the steady-state expectations

$$\bar{h}(\theta) = \begin{pmatrix} L\bar{\alpha}(\theta) \\ \bar{h}_1(\theta) + \bar{\alpha}(\theta)\bar{Z}(\theta) \end{pmatrix}, \qquad \bar{G}(\theta) = \begin{pmatrix} 1 & 0 \\ \bar{Z}(\theta)/L & \bar{G}_1(\theta) \end{pmatrix}$$

of $h_\theta(Y_k)$ and $G_\theta(Y_k)$. For the present case, the entries of $\bar{h}$ and $\bar{G}$ are given by

$$\bar{h}_1(\theta) = \sum_{k=0}^{\infty} \lambda^k \langle P_\theta^k c - \bar{\alpha}(\theta)\underline{1}, \phi_\theta \rangle_\theta,$$

$$\bar{G}_1(\theta) = \langle \phi_\theta, \phi_\theta' \rangle_\theta - (1-\lambda)\sum_{k=0}^{\infty} \lambda^k \langle P_\theta^{k+1}\phi_\theta, \phi_\theta' \rangle_\theta,$$

and $\bar{Z}(\theta) = (1-\lambda)^{-1}\langle \underline{1}, \phi_\theta\rangle_\theta$. As in Assumption 4.8(c), let $\hat{\phi}_\theta = \phi_\theta - \langle \phi_\theta, \underline{1}\rangle_\theta \underline{1}$. Then, $P_\theta\phi_\theta - \phi_\theta = P_\theta\hat{\phi}_\theta - \hat{\phi}_\theta$, and $\bar{G}_1(\theta)$ can also be written as

$$\bar{G}_1(\theta) = \langle \hat{\phi}_\theta, \hat{\phi}'_\theta\rangle_\theta - (1-\lambda)\sum_{k=0}^{\infty}\lambda^k\langle P_\theta^{k+1}\hat{\phi}_\theta, \hat{\phi}'_\theta\rangle_\theta.$$

By an argument similar to the one used for the case of TD(1), we can see that $\bar{G}(\cdot)$ and $\bar{h}(\cdot)$ are bounded and, therefore, part (b) of Assumption A.3 is satisfied.

LEMMA 5.4. *There exists a positive constant C, such that for all $k \geq 0$, θ, x, λ, we have*

(a) $\left|\mathbf{E}_{\theta,x}\big[(c(X_k,U_k) - \bar{\alpha}(\theta))Z_k\big] - \bar{h}_1(\theta)\right| \leq Ck\max(\lambda,\rho)^k L(x)$,

(b) $\left|\mathbf{E}_{\theta,x}\big[Z_k(\phi'_\theta(X_k,U_k) - \phi'_\theta(X_{k+1},U_{k+1}))\big] - \bar{G}(\theta)\right| \leq Ck\max(\lambda,\rho)^k L(x)$.

Proof. We have

$$\begin{aligned}
&\left|\mathbf{E}_{\theta,x}\big[(c(X_k,U_k) - \bar{\alpha}(\theta))Z_k\big] - \bar{h}_1(\theta)\right| \\
&\quad\leq \sum_{l=0}^{k}\lambda^l\left|\mathbf{E}_{\theta,x}\big[(c(X_k,U_k) - \bar{\alpha}(\theta))\phi_\theta(X_{k-l},U_{k-l})\big] - \langle P_\theta^l c - \bar{\alpha}(\theta)\underline{1}, \phi_\theta\rangle_\theta\right| \\
&\qquad + C'\lambda^k \\
&\quad\leq \sum_{l=0}^{k} C'\lambda^l\rho^{k-l}L(x) + C'\lambda^k L(x) \\
&\quad\leq \sum_{l=0}^{k} 2C'\max(\lambda,\rho)^k L(x),
\end{aligned}$$

where the second inequality makes use of Lemma 4.3(e) and the assumption $L(x) \geq 1$. This proves part (a). The proof of part (b) is similar. □

From the previous lemma, it is clear that, for $\theta \in \mathbb{R}^n$ and $y = (x,u,z)$,

$$\begin{aligned}
\hat{h}_\theta(y) &= \sum_{k=0}^{\infty}\mathbf{E}_{\theta,x}\left[(h_\theta(Y_k) - \bar{h}(\theta))\ \middle|Y_0 = y\right], \\
\hat{G}_\theta(y) &= \sum_{k=0}^{\infty}\mathbf{E}_{\theta,x}\left[(G_\theta(Y_k) - \bar{G}(\theta))\ \middle|Y_0 = y\right],
\end{aligned}$$

are well-defined, and it is easy to check that part (a) of Assumption A.3 is satisfied.

To verify part (c) of Assumption A.3, we have the following counterpart of Lemma 5.2.

LEMMA 5.5. *For every $d > 1$, we have* $\sup_k \mathbf{E}[|\hat{Z}_k|^d] < \infty$.

Proof. We have, using Jensen's inequality,

$$\begin{aligned}
|\hat{Z}_k|^d &= \frac{1}{(1-\lambda)^d}\left|(1-\lambda)\sum_{l=0}^{k}\lambda^{k-l}\phi_{\theta_k}(\hat{X}_k,\hat{U}_k)\right|^d \\
&\leq \frac{1}{(1-\lambda)^d}(1-\lambda)\sum_{l=0}^{k}\lambda^{k-l}\left|\phi_{\theta_k}(\hat{X}_k,\hat{U})\right|^d.
\end{aligned}$$

We note that $\mathbf{E}[|\phi_{\theta_k}(\hat{X}_k, \hat{U}_k)|^d]$ is bounded (Lemma 4.3(b)), from which it follows that $\mathbf{E}[|\hat{Z}_k|^d]$ is bounded. □

The verification of parts (d) and (e) of Assumption A.3 is tedious, and we provide only an outline (see [15] for the details). The idea is to write the components of $\hat{h}_\theta(\cdot), \hat{G}_\theta(\cdot)$ that are linear in z in the form

$$\sum_{k=0}^{\infty} \lambda^k \mathbf{E}_{\theta,x}[f_\theta(Y_k) \mid U_0 = u, Z_0 = z]$$

for suitably defined functions f_θ, and show that the map $\theta \mapsto \mathbf{E}_\theta[f_\theta(Y_k) \mid U_0 = u, Z_0 = z]$ is Lipschitz continuous, with Lipschitz constant at most polynomial in k. The "forgetting" factor λ^k dominates the polynomial in k, and thus the sum will be Lipschitz continuous in θ. Assumption A.5 follows from Holder's inequality, the previous lemma and part (b) of Lemma 4.3. For the components that are not linear in z, likelihood ratio methods are used.

Finally, we will verify Assumption A.6 in the following lemma.

LEMMA 5.6. *There exist L and $\epsilon > 0$ such that, for all $\theta \in \mathbb{R}^n$ and $R \in \mathbb{R}^{m+1}$,*

$$R'\bar{G}(\theta)R \geq \epsilon |R|^2.$$

Proof. Recall the definition $\hat{\phi}_\theta = \phi_\theta - \langle \phi_\theta, \underline{1} \rangle_\theta \underline{1}$ of $\hat{\phi}_\theta$. Using Lemma 4.3(e) and the fact $\pi_\theta(\hat{\phi}_\theta) = 0$, we obtain, for some constant C,

$$||P_\theta^k \hat{\phi}_\theta^j||_\theta \leq C\rho^k \qquad \forall \theta,\ k.$$

Therefore, for any $r \in \mathbb{R}^m$, we have

$$\begin{aligned} \left\| P_\theta^k (r'\hat{\phi}_\theta) \right\|_\theta &= \left\| \sum_j r_j P_\theta^k \hat{\phi}_\theta^j \right\|_\theta \\ &\leq \sum_j |r_j| \cdot ||P_\theta^k \hat{\phi}_\theta^j||_\theta \\ &\leq C_1 \rho^k |r|. \end{aligned}$$

We note that the transition operator P_θ is nonexpanding, i.e., $\|P_\theta f\|_\theta \leq \|f\|_\theta$, for every $f \in \mathcal{L}^2(\eta_\theta)$; see, e.g., [21]. Using this property and some algebraic manipulations, we obtain

$$\begin{aligned} r'\bar{G}_1(\theta)r &= r'\langle \hat{\phi}_\theta, \hat{\phi}'_\theta \rangle_\theta r - (1-\lambda) \sum_{k=0}^{\infty} \lambda^k r' \langle P_\theta^k \hat{\phi}_\theta, \hat{\phi}'_\theta \rangle_\theta r \\ &= ||r'\hat{\phi}_\theta||_\theta^2 - (1-\lambda) \sum_{k=0}^{\infty} \lambda^k \langle P_\theta^k (r'\hat{\phi}_\theta), r'\hat{\phi}_\theta \rangle_\theta \\ &\geq ||r'\hat{\phi}_\theta||_\theta^2 - (1-\lambda) \left\{ \sum_{k=0}^{k_0-1} \lambda^k ||r'\hat{\phi}_\theta||_\theta^2 + \sum_{k \geq k_0} C_1 \lambda^k \rho^k ||r'\hat{\phi}_\theta||_\theta |r| \right\} \\ &\geq ||r'\hat{\phi}_\theta||_\theta^2 - (1-\lambda^{k_0}) ||r'\phi_\theta||_\theta^2 - C_1 (\lambda\rho)^{k_0} \frac{(1-\lambda)}{(1-\rho\lambda)} ||(r'\phi_\theta)||_\theta |r| \\ &\geq |r|^2 \lambda^{k_0} \left(a - \frac{C_2 \rho^{k_0} (1-\lambda)}{(1-\rho\lambda)} \right), \end{aligned}$$

where the last step made use of the uniform positive definiteness property (Assumption 4.8(c)). We choose k_0 so that

$$\rho^{k_0} < \frac{a(1-\rho\lambda)}{C_2(1-\lambda)}$$

and conclude that $\bar{G}_1(\theta)$ is uniformly positive definite. From this point on, the proof is identical to the proof of Lemma 5.3. □

Having verified all the hypotheses of Theorem A.7, we have proved the following result.

THEOREM 5.7. *Under Assumptions* 3.3, 4.1, 4.2, 4.4, 4.5, 4.8, and 4.9 *and for any TD critic, the sequence* R_k *is bounded, and* $\lim_k |\bar{G}(\theta_k)R_k - \bar{h}(\theta_k)| = 0$.

6. Convergence of the actor. For every $\theta \in \mathbb{R}^n$ and $(x,u) \in \mathbb{X} \times \mathbb{U}$, let

$$H_\theta(x,u) = \psi_\theta(x,u)\phi'_\theta(x,u), \qquad \bar{H}(\theta) = \langle \psi_\theta, \phi'_\theta \rangle_\theta.$$

Note that H_θ belongs to $\mathcal{D}$, and consequently $\bar{H}(\theta)$ is bounded. Let $\bar{r}(\theta)$ be such that $\bar{h}_1(\theta) = \bar{G}_1(\theta)\bar{r}(\theta)$, so that $\bar{r}(\theta)$ is the limit of the critic parameter r if the policy parameter θ was held fixed. The recursion for the actor parameter θ can be written as

$$\begin{aligned}\theta_{k+1} &= \theta_k - \beta_k H_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1})(r_k\Gamma(r_k)) \\ &= \theta_k - \beta_k \bar{H}(\theta_k)\left(\bar{r}(\theta_k)\Gamma(\bar{r}(\theta_k))\right) \\ &\quad -\beta_k(H_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) - \bar{H}(\theta_k))(r_k\Gamma(r_k)) \\ &\quad -\beta_k \bar{H}(\theta_k)(r_k\Gamma(r_k) - \bar{r}(\theta_k)\Gamma(\bar{r}(\theta_k))).\end{aligned}$$

Let

$$\begin{aligned} f(\theta) &= \bar{H}(\theta)\bar{r}(\theta), \\ e_k^{(1)} &= (H_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) - \bar{H}(\theta_k))r_k\Gamma(r_k), \\ e_k^{(2)} &= \bar{H}(\theta_k)(r_k\Gamma(r_k) - \bar{r}(\theta_k)\Gamma(\bar{r}(\theta_k))).\end{aligned}$$

Using Taylor's series expansion, one can see that

$$\begin{aligned}\bar{\alpha}(\theta_{k+1}) \le \bar{\alpha}(\theta_k) &- \beta_k\Gamma(\bar{r}(\theta))\nabla\bar{\alpha}(\theta_k)\cdot f(\theta_k) - \beta_k\nabla\bar{\alpha}(\theta_k)\cdot e_k^{(1)} \\ &-\beta_k\nabla\bar{\alpha}(\theta_k)\cdot e_k^{(2)} + C\beta_k^2\left|H_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1})(r_k\Gamma(r_k))\right|^2,\end{aligned} \tag{6.1}$$

where C reflects a bound on the Hessian of $\bar{\alpha}(\theta)$.

Note that $\bar{r}(\theta)$ and $f(\theta)$ depend on the parameter λ of the critic. The following lemma characterizes this dependence.

LEMMA 6.1. *If a TD(*λ*) critic is used, with* $0 < \lambda \le 1$, *then* $f(\theta) = \nabla\bar{\alpha}(\theta) + \varepsilon(\lambda,\theta)$, *where* $\sup_\theta |\varepsilon(\lambda,\theta)| \le C(1-\lambda)$, *and where the constant* $C > 0$ *is independent of* λ.

Proof. Consider first the case of a TD(1) critic. By definition, $\bar{r}(\theta)$ is the solution to the linear equation $\bar{G}_1(\theta)\bar{r}(\theta) = \bar{h}_1(\theta)$, or

$$\langle \phi_\theta, \phi'_\theta\bar{r}(\theta)\rangle_\theta = \langle \phi_\theta, Q_\theta\rangle_\theta.$$

Thus, $\phi_\theta' \bar{r}(\theta) - Q_\theta$ is orthogonal to ϕ_θ in $\mathcal{L}^2(\eta_\theta)$. By Assumption 4.8(d), the components of ψ_θ are contained in the subspace spanned by the components of ϕ_θ. It follows that $\phi_\theta' \bar{r}(\theta) - Q_\theta$ is also orthogonal to ψ_θ. Therefore,

$$\bar{H}(\theta)\bar{r}(\theta) = \langle \psi_\theta, \phi_\theta' \rangle_\theta \bar{r}(\theta) = \langle \psi_\theta, Q_\theta \rangle_\theta = \nabla \bar{\alpha}(\theta),$$

where the last equality is the gradient formula in Theorem 4.6.

For $\lambda < 1$, let us write $\bar{G}_1^\lambda(\theta)$ and $\bar{h}_1^\lambda(\theta)$ for $\bar{G}_1(\theta)$ and $\bar{h}_1(\theta)$, defined in section 5.2, to show explicitly the dependence on λ. Let $\hat{\phi}_\theta = \phi_\theta - \langle \phi_\theta, \underline{1} \rangle_\theta \underline{1}$. Then it is easy to see that

$$|\bar{G}_1^\lambda(\theta) - \langle \hat{\phi}_\theta, \hat{\phi}_\theta' \rangle_\theta| = (1-\lambda) \left| \sum_{k=0}^{\infty} \lambda^k \langle P_\theta^k \hat{\phi}_\theta, \hat{\phi}_\theta \rangle_\theta \right| \le C \left(\frac{1-\lambda}{1-\rho\lambda} \right),$$

where the inequality follows from the geometric ergodicity condition (4.3). Similarly, one can also see that $|\bar{h}_1^\lambda(\theta) - \langle Q_\theta, \hat{\phi}_\theta \rangle_\theta| \le C(1-\lambda)$. Let $\bar{r}(\theta)$ and $\bar{r}^\lambda(\theta)$ be solutions of the linear equations $\langle \hat{\phi}_\theta, \hat{\phi}_\theta' r \rangle_\theta = \langle Q_\theta, \phi_\theta \rangle_\theta$ and $\bar{G}_1^\lambda(\theta) r = \bar{h}_1^\lambda(\theta)$, respectively. Then

$$\langle \hat{\phi}_\theta, \hat{\phi}_\theta' \rangle_\theta (\bar{r}(\theta) - \bar{r}^\lambda(\theta)) = (\bar{h}_1(\theta) - \bar{h}_1^\lambda(\theta)) + (\bar{G}_1^\lambda(\theta) - \langle \hat{\phi}_\theta, \hat{\phi}_\theta' \rangle_\theta) \bar{r}^\lambda(\theta),$$

which implies that $|\bar{r}(\theta) - \bar{r}^\lambda(\theta)| \le C(1-\lambda)$. The rest follows from the observation that $\bar{H}(\theta)\bar{r}(\theta) = \nabla \bar{\alpha}(\theta)$. □

LEMMA 6.2 (convergence of the noise terms).

(a) $\sum_{k=0}^{\infty} \beta_k \nabla \bar{\alpha}(\theta_k) \cdot e_k^{(1)}$ *converges w.p.*1.

(b) $\lim_k e_k^{(2)} = 0$ *w.p.*1.

(c) $\sum_k \beta_k^2 |H_{\theta_k}(\hat{X}_k, \hat{U}_k) r_k \Gamma(r_k)|^2 < \infty$ *w.p.*1.

Proof. Since r_k is bounded and $\Gamma(\cdot)$ satisfies the condition (3.3), it is easy to see that $r\Gamma(r)$ is bounded and $|r\Gamma(r) - \hat{r}\Gamma(\hat{r})| < C|r - \hat{r}|$ for some constant C. The proof of part (a) is now similar to the proof of Lemma 2 on page 224 of [3]. Part (b) follows from Theorem 5.7 and the fact that $\bar{H}(\cdot)$ is bounded. Part (c) follows from the inequality

$$|H_{\theta_k}(\hat{X}_k, \hat{U}_k) r_k \Gamma(r_k)| \le C |H_{\theta_k}(\hat{X}_k, \hat{U}_k)|$$

for some $C > 0$ and the boundedness of $\mathbf{E}[|H_{\theta_k}(\hat{X}_k, \hat{U}_k)|^2]$ (from part (b) of Lemma 4.3). □

THEOREM 6.3 (convergence of actor-critic algorithms). *Let Assumptions* 3.3, 4.1, 4.2, 4.4, 4.5, 4.8, and 4.9 *hold.*

(a) *If a TD*(1) *critic is used, then* $\liminf_k |\nabla \bar{\alpha}(\theta_k)| = 0$ *w.p.*1.

(b) *For any* $\epsilon > 0$*, there exists some* λ *sufficiently close to* 1*, so that the algorithm that uses a TD*(λ) *critic (with* $0 < \lambda < 1$*) satisfies* $\liminf_k |\nabla \bar{\alpha}(\theta_k)| < \epsilon$ *w.p.*1.

Proof. The proof is standard [24], and we provide only an outline. Fix some $T > 0$, and define a sequence k_j by

$$k_0 = 0, \qquad k_{j+1} = \min \left\{ k \ge k_j \;\middle|\; \sum_{l=k_j}^{k} \beta_k \ge T \right\} \quad \text{for} \quad j > 0.$$

Using (6.1), we have

$$\bar{\alpha}(\theta_{k_{j+1}}) \le \bar{\alpha}(\theta_{k_j}) - \sum_{k=k_j}^{k_{j+1}-1} \beta_k \left(|\nabla \bar{\alpha}(\theta_k)|^2 - C(1-\lambda) |\nabla \bar{\alpha}(\theta_k)| \right) + \delta_j,$$

where δ_j is defined by

$$\delta_j = \sum_{k=k_j}^{k_{j+1}-1} \left(\beta_k \nabla \bar{\alpha}(\theta_k) \cdot (e_k^{(1)} + e_k^{(2)}) + C\beta_k^2 |H_{\theta_k}(\hat{X}_k, \hat{U}_k) r_k \Gamma(r_k)|^2 \right).$$

Lemma 6.2 implies that δ_j goes to zero. If the result fails to hold, it can be shown that the sequence $\bar{\alpha}(\theta_k)$ would decrease indefinitely, contradicting the boundedness of $\bar{\alpha}(\theta)$. The result follows easily. □

Appendix A. A result on linear stochastic approximation.

We recall the following result from [14]. Consider a stochastic process $\{\hat{Y}_k\}$ taking values in a Polish space $\mathbb{Y}$ with Borel σ-field denoted by $\mathcal{B}(\mathbb{Y})$. Let $\{P_\theta(y, d\bar{y}); \theta \in \mathbb{R}^n\}$ be a parameterized family of transition kernels on $\mathbb{Y}$. Consider the following iterations to update a vector $R \in \mathbb{R}^m$ and the parameter $\theta \in \mathbb{R}^n$:

$$\begin{aligned} R_{k+1} &= R_k + \gamma_k (h_{\theta_k}(\hat{Y}_{k+1}) - G_{\theta_k}(\hat{Y}_{k+1}) R_k + \xi_{k+1} R_k), \\ \theta_{k+1} &= \theta_k + \beta_k H_{k+1}. \end{aligned} \tag{A.1}$$

In the above iteration, $\{h_\theta(\cdot), G_\theta(\cdot) : \theta \in \mathbb{R}^n\}$ is a parameterized family of m-vector-valued and $m \times m$-matrix-valued measurable functions on $\mathbb{Y}$. We introduce the following assumptions.

Assumption A.1. The step-size sequence $\{\gamma_k\}$ is deterministic and nonincreasing and satisfies

$$\sum_k \gamma_k = \infty, \qquad \sum_k \gamma_k^2 < \infty.$$

Let $\mathcal{F}_k$ be the σ-field generated by $\{\hat{Y}_l, H_l, r_l, \theta_l, l \le k\}$.

Assumption A.2. For a measurable set $A \subset \mathbb{Y}$,

$$\mathbf{P}(\hat{Y}_{k+1} \in A \mid \mathcal{F}_k) = \mathbf{P}(\hat{Y}_{k+1} \in A \mid \hat{Y}_k, \theta_k) = P_{\theta_k}(\hat{Y}_k, A).$$

For any measurable function f on $\mathbb{Y}$, let $P_\theta f$ denote the measurable function $y \mapsto \int P_\theta(y, d\bar{y}) f(\bar{y})$, and for any vector r, let $|r|$ denote its Euclidean norm.

Assumption A.3 (existence and properties of solutions to the Poisson equation). For each θ, there exist functions $\bar{h}(\theta) \in \mathbb{R}^m$, $\bar{G}(\theta) \in \mathbb{R}^{m \times m}$, $\hat{h}_\theta : \mathbb{Y} \to \mathbb{R}^m$, and $\hat{G}_\theta : \mathbb{Y} \to \mathbb{R}^{m \times m}$ that satisfy the following:

(a) For each $y \in \mathbb{Y}$,

$$\begin{aligned} \hat{h}_\theta(y) &= h_\theta(y) - \bar{h}(\theta) + (P_\theta \hat{h}_\theta)(y), \\ \hat{G}_\theta(y) &= G_\theta(y) - \bar{G}(\theta) + (P_\theta \hat{G}_\theta)(y). \end{aligned}$$

(b) For some constant C and for all θ, we have

$$\max(|\bar{h}(\theta)|, |\bar{G}(\theta)|) \le C.$$

(c) For any $d > 0$, there exists $C_d > 0$ such that

$$\sup_k \mathbf{E}[|f_{\theta_k}(\hat{Y}_k)|^d] \le C_d,$$

where $f_\theta(\cdot)$ represents any of the functions $\hat{h}_\theta(\cdot), h_\theta(\cdot), \hat{G}_\theta(\cdot), G_\theta(\cdot)$.

(d) For some constant $C > 0$ and for all $\theta, \bar{\theta} \in \mathbb{R}^n$,

$$\max(|\bar{h}(\theta) - \bar{h}(\bar{\theta})|, |\bar{G}(\theta) - \bar{G}(\bar{\theta})|) \leq C|\theta - \bar{\theta}|.$$

(e) There exists a positive measurable function $C(\cdot)$ on $\mathbb{Y}$ such that, for each $d > 0$,

$$\sup_k \mathbf{E}[C(\hat{Y}_k)^d] < \infty$$

and

$$|P_\theta f_\theta(y) - P_{\bar{\theta}} f_{\bar{\theta}}(y)| \leq C(y)|\theta - \bar{\theta}|,$$

where $f_\theta(\cdot)$ is any of the functions $\hat{h}_\theta(\cdot)$ and $\hat{G}_\theta(\cdot)$.

Assumption A.4 (slowly changing environment). The (random) process $\{H_k\}$ satisfies

$$\sup_k \mathbf{E}\left[|H_k|^d\right] < \infty$$

for all $d > 0$. Furthermore, the sequence $\{\beta_k\}$ is deterministic and satisfies

$$\sum_k \left(\frac{\beta_k}{\gamma_k}\right)^d < \infty$$

for some $d > 0$.

Assumption A.5. The sequence $\{\xi_k\}$ is an $m \times m$-matrix-valued $\mathcal{F}_k$-martingale difference, with bounded moments, i.e.,

$$\mathbf{E}\left[\xi_{k+1}|\mathcal{F}_k\right] = 0, \qquad \sup_k \mathbf{E}\left[|\xi_{k+1}|^d\right] < \infty$$

for each $d > 0$.

Assumption A.6 (uniform positive definiteness). There exists $a > 0$ such that, for all $r \in \mathbb{R}^m$ and $\theta \in \mathbb{R}^n$,

$$r'\bar{G}(\theta)r \geq a|r|^2.$$

THEOREM A.7. *If Assumptions* A.1–A.6 *are satisfied, then the sequence R_k is bounded and*

$$\lim_k \left|R_k - \bar{G}(\theta_k)^{-1}\bar{h}(\theta_k)\right| = 0.$$

REFERENCES

[1] K. B. ATHREYA AND P. NEY, *A new approach to the limit theory of recurrent Markov chains*, Trans. Amer. Math. Soc., 245 (1978), pp. 493–501.

[2] A. BARTO, R. SUTTON, AND C. ANDERSON, *Neuron-like elements that can solve difficult learning control problems*, IEEE Transactions on Systems, Man and Cybernetics, 13 (1983), pp. 835–846.

[3] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin, Heidelberg, 1990.

[4] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, Athena Scientific, Belmont, MA, 1995.

[5] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.

[6] V. S. Borkar, *Stochastic approximation with two time scales*, Systems Control Lett., 29 (1997), pp. 291–294.

[7] X. R. Cao and H. F. Chen, *Perturbation realization, potentials, and sensitivity analysis of Markov processes*, IEEE Trans. Automat. Control, 42 (1997), pp. 1382–1393.

[8] P. W. Glynn, *Stochastic approximation for Monte Carlo optimization*, in Proceedings of the 1986 Winter Simulation Conference, Washington, DC, 1986, pp. 285–289.

[9] P. W. Glynn and P. L'Ecuyer, *Likelihood ratio gradient estimation for stochastic recursions*, Adv. Appl. Probab., 27 (1995), pp. 1019–1053.

[10] T. Jaakkola, S. P. Singh, and M. I. Jordan, *Reinforcement learning algorithms for partially observable Markov decision problems*, in Advances in Neural Information Processing Systems 7, G. Tesauro and D. Touretzky, eds., Morgan Kaufman, San Francisco, CA, 1995, pp. 345–352.

[11] V. R. Konda, *Actor-Critic Algorithms*, Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2002.

[12] V. R. Konda and V. S. Borkar, *Actor-critic–type learning algorithms for Markov decision processes*, SIAM J. Control Optim., 38 (1999), pp. 94–123.

[13] V. R. Konda and J. N. Tsitsiklis, *Actor-critic algorithms*, in Advances in Neural Information Processing Systems 12, S. A. Solla, T. K. Leen, and K.-R. Muller, eds., MIT Press, Cambridge, MA, 2000, pp. 1008–1014.

[14] V. R. Konda and J. N. Tsitsiklis, *Linear stochastic approximation driven by slowly varying Markov chains*, 2002, submitted.

[15] V. R. Konda and J. N. Tsitsiklis, *Appendix to "On Actor-critic algorithms,"* http://web.mit.edu/jnt/www/Papers.html/actor-app.pdf, July 2002.

[16] P. Marbach and J. N. Tsitsiklis, *Simulation-based optimization of Markov reward processes*, IEEE Trans. Automat. Control, 46 (2001), pp. 191–209.

[17] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.

[18] E. Nummelin, *A splitting technique for Harris recurrent chains*, Z. Wahrscheinlichkeitstheorie and Verw. Geb., 43 (1978), pp. 119–143.

[19] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.

[20] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, *Policy gradient methods for reinforcement learning with function approximation*, in Advances in Neural Information Processing Systems 12, S. A. Solla, T. K. Leen, and K.-R. Muller, eds., MIT Press, Cambridge, MA, 2000, pp. 1057–1063.

[21] J. N. Tsitsiklis and B. Van Roy, *An analysis of temporal-difference learning with function approximation*, IEEE Trans. Automat. Control, 42 (1997), pp. 674–690.

[22] J. N. Tsitsiklis and B. Van Roy, *Average cost temporal-difference learning*, Automatica J. IFAC, 35 (1999), pp. 1799–1808.

[23] R. Williams, *Simple statistical gradient following algorithms for connectionist reinforcement learning*, Machine Learning, 8 (1992), pp. 229–256.

[24] B. T. Polyak, *Pseudogradient adaptation and training algorithms*, Autom. Remote Control, 34 (1973), pp. 377–397.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1167–1184

ON THE PARETO CONTROL AND NO-REGRET CONTROL FOR DISTRIBUTED SYSTEMS WITH INCOMPLETE DATA*

O. NAKOULIMA†, A. OMRANE†, AND J. VELIN†

Abstract. We discuss the control of distributed systems with incomplete data following the notion of no-regret control (or, equivalently, Pareto control) used by Lions in [*C. R. Acad. Sci. Paris Ser. I Math.*, 302 (1986), pp. 223–227] and [*C. R. Acad. Sci. Paris Ser. I Math.*, 302 (1992), pp. 1253–1257]. We associate with the no-regret control a sequence of low-regret controls defined by a quadratic perturbation previously used by Nakoulima, Omrane, and Velin in [*C. R. Acad. Sci. Paris Ser. I Math.*, 330 (2000), pp. 801–806].

In the first part, we prove that the perturbed system corresponds to a sequence of standard control problems and converges to the no-regret (or Pareto) control for which we obtain a singular optimality system. We give also some applications.

In the second part, we show how the method can be extended to the evolution case. Equations of parabolic type, Petrowsky type, or hyperbolic type are considered.

Key words. Pareto control, no-regret control, low-regret control, systems with incomplete data, cost function, quadratic perturbation

AMS subject classifications. 49K40, 35B37, 35K55, 90C30, 93A15, 93D09

DOI. 10.1137/S0363012900380188

1. Introduction.

Let $\mathcal{V}$ be a real Hilbert space of dual $\mathcal{V}'$, $A \in \mathcal{L}(\mathcal{V};\mathcal{V}')$ an elliptic (parabolic or hyperbolic in the sections below) differential operator modeling a distributed system, $\mathcal{U}$ the Hilbert space of controls, and $B \in \mathcal{L}(\mathcal{U};\mathcal{V}')$. Let G be a nonempty closed vector subspace of the Hilbert space of uncertainties F, and $\beta \in \mathcal{L}(F,\mathcal{V}')$.

For $f \in \mathcal{V}'$, the state equation related to the control $v \in \mathcal{U}$ and to the uncertainty $g \in G$ is given by

$$A\,y(v,g) \;=\; f \;+ B\,v + \beta\,g. \tag{1.1}$$

Supposing that A is an isomorphism from $\mathcal{V}$ to $\mathcal{V}'$, (1.1) is well posed in $\mathcal{V}$. Denote by $y = y(v,g)$ the unique solution to (1.1). For every $g \in G$ we have then a possible state for which we rely on a cost function given by

$$J(v,g) = \Big\|C\,y - z_d\Big\|^2_{\mathcal{H}} + N\Big\|v\Big\|^2_{\mathcal{U}}, \tag{1.2}$$

where $C \in \mathcal{L}(\mathcal{V};\mathcal{H})$, $\mathcal{H}$ is a Hilbert space, $z_d \in \mathcal{H}$ fixed, $N > 0$, and $\|.\|_X$ is the norm on the real Hilbert space X. We are concerned with the optimal control of the problem (1.1)–(1.2); i.e., we want to solve

$$\inf_{v\in\mathcal{U}} J(v,g) \qquad \forall g \in G,$$

which clearly makes no sense when $G \neq \{0\}$ (G being an infinite space).

*Received by the editors October 30, 2000; accepted for publication (in revised form) February 24, 2003; published electronically August 6, 2003.
http://www.siam.org/journals/sicon/42-4/38018.html

†Laboratoire de Mathématiques et Informatique, Université des Antilles et de La Guyane, Campus Fouillole 97159 Pointe-à-Pitre, Guadeloupe (FWI) (onakouli@univ-ag.fr, aomrane@univ-ag.fr, jvelin@univ-ag.fr).

The idea is to look for a solution for the following minimization problem:

$$\inf_{v\in\mathcal{U}}\left(\sup_{g\in G} J(v,g)\right),$$

but J is not upper bounded as $\sup_{g\in G} J(v,g) = +\infty$.

Lions used the notions of Pareto control [12] and no-regret control [14] in application to control the system (1.1)–(1.2).

The concept of Pareto[1] is motivated by a number of applications in economics, and also in ecology. In his book [9], Kotarski discussed the Pareto optimum problem, where some results of geometrical and numerical interest are obtained in the case of optimal control. In [8], he generalized the well-known Dubovicki–Milutin theorem (on the feasible sets of Pareto) and applied it to obtain a necessary condition for the Pareto *minimum*, or necessary and sufficient conditions on the Pareto optimum, extending a geometrical work of Censor [3]. Lions [12], [13] used the concept to obtain controls for distributed problems with incomplete data.

The no-regret concept was introduced many years later in statistics by Savage [18]. In several works, Lions applied this notion and a related idea called "low-regret" control to problems of incomplete data (see [14], [15], [16], [7], [6]) for various applications. In [14], for example, he extends the results of the work of Allwright [1] to the infinite dimension case. In [7] with Gabay, a decision criterion is added to the uncertainties closed subspace; it improves by extending the notion of low-regret to many agents: each agent wishes to act with least-regret and all agents wish to have minimum exchanges of information, in order to make things as local as possible. The low-regret control is applied to systems where there are *controls* and *unknown perturbations*. One then looks for the control not making things worse with respect to a nominal control u_0 (or to then doing nothing, $u_0 = 0$), independently of the perturbations which may be of infinite number.

We will see in section 2 that Pareto controls and no-regret controls are actually the same.

In this work, we give a characterization of the no-regret (or the Pareto) control for problems of incomplete data, in both the stationary and evolution cases. We improve the results of the work in [14] (and also [12]) of Lions by giving the precise optimality system for the low-regret control and by describing a number of applications. Thanks to a quadratic perturbation used by the authors in [17], the optimality system for the no-regret control appears clearly as the limit of a standard control problem.

Some of the results in this paper are summarized in [17]. The proofs in the present paper and the treatment of the evolution cases are new.

The paper is organized as follows. In section 2, we see the main definitions, we verify the equivalence between the two approaches of Lions, and we introduce the low-regret control method. We then give the optimality system for the perturbed problem and prove that the optimal controls for the perturbed problem converge to the no-regret (Pareto) control of original problem. Moreover, by passing to the limit in the associated optimality system of the perturbed problem, we obtain a singular optimality system for the no-regret (Pareto) control. In section 3, we give several examples of elliptic type. Section 4 is devoted to the evolution case. Here, we give

[1]Wifredo Pareto (1848–1923) was an Italian economist and a political sociologist. He defined the efficient optimum, and in particular was the one who devised the law of *trivial many* and *critical few* known as the 80:20 rule.

the theoretical results for parabolic and hyperbolic distributed systems. An example involving a parabolic system is considered in the last section.

2. No-regret control for stationary problems.

2.1. Definitions and preliminary results. We give definitions for the Pareto and no-regret controls related to a given control here, as well as the preliminary results.

DEFINITION 2.1. *We say that $u \in \mathcal{U}$ is a Pareto control for* (1.1)–(1.2) *(cf. Lions* [12]*) iff* $J(u,g) \leq J(v,g)\ \forall v \in \mathcal{U}, \forall g \in G$, *and if there exists at least* $g_0 \in G$ *such that* $J(u,g_0) < J(v,g_0)\ \forall v \in \mathcal{U}$.

DEFINITION 2.2. *Let $u \in \mathcal{U}$ be a Pareto control. We say that u is related to a control $u_0 \in \mathcal{U}$ if*

$$J(u,g) \leq J(u_0,g) \qquad \forall g \in G.$$

DEFINITION 2.3. *We say that $u \in \mathcal{U}$ is a no-regret control for* (1.1)–(1.2) *related to a control u_0 if u is a solution to the following problem:*

$$\inf_{v \in \mathcal{U}} \sup_{g \in G} \left(J(v,g) - J(u_0,g) \right). \tag{2.1}$$

When $u_0 = 0$, Definition 2.3 reduces to the definition of no-regret control of Lions [14].

LEMMA 2.4. *For any $u_0 \in \mathcal{U}$ and $v \in \mathcal{U}$ we have*

$$J(v,g) - J(u_0,g) = J(v,0) - J(u_0,0) + 2\langle \beta^* \zeta(v-u_0), g \rangle_{G',G} \qquad \forall g \in G,$$

where $\zeta(v) \in \mathcal{V}$ is defined for $v \in \mathcal{U}$ by

$$A^* \zeta(v) = C^* C(y(v,0) - y(0,0)),$$

A^* (resp., β^*) being the adjoint of A (resp., β).

Proof. We have in fact

$$J(v,g) - J(u_0,g) = J(v,0) - J(u_0,0) + 2\langle C(y(v-u_0,0) - y(0,0)), C(y(0,g) - y(0,0)) \rangle_{\mathcal{H},\mathcal{H}}$$

$\forall g \in G$. We then introduce $\zeta(v) \in \mathcal{V}$ defined by $A^* \zeta(v) = C^* C(y(v,0) - y(0,0))$, where A^* is the adjoint of A. Then

$$\begin{aligned} \langle C(y(v,0) - y(0,0)), C(y(0,g) - y(0,0)) \rangle_{\mathcal{H},\mathcal{H}} &= \langle A^* \zeta(v), y(0,g) - y(0,0) \rangle_{\mathcal{V}',\mathcal{V}}, \\ &= \langle \zeta(v), \beta g \rangle_{\mathcal{V},\mathcal{V}'} = \langle \beta^* \zeta(v), g \rangle_{G',G} \end{aligned}$$

(notice that $A(y(0,g) - y(0,0)) = \beta\, g$).

Remark 1. For sake of simplicity, we denote by $S(v) = \beta^* \zeta(v)$ the linear function for $v \in \mathcal{U}$. Then we have

$$J(v,g) - J(u_0,g) = J(v,0) - J(u_0,0) + 2\langle S(v-u_0), g \rangle_{G',G} \qquad \forall g \in G. \tag{2.2}$$

In the applications below $\beta = Id$, and we have $S(v) = \zeta(v)\ \forall v \in \mathcal{U}$.

Remark 2. Of course the problem (2.1) is defined only for the controls $v \in \mathcal{U}$ such that

$$\sup_{g \in G} \left(J(v,g) - J(u_0,g) \right) < \infty.$$

From (2.2) this is realized for the no-regret control (and the Pareto control) v iff $v \in K + u_0$, where $K = \{w \in \mathcal{U}, \langle S(w), g \rangle = 0 \ \forall g \in G\}$.

PROPOSITION 2.5 (cf. Lions [12]). *Let be $u_0 \in \mathcal{U}$. Then there exists a unique Pareto control related to u_0. Moreover, it is the unique element of the set $K + u_0$, which minimizes the functional $J(v, 0)$ on $K + u_0$.*

We can now prove the following.

THEOREM 2.6. *Let $u_0 \in \mathcal{U}$ be a given control. Then we have the following: a control $u \in \mathcal{U}$ is a Pareto control related to u_0 iff u is a no-regret control related to u_0.*

Proof. Let u be a Pareto control related to u_0, and let be $v \in K + u_0$. Then $\langle S(u - u_0), g\rangle = 0 = \langle S(v - u_0), g\rangle \ \forall g \in G$, and we have $J(u, 0) \leq J(v, 0)$ according to Proposition 2.5. Hence, using (2.2)

$$J(u, 0) - J(u_0, 0) + 2\langle S(u - u_0), g\rangle \leq \sup_{g\in G} (J(v, g) - J(u_0, g)) ;$$

that is, $\sup_{g\in G} (J(u, g) - J(u_0, g)) \leq \sup_{g\in G} (J(v, g) - J(u_0, g))$. So,

$$\sup_{g\in G} (J(u, g) - J(u_0, g)) = \inf_{v\in K+u_0} \Big(\sup_{g\in G} (J(v, g) - J(u_0, g)) \Big).$$

Now, let be $v \in \mathcal{U}\setminus \{K + u_0\}$. There is at least one $g_0 \in G$ such that $\langle S(v-u_0), g_0\rangle \neq 0$. Then we have

$$\sup_{g\in G} (J(v, g) - J(u_0, g)) = J(v, 0) - J(u_0, 0) + 2 \sup_{g\in G} \langle S(v - u_0), g\rangle = +\infty.$$

(Note that G is a vector space, and henceforth we have the only two possibilities: $\sup_{g\in G}\langle S(w), g\rangle = 0$ or $\sup_{g\in G}\langle S(w), g\rangle = +\infty$. Indeed, $\lim_{t\to+\infty}\langle S(v - u_0), tg_0\rangle = +\infty$.)

From another side, as u is a Pareto control we have $J(u, g) - J(u_0, g) \leq 0 \ \forall g \in G$; hence

$$J(u, g) - J(u_0, g) \leq 0 \leq \sup_{g\in G} (J(v, g) - J(u_0, g)) \qquad \forall g \in G.$$

Finally,

$$\sup_{g\in G} (J(u, g) - J(u_0, g)) = \inf_{v\in \mathcal{U}\setminus(K+u_0)} \Big(\sup_{g\in G} (J(v, g) - J(u_0, g)) \Big).$$

In conclusion, u is a no-regret control related to u_0.

Conversely, let u be a no-regret control related to u_0. We have

$$\sup_{g\in G} (J(u, g) - J(u_0, g)) \leq \sup_{g\in G} (J(v, g) - J(u_0, g)) \qquad \forall v \in \mathcal{U}.$$

Then for $v = u_0$,

$$J(u, 0) + \sup_{g\in G} \langle S(u - u_0), g\rangle \leq J(u_0, 0) = c \quad \text{constant}.$$

As $J(u, 0) \geq 0$, we have $\sup_{g\in G}\langle S(u - u_0), g\rangle \leq c$. We deduce that $\sup_{g\in G}\langle S(u - u_0), g\rangle = 0$. Consequently, $\langle S(u - u_0), g\rangle \leq 0 \ \forall g \in G$, and hence $\langle S(u - u_0), g\rangle = 0$. So, $u \in K + u_0$, and we have

$$J(u, 0) \leq J(v, 0) \qquad \forall v \in K + u_0.$$

In conclusion, u is a Pareto control related to u_0. ∎

Remark 3. By Proposition 2.5, we know that there exists a unique Pareto control related to u_0, and that is the only one which minimizes the functional $\inf_{v\in K+u_0} J(v,0)$. In the second part of Theorem 2.6, it is proved that the no-regret control related to u_0—if it exists—also minimizes this functional. As a matter of fact, that suffices to show the existence of a unique no-regret control related to u_0 and that the Pareto control and no-regret control for the problem (1.1)–(1.2) are actually the same.

We are interested in the existence and the characterization of the no-regret (or Pareto) control related to u_0. We follow the lines of [14] where Lions introduced the method of low-regret control.

2.2. Low-regret control. As in [17], we define the low-regret control by relaxing the problem (2.1) as follows:

$$\inf_{v\in\mathcal{U}} \sup_{g\in G} \left[J(v,g) - J(u_0,g) - \gamma\|g\|_G^2\right], \tag{2.3}$$

where $u_0 \in \mathcal{U}$ is a given control, and where γ is a strictly positive parameter. The solution to problem (2.3), if it exists, will be the low-regret control related to u_0, of the problem (1.1)–(1.2).

Remark 4 (cf. Lions [15]). With the "low-regret control," we admit the possibility of making a choice of controls v "slightly worse" ($J(v,g) - J(u_0,g) \le \gamma\|g\|_G^2$ and not $J(v,g) - J(u_0,g) \le 0$ as for the no-regret control) than by doing better than u_0—but not much better—if we choose γ small enough (compared to the worst things that could happen with the "pollution" g).

The best possible choice of v is then given by (2.3).

From (2.2) the problem (2.3) also writes

$$\inf_{v\in\mathcal{U}} \left[J(v,0) - J(u_0,0) + \sup_{g\in G} \left(\langle 2S(v-u_0), g\rangle - \gamma\|g\|_G^2\right)\right].$$

Remark 5. By the perturbation (2.3) we have explicitly the conjugate

$$\sup_{g\in G} \left(\langle 2S(v-u_0), g\rangle - \gamma\|g\|_G^2\right)$$

as we find

$$\sup_{g\in G} \left(\langle 2S(v-u_0), g\rangle - \gamma\|g\|_G^2\right) = \frac{1}{\gamma}\left\|S(v-u_0)\right\|_{G'}^2.$$

With this, if we identify G and G', the problem (2.3) takes the form

$$\inf_{v\in\mathcal{U}} \mathcal{J}^\gamma(v), \tag{2.4}$$

where

$$\mathcal{J}^\gamma(v) = J(v,0) - J(u_0,0) + \frac{1}{\gamma}\left\|S(v-u_0)\right\|_G^2. \tag{2.5}$$

We recognize then a standard optimization problem of a quadratic cost functional.

2.3. Approached optimality system. Now we give the optimality system for the low-regret control u_γ.

PROPOSITION 2.7. *The problem* (2.4)–(2.5) *admits a unique solution* u_γ *called the low-regret control related to* u_0.

Proof. We have $\mathcal{J}^\gamma(v) \geq -J(u_0,0)\ \forall v \in \mathcal{U}$. Then $d_\gamma = \inf_{v\in\mathcal{U}} \mathcal{J}^\gamma(v)$ exists. Let then $v_n = v_n(\gamma)$ be a minimizing sequence such that $d_\gamma = \lim_{n\to\infty} \mathcal{J}^\gamma(v_n)$. We have

$$-J(u_0,0) \leq \mathcal{J}^\gamma(v_n) = J(v_n,0) - J(u_0,0) + \frac{1}{\gamma}\Big\|S(v_n)\Big\|_G^2 \leq d_\gamma + 1.$$

Then we deduce the bounds

$$\Big\|v_n\Big\|_{\mathcal{U}} \leq c_\gamma, \qquad \frac{1}{\sqrt{\gamma}}\Big\|S(v_n - u_0)\Big\|_G \leq c_\gamma, \qquad \Big\|Cy(v_n,0) - z_d\Big\|_{\mathcal{H}} \leq c_\gamma,$$

where the constant c_γ (independent of n) is not the same each time.

There exists $u_\gamma \in \mathcal{U}$ such that $v_n \rightharpoonup u_\gamma$ weakly in the Hilbert space $\mathcal{U}$. Also, $y(v_n,0) \to y(u_\gamma,0)$ (continuity w.r.t the data). We also deduce from the strict convexity of the cost function $\mathcal{J}^\gamma$ that u_γ is unique. ∎

THEOREM 2.8. *The solution* u_γ *of the relaxed problem* (2.4)–(2.5) *weakly converges in* $\mathcal{U}$ *as* $\gamma \to 0$ *to the unique no-regret control related to* u_0.

Proof. Let u_γ be the solution to (2.4)–(2.5). Then

$$J(u_\gamma,0) - J(u_0,0) + \frac{1}{\gamma}\Big\|S(u_\gamma - u_0)\Big\|_G^2 \leq J(v,0) - J(u_0,0) + \frac{1}{\gamma}\Big\|S(v-u_0)\Big\|_G^2 \quad \forall v \in \mathcal{U}.$$

Particularly for $v = u_0$, we have

$$J(u_\gamma,0) - J(u_0,0) + \frac{1}{\gamma}\Big\|S(u_\gamma - u_0)\Big\|_G^2 \leq 0,$$

and the structure of $J(u_\gamma,0)$ in (1.2) gives

$$\Big\|Cy(u_\gamma,0) - z_d\Big\|_{\mathcal{H}}^2 + N\Big\|u_\gamma\Big\|_{\mathcal{U}}^2 + \frac{1}{\gamma}\Big\|S(u_\gamma - u_0)\Big\|_G^2 \leq J(u_0,0). \tag{2.6}$$

We deduce that $\|u_\gamma\|_{\mathcal{U}} \leq c$. Then we can extract a subsequence u_γ which weakly converges towards $u \in \mathcal{U}$, the solution to (2.4).

Now for $v \in \mathcal{U}$ we have

$$J(v,g) - J(u_0,g) - \gamma\|g\|^2 \leq J(v,g) - J(u_0,g) \quad \forall g \in G.$$

Then

$$J(u_\gamma,g) - J(u_0,g) - \gamma\|g\|^2 \leq \sup_{g\in G}\left(J(v,g) - J(u_0,g)\right) \quad \forall g \in G,$$

and passing to the limit in γ we obtain

$$J(u,g) - J(u_0,g) \leq \sup_{g\in G}\left(J(v,g) - J(u_0,g)\right) \qquad \forall g \in G.$$

We deduce easily that u is a no-regret control related to u_0. ∎

Now we give the optimality system for the low-regret control.

PROPOSITION 2.9. *The low-regret control u_γ solution to* (2.4)–(2.5) *is characterized by the unique solution $\{y_\gamma, \zeta_\gamma, \rho_\gamma, p_\gamma\}$ of the optimality system*

$$\begin{cases} A\, y_\gamma & = f + Bu_\gamma, \quad A^*\zeta_\gamma = C^*C(y_\gamma - y(0,0)), \\ A\, \rho_\gamma & = \frac{1}{\gamma}\beta\beta^*\zeta_\gamma, \quad A^*p_\gamma = C^*(Cy_\gamma - z_d) + C^*C\rho_\gamma, \\ B^*p_\gamma + Nu_\gamma & = 0 \quad in \quad \mathcal{U}. \end{cases}$$

Proof. Let u_γ be the solution of (2.4)–(2.5) on $\mathcal{U}$. The Euler–Lagrange necessary condition gives for every $w \in \mathcal{U}$

$$\langle C^*\left(Cy(u_\gamma,0) - z_d\right), y(w,0) - y(0,0)\rangle_{\mathcal{H}\times\mathcal{H}} + \langle Nu_\gamma, w\rangle_{\mathcal{U}\times\mathcal{U}} + 2\left\langle \frac{1}{\gamma}S(u_\gamma), S(w)\right\rangle_{G\times G} \geq 0.$$

Denoting $y_\gamma = y(u_\gamma, 0)$ and $\xi_\gamma(v) = \beta S(v)$ we have

$$A^*\xi_\gamma = C^*C(y_\gamma - y(0,0)).$$

Let ρ_γ be the solution to

$$A\rho_\gamma = \frac{1}{\gamma}\beta\beta^*\xi.$$

And as it is classical, we introduce the adjoint state p_γ defined by

$$B^*p_\gamma = C^*(Cy_\gamma - z_d) + C^*C\rho_\gamma$$

so that we obtain

$$\langle B^*p_\gamma + Nu_\gamma, w\rangle \geq 0 \qquad \forall w \in \mathcal{U}.$$

But also we have $\langle B^*p_\gamma + Nu_\gamma, w\rangle \leq 0 \ \forall w \in \mathcal{U}$. The optimality system follows. □

2.4. Singular optimality system. Now, we give the optimality system for the no-regret control.

As in [12] let $\mathcal{R}$ be an operator defined as follows.

We solve first

$$A\,\rho = \beta\, g, \quad g \in G, \quad \rho \in \mathcal{V},$$

then

$$A^*\sigma = C^*C\rho, \quad \sigma \in \mathcal{V},$$

and we set $\mathcal{R}\, g = B^*\,\sigma$. We suppose that

$$\left\|\mathcal{R}\, g\right\|_{\widehat{G}} \geq c\left\|g\right\|_G, \quad c > 0, \quad \text{for any} \quad g \in G, \tag{2.7}$$

where $\widehat{G}$ is the completion of G in F, containing the elements $\mathcal{R}g$.

Remark 6. The space $\widehat{G}$ is in fact the completion of G for a subspace $(H, \|.\|_{\|.\|})$ of F which can be bigger than G. This will be made precise in the applications below.

Remark 7. The hypothesis (2.7) is very useful theoretically but is not necessary in practice. We need only to make sure that the adjoint state p_γ of Proposition 2.9 is bounded in a suitable Hilbert space, which is the case in the applications given below.

THEOREM 2.10. *Suppose that* (2.7) *holds true. Then the no-regret control* u *related to* u_0*, solution to* (2.1)*, is characterized by the unique* $\{y, \lambda, \rho, p\}$ *solution to the singular optimality system*

$$\begin{cases} A\,y & = f + Bu, \\ A\,\rho & = \lambda, \\ B^*p + Nu & = 0, \end{cases} \qquad A^*p = C^*(Cy - z_d) + C^*C\rho,$$

with $\lambda \in \widehat{G}$.

Proof. From the relation (2.6) and Theorem 2.8, the sequence $\{u_\gamma\}$ weakly converges in $\mathcal{U}$ to u the unique no-regret control related to u_0. The operator B being continuous from $\mathcal{U}$ to $\mathcal{V}'$, $\{Bu_\gamma\}$ weakly converges in $\mathcal{V}'$ to Bu. Now, from the above optimality system of Proposition 2.9, the sequence $\{Ay_\gamma\}$ is bounded in $\mathcal{V}'$ and, as A is an isomorphism, weakly converges to Ay in $\mathcal{V}'$. Passing to the limit in the first equation we obtain $Ay = f + B\,u$. We also deduce from Proposition 2.9 that $B^*p_\gamma = -Nu_\gamma$ is bounded in $\mathcal{V}'$. According to the hypothesis (2.7), let $\mathcal{R}$ be the operator such that $\mathcal{R}(\frac{1}{\gamma}\beta^*\xi_\gamma) = B^*p_\gamma$. We deduce under (2.7) that $\{\frac{1}{\gamma}\beta^*\xi_\gamma\}$ is bounded in G subset of the Hilbert space F. Then it converges to $\lambda \in \widehat{G} \subset F$. Hence, $A\rho_\gamma = \frac{1}{\gamma}\beta^*\xi_\gamma$ is bounded, and then $\{\rho_\gamma\}$—also bounded thanks to the isomorphism of A—weakly converges to $\rho \in \mathcal{V}$. Consequently, $A\rho_\gamma \rightharpoonup A\rho$.

From the boundness of $\{\rho_\gamma\}$ and $\{y_\gamma\}$ we obtain that A^*p_γ is bounded. Then $\{p_\gamma\}$ converges to p. The optimality system follows. □

Remark 8. The situation described by Theorem 2.10, as indicated by Lions in [12], is completely general, but with λ which should be in the completion of G. This will be made precise in the following applications.

3. Application. In this section, we apply the above method throughout the examples given below in different situations: control and uncertainty given in the interior domain, as well as on the boundary.

Example 1. A distributed control, uncertain boundary values, and a boundary cost function.

Let Ω be a bounded open domain of $\mathbb{R}^N$ of regular boundary Γ. We consider the distributed system

$$\begin{cases} -\Delta y + y = f + v & \text{in } \Omega, \\ \dfrac{\partial y}{\partial \nu} = g & \text{on } \Gamma, \end{cases} \tag{3.1}$$

where $v \in \mathcal{U} = L^2(\Omega)$, and where $g \in G \subset F = L^2(\Gamma)$, G a closed subspace of F. If $f \in L^2(\Omega)$, there exists a unique $y(v, g) \in H^{3/2}(\Omega)$ solution to (3.1).

We associate with the state $y(v, g)$ the cost function

$$J(v, g) = \Big|y(v, g) - z_d\Big|^2_{L^2(\Gamma)} + N\Big\|v\Big\|^2_{L^2(\Omega)}. \tag{3.2}$$

For $u_0 \in \mathcal{U}$, there exists a unique no-regret control u related to u_0. For simplicity, take $u_0 = 0$. The problem now is to give the optimality system for the no-regret control u.

Notice that

$$J(v,g) - J(0,g) = J(v,0) - J(0,0) + 2\left(y(v,0) - y(0,0), y(0,g) - y(0,0)\right)_{L^2(\Gamma)}$$

and that the function $v \longmapsto y(v,0) - y(0,0)$ (resp., $g \longmapsto y(0,g) - y(0,0)$) is linear w.r.t v (resp., g) and is the solution to

$$\begin{cases} A\,z = v & \text{in } \Omega \\ \frac{\partial z}{\partial \nu} = 0 & \text{on } \Gamma \end{cases} \qquad \left(\text{resp.,} \quad \begin{cases} A\,z = 0 & \text{in } \Omega, \\ \frac{\partial z}{\partial \nu} = g & \text{on } \Gamma \end{cases}\right),$$

where $A = -\Delta + I$. Using the Green formula

$$(\varphi, A\psi)_{L^2(\Omega)} - (\psi, A\varphi)_{L^2(\Omega)} = \int_\Gamma \varphi \frac{\partial \psi}{\partial \nu_A} d\gamma - \int_\Gamma \psi \frac{\partial \varphi}{\partial \nu_A} d\gamma, \tag{3.3}$$

we find

$$0 = \int_\Gamma \left(y(0,g) - y(0,0)\right)\left(y(v,0) - y(0,0)\right) d\gamma - \int_\Gamma S(v)\, g\, d\gamma,$$

where $v \longmapsto S(v)$ is a linear function so that $AS = 0$, $\frac{\partial S}{\partial \nu} = y(v,0) - y(0,0)$.

Moreover, the following regularity result holds: We have $y(0,g) - y(0,0) \in H^{3/2}(\Omega)$ as $\frac{\partial}{\partial \nu_A}\left(y(0,g) - y(0,0)\right) \in L^2(\Gamma)$, and as $S(v) \in H^2(\Omega)$ we have also $\frac{\partial S}{\partial \nu} = y(v,0) - y(0,0) \in H^{3/2}(\Omega)$.

From section 2, the low-regret control method associated with (3.1)–(3.2) is defined by

$$\mathcal{J}^\gamma(v) = J(v,0) - J(0,0) + \frac{1}{\gamma}\left\|S(v)\right\|^2_{L^2(\Gamma)}, \tag{3.4}$$

where $S(v) = \zeta(v)$ is the solution of

$$\begin{cases} A\,S(v) = 0 & \text{in} \quad \Omega, \\ \frac{\partial S}{\partial \nu_A} = y(v,0) - y(0,0) & \text{on} \quad \Gamma. \end{cases}$$

The problem

$$\inf_{v \in \mathcal{U}} \mathcal{J}^\gamma(v) \tag{3.5}$$

admits a unique solution $v = u_\gamma$. Then the necessary condition of first order of Euler on $\mathcal{U}$ for every $w \in \mathcal{U}$ writes

$$\left(y(u_\gamma,0) - z_d, y(w,0) - y(0,0)\right) + (Nu_\gamma, w) + \left(\frac{1}{\gamma} S(u_\gamma), S(w)\right) \geq 0. \tag{3.6}$$

We have the following proposition.

Proposition 3.1. *The low-regret control u_γ solution to* (3.4)–(3.5) *is characterized by the unique $\{y_\gamma, \zeta_\gamma, \rho_\gamma, p_\gamma\}$ solution to*

$$\begin{cases} A\,y_\gamma = f + u_\gamma, & A\zeta_\gamma = 0, & A\,\rho_\gamma = 0, & A\,p_\gamma = 0, \\ \frac{\partial y_\gamma}{\partial \nu} = 0, & \frac{\partial \zeta_\gamma}{\partial \nu} = y_\gamma - y(0,0), & \frac{\partial \rho_\gamma}{\partial \nu} = \frac{1}{\gamma}\zeta_\gamma, & \frac{\partial p_\gamma}{\partial \nu} = y_\gamma - z_d + \rho_\gamma, \\ p_\gamma + Nu_\gamma = 0 & & & \text{in} \quad L^2(\Omega), \end{cases}$$

with

$$u_\gamma \in L^2(\Omega) \qquad \textit{and} \quad y_\gamma, \zeta_\gamma, \rho_\gamma, p_\gamma \in H^{3/2}(\Omega).$$

Proof. Denote $y_\gamma = y(u_\gamma, 0)$, and let $\zeta(u_\gamma)$ be the solution of $A\zeta_\gamma = 0$, $\frac{\partial \zeta_\gamma}{\partial \nu} = y_\gamma - y(0,0)$. Let now ρ_γ be the solution of $A\rho_\gamma = 0$, $\frac{\partial \rho_\gamma}{\partial \nu} = \frac{1}{\gamma}\zeta_\gamma$. Then by the above Green formula

$$\begin{aligned}\left(\frac{1}{\gamma}\zeta_\gamma(u_\gamma), \zeta_\gamma(w)\right)_{L^2(\Gamma)} &= (A\rho_\gamma, \zeta)_{L^2(\Omega)} - (\rho_\gamma, A\zeta_\gamma)_{L^2(\Omega)} + \left(\rho_\gamma, \frac{\partial \zeta_\gamma}{\partial \nu}\right)_{L^2(\Gamma)} \\ &= \left(\rho_\gamma, \frac{\partial \zeta_\gamma}{\partial \nu}\right)_{L^2(\Gamma)}.\end{aligned}$$

The inequality (3.6) becomes

$$(y_\gamma - z_d + \rho_\gamma, y(w,0) - y(0,0)) + (Nu_\gamma, w) \le 0.$$

Now, and as it is classical, we calculate the adjoint state p_γ such that $Ap_\gamma = 0$, $\frac{\partial p_\gamma}{\partial \nu} = y_\gamma - z_d + \rho_\gamma$.

This is for any w in the vector space $\mathcal{U}$. Then we have

$$p_\gamma \; + \; Nu_\gamma \; = 0. \qquad \square$$

Remark 9. The passage to the limit on γ for the no-regret control is an adaptation of the proof of the Theorem 2.10. Let us note that we do not need the hypothesis (2.7) as we have $B^* = B = Id$.

We obtain the following theorem.

THEOREM 3.2. *The no-regret control u related to $u_0 = 0$ of the problem* (3.1)–(3.2) *is characterized by the unique solution $\{y, \lambda, \rho, p\}$ of the optimality system*

$$\begin{cases} A\,y & = f + u, \quad A\,\rho = 0, \quad A\,p = 0 & \textit{in } \Omega, \\ \dfrac{\partial y}{\partial \nu} & = 0, \qquad \dfrac{\partial \rho}{\partial \nu} = \lambda, \quad \dfrac{\partial p}{\partial \nu} = y - z_d + \rho & \textit{on } \Gamma, \\ p + Nu & = 0 & \textit{in } L^2(\Omega), \end{cases}$$

with

$$\begin{cases} u \in L^2(\Omega),\ y \in H^{3/2}(\Omega), & p \in L^2(\Omega), \\ \lambda \in \widehat{G} \qquad \textit{completion of } G \textit{ for the norm } H^{-5/2}(\Gamma),\ \rho \in H^{-1}(\Omega). \end{cases} \qquad \square$$

Example 2. A boundary control, boundary uncertainty, boundary cost function.

Let Ω be an open domain from $\mathbb{R}^N$ of boundary $\partial\Omega = \Gamma_0 \cup \Gamma_1$, with Γ_0 and Γ_1 being two regular boundaries such that $\Gamma_0 \cap \Gamma_1 = \emptyset$.

We consider the distributed parameter system

$$\text{(3.7)} \qquad \begin{cases} -\Delta y + y = 0 & \text{in } \Omega, \\ \dfrac{\partial y}{\partial \nu} = v & \text{on } \Gamma_0, \\ \dfrac{\partial y}{\partial \nu} = g & \text{on } \Gamma_1. \end{cases}$$

For $v \in \mathcal{U} = L^2(\Gamma_0)$ and $g \in G \subset L^2(\Gamma_1)$, (3.7) admits a unique solution $y(v,g) \in H^{3/2}(\Omega)$.

We associate with the state $y(v,g)$ the cost function

$$J(v,g) = \Big|y(v,g) - z_d\Big|^2_{L^2(\Gamma_0)} + N\Big|v\Big|^2_{L^2(\Gamma_0)}. \tag{3.8}$$

For u_0 fixed in $\mathcal{U}$, there exists a unique no-regret control u related to u_0. We suppose that $u_0 = 0$.

The low-regret control associated is defined by the following cost function:

$$\mathcal{J}^\gamma(v) = J(v,0) - J(0,0) + \frac{1}{\gamma}\Big\|S(v)\Big\|^2_{L^2(\Gamma_1)}, \tag{3.9}$$

where $S(v) = \zeta(v)$ is the solution to

$$\begin{cases} A\,S(v) = 0 & \text{in } \Omega, \\ \dfrac{\partial S}{\partial \nu} = y(v,0) & \text{on } \Gamma_0, \\ \dfrac{\partial S}{\partial \nu} = 0 & \text{on } \Gamma_1 \end{cases} \tag{3.10}$$

and where $A = -\Delta + I$.

Indeed,

$$J(v,g) - J(0,g) = J(v,0) - J(0,0) + 2(y(v,0), y(0,g))_{L^2(\Gamma_0)\times L^2(\Gamma_0)}.$$

Then by the Green formula we obtain

$$(y(v,0), y(0,g))_{L^2(\Gamma_0)\times L^2(\Gamma_0)} = (S(v), g)_{L^2(\Gamma_0)\times L^2(\Gamma_0)},$$

with $S(.)$ the solution to (3.10). The problem

$$\inf_{v\in\mathcal{U}} \mathcal{J}^\gamma(v) \tag{3.11}$$

admits a unique solution u_γ called the low-regret control.

PROPOSITION 3.3. *The low-regret control u_γ solution to* (3.9)–(3.11) *is characterized by the unique solution $\{y_\gamma, \zeta_\gamma, \rho_\gamma, p_\gamma\}$ of the optimality system*

$$\begin{cases} A\,y_\gamma = 0, & A\zeta_\gamma = 0, & A\,\rho_\gamma = 0, & A\,p_\gamma = 0 & \text{in} \quad \Omega, \\ \dfrac{\partial y_\gamma}{\partial \nu} = u_\gamma, & \dfrac{\partial \zeta_\gamma}{\partial \nu} = y_\gamma, & \dfrac{\partial \rho_\gamma}{\partial \nu} = 0, & \dfrac{\partial p_\gamma}{\partial \nu} = y_\gamma - z_d + \rho_\gamma & \text{on} \quad \Gamma_0, \\ \dfrac{\partial y_\gamma}{\partial \nu} = 0, & \dfrac{\partial \zeta_\gamma}{\partial \nu} = 0, & \dfrac{\partial \rho_\gamma}{\partial \nu} = \frac{1}{\gamma}\zeta_\gamma, & \dfrac{\partial p_\gamma}{\partial \nu} = 0 & \text{on} \quad \Gamma_1, \\ p_\gamma + N u_\gamma = 0 & & & & \text{on} \quad \Gamma_0, \end{cases}$$

with, $u_\gamma \in L^2(\Gamma_0)$, and $y_\gamma \in H^{3/2}(\Omega), \zeta_\gamma \in H^{5/2}(\Omega), \rho_\gamma \in H^{7/2}(\Omega), p_\gamma \in H^{1/2}(\Omega)$.

Proof. The Euler condition gives

$$(y_\gamma - z_d, y(w,0))_{L^2(\Gamma_1)\times L^2(\Gamma_1)} + N\,(u_\gamma, w)_{L^2(\Gamma_0)\times L^2(\Gamma_0)} + \left(\frac{1}{\gamma}\xi(u_\gamma), \xi(w)\right)_{L^2(\Gamma_1)\times L^2(\Gamma_1)} \geq 0. \tag{3.12}$$

We first solve for ρ_γ: $A\,\rho_\gamma = 0$, with $\frac{\partial \rho_\gamma}{\partial \nu} = 0$ on Γ_0, and $\frac{\partial \rho_\gamma}{\partial \nu} = \frac{1}{\gamma}\xi(u_\gamma)$ on Γ_1. Hence

$$\left(\frac{1}{\gamma}\xi(u_\gamma), \xi(w)\right)_{L^2(\Gamma_1)\times L^2(\Gamma_1)} = (\rho_\gamma, y(w,0))_{L^2(\Gamma_0)\times L^2(\Gamma_0)}$$

so that (3.12) becomes

$$(y_\gamma - z_d + \rho_\gamma, y(w,0))_{L^2(\Gamma_0)\times L^2(\Gamma_0)} + N\,(u_\gamma, w)_{L^2(\Gamma_0)\times L^2(\Gamma_0)} \geq 0.$$

Let now p_γ be the solution of $A\,p_\gamma = 0$, with $\frac{\partial p_\gamma}{\partial \nu} = y_\gamma - z_d$ on Γ_0, and $\frac{\partial p_\gamma}{\partial \nu} = 0$ on Γ_1.

We have then

$$(y_\gamma - z_d + \rho_\gamma, y(w,0))_{L^2(\Gamma_0)\times L^2(\Gamma_0)} = (p_\gamma, w)_{L^2(\Gamma_0)\times L^2(\Gamma_0)}.$$

Finally, as $L^2(\Gamma_0)$ is a vector space, we have

$$p_\gamma \;+\; N u_\gamma \;= 0 \qquad \forall w \in L^2(\Gamma_0). \qquad \square$$

The passage to the limit on γ leads to the following theorem.

THEOREM 3.4. *The no-regret control u of the system* (3.7)–(3.8) *is characterized by the unique solution $\{y, \lambda, \rho, p\}$ of the optimality system*

$$\begin{cases} A\,y = 0, & A\,\rho = 0, & A\,p = 0 & \text{in} \quad \Omega, \\ \dfrac{\partial y}{\partial \nu} = u, & \dfrac{\partial \rho}{\partial \nu} = 0, & \dfrac{\partial p}{\partial \nu} = y - z_d + \rho & \text{on} \quad \Gamma_0, \\ \dfrac{\partial y}{\partial \nu} = 0, & \dfrac{\partial \rho}{\partial \nu} = \lambda, & \dfrac{\partial p}{\partial \nu} = 0 & \text{on} \quad \Gamma_1, \\ p + Nu = 0 & & & \text{in} \quad L^2(\Gamma_0), \end{cases}$$

with

$$\begin{cases} u \in L^2(\Omega), & y \in H^{3/2}(\Omega), & p \in H^{1/2}(\Omega), \\ \lambda \in \widehat{G} \text{ completion of } G \text{ in } H^{-2}(\Gamma), & \rho \in H^{-1/2}(\Omega). \end{cases} \qquad \square$$

4. The evolution case.

4.1. No-regret control for systems of parabolic type. In this section, $A \in \mathcal{L}(\mathcal{V}; \mathcal{V}')$ is an elliptic differential operator

$$(\,A\,v\,,\,v\,) \geq \alpha \left\|v\right\|^2, \quad \alpha > 0, \quad \left\|\;\right\| = \text{norm in} \quad \mathcal{V},$$

$B \in \mathcal{L}\left(\mathcal{U}; L^2(0,T;\mathcal{V}')\right)$, and F is the real Hilbert space of uncertainties such that

$$\mathcal{V} \;\subset\; F \;\subset\; \mathcal{V}'.$$

Let then G be the closed vector subspace of F.

For $f \in L^2(0,T;\mathcal{V}')$, the state equation that we consider is

$$\frac{\partial y}{\partial t} \;+\; A\,y \;=\; f \;+ B\,v, \tag{4.1}$$

with

$$y(t=0,v,g) = y_0 + g, \tag{4.2}$$

where y_0 is a given data in F and where $g \in G$.

For chosen v and g, the problem (4.1)–(4.2) admits a unique solution noted $y(v,g) \in L^2(0,T;\mathcal{V})$.

For a fixed $t \in (0,T)$, and for any $g \in G$ we have then a possible state for which we attach a cost function given by

$$J(v,g) = \int_0^T \|Cy(v,g) - z_d\|^2_{\mathcal{H}} dt + N \int_0^T \|v\|^2_{\mathcal{U}}\, dt, \tag{4.3}$$

where

$$C \in \mathcal{L}\,(\,L^2(0,T;\mathcal{V});\mathcal{H}\,), \tag{4.4}$$

the set $\mathcal{H}$ is a Hilbert space, $z_d \in \mathcal{H}$ fixed, $N > 0$, and $\|.\|_X$ represents the norm defined on the Hilbert space X.

When $G = \{0\}$, a standard control problem is to find

$$\inf_{v \in \mathcal{U}} J(v,0). \tag{4.5}$$

We now develop the approach of the first part to this evolution case, when $G \neq \{0\}$.

4.1.1. Least regret control. Approached optimality system. Following the lines of [13] and using the notations in [17], we have then

$$J(v,g) - J(u_0,g) = J(v,0) - J(u_0,0) + 2\langle\, \xi(v-u_0), g\,\rangle_{G' \times G}, \tag{4.6}$$

where

$$S(v) = \zeta(t=0,v) \tag{4.7}$$

and where ζ is the solution to the backwards problem

$$\begin{cases} -\zeta' + A^*\,\zeta \;= C^*C(y(v,0) - y(0,0)), \\ \zeta\,(t=T,v) = 0, \end{cases} \tag{4.8}$$

with $\zeta' = \frac{\partial \zeta}{\partial t}$.

Then the low-regret control associated with the problem (4.1)–(4.3) is defined by

$$\inf_{v \in \mathcal{U}} \mathcal{J}^\gamma(v), \tag{4.9}$$

$$\mathcal{J}^\gamma(v) = J(v,0) - J(u_0,0) + \frac{1}{\gamma}\Big\|\zeta(0,v-u_0)\Big\|^2_{G'}, \tag{4.10}$$

where G' is the dual of G which can be identified to G. The problem (4.9)–(4.10) has a unique solution u_γ called low-regret control.

PROPOSITION 4.1. *The low-regret control u_γ solution to* (4.9)–(4.10) *is characterized by the unique solution $\{y_\gamma, \zeta_\gamma, \rho_\gamma, p_\gamma\}$ of the optimality system*

$$\begin{cases} y'_\gamma \,+\, A\,y_\gamma & = f + Bu_\gamma, & -\zeta'_\gamma \,+\, A^*\zeta_\gamma \,=\, C^*C(y_\gamma - y(0,0)), \\ \rho'_\gamma \,+\, A\rho_\gamma & = 0, & -p'_\gamma \,+\, A^*p_\gamma \,=\, C^*(Cy_\gamma - z_d) + C^*C\rho_\gamma, \\ y_\gamma(t=0) = y_0, & \rho_\gamma(0) = \frac{1}{\gamma}\zeta_\gamma, & \zeta_\gamma(T) = 0, \qquad p_\gamma(T) = 0, \\ B^*p_\gamma + Nu_\gamma & = 0 \quad \textit{in} \quad \mathcal{U}. \end{cases}$$

Proof. Let u_γ be the solution of the problem (4.9)–(4.10). The Euler first order condition gives the following optimality system:

(4.11)

$$(\,Cy(u_\gamma,0)-z_d, C(y(w,0)-y(0,0))\,)_{\mathcal{H}\times\mathcal{H}} + (\,Nu_\gamma, w\,)_{\mathcal{U}\times\mathcal{U}} + \left(\frac{1}{\gamma}\zeta(0,u_\gamma-u_0),\zeta(0,w)\right)_{G\times G} \geq 0.$$

With this in mind, let $y_\gamma = y(u_\gamma,0)$, and look for $\zeta_\gamma = \zeta(0,u_\gamma-u_0)$ to be the solution of (4.8) and $\rho_\gamma \in \mathcal{V}$ the solution of

$$\begin{cases} \rho'_\gamma + A\rho_\gamma = 0, \\ \rho_\gamma(t=0) = \frac{1}{\gamma}\zeta_\gamma. \end{cases}$$

As it is classical, we introduce the adjoint state p_γ defined by

$$-p'_\gamma + A^* p_\gamma = C^*(Cy_\gamma - z_d) + C^*C\rho_\gamma, \qquad \text{with} \quad p_\gamma(T)=0.$$

Hence we deduce from (4.11)

$$B^* p_\gamma + Nu_\gamma = 0 \qquad \text{in} \quad \mathcal{U}. \tag{4.12}$$

This ends the proof. ∎

4.1.2. Singular optimality system. We now give the optimality system for the no-regret control. We need a supplementary hypothesis. Let $\rho \in L^2(0,T;V)$ be defined by

$$\rho' + A\,\rho = 0, \quad \rho(0) = g, \quad g \in G,$$

and $\sigma \in L^2(0,T;V)$ as

$$-\sigma' + A^*\sigma = C^*C\,\rho, \quad \sigma(T)=0.$$

Setting $R\,g = B^*\sigma$, then we define the continuous operator $g \mapsto Rg$ from F to $\mathcal{U}$, and we do the hypothesis

$$\Big\|R\,g\Big\|_{\mathcal{U}} \geq c\,\Big\|g\Big\|_{F} \quad c>0 \quad \forall g \in G. \tag{4.13}$$

THEOREM 4.2. *We suppose that* (4.13) *holds true. Then the no-regret control* u *related to* u_0*, for the system* (4.1)–(4.3)*, is characterized by the unique solution* $\{y,\lambda,\rho,p\}$ *to the optimality system*

$$\begin{cases} y' + Ay = f + Bu, & -\zeta' + A^*\zeta = C^*C(y-y(0,0)), \\ \rho' + A\rho = 0, & -p' + A^*p = C^*(Cy-z_d) + C^*C\rho, \\ y(0)=y_0, \quad \rho(0)=\lambda, & \zeta(T,u)=0, \quad p(T)=0, \\ B^*p + Nu = 0 \quad in \quad \mathcal{U}, & \end{cases}$$

with $\lambda \in \widehat{G}$.

Proof. The proof holds from the approached optimality system of Proposition 4.1 for which a priori estimates allow us to pass to the limit when $\gamma \to 0$ as in section 2. ∎

4.2. No-regret control for well-posed systems of Petrowsky type. We now consider an elliptic differential operator A such as

$$A^* = A,$$

and to simplify we consider the state equation

$$y'' + A\,y = v, \tag{4.14}$$

with

$$y \in L^\infty(0,T;V), \qquad y' \in L^\infty(0,T;F), \tag{4.15}$$

$$y(0) = y_0 + g_0, \qquad y'(0) = y_1 + g_1, \tag{4.16}$$

where $\{y_0, y_1\}$ is bounded in $\mathcal{V} \times F$ and where

$$\begin{cases} g_0 \in G_0, & G_0 = \text{ closed vector subspace of } \mathcal{V}, \\ g_1 \in G_1, & G_1 = \text{ closed vector subspace of } F. \end{cases} \tag{4.17}$$

Let $y(v,g)$ be the solution of (4.14)–(4.16), $g = (g_0, g_1)$. Let C be defined by (4.4) and the cost function $J(v,g)$ be defined by (4.3). We look for the no-regret control related to $u_0 = 0$. We define $y = y(v,0)$ and $\zeta(t,v)$ (or $\zeta(t)$), respectively, by

$$y'' + A\,y = v, \qquad y(t=0) = y_0, \qquad y'(t=0) = y_1, \tag{4.18}$$

$$\zeta'' + A\,\zeta = C^*C\,y(v,0), \qquad \zeta(T) = 0, \qquad \zeta'(T) = 0. \tag{4.19}$$

Set $z = y(0,g) - y(0,0)$. Then z is the solution of

$$\begin{cases} z'' + A\,z = 0, \\ z(0) = g_0, \\ z'(0) = g_1. \end{cases}$$

Then by the Green formula we obtain

$$\begin{aligned} J(v,g) - J(0,g) &= J(v,0) - J(0,0) + 2\int_0^T (\,\zeta'' + \Delta\zeta, z\,)\,dt \\ &= J(v,0) - J(0,0) + 2\,(\,\zeta(0), g_1\,)_{G_0,G_1} - 2\,(\,\zeta'(0), g_0\,)_{G_1,G_0}. \end{aligned}$$

As the low-regret control solution is defined by the

$$\inf_{v\in\mathcal{U}} \left(\sup_{g\in G_0\times G_1} \left(J(v,g) - J(0,g) + \gamma \left\| g_0 \right\|_{G_0}^2 - \gamma \left\| g_1 \right\|_{G_1}^2 \right) \right),$$

the low-regret control method reads

$$\inf_{v\in\mathcal{U}} \mathcal{J}^\gamma(v) = \inf_{v\in\mathcal{U}} \left(J(v,0) - J(0,0) + \frac{1}{\gamma} \left\| \zeta(0) \right\|_{G_1}^2 - \frac{1}{\gamma} \left\| \zeta'(0) \right\|_{G_0}^2 \right). \tag{4.20}$$

And we have for the Petrowsky systems the following result.

THEOREM 4.3. *The no-regret control u related to $u_0 = 0$ is characterized by the unique solution $\{y, \lambda_0, \lambda_1, \zeta, \rho, p\}$ to the optimality system*

$$\begin{cases} y'' + A\,y = 0, & \zeta'' + A\,\zeta = 0, & \rho'' + A\,\rho = 0, & p'' + A\,p = 0, \\ y(0) = y_0, & \zeta(T) = 0, & \rho(0) = \lambda_0, & p(T) = 0, \\ y'(0) = y_1, & \zeta'(T) = 0, & \rho'(0) = \lambda_1, & p'(T) = 0, \\ p + Nu = 0, & & & \end{cases}$$

with

$$\begin{cases} \lambda_0 = -\lim\limits_{\gamma\to 0} \dfrac{1}{\gamma}\zeta'(0),\ \lambda_0 \in \widehat{G_0} & \textit{completion of} \quad G_0 \quad \textit{for the norm} \quad \left\|.\right\|_{G_0}, \\ \lambda_1 = \lim\limits_{\gamma\to 0} \dfrac{1}{\gamma}\zeta(0), \quad \lambda_1 \in \widehat{G_1} & \textit{completion of} \quad G_1 \quad \textit{for the norm} \quad \left\|.\right\|_{G_1}. \end{cases} \qquad \square$$

Remark 10. These results are also valid for well-posed problems of hyperbolic type.

5. Application. Hereafter, we discuss an example of parabolic type with boundary control, boundary uncertainty, and cost function.

Let Ω be an open set of $\mathbb{R}^N$ of boundary $\Gamma_0 \cup \Gamma_1$, with Γ_0 and Γ_1 being two regular boundaries of empty set intersection. We consider the distributed system

$$\begin{cases} y' - \Delta\,y + y = 0 & \text{in } \Omega, \\ y(0,x,v,g) = 0 & \text{on } \{0\}\times\Omega, \\ \dfrac{\partial y}{\partial \nu} = v & \text{on }]0,T[\times\Gamma_0 = \Sigma_0, \\ \dfrac{\partial y}{\partial \nu} = g & \text{on }]0,T[\times\Gamma_1 = \Sigma_1. \end{cases} \tag{5.1}$$

For $v \in \mathcal{U} = L^2(\Sigma_0)$, $g \in G$, a vector closed subspace of $L^2(\Sigma_1)$, (5.1) has a unique solution $y(t,x,v,g)$ noted $y(v,g)$. We associate with the state $y(v,g)$ the cost function

$$J(v,g) = \Big|y(v,g) - z_d\Big|^2_{L^2(\Sigma_0)} + N\Big|v\Big|^2_{L^2(\Sigma_0)}. \tag{5.2}$$

For u_0 fixed in $\mathcal{U}$, there exists a unique control u related to u_0. Take $u_0 = 0$. Then the associated low-regret control is defined by the following cost function:

$$\mathcal{J}^\gamma(v) = J(v,0) - J(0,0) + \frac{1}{\gamma}\Big|\zeta(v)\Big|^2_{L^2(\Sigma_1)}, \tag{5.3}$$

where ζ is the solution of

$$\begin{cases} -\zeta' + A\,\zeta = 0 & \text{in } \Omega, \\ \zeta(T,v) = 0 & \text{on } \{T\}\times\Omega, \\ \dfrac{\partial \zeta}{\partial \nu_A} = y(v,0) & \text{on } \Sigma_0, \\ \dfrac{\partial \zeta}{\partial \nu_A} = 0 & \text{on } \Sigma_1 \end{cases}$$

and where $A = -\Delta + I = A^*$.

The problem

$$\inf_{v\in\mathcal{U}} \mathcal{J}^{\gamma}(v) \tag{5.4}$$

has a unique solution: the low-regret control u_γ. We set $y_\gamma = y(u_\gamma, 0)$ and $\zeta_\gamma = \zeta(u_\gamma)$. We then have immediately the following proposition.

PROPOSITION 5.1. *The low-regret control u_γ is characterized by the unique solution $\{y_\gamma, \zeta_\gamma, \rho_\gamma, p_\gamma\}$ of the optimality system*

$$\begin{cases} y_\gamma' + A\,y_\gamma = 0, & -\zeta_\gamma' + A\,\zeta = 0, & \rho_\gamma + A\,\rho_\gamma = 0, & -p_\gamma' + A\,p_\gamma = 0, \\ y_\gamma(0) = 0, & \zeta_\gamma(T) = 0, & \rho_\gamma(0) = 0, & p_\gamma(T) = 0, \\ \text{on}\quad \Sigma_0, \quad \dfrac{\partial y_\gamma}{\partial \nu} = u_\gamma, & \dfrac{\partial \zeta_\gamma}{\partial \nu} = y(v,0), & \dfrac{\partial \rho_\gamma}{\partial \nu} = 0, & \dfrac{\partial p_\gamma}{\partial \nu} = y_\gamma - z_d + \rho_\gamma, \\ \text{on}\quad \Sigma_1, \quad \dfrac{\partial y_\gamma}{\partial \nu} = 0, & \dfrac{\partial \zeta_\gamma}{\partial \nu} = 0, & \dfrac{\partial \rho_\gamma}{\partial \nu} = \frac{1}{\gamma}\zeta_\gamma, & \dfrac{\partial p_\gamma}{\partial \nu} = 0, \\ p_\gamma + N u_\gamma = 0 & & & \text{in}\quad L^2(\Sigma_0), \end{cases}$$

with,
$u_\gamma \in L^2(\Sigma_0)$, *and* $y_\gamma \in L^2((0,T); H^{3/2}(\Omega))$, $\zeta_\gamma \in L^2((0,T); H^{5/2}(\Omega))$, $\rho_\gamma \in L^2((0,T); H^{7/2}(\Omega))$, $p_\gamma \in L^2((0,T); H^{1/2}(\Omega))$.

For the proof, we use the same technique as detailed for the stationary Example 2 in section 3. □

We also deduce easily the following theorem.

THEOREM 5.2. *The no-regret control u related to $u_0 = 0$ of the system* (5.1)–(5.2) *is characterized by the unique solution $\{y, \lambda, \rho, p\}$ of the optimality system*

$$\begin{cases} y' + A\,y = 0, & -\zeta' + A\zeta = 0, & \rho' + A\rho = 0, & -p' + A\,p = 0 & \text{in}\quad \Omega, \\ y(0) = 0, & \zeta(T) = 0, & \rho(0) = 0, & p(T) = 0, & \\ \dfrac{\partial y}{\partial \nu} = u, & \dfrac{\partial \zeta}{\partial \nu} = y, & \dfrac{\partial \rho}{\partial \nu} = 0, & \dfrac{\partial p}{\partial \nu} = y - z_d + \rho & \text{on}\quad \Sigma_0, \\ \dfrac{\partial y}{\partial \nu} = 0, & \dfrac{\partial \zeta}{\partial \nu} = 0, & \dfrac{\partial \rho}{\partial \nu} = \lambda, & \dfrac{\partial p}{\partial \nu} = 0 & \text{on}\quad \Sigma_1, \\ p + N u = 0 & & & & \text{in}\quad L^2(\Sigma_0), \end{cases}$$

with

$$\begin{cases} u \in L^2((0,T); L^2(\Omega)), \quad y \in L^2((0,T); H^{3/2}(\Omega)), \\ \lambda \in \widehat{G}\ \textit{completion of}\ G\ \textit{in}\ L^2((0,T); H^{-2}(\Sigma_1)), \\ \rho \in L^2((0,T); H^{-1/2}(\Omega)), \quad p \in L^2((0,T); H^{1/2}(\Omega)). \end{cases}$$ □

Conclusion. As we have seen, the low-regret control method allows us to transform systematically a problem with uncertainty to a standard control problem. It is then easier to obtain optimality systems applying the Euler–Lagrange formula.

This method can be used for the control of singular distributed systems as in [4] (see also [5]). Here, the singularity of the backward heat equation is taken off by adding the needed data which may belong to the unknown vector closed subspace G of a given Hilbert space of uncertainties. The system becomes regular, but it contains incomplete data. We then give an optimality system to the no-regret control. In [4], the comparison with the classical penalization method for the control of the backward heat equation in Lions [11] is discussed.

REFERENCES

[1] J. C. Allwright, *Deterministic optimal control,* J. Optim. Theory Appl., 32 (1980), pp. 327–344.

[2] J. P. Aubin, *L'analyse non linéaire et ses motivations économiques*, Masson, Paris-New York, 1984.

[3] Y. Censor, *Optimality in multi-objective problems,* Appl. Math. Optim., 149 (1977), pp. 41–59.

[4] R. Dorville, *Sur le contrôle de quelques problèmes mal posés associés à l'équation de la chaleur*, Ph.D. thesis, Université des Antilles et de la Guyane, Guadeloupe (French West Indies), to appear.

[5] R. Dorville, O. Nakoulima, and A. Omrane, *Low-regret control for singular distributed systems: The backwards heat ill-posed problem*, Appl. Math. Lett., to appear.

[6] D. Gabay, *private communication*, Almeria, 1992.

[7] D. Gabay and J. L. Lions, *Décisions stratégiques à moindres regrets*, C. R. Acad. Sci. Paris Ser. I Math., 319 (1994), pp. 1249–1256.

[8] W. Kotarski, *Characterization of Pareto optimal points in problems with multi-equality constraints,* Optimization, 20 (1989), pp. 93–106.

[9] W. Kotarski, *Some Problems of Optimal and Pareto Optimal Control for Distributed Parameter Systems,* Pr. Nauk. Uniw. Sl. Katow. 1668, Wydawnictwo Universytetu Slaskiego, Katowice, Poland, 1997.

[10] J. L. Lions, *Contrôle optimal des systèmes gouvernés par des équations aux dérivées partielles,* Dunod, Paris, 1969.

[11] J. L. Lions, *Contrôle optimal pour les systèmes distribués singuliers,* Guathiers-Villard, Paris, 1983.

[12] J. L. Lions, *Contrôle de Pareto de systèmes distribués. Le cas stationnaire,* C. R. Acad. Sci. Paris Ser. I Math., 302 (1986), pp. 223–227.

[13] J. L. Lions, *Contrôle de Pareto de systèmes distribués. Le cas d'évolution,* C. R. Acad. Sci. Paris Ser. I Math., 302 (1986), pp. 413–417.

[14] J. L. Lions, *Contrôle à moindres regrets des systèmes distribués,* C. R. Acad. Sci. Paris Ser. I Math., 315 (1992), pp. 1253–1257.

[15] J. L. Lions, *No-regret and low-regret control,* Environment, Economics and Their Mathematical Models, Masson, Paris, 1994.

[16] J. L. Lions, *Duality Arguments for Multi Agents Least-Regret Control,* Collège de France, Paris, 1999.

[17] O. Nakoulima, A. Omrane, and J. Velin, *Perturbations à moindres regrets dans les systèmes distribués à données manquantes,* C. R. Acad. Sci. Paris Ser. I Math., 330 (2000), pp. 801–806.

[18] L. J. Savage, *The Foundations of Statistics,* 2nd ed., Dover, New York, 1972.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1185–1197

STABILITY CRITERIA FOR LINEAR SYSTEMS WITH PARTLY UNCERTAIN PERIODIC COEFFICIENTS*

ALEXANDR ZEVIN† AND MARK PINSKY‡

Abstract. Stability criteria of partly uncertain linear systems with variable coefficients play an important part in various fields of robust control. In this paper we consider linear systems with periodic coefficients that may vary within their known lower and upper bounds depending upon time as well. Stability analysis of systems with variable coefficients is often based on the Lyapunov functions method leading in many cases to rigid sufficient stability conditions. A new approach to stability analysis of such systems is described in this paper. Using this method we obtain an upper bound for the real parts of the characteristic exponents of the system solutions yielding sufficient stability criteria and find the cases in which these bounds are reached. This explicitly determines the "worst" possible periodic coefficients yielding the largest value of the maximal real part of the characteristic exponents. Certain results are extended to the case of nonperiodic time-dependent bounded coefficients.

Key words. linear systems, uncertain periodic coefficients, stability criteria

AMS subject classifications. 34D08, 93C05

DOI. 10.1137/S0363012901385526

1. Introduction.

We consider a linear system with uncertain periodic coefficients

$$\begin{gathered}\dot{x} = A(t)x,\\ x \in R^n,\ A(t) = [a_{ik}(t)]_{i,k=1}^n,\quad a_{ik}(t) = a_{ik}(t+T).\end{gathered} \tag{1.1}$$

It is assumed that the coefficients are piecewise continuous and lie within the prescribed limits:

$$a_{ik}^-(t) \le a_{ik}(t) \le a_{ik}^+(t). \tag{1.2}$$

Sometimes it is more convenient to represent the matrix $A(t)$ as a sum of fixed and uncertain matrices:

$$\begin{gathered}A(t) = A^0(t) + \Delta(t),\\ A^0(t) = [a_{ik}^0(t)]_{i,k=1}^n, \qquad \Delta(t) = [\delta_{ik}(t)]_{i,k=1}^n,\\ a_{ik}^0(t) = \tfrac{1}{2}\left[a_{ik}^-(t) + a_{ik}^+(t)\right],\quad |\delta_{ik}(t)| \le \tfrac{1}{2}\left[a_{ik}^+(t) - a_{ik}^-(t)\right].\end{gathered} \tag{1.3}$$

In the field of robust stability (see, for example, [1]) restrictions on uncertain entries of the matrix are often imposed on an upper bound δ_0 for the norm of the uncertain matrix $\Delta(t)$. Both of these assumptions are considered in this paper.

Stability problems for linear systems with partly uncertain periodic coefficients arise in numerous applications as well as in stability analysis of equilibrium or periodic solutions of nonlinear systems:

$$\begin{gathered}\dot{y} = f(y,t),\\ f(y,t) = f(y,t+T),\ f(0,t) = 0,\ y \in R^n,\end{gathered} \tag{1.4}$$

*Received by the editors February 23, 2001; accepted for publication (in revised form) January 17, 2003; published electronically August 6, 2003.
http://www.siam.org/journals/sicon/42-4/38552.html

†Transmag Research Institute, Academy of Sciences of Ukraine, 49005 Dnepropetrovsk, Piesarzhevsky 5, Ukraine (zevin@npkista.dp.ua).
‡Mathematics Department, University of Nevada, Reno, Reno, NV 89511 (pinsky@unr.edu).

where the function $f(y,t)$ is not precisely known. Stability of equilibrium $y(t) \equiv 0$ is determined by variational equation (1.1) where $a_{ik}(t) = \partial f_i(y,t)/\partial y_k$ for $y = 0$. This problem turns out to be the one stated above if bilateral bounds for the elements $a_{ik}(t)$ are known.

A variation of the previous problem could be described as follows. Assume that system (1.4) is precisely defined and existence of its periodic solution within a compact region Ω is established, but the solution itself is unknown. Then one may find bilateral bounds for the elements of the Jacobian $f_y(y,t)$ in Ω which imply the considered problem.

Stability analysis of time-invariant linear systems with partly uncertain coefficients attracted a flood of publications in the last two decades, and a few approaches for estimating of robust stability of such systems have been developed. (See, for example, monograph [2] for a recent review on this subject.) Indeed, these techniques cannot be directly generalized on time-varying (in particular, periodic) uncertain systems since this problem turns out to be considerably more complicated. Mentioning that known results on stability of uncertain time-varying systems (see, e.g., [3], [4], and [5]) are based on applications of the Lyapunov functions method often yielding rigid sufficient stability criteria.

A new approach to stability analysis of partly uncertain linear periodic (and more general time-varying) systems is developed in this paper. This approach does not rely on availability of suitable Lyapunov functions and often yields more accurate and informative stability criteria leading to a robust stability control. Let us introduce briefly the main idea of this approach.

Let $W(t) = [x_1(t), \dots, x_n(t)]$ be a transition matrix of (1.1) with a periodic matrix $A(t) = A(t+T)$. ($W(0) = I$, where I is the unit matrix.) It is known (see, for example, [6]) that the stability of this system is determined by the eigenvalue (Floquet multiplier) ρ_{n} of the monodromy matrix $W(T)$ with the largest modulus. Namely, if $|\rho_n| < 1$ and, therefore, the real part of the characteristic exponent $\alpha_n = (\ln \rho_n)/T$ is negative, the system is asymptotically stable; if $|\rho_n| > 1$ ($\mathrm{Re}\alpha_n > 0$), the system is unstable.

Denote by $U(t)$ the set of $n \times n$ matrices $A(t)$ satisfying conditions (1.2); let $\alpha_+ = \sup \mathrm{Re}\alpha_{\mathrm{n}}(A)$ for $A(t) \in U(t)$. Equation (1.1) is asymptotically stable for any $A(t) \in U(t)$ if $\alpha_+ < 0$.

Our approach is based on deriving an upper bound ρ^* for the largest in modulus multiplier and, hence, the upper bound α^* for α_+. In contrast to the Lyapunov functions method, our approach also estimates the degree of system stability or instability of the initial system. In fact, if $\alpha^* < 0$, then the Euclidean norm of any solution of (1.1), $\|x(t)\| < C\exp(\alpha^* t)$ for some C and $t > 0$. The condition of $\alpha^* > 0$ leads to a sufficient stabilizing feedback control $u = C(t)x$; for example, the system $\dot{x}(t) = A(t)x - aIx$ with $a > \alpha^*$ is certainly asymptotically stable.

Theorem 1 determines an upper bound for the value $\mathrm{Re}\alpha_n(A)$ of a given matrix $A(t)$. This bound is defined by the largest real characteristic exponent $\alpha_n(B)$ of the system $\dot{x} = B(t)x$ with a matrix $B(t+T) = B(t)$ such that the entries $a_{ik}(t)$ and $b_{ik}(t)$ of matrices $A(t)$ and $B(t)$ satisfy inequalities (2.2). Its proof (section 4) is based on the following steps.

Let ρ_i be a real multiplier of (1.1); then there exists the real solution $x_i(t) = \exp(\alpha_i t)u_i(t)$, where $\alpha_i = (\ln \rho_i)/T$ and $u_i(t)$ is a T-periodic function. Using the corresponding Green's function, we transform (1.1) to the integral equation $u = R(A)u$. In the analogous equation for (2.1), the corresponding integral operator $R(B)$ is strongly

positive (transforms any nonnegative function $u(t)$ into positive one), which enables us to utilize some profound results in the theory of positive operators. (See [7] and [8] for additional information.) As a result, it is proved that the multiplier with largest modulus of the equation $\dot{x} = B(t)x$ is real and exceeds the multiplier ρ_i. Next, we extend this inequality to the value $|\rho_i|$ of any complex multiplier ρ_i. Using lower and upper bounds $a_{ik}^-(t)$ and $a_{ik}^+(t)$ of $a_{ik}(t)$, we construct a matrix $B(t)$ which satisfies inequalities (2.2) for any $A(t) \in U(t)$ and, thus, yields a required bound for the value α_+. We found cases in which the corresponding matrix $B(t) \in U(t)$. This implies that $A(t) = B(t)$ is the "worst" admissible matrix with the corresponding value $\alpha_n(A) = \alpha_+$.

Theorem 2 presents an upper bound for α_+ directly through the largest eigenvalues of symmetric components of the matrices $A^0(t)$ and $\Delta(t)$. In its proof we employ the known bound of the value $\mathrm{Re}\alpha_n$ of system (1.1) with a given matrix $A(t)$ (see [6]):

$$\mathrm{Re}\alpha_n \le \frac{1}{2T}\int_0^T q(t)\mathrm{d}t, \tag{1.5}$$

where $q(t)$ is the largest eigenvalue of the matrix $A(t)+A^\tau(t)$ ($^\tau$ means transposition).

Theorem 3 determines asymptotic stability of systems (1.1) and (1.3) via the norm of the transition matrix of the unperturbed system ($A(t) \equiv A^0(t)$) and an upper bound for the norm of the uncertain matrix $\Delta(t)$. These values also enable us to establish an upper bound for the value α_+ (Theorem 4).

The results of Theorem 1 are extended to the case in which the coefficients $a_{ik}(t)$ are bounded but not necessarily periodic (Theorem 5). In this case one may choose a constant matrix B satisfying inequalities (2.2) for all t; its largest real eigenvalue gives an upper bound for the supremal Lyapunov exponent of system (1.1). Theorem 6 extends Theorem 2 on the nonperiodic case.

This paper is organized as follows. The main results are formulated in section 2. In section 3 we include examples and a discussion of our theorems. All proofs are presented in section 4.

2. Main results. Consider the equation

$$\begin{gathered}\dot{x} = B(t)x,\\ B(t) = [b_{ik}(t)]_{i,k=1}^n, \quad b_{ik}(t) = b_{ik}(t+T).\end{gathered} \tag{2.1}$$

Suppose that for $t \in [0,T]$,

$$|a_{ik}(t)| \le b_{ik}(t), \quad i \ne k, \quad a_{ii}(t) \le b_{ii}(t), \quad i,k = 1,\dots,n. \tag{2.2}$$

Moreover, we assume that the matrix $B(t)$ is nondecomposable at some $t = t_0$, which means that system (2.1) with $B(t) \equiv B(t_0)$ does not contain an independent subsystem of order $p < n$. (Note that this condition may be reached by an arbitrary small perturbation of the matrix $B(t)$.) Since (2.2), the off-diagonal elements of the matrix $B(t)$ are nonnegative, so the largest in modulus multiplier $\rho_n(B)$ of system (2.1) is real, positive, and simple. (See, for example, Theorem 4.7 in [7] for more details.)

The following theorem compares the largest in modulus multipliers of (1.1) and (2.1).

THEOREM 1. *Let* (2.2) *be true; then*

$$|\rho_n(A)| \le \rho_n(B). \tag{2.3}$$

Thus, the corresponding characteristic exponent assumes the inequality

$$\mathrm{Re}\alpha_n(A) \le \alpha_n(B) = \frac{1}{T}\ln\rho_n(B). \tag{2.4}$$

Moreover, if (2.2) is true for any $A(t) \in U(t)$, then the corresponding value is

$$\alpha_+ \le \frac{1}{T}\ln\rho_n(B). \tag{2.5}$$

Suppose now that the matrix $A(t)$ is represented in the form (1.3). Let us define

$$\begin{aligned} &A_s^0(t) = \left[A^0(t) + A^{0\tau}(t)\right], \qquad \Delta_s(t) = \left[\Delta_+(t) + \Delta_+^\tau(t)\right], \\ &\Delta_+(t) = \left[\delta_{ik}^+(t)\right]_{i,k=1}^n, \qquad \delta_{ik}^+(t) = \tfrac{1}{2}\left[a_{ik}^+(t) - a_{ik}^-(t)\right]. \end{aligned} \tag{2.6}$$

Let $\beta_A(t)$ and $\beta_\Delta(t)$ be the largest eigenvalues of the matrices $A_s^0(t)$ and $\Delta_s(t)$. Note that all the eigenvalues of these matrices are real due to their symmetry. The next theorem gives an upper bound for α_+ directly through $\beta_A(t)$ and $\beta_\Delta(t)$.

THEOREM 2. *In problem* (1.1)–(1.3)

$$\alpha_+ \le \frac{1}{2T}\int_0^T [\beta_A(t) + \beta_\Delta(t)]\mathrm{d}t. \tag{2.7}$$

Consider now stability of system (1.1), (1.3) assuming that only an upper bound δ_0 for the norm $\|\Delta(t)\|$ of the uncertain matrix is known. Symbol $\|\Delta(t)\|$ assigns any matrix norm which satisfies, in addition to the standard norm conditions, the inequality

$$\|\Delta_1\Delta_2\| \le \|\Delta_1\|\,\|\Delta_2\|. \tag{2.8}$$

In particular, the norm could assume one of the following forms:

$$\|\Delta(t)\| = \sqrt{\sum_{i,k=1}^n \delta_{ik}^2(t)}, \tag{2.9}$$

$$\|\Delta(t)\| = \max_{1\le i\le n}\sum_{k=1}^n |\delta_{ik}|, \tag{2.10}$$

$$\|\Delta(t)\| = \max_{1\le k\le n}\sum_{i=1}^n |\delta_{ik}|. \tag{2.11}$$

We note that if the norm is determined by (2.9), Theorem 2 is extended to this case by exchanging in (2.7) $\beta_\Delta(t)$ by δ_0. In fact, the eigenvalues $\beta_i(t), i = 1,\dots n$, of the matrix $\Delta(t)$ satisfy the Schur inequality [9]

$$\sum_{i,k=1}^n \delta_{ik}^2(t) \ge \sum_{i=1}^n \beta_i^2(t). \tag{2.12}$$

Clearly, $\|\Delta_s(t)\| \le 2\|\Delta(t)\|$, so $2\delta_0 \ge \beta_\Delta(t)$.

Consider (1.1) in the absence of uncertain perturbations:

$$\dot{x} = A^0(t)x. \tag{2.13}$$

We assume that it is asymptotically stable, so the corresponding Floquet multipliers ρ_i^0, $i = 1, \dots, n$, lie within the unit circle (see [6]). Let $W^0(t,s)$ be the transition matrix of (2.1) ($W^0(s,s) = I$) and $w(t,s) = \|W^0(t,s)\|$. By supposition, $|\rho_i^0| < 1$, so $\mathrm{Re}\alpha_i^0 < 0$, where $\alpha_i^0 = (\ln \rho_i^0)/T$ are the characteristic exponents of system (2.13). In general, the elements of $W^0(t,s)$ tend to zero as $\exp(\mu t)P(t)$, where $\mu = \max_i \mathrm{Re}\alpha_i^0$ and $P(t)$ is a polynomial of order $k < p$. (p is the multiplicity of the corresponding multiplier.) Therefore, there exists

$$w_0 = \lim_{t\to\infty} \int_0^t w(t,s)\mathrm{d}s. \tag{2.14}$$

The following theorem gives sufficient stability condition for (1.3) expressed in the upper bound δ_0 for the norm of the uncertain matrix $\Delta(t)$.

THEOREM 3. *Equation* (1.3) *is asymptotically stable if*

$$\delta_0 < 1/w_0. \tag{2.15}$$

The next theorem gives an upper bound for the value $\alpha_+ = \sup \mathrm{Re}\alpha_{\mathrm{n}}(A)$ for $A(t) \in U(t)$.

THEOREM 4. *The value* α_+ *satisfies the inequality*

$$\alpha_+ \le \frac{\delta_0^2 w_*^2}{2}, \tag{2.16}$$

where

$$w_* = \lim_{t\to\infty} \left[\int_0^t w^2(t,s)\mathrm{d}s\right]^{1/2}.$$

Consider now the nonperiodic case, i.e., assume that the entries $a_{ik}(t)$ of the matrix $A(t)$ vary arbitrarily within bounds (1.2). Let $x(t, x_0, A)$ be the solution of (1.1) for some $A(t)$ with initial condition $x(0, x_0, A) = x_0$. The corresponding Lyapunov exponent is defined by

$$\lambda(x_0, A) = \overline{\lim_{t\to\infty}} \frac{1}{t} \ln \|x(t, x_0, A)\|. \tag{2.17}$$

Stability of such a system is determined by the supremal Lyapunov exponent

$$\lambda_+ = \sup_{x_0\neq 0,\ A(t)\in U(t)} \lambda(x_0, A). \tag{2.18}$$

If $\lambda_+ < 0$, then system (1.1) is exponentially stable for any admissible $A(t)$.

Let us consider the extensions of Theorems 1 and 2.

Suppose that a constant matrix B satisfies inequality (2.2) for any $t > 0$. Let $\lambda(B)$ be the largest real eigenvalue of the matrix B. The next theorem extends Theorem 1 to the nonperiodic case.

THEOREM 5. *Assume that* (2.2) *is true; then*

$$\lambda_+ \le \lambda(B). \tag{2.19}$$

Denote by Δ_m the matrix with the elements $\delta_{ik}^m = 0.5 \sup_t [a_{ik}^+(t) - a_{ik}^-(t)]$, $i, k = 1, \dots, n$; let β_m be the largest eigenvalue of the matrix $\Delta_m + \Delta_m^\tau$.

THEOREM 6. *The value λ_+ satisfies the inequality*

$$\lambda_+ \le \lim_{T\to\infty} \left[\frac{1}{2T} \int_0^T [\beta_A(t) + \beta_m] \mathrm{d}t \right]. \tag{2.20}$$

3. Discussion and example. The determination of an upper bound (2.3) for multipliers of system (1.1), (1.2) using Theorem 1 requires choosing a matrix $B(t)$ which assumes inequality (2.2) for any $A(t) \in U$ and the calculation of the largest real eigenvalue $\rho_n(B)$ of the corresponding monodromy matrix. In particular, one may choose

$$b_{ik}(t) \equiv \max |a_{ik}(t)| \quad \text{for } i \ne k, \quad b_{ii}(t) \equiv \max a_{ii}(t) = a_{ii}^+(t), \ i, k = 1, \dots, n. \tag{3.1}$$

If in the form (1.3) the off-diagonal elements of the matrix $A^0(t)$ are nonnegative, then $\max |a_{ik}(t)| = a_{ik}^+$ for $i \ne k$. Thus one may take

$$b_{ik}(t) \equiv a_{ik}^+(t), \quad i, k = 1, \dots, n, \tag{3.2}$$

which implies that the corresponding matrix $B(t) \in U(t)$. It follows that in this case the required bound α_+ is reached if all coefficients of the matrix $A(t)$ take their maximal admissible values.

The upper bound is also reached if for every $i = 1, \dots, n$, all off-diagonal elements of the ith row and column of the matrix $A^0(t)$ are both positive or negative for all t. Really, the change of the variable $x_i \to -x_i$ implies the change of the sign in the elements of these rows and columns. Hence, the system could be reduced to a system with nonnegative off-diagonal elements which does not affect the multipliers of the original system. So, the "worst" admissible matrix is such that the coefficients $a_{ik}(t)$ and $a_{ki}(t)$, $i, k = 1, \dots, n$; $k \ne i$, take their maximal values $a_{ik}^+(t)$ and $a_{ki}^+(t)$ when the elements $a_{ik}^0(t)$ and $a_{ki}^0(t)$ are positive; if they are negative, then $a_{ik}(t) \equiv a_{ik}^-$ and $a_{ki}(t) \equiv a_{ki}^-$; in both cases, $a_{ii}(t) \equiv a_{ii}^+(t)$.

Conditions of Theorem 2 avoid calculation of a monodromy matrix but require calculations of the eigenvalues $\beta_A(t)$ and $\beta_\Delta(t)$ of the matrices $A_s(t)$ and $\Delta_s(t)$ for every t. Though upper bound (2.7) is not generically reached for any admissible $A(t)$, it may provide a reasonable estimate, especially if the matrices $A^0(t)$ and $\Delta(t)$ are near-symmetric. (Note that for a constant symmetric matrix, the right-hand side of (2.7) equals its largest eigenvalue.)

Consider now Theorem 3. First of all we note that for scalar equation (1.3), stability condition (2.15) is precise. Really, let $A(t) = -h < 0$; then $w(t) = \exp(-ht)$, $w_0 = 1/h$ and inequality (2.15) becomes $|\Delta(t)| < h$. For $\Delta(t) \equiv h$, (1.3) is not asymptotically stable and condition (2.15) cannot be improved.

Systems in the following form are often found in various applications:

$$\ddot{y} + H(t)\dot{y} + A(t)y + \Delta(t)y = 0, \tag{3.3}$$

where $y \in R^k$, the matrices $H(t)$, $A(t)$, and $\Delta(t)$ are T-periodic, and $\Delta(t)$ is uncertain with $\|\Delta(t)\| \le \delta_0$. Suppose that for $\Delta(t) = 0$, the system is asymptotically stable. Let us derive a sufficient stability condition for (3.3) using reasoning similar to the proof of Theorem 3.

Setting $y = y_1$, $\dot{y} = y_2$, and $x = (y_1, y_2)$, one may reduce (3.3) to (1.1). Representing the corresponding transition matrix in the form

$$W(t,s) = \begin{bmatrix} W_{11}(t,s) & W_{12}(t,s) \\ W_{21}(t,s) & W_{22}(t,s) \end{bmatrix}, \tag{3.4}$$

we obtain

$$y_k(t) = W_{11}(t,0)y_k(0) + W_{12}(t,0)\dot{y}_k(0) - \int_0^t W_{12}(t,s)\Delta(s)y_k(s)\mathrm{d}s.$$

Assuming

$$y_k(t) = \exp(\alpha_k t)u_k(t), \quad u_k(t) = u_k(t+T), \quad \|u_k(0)\| = 1 \geq \|u_k(t)\|,$$

one can, analogously to the proof of Theorem 3, find that (3.3) is asymptotically stable if

$$\delta_0 < \frac{1}{w_{12}}, \quad w_{12} = \lim_{t\to\infty} \int_0^t \|W_{12}(t,s)\|\,\mathrm{d}s. \tag{3.5}$$

Theorem 5 provides a sufficient stability condition for system (1.1) with an uncertain nonperiodic matrix $A(t)$. Its utility requires finding a constant matrix B satisfying inequalities (2.2) for all t, which could be made if one takes

$$b_{ik} = \sup_t |a_{ik}(t)| \quad \text{for } i \neq k, \quad b_{ii} \equiv \sup_t a_{ii}(t), \quad i,k = 1,\ldots,n. \tag{3.6}$$

Note that the accuracy of the above bounds depends upon matrix $A(t)$. Hence it is worthwhile to find all of the bounds and to choose the best one in analysis of a specific system.

To illustrate the obtained results, consider a plane system, (1.1) ($n = 2$), with coefficients $a_{ik}(t) = a_{ik}(t+T)$ contained within constant intervals $[a_{ik}^-, a_{ik}^+]$, $i,k = 1,2$. To apply Theorem 1, we form the matrix B with the elements

$$b_{ik} = \max\left(\left|a_{ik}^-\right|, \left|a_{ik}^+\right|\right), \ i \neq k, \ b_{ii} = a_{ii}^+, \ i = 1,2. \tag{3.7}$$

Since B is a constant matrix, its largest real eigenvalue $\lambda(B)$ gives an upper bound for the largest characteristic exponent $\mathrm{Re}\alpha_n$ of system (1.1) with any admissible matrix $A(t)$. By Theorem 5, in the nonperiodic case, $\lambda(B)$ may serve as an upper bound for the supremal Lyapunov exponent λ_+. At the beginning of this section it is shown that this bound is reached if the signs of the elements $a_{12}^0 = 0.5(a_{12}^- + a_{12}^+)$ and $a_{21}^0 = 0.5(a_{21}^- + a_{21}^+)$ are identical. If they both are positive (negative), the coefficients $a_{12}(t)$ and $a_{21}(t)$ of the "worst" matrix $A(t)$ reach their maximal (minimal) admissible values; in the both cases, the diagonal elements $a_{ii}(t) \equiv a_{ii}^+$.

To apply Theorem 2, we form the matrices A^0 and Δ_+ with the elements

$$a_{\mathrm{ik}}^0 = \frac{1}{2}(a_{ik}^- + a_{ik}^+), \qquad \delta_{ik}^+ = \frac{1}{2}(a_{ik}^+ - a_{ik}^-), \qquad i,k = 1,2. \tag{3.8}$$

For any admissible matrix $A(t)$, $\mathrm{Re}\alpha_n \leq 0.5(\beta_A + \beta_\Delta)$, where β_A and β_Δ are the largest eigenvalues of the matrices $A^0 + A^{0\tau}$ and $\Delta_+ + \Delta_+^\tau$.

For example, assume $a_{11} = 0$, $a_{12} = 1$, $a_{2i} \in [a_{2i}^-, a_{2i}^+]$, $i = 1,2$. Setting $b_{21} = \max(|a_{21}^-|, |a_{21}^+|)$, $b_{22} = a_{22}^+$ and using Theorem 5, we find that $\lambda_+ \leq \lambda(B) = 0.5(a_{22}^+ +$

$\sqrt{a_{22}^{+2} + 4b_{21}}$). For $a_{21}^+ \geq |a_{21}^-|$, this bound is precise and is reached if $a_{21}(t) \equiv a_{21}^+$, $a_{22}(t) \equiv a_{22}^+$.

If $a_{21}^+ < |a_{21}^-|$, this upper bound could become overly conservative. In particular, since $\lambda(B) > 0$, the system could be stable if $a_{21}^+ < 0$ and $a_{22}^+ < 0$, but the upper bound derived using Theorem 5 is positive. In the periodic case ($a_{2i}(t+T) = a_{2i}(t)$), a better estimate could be found using Theorem 3. Assume $a_{22}(t) \equiv -2h$; setting $x_1 = x$, $x_2 = \dot{x}$, we write the considered system in the form

$$\begin{gathered} \ddot{x} + 2h\dot{x} + \omega_0^2 x + \delta(t)x = 0, \\ \omega_0^2 = 0.5(a_{21}^- + a_{21}^+),\ |\delta(t)| \leq \delta_0 = 0.5(a_{21}^+ - a_{21}^-). \end{gathered} \tag{3.9}$$

Next, the elements of matrix (3.4) are

$$W_{12}(t,s) = W_{12}(t-s) = \frac{1}{\omega^*} \exp[-h(t-s)] \sin \omega^*(t-s), \qquad \omega^* = \sqrt{\omega_0^2 - h^2},$$

and from (3.5) we get

$$\begin{aligned} w_{12} &= \lim_{t\to\infty} \int_0^t |W_{12}(t-s)|\, \mathrm{d}s = \int_0^\infty |W_{12}(v)|\, \mathrm{d}v \\ &= \frac{1}{\omega_0^2} \left[1 + 2 \sum_{k=1}^{\infty} \exp(k\beta) \right] = \frac{1 + \exp\beta}{\omega_0^2(1 - \exp\beta)}, \end{aligned} \tag{3.10}$$

where $\beta = -\pi h/\omega^*$.

Hence, according to (3.5), equation (3.9) is asymptotically stable if

$$\delta_0 < \delta_0^* = \frac{\omega_0^2(1 - \exp\beta)}{1 + \exp\beta}. \tag{3.11}$$

Note that obtained stability criteria could be used to check absolute stability of linear systems controlled by a nonlinear feedback function $f(\sigma, t)(\sigma = (c, x))$ (see, e.g., [11], [12], [13]), which might take any values within a prescribed sector, i.e.,

$$0 \leq f(\sigma, t)\sigma \leq K\sigma^2. \tag{3.12}$$

Really, this problem is equivalent to stability of the corresponding linear uncertain system with $f(\sigma, t) = u(t)\sigma$, where $0 \leq u(t) \leq K$.

In particular, inequality (3.11) provides sufficient criterion for absolute stability of the system

$$\begin{gathered} \ddot{x} + 2h\dot{x} + \omega_0^2 x + \varphi(x,t) = 0, \\ -\delta_0 x^2 \leq \varphi(x,t)x \leq \delta_0 x^2. \end{gathered} \tag{3.13}$$

Putting $f(x,t) = \varphi(x,t) + \delta_0 x$ and observing that $\omega_0^2 - \delta_0 = -a_{21}^+$, we reduce (3.13) to the form

$$\begin{gathered} \ddot{x} + 2h\dot{x} - a_{21}^+ x + f(x,t) = 0, \\ 0 \leq f(x,t)x \leq Kx^2,\ K = 2\delta_0. \end{gathered} \tag{3.14}$$

Suppose that $a_{21}^+ < 0$. (For $a_{21}^+ \geq 0$, system (3.14) with the admissible function $f(x,t) \equiv 0$ is not asymptotically stable.) Let us compare the accuracy of (3.11) with

TABLE

h	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
δ_0^P	0.20	0.39	0.57	0.73	0.87	0.96	1.00	0.96	0.08	0.0
δ_0^*	0.16	0.31	0.46	0.60	0.72	0.83	0.92	0.97	0.99	1.00

the classical Popov criteria [13]. The latter guarantees absolute stability of system (3.14) if for all $\omega \in [0, \infty)$,

$$\frac{1}{K} > \frac{\omega^2 + a_{21}^+}{(\omega^2 + a_{21}^+)^2 + 4\omega^2 h^2}. \tag{3.15}$$

The right-hand side of (3.15) reaches its maximal in ω value for $\omega^2 = -a_{12}^+ + 2h\sqrt{-a_{12}^+}$; thus (3.15) holds if

$$K < 4h\sqrt{-a_{21}^+} + 4h^2. \tag{3.16}$$

Putting here $-a_{21}^+ = \omega_0^2 - K/2$, we get

$$K < K_P = 2\delta_0^P = 4h\sqrt{\omega_0^2 - h^2}. \tag{3.17}$$

For $\omega_0^2 = 1$, the limit values $\delta_0^P(h)$ and $\delta_0^*(h)$ obtained by the Popov and developed criteria (formulas (3.17) and (3.11)) are compared in the table.

As is seen, the Popov condition is slightly less conservative for relatively small h, while for larger h, our approach provides better results. For $h \to 1$, $\delta_0^P(h) \to 0$, while $\delta_0^*(h)$ approaches the precise bound $\delta_0^*(1) = 1$. (Really, for $h = \omega_0 = 1$ and $\varphi = -x$, equation (3.13) admits the solution $x(t) = const$ and, thus, is not asymptotically stable.) Note that Theorem 3, unlike the Popov condition, could be equally applied to controlled systems with a periodic linear part.

4. Proof of Theorem 1. Suppose first that the largest in the modulus multiplier ρ_n of (1.1) is real; let $x_n(t)$ be the corresponding solution ($x_n(t+T) = \rho_n x(t)$). As is known, it may be represented in the form

$$\begin{gathered} x_n(t) = \exp(\alpha_n t) u_n(t), \qquad u_n(t) = u_n(t+T), \\ \alpha_n = (\ln \rho_n)/T. \end{gathered} \tag{4.1}$$

Substituting (4.1) in (1.1), we find that $u_n(t)$ satisfies the equation

$$\dot{u} + \alpha_n u = A(t)u, \tag{4.2}$$

which we rewrite it in the form

$$\dot{u} + (\alpha_n + r)u = (A(t) + rI)u, \tag{4.3}$$

where the constant r is chosen so that

$$\alpha_n + r > 0 \text{ and } r \geq b_{ii}(t),\ i = 1, \dots, n,\ t \in [0, T]. \tag{4.4}$$

Using Green's function $g(t, s, \alpha)$ for the problem

$$\dot{y} + (\alpha + r)y = p(t), \qquad y(0) = y(T), \tag{4.5}$$

we reduce (4.3) to the integral equation

$$u = R(\alpha_n, A)u,$$

$$R(\alpha_n, A)u = \int_0^T G(t, s, \alpha_n)[A(s) + rI]u(s)\mathrm{d}s,\ G(t, s, \alpha_n) = \mathrm{diag}(g(t, s, \alpha_n)),$$

$$g(t, s, \alpha) = \frac{\exp[\beta(\mathrm{s} - T)]}{\exp(\beta T) - 1} \text{ for } t < s; \tag{4.6}$$

$$g(t, s, \alpha) = \frac{\exp[\beta(\mathrm{s} + T - t)]}{\exp(\beta T) - 1} \text{ for } t > s, \qquad \beta = \alpha + r.$$

Observing that $g(t, s, \alpha_n) > 0$ for $t; s \in [0, T]$, from (2.2), (4.4), and (4.6) we find

$$|u_n(t)| \le \int_0^T G(t, s, \alpha_n)\,|A(s) + rI|\,|u_n(s)|\,\mathrm{d}s \le R(\alpha_n, B)\,|u_n(t)|\,,$$

$$R(\alpha_n, B)u_n = \int_0^T G(t, s, \alpha_n)[B(s) + rI]u_n(s)\mathrm{d}s. \tag{4.7}$$

In view of (2.1) and (4.4), the matrix $B(s) + rI$ is nonnegative; since, moreover, $B(t_0)$ is nondecomposable, the operator $R(\alpha_n, B)$ is strongly positive [7] (i.e., it transforms any nonnegative function $u(t)$ into a positive one). Hence, the eigenvalue problem

$$\lambda u = R(\alpha_n, B)u \tag{4.8}$$

possesses a unique positive eigenvalue $\lambda_0(\alpha_n)$ such that the corresponding solution $u_0(t) > 0$ for $t \in [0, T]$; any other eigenvalue of this problem obeys the inequality $|\lambda_i| < \lambda_0$ [7, Theorem 4.3]. For $\lambda > \lambda_0$ and any $v(t) \ge 0$, inequality $R(\alpha_n, B)v \ge \lambda v$ fails [8, Theorem 2.18]. Thus, inequality (4.7) implies that

$$1 \le \lambda_0(\alpha_n). \tag{4.9}$$

Let $x^*(t) = \exp(\alpha^* t)u^*(t)(\alpha^* = (\ln \rho^*)/T, u^*(t) = u^*(t + T))$ be the solution of (2.2) corresponding to the largest in modulus multiplier $\rho^* = \rho_n(B)$. Since the off-diagonal elements of the matrix $B(t)$ are nonnegative, $u^*(t) > 0$ [7, Theorem 4.7]. Analogously, we find that $u^*(t)$ satisfies the equation

$$\lambda u = R(\alpha^*, B)u \tag{4.10}$$

with $\lambda = 1$. Because $u^*(t) > 0$, $\lambda = 1$ is the largest in modulus eigenvalue of the operator $R(\alpha^*, B)$ [8], i.e.,

$$\lambda_0(\alpha^*) = 1. \tag{4.11}$$

It is follows from (4.6) that $g(t, s, \alpha)$ and thus—see [8]—the operator $R(\alpha, B)$ and the eigenvalue $\lambda_0(\alpha)$ decrease as α increases. Hence the required inequality $\alpha^* \ge \alpha_n$ follows immediately from (4.9) and (4.11).

Let us assume now that a multiplier $\rho_n = \exp[(\alpha_n + i\omega_n)T]$ is complex. Suppose first that $\omega_n T/(2\pi) = p/q$, where p and q are integers. Since (1.1) is T-periodic, the matrix $W(qT) = [W(T)]^q$ and its eigenvalues equal $\rho_i^q, i = 1, \ldots, n$, and, thus, ρ_n^q is real. Correspondingly it follows that $|\alpha_n^q| \le \alpha^{*q}$, and the required inequality holds true.

If the number $\omega_n T/(2\pi)$ is irrational, we put $\omega_n(\varepsilon) = \omega_n + \varepsilon$. The value ω_n appears in the matrix of fundamental solutions $W(t)$ in the form $\exp(\omega_n t)$; thus $W(t,\varepsilon)$ and the corresponding matrix $A(t,\varepsilon) = \dot{W}(t,\varepsilon)W^{-1}(t,\varepsilon)$ are continuous in ε. Setting $\varepsilon \to 0$ and observing that for rational values $(\omega_n + \varepsilon)T/(2\pi)$, the multiplier $\rho_n(\varepsilon)$ satisfies inequality (2.3), we find that the last is also true for $\varepsilon = 0$, which finally proves the theorem.

Proof of Theorem 2. The real part of a characteristic exponent α_n of (1.1) satisfies the inequality (1.5) (see [6]), where $q(t)$ is the largest eigenvalue of the matrix $A(t) + A^{\tau}(t)$. Using representation (1.3) we find that $q(t) \le \beta_A(t) + \beta_s(t)$, where $\beta_s(t)$ is the largest eigenvalue of the matrix $\Delta(t) + \Delta^{\tau}(t)$. Clearly, its elements satisfy the inequality $|\delta_{ik}(t) + \delta_{ki}(t)| \le 2\delta_{ik}^{+}(t)$; hence, according to a Wielandt lemma [10], $\beta_s(t) \le \beta_{\Delta}(t)$. The theorem is proved.

Proof of Theorem 3. If δ_0 is small, (1.1) is asymptotically stable. For the critical value, δ_0^*, there exists a matrix $\Delta(t)$ with $\|\Delta(t)\| \le \delta_0^*$ such that $|\rho_i| \le 1, i = 1, \dots, n$; $|\rho_{\mathrm{k}}| = 1$ for a certain k, where ρ_i are the multipliers of (1.1). The corresponding solution of (1.1) is

$$x_k(t) = \exp(\alpha_k t)u_k(t), \quad u_k(t) = u_k(t+T), \tag{4.12}$$

where the value α_k is purely imaginary. Without loss of generality, we assume that

$$\|x_k(0)\| = \|u_k(0)\| = 1 \ge \|x_k(t)\|. \tag{4.13}$$

Clearly, $x_k(t)$ may be represented in the form

$$x_k(t) = W^0(t,0)x_k(0) + \int_0^t W^0(t,s)\Delta(s)x_k(s)\mathrm{d}s. \tag{4.14}$$

Taking into account (4.13), we find that for any integer p,

$$\begin{aligned} \|x_k(pT)\| = 1 &\le \|W^0(pT,0)\| + \int_0^{pT} \|W^0(pT,s)\|\|\Delta(s)\|\|x_k(s)\|\mathrm{d}s \\ &< \|W^0(pT,0)\| + w_0\delta_0^*. \end{aligned} \tag{4.15}$$

Noticing that $\|W(pT,0)\| \to 0$ as $p \to \infty$, we derive from (4.15) that $w_0\delta_0^* \ge 1$. Thus, for $\delta_0 < 1/w_0$, a value α_k cannot be purely imaginary, and all multipliers of (1.1) lie within the unit circle.

Proof of Theorem 4. Representing a solution $x_k(t) = \exp(\alpha_k t + ivt)u_k(t)$ in the form (2.6) and assuming $\|u_k(0)\| = 1 \ge \|u_k(t)\|$, we get

$$\begin{aligned} \|x_k(pT)\| &= \exp(\alpha_{\mathrm{k}} pT) \\ &\le \left\|W^0(pT,0)\right\| + \int_0^{pT} \left\|W^0(pT,s)\right\| \|\Delta(s)\| \, \|x_k(s)\| \, \mathrm{d}s \\ &\le \left\|W^0(pT,0)\right\| + \delta_0 \left[\int_0^{pT} w^2(pT,s)\mathrm{d}s \int_0^{pT} \left[\exp(\alpha_k s)\right]^2 \mathrm{d}s\right]^{1/2} \\ &< \left\|W^0(pT,0)\right\| + \frac{\delta_0 w_* \exp(\alpha_{\mathrm{k}} pT)}{(2\alpha_k)^{1/2}}. \end{aligned} \tag{4.16}$$

Since $\|W(pT,0)\| \to 0$ as $p \to \infty$, we derive from (4.16) the required inequality (2.16).

Proof of Theorem 5. We rewrite (1.1) in the form

$$\dot{x} + rx = [A(t) + rI]x, \tag{4.17}$$

where the constant r is chosen to satisfy the following inequality:

$$r + a_{ii}(t) > 0 \text{ for } i = 1, \dots, n,\ t > 0. \tag{4.18}$$

Let $x(t, x_0, A)(x(0, x_0, A) = x_0)$ be a solution of (4.17); clearly, it satisfies the equation

$$x(t) = x_0 + \frac{1}{r}\int_0^t \exp[-r(t-s)][A(s) + rI]x(s)\mathrm{d}s. \tag{4.19}$$

Let $x^+(t, x_0^+)(x^+(0, x_0^+) = x_0^+)$ be a solution of the equation $\dot{x} + rx = (B + rI)x$ and, analogously,

$$x^+(t) = x_0^+ + \frac{1}{r}\int_0^t \exp[-r(t-s)](B + rI)x^+(s)\mathrm{d}s. \tag{4.20}$$

We take $x_0^+ > |x_0|$; then $x_0^+(t) > |x_0(t)|$ for small t. We now show that this inequality holds for any $t > 0$. Really, suppose, by contradiction, that it fails for some $t = t^*$; then due to (4.19) and (4.20),

$$\begin{aligned} x^+(t^*) - |x(t^*)| > x_0^+ - |x_0| \\ + \frac{1}{r}\int_0^{t^*} \exp[-r(t^* - s)][(B + rI)x^+(s) - |A(s) + rI||x(s)|]\mathrm{d}s. \end{aligned} \tag{4.21}$$

The integrand in (4.21) is positive since

$$x^+(s) > |x(s)| \text{ for } s \in [0, t^*),\ b_{ik} \geq |a_{ik}(s)|,\ i \neq k,\ b_{ii} + r \geq |a_{ii}(s) + r|\,,$$

and, therefore, $x_0^+(t^*) > |x_0(t^*)|$. This contradiction implies that $x_0^+(t) > |x_0(t)|$ for all $t > 0$, and the corresponding Lyapunov exponents satisfy the inequality

$$\lambda(x_0^+, B) \geq \lambda(x_0, A). \tag{4.22}$$

Hence, for any solution of (1.1), there exists a solution of (2.1) such that the corresponding Lyapunov exponents obey inequality (4.22), which implies that this inequality holds true for the supremal Lyapunov exponents.

For a constant matrix B, the supremal Lyapunov exponent $\lambda_+(B)$ equals the largest real part of the eigenvalues. Due to the Perron theorem [9], for the positive matrix $B + rI$, the largest in modulus eigenvalue is positive and real. Since the real parts of the eigenvalues of the matrices $B + rI$ and B differ by the same value r, the value $\lambda_+(B)$ is the largest real eigenvalue $\lambda(B)$.

Proof of Theorem 6. For any solution $x(t)$ of (1.1), one has

$$\frac{\mathrm{d}}{\mathrm{d}t}(x^\tau x) = [x^\tau(A^\tau(t) + A(t))x] \leq \lambda_m(t)(x^\tau x), \tag{4.23}$$

where λ_m is the largest eigenvalue of the matrix $A^\tau(t) + A(t)$. Hence,

$$x^\tau(t)x(t) = \|x(t)\|^2 \leq \|x(0)\|^2 \exp\int_0^t \lambda_m(s)\mathrm{d}s. \tag{4.24}$$

Taking into account that the elements of the matrices $\Delta(t)$ and Δ_m satisfy the inequality $|\delta_{ik}(t)| \le \delta_{ik}^m$, $i,k = 1,\dots,n$, and using a Wielandt lemma [10], we find that the largest eigenvalue of the matrix $\Delta_m + \Delta_m^\tau$, β_m exceeds that of the matrix $\Delta(t)+\Delta^\tau(t)$, and, hence, $\lambda_m(t) \le \beta_A(t)+\beta_m$. Now required inequality (2.20) follows from (4.24).

REFERENCES

[1] Z. Bubnicki, *General approach to stability and stabilization for a class of uncertain discrete non-linear systems*, Internat. J. Control, 73 (2000), pp. 1298–1306.

[2] Z. Qu, *Robust Control of Nonlinear Uncertain Systems*, John Wiley, New York, 1998.

[3] D. D. Šiljak, *Large Scale Dynamics Systems, Stability and Structure*, North–Holland, Amsterdam, 1978.

[4] H. L. S. Almeida, A. Bhaya, D. M. Falcão, and E. A. Kaszkurewicz, *Team algorithm for robust stability analysis and control design of uncertain time-varying linear systems using piecewise quadratic Liapunov functions*, in Proceedings of the 37th IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 1998, pp. 4410–4415.

[5] A. Bhaya and E. Kaszkurewicz, *A note on a robust stability problem*, Internat. J. Control, 73 (2000), pp. 1346–1348.

[6] V. A. Yakubovich and V. M. Starshinsky, *Linear Differential Equations with Periodic Coefficients*, John Wiley, New York, 1980.

[7] M. A. Krasnosel'skii, *The Operator of Translation Along the Trajectories of Differential Equations*, Transl. Math. Monogr. 19, AMS, Providence, RI, 1968.

[8] M. A. Krasnosel'skii, *Positive Solutions of Operator Equations*, Noordhoff, Groningen, The Netherlands, 1968.

[9] R. Bellman, *Introduction to Matrix Analysis*, McGraw–Hill, New York, 1960.

[10] H. Wielandt, *Unzerlegbare, night negative Matrizen*, Math. Z., 52 (1950), pp. 642–648.

[11] A. I. Lur'e, *Some nonlinear problems of automatic control theory*, Gostehizdat, Moscow, 1951.

[12] K. S. Narendra and J. F. Taylor, *Frequency Domain Criteria for Absolute Stability*, Academic Press, New York, 1973.

[13] V. M. Popov, *Absolute stability of nonlinear systems of automatic control*, Automat. Remote Control, 22 (1962), pp. 857–875.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1198–1221

OPTIMALITY CONDITIONS FOR SIMULTANEOUS TOPOLOGY AND SHAPE OPTIMIZATION*

J. SOKOŁOWSKI† AND A. ŻOCHOWSKI‡

Abstract. New optimality conditions are derived for a class of shape optimization problems. The conditions are established on the boundary by an application of the boundary variations technique and in the interior of an optimal domain by exploiting the topological derivative method. An example is provided for which the classical second order sufficient optimality conditions are verified for an optimal simply connected domain. However, the value of the cost can be improved by the topology variations, and therefore, the optimal solution can be substantially changed by applying the topology optimization.

Key words. shape optimization, shape derivative, topological derivative, asymptotic expansion, domain differential

AMS subject classifications. Primary, 49Q10, 49Q12; Secondary, 35J05, 35J50, 35B37

DOI. 10.1137/S0363012901384430

1. Introduction.

In classical theory of shape optimization the first order necessary optimality conditions account for boundary variations of an optimal domain. On the other hand the relaxed formulation based on the homogenization technique is used [1], [4], [19] in the topology optimization of energy functionals, the so-called compliances in structural optimization. For such a formulation the coefficients of an elliptic operator are selected in an optimal way and the resulting optimal design takes the form of a composite microstructure rather than any geometrical domain. The methods of topology optimization based on asymptotic analysis and related to the topological derivatives or topological asymptotics include the so-called *bubble method*, used for the topology optimization in structural mechanics [7], [26]. Numerical results based on topological asymptotics can be found in [8]. We refer also to [11], [12] for an application to inverse problems and to [18], [17] for the related results in the case of the cavity of arbitrary shape. Further applications in mechanics can be found in [6], [13], [15], [16]. It seems that in the literature on the subject there is a lack of general method or technique that can be applied in the process of optimization of an arbitrary shape functional for simultaneous boundary and topology variations. Such an approach would be very useful for numerical solution of, e.g., optimum design problems in structural mechanics. In the paper [29] the so-called topological derivative (TD) of an arbitrary shape functional is introduced. Such a derivative is evaluated by an application of the asymptotic analysis in singularly perturbed geometrical domains [10], [20], [21] for a class of elliptic equations including the two-dimensional (2D) elasticity system [33] and three-dimensional (3D) elasticity system [32]. TD determines

*Received by the editors February 6, 2001; accepted for publication (in revised form) February 2, 2003; published electronically August 6, 2003. Research for this paper was partially supported by the grant 4 T11A 01524 of the State Committee for the Scientific Research of the Republic of Poland and by the French-Polish research programme POLONIUM between Institut Elie Cartan and INRIA-Lorraine and the Systems Research Institute of the Polish Academy of Sciences.

http://www.siam.org/journals/sicon/42-4/38443.html

†Institut Elie Cartan, Laboratoire de Mathématiques, Université Henri Poincaré Nancy I, B.P. 239, 54506 Vandoeuvre lès Nancy Cedex, France and Systems Research Institute of the Polish Academy of Sciences, ul. Newelska 6, 01-447 Warszawa, Poland (sokolows@iecn.u-nancy.fr).

‡Systems Research Institute of the Polish Academy of Sciences, ul. Newelska 6, 01-447 Warszawa, Poland (zochowsk@ibspan.waw.pl).

whether a change of topology by nucleation of a small hole, or in the similar setting of a small inclusion at a given point $\boldsymbol{x} \in \Omega$, would result in improving the value of a given shape functional or not. In the case of cavities in the form of 2D circles or 3D balls (in the case of Laplace equations in the form of balls in $\mathbb{R}^n$ for an arbitrary space dimension $n \geq 2$) the constructive results are obtained [29], [31], [32] by using the shape calculus combined with the asymptotic approximations of solutions to elliptic BVPs.

Asymptotic approximations of solutions for problems with small geometrical singularities are constructed by the method of matched and compound asymptotic expansions in [10], [14], [20], [21], [22], [23], [24]. We refer the reader to [25] for proofs of results on asymptotic analysis of elliptic BVPs and associated shape functionals in weighted Hölder spaces. The approach allows for the pointwise estimates of remainders during the process of constructions of asymptotic approximations. The results on asymptotic analysis of volume and surface shape functionals [25] are established in terms of the so-called *polarization tensor*—an integral attribute of cavities and inclusions [24], in particular for the anisotropic elasticity in three dimensions. Let us point out that in the framework of Lagrangian formalism in shape optimization [5], the fixed domain setting for elliptic BVPs is required in order to derive the first order optimality conditions. The topology variations of geometrical domains are defined as *functions* of small parameter ρ. Such a fixed domain setting with respect to the small parameter ρ can be introduced by an application of the theory of self-adjoint extensions of elliptic operators defined in domains with geometrical singularities [22]; we refer the reader to [25] for the related results in the case of shape functionals. The Lagrangian formalism in the fixed domain setting combined with the formal asymptotics of solutions is used, e.g., in [8] for the purposes of numerical methods in topology optimization.

In the present paper the approach of [29] is extended to the case of a finite number of circular holes treated by means of TD combined with simultaneous boundary variations by an application of the speed method [28] to Frechet differentiable shape functionals. We combine, for the first time, the topology variations with the boundary variations in order to derive the first order optimality conditions for shape optimization problems. Therefore, the first order necessary optimality conditions are established for a class of shape optimization problems in a more general setting compared to the classical theory [28], [5]. We restrict ourselves, for simplicity, to a scalar model problem and to a simple class of topology variations; however, the same procedure will result in the first order optimality conditions for more general shape and topology optimization problems.

To deal with various types of domain modification we introduce the following general notation for different types of variations of shape functionals and of solutions to partial differential equations:

Shape derivative is used in order to determine the variations of solutions to BVPs resulting from the boundary variations of geometrical domains. In this framework, first, the Frechet differentiability of shape functionals is established, and then the speed method is applied to determine the shape derivative. We refer to, e.g., [28] and the recent book [5] for a general description of the speed method and the related results on Frechet differentiability of shape functionals.

Topological derivative (also *topological differential*) accounts for variations of shape functionals resulting from the emerging of one or several small holes, cavities, or inclusions in the interior of the geometrical domain. We refer to

section 4 for a detailed description of TDs of shape functionals in the case of the Laplace equation.

Domain differential, introduced here, unifies the influence on shape functionals of boundary variations and, at the same time, of the nucleation of internal holes or cavities. We refer to section 6 for the definition and properties of the domain differential in the case of the Laplace equation and to section 7 for an application of our method to the derivation of the necessary optimality conditions for a shape optimization problem.

Let us recall briefly the definitions used in the shape calculus. In the formulae given below, $J(\Omega)$ is an integral functional depending on the solution of the BVP defined on Ω. We assume that $\Omega \subset \mathbb{R}^n$ is a domain with piecewise smooth boundary, $n \geq 2$, and we assume for simplicity that $n = 2$; however, most of the results of the paper are valid for $n = 3$. The elliptic equations with the so-called polynomial property [23] are well-suited BVPs for our analysis. We introduce the mapping $T_\tau : \Omega \to \Omega_\tau$ associated with the vector field $\boldsymbol{V}(\cdot,\cdot) \in C^1(0,\delta; C_0^2(\mathbb{R}^2,\mathbb{R}^2))$ supported in a small neighborhood of $\Gamma = \partial\Omega$. The domain Ω_τ is defined by [28]

$$\Omega_\tau = \{\, \boldsymbol{x} = \boldsymbol{x}(\tau, \boldsymbol{X}) \mid \boldsymbol{X} \in \Omega \,\},$$

where $\boldsymbol{x} = \boldsymbol{x}(\tau, \boldsymbol{X})$ denotes the solution to the system

$$\frac{d\boldsymbol{x}}{d\tau}(\tau) = \boldsymbol{V}(\tau, \boldsymbol{x}), \qquad \boldsymbol{x}(0) = \boldsymbol{X}.$$

Then the shape derivative of $J(\Omega)$ at Ω^* and in the direction $\boldsymbol{V}$ is given by

$$\mathcal{S}J(\Omega^*; \boldsymbol{V}) = \lim_{\tau \to 0} \frac{1}{\tau} \left[J(T_\tau(\Omega^*)) - J(\Omega^*) \right]. \tag{1.1}$$

Now, let us create a small hole $B(\boldsymbol{x}, \rho) = \{\boldsymbol{y} \in \mathbb{R}^2 |\ |\boldsymbol{x} - \boldsymbol{y}| < \rho\}$ at the point $\boldsymbol{x}$ in the interior of the domain Ω, and let us prescribe the Neumann homogeneous conditions on its boundary $\partial B(\boldsymbol{x}, \rho)$. The meaning of the Neumann condition can be made precise for a general elliptic BVP by using the Green formula associated with the problem. We denote by $\overline{B(\boldsymbol{x}, \rho)}$ the closure of the ball $B(\boldsymbol{x}, \rho)$ (circle for $n = 2$), and we define $\Omega_\rho = \Omega \setminus \overline{B(\boldsymbol{x}, \rho)}$. According to [29], the TD $\mathcal{T}J(\Omega^*; \boldsymbol{x})$ of $J(\Omega)$ at $\Omega^* \subset \mathbb{R}^2$ is a function depending on the center $\boldsymbol{x}$ of the small hole and in two dimensions is given by the following limit, if the limit exists:

$$\mathcal{T}J(\Omega^*; \boldsymbol{x}) = \lim_{\rho \to 0+} \frac{J(\Omega^* \setminus \overline{B(\boldsymbol{x}, \rho)}) - J(\Omega^*)}{\pi \rho^2}. \tag{1.2}$$

Remark 1.1. The following formula [32] is useful in applications, since it uses only the first order shape derivatives of the shape functional $J(\Omega)$ in order to evaluate the TDs of $J(\Omega)$ at $\Omega^* \subset \mathbb{R}^n$, $n \geq 2$:

$$\mathcal{T}J(\Omega^*; \boldsymbol{x}) = \lim_{\rho \downarrow 0} \frac{dJ(\Omega^* \setminus \overline{B(\boldsymbol{x}, \rho)})}{d(|B(\boldsymbol{x}, \rho)|)}, \tag{1.3}$$

where $|B(\boldsymbol{x}, \rho)|$ is the n-dimensional measure of the ball $B(\boldsymbol{x}, \rho)$.

Remark 1.2. It is an interesting and, it seems, difficult problem for further studies in asymptotic analysis to introduce, in addition to the *interior* topology variations, the *exterior* topology variations defined on the boundary of geometrical domain. Such exterior topology variations will, in particular, replace the boundary variations in the

procedure of derivation of necessary optimality conditions for the shape and topology optimization problems.

In section 6 we show that under the volume constraints $|\Omega|=const$ the necessary optimality conditions for the problem of minimization of the shape functional $J(\Omega)$ takes the form (6.6). Such first order necessary optimality conditions seem to be new in shape optimization.

In section 7 we present an example with the optimal solution known explicitly. For such a problem all the computations can be carried out analytically, in particular, for the TDs. The optimal shape in the form of a disk corresponds to the fixed topology, i.e., when only simply connected domains are admissible. When the requirement on the fixed topology is relaxed and still the symmetry of admissible domains is required, the disk remains optimal under the volume constraints. Finally, for the admissible domains which are no longer symmetric, the optimality conditions including the TD are not satisfied. Therefore, a topology variation in the precisely indicated region of the disk improves the value of the cost functional. It shows that the domain differential allows us to distinguish all the cases listed above. Thus, the optimality conditions we establish are more precise compared to the classical boundary variation technique. The similar results can be expected, as well, for the optimization problems with respect to the geometrical domain involving eigenvalues.

2. Preliminaries. We are going to analyze a class of shape optimization problems for the Laplace equation. Therefore, we need the precise results on asymptotic approximations of solutions to the equation with respect to small parameter $\rho > 0$ which defines the size of geometrical singularities. We consider in the paper a model problem; therefore the arguments are elementary, and we provide the proofs for the convenience of the reader.

Let us denote by $C(\boldsymbol{x}^0;\rho)$ a circle $C(\boldsymbol{x}^0;\rho) = \{\ \boldsymbol{x}\ |\ \ |\boldsymbol{x}-\boldsymbol{x}^0| = \rho\ \}$, by $P(\boldsymbol{x}^0;\rho,R)$ a ring $P(\boldsymbol{x}^0;\rho,R) = \{\ \boldsymbol{x}\ |\ \ \rho < |\boldsymbol{x}-\boldsymbol{x}^0| < R\ \}$, and finally $B(\boldsymbol{x}^0;R) = \{\ \boldsymbol{x}\ |\ \ |\boldsymbol{x}-\boldsymbol{x}^0| < R\ \}$, where the bar denotes the closure $\overline{B(\boldsymbol{x}^0;R)} = \{\ \boldsymbol{x}\ |\ \ |\boldsymbol{x}-\boldsymbol{x}^0| \le R\ \}$. If $\boldsymbol{x}^0 = 0$, we simplify the notation to $C(\rho)$, $P(\rho,R)$, and $B(R)$. We formulate the BVP

$$\Delta w_\rho = 0 \quad \text{in} \quad P(\rho,R), \tag{2.1a}$$

$$w_\rho = 0 \quad \text{on} \quad C(R), \tag{2.1b}$$

$$\frac{\partial}{\partial n} w_\rho = h_\rho \quad \text{on} \quad C(\rho). \tag{2.1c}$$

Then the following estimates can be established for solutions to (2.1). The estimates are classical, we provide the proof for convenience of the reader.

LEMMA 2.1. *Let $R_0 < R$ be fixed and h_ρ be the trace on $C(\rho)$ of the function $h \in C^1(\overline{B(R_0)})$. Then the solution w_ρ of (2.1) satisfies on $C(\rho)$, for $\rho < R_0$, the estimates*

$$\begin{aligned} |w_\rho| &\le \Lambda\rho|\log\rho|, \\ |\nabla w_\rho| &\le \Lambda \end{aligned}$$

and on $P(R_0,R)$

$$|w_\rho| \le \Lambda\rho.$$

Here Λ is a generic constant depending only on the norm of h in $C^1(\overline{B(R_0)})$ and on parameters R_0, R.

Proof. Let us recall that the Laplacian takes the following form in the polar coordinate system,

$$\Delta w = w_{rr} + \frac{1}{r} w_r + \frac{1}{r^2} w_{\phi\phi}, \tag{2.2}$$

and if w is radially symmetric,

$$\Delta w = w_{rr} + \frac{1}{r} w_r. \tag{2.3}$$

The function h_ρ admits the Fourier series expansion

$$h_\rho = c_\rho + \sum_{k=1}^{\infty} (a_{\rho,k} \sin k\phi + b_{\rho,k} \cos k\phi),$$

where in addition

$$c_\rho^2 + \sum_{k=1}^{\infty} (a_{\rho,k}^2 + b_{\rho,k}^2) \le \Lambda_1^2. \tag{2.4}$$

We determine the solution w_ρ of (2.1) in the form of a series.

The first term, corresponding to c_ρ, is radially symmetric and, taking (2.3) into account, has the following representation:

$$w_\rho^0 = A + B \ln r.$$

From the boundary condition

$$\frac{\partial w_\rho^0}{\partial n} = c_\rho \quad \text{on} \quad C(\rho),$$

it follows that

$$A + B \ln R = 0,$$
$$B \frac{1}{\rho} = c_\rho;$$

hence

$$B = \rho c_\rho, \qquad A = -\rho c_\rho \ln R,$$

where $|c_\rho| \le \Lambda_1$. Finally,

$$w_\rho^0 = \rho c_\rho \ln r - \rho c_\rho \ln R. \tag{2.5}$$

Consider now the terms corresponding to the boundary condition

$$\frac{\partial w_\rho^k}{\partial n} = a_{\rho,k} \cdot \sin k\phi \quad \text{on} \quad C(\rho)$$

for $k \ge 1$ (cosine terms can be treated in the same way). We seek the solution in the form

$$w_\rho^k = v(r) \cdot \sin k\phi,$$

and, in view of (2.2), we have the representation

$$v(r) = Ar^k + B\frac{1}{r^k}.$$

Again, taking into account the boundary conditions it follows that

$$AR^k + B\frac{1}{R^k} = 0,$$
$$kA\rho^{k-1} - kB\frac{1}{\rho^{k+1}} = a_{\rho,k}.$$

We obtain

$$w_\rho^k = \frac{\rho^{k+1}\cdot a_{\rho,k}}{k(R^{2k}+\rho^{k-1})}\left(r^k - \frac{R^{2k}}{r^k}\right)\sin k\phi, \tag{2.6}$$

where $|a_{\rho,k}| \le \Lambda_1$.

Substituting $r := \rho$ in (2.5), (2.6) we get

$$|w_\rho^0(\rho)| \le \Lambda_2(\Lambda_1)\rho|\ln\rho|,$$
$$|w_\rho^k(\rho)| \le \Lambda_3(\Lambda_1)\rho.$$

The convergence of the series for the solution w_ρ follows from (2.4) for sufficiently small ρ. □

Now we consider the bounded domain Ω which satisfies the following assumption.

(H1) Ω is a bounded domain with the boundary $\Gamma = \partial\Omega$ consisting of a finite number of smooth arcs; see Figure 1.

We define the BVP in Ω:

$$\Delta u_0 = f \quad \text{in} \quad \Omega, \tag{2.7a}$$
$$u_0 = g_D \quad \text{on} \quad \Gamma^D, \tag{2.7b}$$
$$\frac{\partial}{\partial n}u_0 = g_N \quad \text{on} \quad \Gamma^N, \tag{2.7c}$$

where $\Gamma^N \cup \Gamma^D = \Gamma = \partial\Omega$, and the data of the problem satisfy the following assumption: m is an integer which is selected in such a way that the solution to (2.7) enjoys the required regularity for our purposes.

(H2) $f \in H^m(\Omega)$, $g_D \in H^{m+3/2}(\Gamma^D)$, $g_N \in H^{m+1/2}(\Gamma^N)$, and the inclusions hold for every arc in Γ^D, Γ^N.

For such a BVP we shall use the regularity results obtained for the solutions to BVPs in nonsmooth domains given, e.g., in [9], [24]. The elliptic regularity of solutions is investigated in [24] in the weighted Sobolev spaces, with the weights defined as functions of the distance to the isolated singular boundary points (corners in the 2D case). For the sake of simplicity, we do not introduce here the specific notation required for the regularity analysis in such spaces. In the case of the scalar equation the related result is given in [24, Theorem 3.1, p. 32] for the Dirichlet problem and in [24, Theorem 4.2, p. 39] for the Neumann problem. The general case of elliptic BVPs in bounded domains is covered by [24, Theorem 3.2, p. 309]. The results established in [24] can be formulated for our purposes in the following theorem.

THEOREM 2.2. *Let* $\boldsymbol{p}^1 \dots \boldsymbol{p}^k$ *be vertices of* Γ *(endpoints of smooth arcs),*

$$\Omega_\delta = \Omega \setminus \bigcup_{i=1}^k \overline{B(\boldsymbol{p}^i;\delta)}, \qquad \Gamma_\delta = \Gamma \setminus \bigcup_{i=1}^k \overline{B(\boldsymbol{p}^i;\delta)}.$$

Then for a given, sufficiently small $\delta > 0$, the solution of (2.7) satisfies the estimate

$$\|u_0\|_{H^{m+2}(\Omega_\delta)} \leq \Lambda(\delta)\left(\|f\|_{H^m(\Omega_{\delta/2})} + \|g_D\|_{H^{m+3/2}(\Gamma_{\delta/2})} + \|g_N\|_{H^{m+1/2}(\Gamma_{\delta/2})}\right).$$

We can replace $\Omega_{\delta/2}, \Gamma_{\delta/2}$ by Ω, Γ in the right-hand side of the above inequality, provided the norms in the right-hand side over the specific sets are finite. □

In the rest of the paper we assume that all the regularity assumptions concerning the data f, g_D, g_N of the problem (2.7) are satisfied with the parameter m specified separately in every particular case.

3. Double asymptotic expansion. We consider the case of two distinct holes. The finite number of distinct holes can be treated in the same way. We establish the asymptotic approximations of solutions to the Laplace equation with respect to the radii of the holes. This is an extension of the results given [30] in the case of a single hole.

Let us select two different points $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$ in the interior of Ω. Next we remove from Ω two balls, $B(\boldsymbol{x}^1;\rho_1)$ and $B(\boldsymbol{x}^2;\rho_2)$. There always exists $\delta > 0$ such that the set of pairs (ρ_1,ρ_2) satisfying the condition

$$0 < \rho_1, \rho_2 < \frac{1}{2}\max\left\{\ \mathrm{dist}(\boldsymbol{x}^1, \partial\Omega_\delta),\ \mathrm{dist}(\boldsymbol{x}^2, \partial\Omega_\delta),\ \frac{1}{2}|\boldsymbol{x}^2 - \boldsymbol{x}^1|\ \right\} \tag{3.1}$$

is nonempty. We define

$$\Omega_{\rho_1\rho_2} = \Omega \setminus (\overline{B(\boldsymbol{x}^1;\rho_1)} \cup \overline{B(\boldsymbol{x}^2;\rho_2)})$$

and denote by $u_{\rho_1\rho_2}$ the solution of the following BVP, which is a modification of (2.7) taking into account the existence of holes:

$$\Delta u_{\rho_1\rho_2} = f \quad \text{in} \quad \Omega_{\rho_1\rho_2}, \tag{3.2a}$$

$$u_{\rho_1\rho_2} = g_D \quad \text{on} \quad \Gamma^D, \tag{3.2b}$$

$$\frac{\partial}{\partial n}u_{\rho_1\rho_2} = g_N \quad \text{on} \quad \Gamma^N, \tag{3.2c}$$

$$\frac{\partial}{\partial n}u_{\rho_1\rho_2} = 0 \quad \text{on} \quad C(\boldsymbol{x}^1;\rho_1) \cup C(\boldsymbol{x}^2;\rho_2). \tag{3.2d}$$

Furthermore, we assume that the number m defining the regularity of the data as specified in Theorem 2.2 satisfies $m \geq 1$.

In view of condition (3.1) satisfied by the radii ρ_1, ρ_2 there exist constants R_0 and R, such that $R_0 > \max\{\rho_1,\rho_2\}$ and

$$\Omega \subset B(\boldsymbol{x}^1;R) \cap B(\boldsymbol{x}^2;R),$$

$$B(\boldsymbol{x}^1;R_0) \cap \partial\Omega_\delta = B(\boldsymbol{x}^2;R_0) \cap \partial\Omega_\delta = B(\boldsymbol{x}^1;R_0) \cap B(\boldsymbol{x}^2;R_0) = \emptyset.$$

All the geometrical objects introduced so far are presented in Figure 1. In order to formulate the main result of this section, we define the functions $s_{\rho_1},\ s_{\rho_2}$ of the form

$$s_{\rho_1}(\boldsymbol{x}) = \frac{\partial u_0}{\partial x_1}(\boldsymbol{x}^1)\cdot\frac{\rho_1^2}{r_1}\cos\theta_1 + \frac{\partial u_0}{\partial x_2}(\boldsymbol{x}^1)\cdot\frac{\rho_1^2}{r_1}\sin\theta_1,$$

$$s_{\rho_2}(\boldsymbol{x}) = \frac{\partial u_0}{\partial x_1}(\boldsymbol{x}^2)\cdot\frac{\rho_2^2}{r_2}\cos\theta_2 + \frac{\partial u_0}{\partial x_2}(\boldsymbol{x}^2)\cdot\frac{\rho_2^2}{r_2}\sin\theta_2.$$

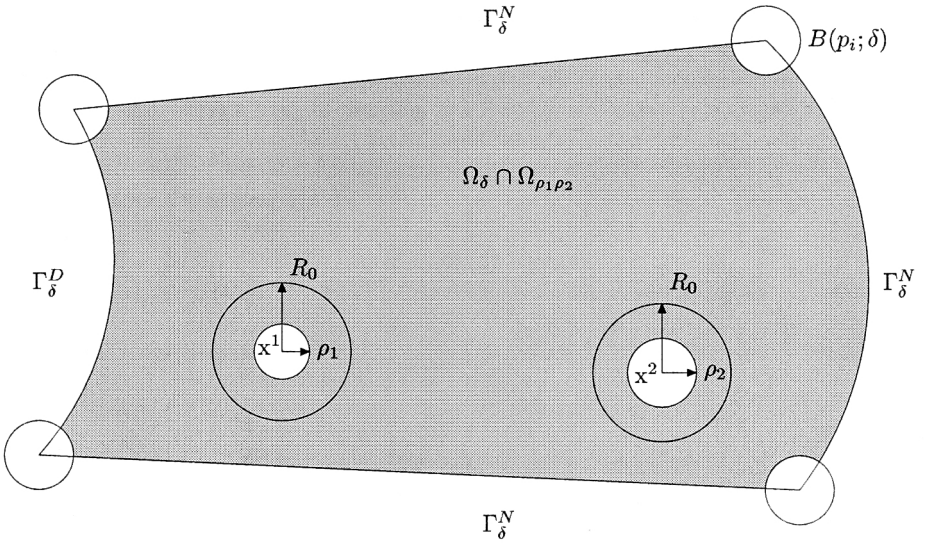


FIG. 1. *The domains Ω_δ, $\Omega_{\rho_1\rho_2}$ and related geometrical objects.*

Here u_0 is the solution of (2.7), while $(r_1 = |\boldsymbol{x} - \boldsymbol{x}^1|, \theta_1)$ and $(r_2 = |\boldsymbol{x} - \boldsymbol{x}^2|, \theta_2)$ are polar coordinate systems around $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$, respectively.

We prove, modifying the argument given in [30] for a single hole, that the following asymptotic approximation is obtained in the case of two holes.

LEMMA 3.1. *The solution of* (3.2) *may be expressed as follows:*

$$u_{\rho_1\rho_2} = u_0 + s_{\rho_1} + s_{\rho_2} + z_{\rho_1\rho_2},$$

where $z_{\rho_1\rho_2}$ satisfies on $C(\boldsymbol{x}^1;\rho_1) \cup C(\boldsymbol{x}^2;\rho_2)$ the estimates

$$\begin{aligned} |z_{\rho_1\rho_2}| &\le \Lambda\left(\rho_1^2|\log\rho_1| + \rho_2^2|\log\rho_2|\right), \\ |\nabla z_{\rho_1\rho_2}| &\le \Lambda(\rho_1+\rho_2), \end{aligned}$$

and, in addition, on the set $P(\boldsymbol{x}^1;R_0,R) \cap P(\boldsymbol{x}^2;R_0,R) \cap \Omega$

$$|z_{\rho_1\rho_2}| \le \Lambda(\rho_1^2+\rho_2^2),$$

with $\Lambda = \Lambda(\Omega,\delta,R_0,R)$.

Proof. Let us set $\rho = \sqrt{\rho_1^2+\rho_2^2}$. We expand $z_{\rho_1\rho_2}$ in the form of the series

$$z_{\rho_1\rho_2} = p^1_{\rho_1} + p^1_{\rho_2} + q^1_\rho + \rho^2(p^2_{\rho_1} + p^2_{\rho_2} + q^2_\rho) + \cdots,$$

where the consecutive terms are defined by the BVPs specified below.

For $p^1_{\rho_1}$ we have

$$\begin{aligned} \Delta p^1_{\rho_1} &= 0 && \text{in} \quad P(\boldsymbol{x}^1;\rho_1,R), \\ p^1_{\rho_1} &= 0 && \text{on} \quad C(\boldsymbol{x}^1;R), \\ \frac{\partial}{\partial n}p^1_{\rho_1} &= -\frac{\partial}{\partial n}(u_0 + s_{\rho_1}) && \text{on} \quad C(\boldsymbol{x}^1;\rho_1), \end{aligned}$$

and for $p^1_{\rho_2}$ we have

$$\begin{aligned} \Delta p^1_{\rho_2} &= 0 && \text{in} \quad P(\boldsymbol{x}^2;\rho_2,R),\\ p^1_{\rho_2} &= 0 && \text{on} \quad C(\boldsymbol{x}^2;R),\\ \frac{\partial}{\partial n}p^1_{\rho_2} &= -\frac{\partial}{\partial n}(u_0+s_{\rho_2}) && \text{on} \quad C(\boldsymbol{x}^2;\rho_2), \end{aligned}$$

while

$$\begin{aligned} \Delta q^1_{\rho} &= 0 && \text{in} \quad \Omega,\\ q^1_{\rho} &= -s_{\rho_1}-s_{\rho_2}-p^1_{\rho_1}-p^1_{\rho_2} && \text{on} \quad \Gamma. \end{aligned}$$

The idea of such a construction can be explained as follows.

Functions $p^1_{\rho_1}$ and $p^1_{\rho_2}$ correct the normal derivatives of $z_{\rho_1\rho_2}$ on $C(\boldsymbol{x}^1;\rho_1)$ and $C(\boldsymbol{x}^2;\rho_2)$, respectively. However, $p^1_{\rho_1}$ introduces the discrepancy around $C(\boldsymbol{x}^2;\rho_2)$ and $p^1_{\rho_2}$ introduces the discrepancy around $C(\boldsymbol{x}^1;\rho_1)$. In addition, $p^1_{\rho_1}$ and $p^1_{\rho_2}$ disturb the Dirichlet boundary conditions on Γ. Therefore, q^1_{ρ} is introduced to correct the boundary conditions on Γ. In this way, further terms of the series for the expansion of $z_{\rho_1\rho_2}$ are introduced.

Now observe that due to the regularity of u_0 (see Theorem 2.2), the gradient ∇u_0 is continuous on the sets $\overline{B(\boldsymbol{x}^1;R_0)}$ and $\overline{B(\boldsymbol{x}^2;R_0)}$. Thus, the norm of the gradient $|\nabla u_0|$ is uniformly bounded on these sets, and we have

$$\left|\frac{\partial}{\partial n}(u_0+s_{\rho_1})\right| \le \Lambda\rho_1 \qquad \text{on} \quad C(\boldsymbol{x}^1;\rho_1),$$

$$\left|\frac{\partial}{\partial n}(u_0+s_{\rho_2})\right| \le \Lambda\rho_2 \qquad \text{on} \quad C(\boldsymbol{x}^2;\rho_2).$$

Therefore, by Lemma 2.1,

$$\begin{aligned} |p^1_{\rho_1}| &\le \Lambda\rho_1^2|\log\rho_1| && \text{on} \quad C(\boldsymbol{x}^1;\rho_1),\\ |p^1_{\rho_1}|+|\nabla p^1_{\rho_1}| &\le \Lambda\rho_1^2 && \text{on} \quad P(\boldsymbol{x}^1;R_0,R),\\ |p^1_{\rho_2}| &\le \Lambda\rho_2^2|\log\rho_2| && \text{on} \quad C(\boldsymbol{x}^2;\rho_2),\\ |p^1_{\rho_2}|+|\nabla p^1_{\rho_2}| &\le \Lambda\rho_2^2 && \text{on} \quad P(\boldsymbol{x}^2;R_0,R). \end{aligned}$$

Hence, $|q^1_{\rho}| \le \Lambda\rho^2$ on Γ, and also

$$|\nabla q^1_{\rho}| \le \Lambda\rho^2 \qquad \text{on} \quad P(\boldsymbol{x}^1;\rho_1,R_0)\cup P(\boldsymbol{x}^2;\rho_2,R_0).$$

In the next step we introduce the corrections of higher order with respect to ρ. As we have seen above, the Neumann boundary conditions on $C(\boldsymbol{x}^1;\rho_1)$ are perturbed by the functions $p^1_{\rho_2}$ and q^1_{ρ}. Therefore, the next set of functions $p^2_{\rho_1}, p^2_{\rho_2}, q^2_{\rho}$ is defined by solutions to the following BVPs.

For $p^2_{\rho_1}$

$$\begin{aligned} \Delta p^2_{\rho_1} &= 0 && \text{in} \quad P(\boldsymbol{x}^1;\rho_1,R),\\ p^2_{\rho_1} &= 0 && \text{on} \quad C(\boldsymbol{x}^1;R),\\ \frac{\partial}{\partial n}p^2_{\rho_1} &= -\frac{1}{\rho^2}\frac{\partial}{\partial n}(q^1_{\rho}+p^1_{\rho_2}) && \text{on} \quad C(\boldsymbol{x}^1;\rho_1), \end{aligned}$$

and for $p^2_{\rho_2}$

$$\begin{aligned} \Delta p^2_{\rho_2} &= 0 && \text{in} \quad P(\boldsymbol{x}^2;\rho_2,R),\\ p^2_{\rho_2} &= 0 && \text{on} \quad C(\boldsymbol{x}^2;R),\\ \frac{\partial}{\partial n}p^1_{\rho_2} &= -\frac{1}{\rho^2}\frac{\partial}{\partial n}(q^1_\rho + p^1_{\rho_1}) && \text{on} \quad C(\boldsymbol{x}^2;\rho_2), \end{aligned}$$

while q^2_ρ satisfies

$$\begin{aligned} \Delta q^2_\rho &= 0 && \text{in} \quad \Omega,\\ q^2_\rho &= -\frac{1}{\rho^2}(p^2_{\rho_1} + p^2_{\rho_2}) && \text{on} \quad \Gamma. \end{aligned}$$

In addition, from properties of $p^1_{\rho_1}, p^1_{\rho_2}, q^1_\rho$ it follows that

$$\left|\frac{\partial}{\partial n}(q^1_\rho + p^1_{\rho_2})\right| \le \Lambda\rho^2 \qquad \text{on} \quad C(\boldsymbol{x}^1;\rho_1),$$

$$\left|\frac{\partial}{\partial n}(q^1_\rho + p^1_{\rho_1})\right| \le \Lambda\rho^2 \qquad \text{on} \quad C(\boldsymbol{x}^2;\rho_2).$$

By Lemma 2.1 the terms $\rho^2(p^2_{\rho_1} + p^2_{\rho_2} + q^2_\rho)$ constitute the correction of the order $\rho^3|\log\rho|$. Finally, we again use Lemma 2.1 to obtain the estimates for $p^1_{\rho_1} + p^1_{\rho_2} + q^1_\rho$ and, therefore, for $z_{\rho_1\rho_2}$. □

4. Topological differential with respect to multiple holes. We shall restrict our analysis to the case of two holes, since the generalization to the case of a finite number of holes can be performed in the same way. Let us define the domain functionals I_u and I_g as follows:

$$I_u(\rho_1,\rho_2) = \int_{\Omega_{\rho_1\rho_2}} F(u_{\rho_1\rho_2})\,dx, \tag{4.1}$$

$$I_g(\rho_1,\rho_2) = \int_{\Omega_{\rho_1\rho_2}} |\nabla u_{\rho_1\rho_2}|^{2p}\,dx, \tag{4.2}$$

$p = 1, 2$, where $u_{\rho_1\rho_2}$ denotes the solution to (3.2) and $\Omega_{\rho_1\rho_2} = \Omega \setminus (\overline{B(\boldsymbol{x}^1;\rho_1)} \cup \overline{B(\boldsymbol{x}^2;\rho_2)})$. Our main result on the asymptotic approximation of the shape functionals with respect to multiple holes is given below.

THEOREM 4.1. *Assume that* (H1) *and* (H2) *are satisfied, with* $m \ge 1$, F *is a* C^2 *function, and* $\rho_1, \rho_2 > 0$ *are small enough.*

Then, the following representation is obtained for the topological variations of the shape functionals:

$$I_u(\rho_1,\rho_2) = I_u(0,0) + \mathcal{T}I_u(\boldsymbol{x}^1)\cdot|B(\boldsymbol{x}^1;\rho_1)| + \mathcal{T}I_u(\boldsymbol{x}^2)\cdot|B(\boldsymbol{x}^2;\rho_2)| + o(\rho^2), \tag{4.3}$$

$$I_g(\rho_1,\rho_2) = I_g(0,0) + \mathcal{T}I_g(\boldsymbol{x}^1)\cdot|B(\boldsymbol{x}^1;\rho_1)| + \mathcal{T}I_g(\boldsymbol{x}^2)\cdot|B(\boldsymbol{x}^2;\rho_2)| + o(\rho^2). \tag{4.4}$$

We omit here the dependence of the TDs $\mathcal{T}I_u(\cdot), \mathcal{T}I_g(\cdot)$ on the domain Ω (see (1.2)), because Ω is fixed in all subsequent considerations in this section. Formulae (4.3) and (4.4) define *the topological gradient* for the functional I_u,

$$\mathcal{T}I_u(\Omega; \boldsymbol{x}^1, \boldsymbol{x}^2) \;=\; \left[\mathcal{T}I_u(\Omega; \boldsymbol{x}^1), \mathcal{T}I_u(\Omega; \boldsymbol{x}^2)\right],$$

which contains the TDs evaluated at the centers $\boldsymbol{x}^1, \boldsymbol{x}^2$ of the holes.

The topological gradient contain TDs corresponding to the holes, and, therefore, it looks like the classical gradient of a multivariable function consisting of all partial derivatives. The topological gradient allows us to express the related variations of the shape functionals in terms of the product of volumes of the balls centered at $\boldsymbol{x}^1, \boldsymbol{x}^2$ multiplied by the TDs evaluated at the points $\boldsymbol{x}^1, \boldsymbol{x}^2$. (The procedure is the same, in general, only for a finite number of such points.)

We use the notation $\nabla u \cdot \nabla v = \sum_{i=1}^{2} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i}$ for the scalar product of two vectors and $\boldsymbol{n}^\top K \nabla u = \sum_{i,j=1}^{2} n_i K_{i,j} \frac{\partial u}{\partial x_j}$ for the product of a matrix with two vectors.

Proof. We evaluate the directional derivative of $I_u(\rho_1, \rho_2)$ at $\rho = 0^+$. Let us recall that in the case of a single hole the formula for $\mathcal{T}I_u(\boldsymbol{x}^0)$ has been derived in [29] and takes the form

$$\mathcal{T}I_u(\boldsymbol{x}^0) = -\left[F(u_0) + f w_0 + 2\nabla u_0 \cdot \nabla w_0\right]_{\boldsymbol{x}=\boldsymbol{x}^0}, \tag{4.5}$$

where w_0 is the adjoint function which satisfies (4.8).

Let us take $c_1, c_2 > 0$, $c_1^2 + c_2^2 = 1$, and the radii of balls in the form $\rho_1 = c_1\rho,\ \rho_2 = c_2\rho$. Then $\rho_1^2 + \rho_2^2 = \rho^2$ and $|B(\boldsymbol{x}^1; \rho_1)| + |B(\boldsymbol{x}^2; \rho_2)| = \pi\rho^2$. By varying the parameter ρ we change the boundaries of both holes simultaneously. Finally, we introduce the space

$$H_D^1(\Omega) = \{\ \phi \in H^1(\Omega) \mid \phi = 0 \quad \text{on} \quad \Gamma^D\ \}$$

and the convex set

$$H_g^1(\Omega) = \{\ \phi \in H^1(\Omega) \mid \phi = g_D \quad \text{on} \quad \Gamma^D\ \}.$$

In the first step we evaluate the derivative

$$\begin{aligned} \frac{d}{d\rho} I_u(\rho_1, \rho_2) = &\int_{\Omega_{\rho_1\rho_2}} F_u(u_{\rho_1\rho_2}) u'_{\rho_1\rho_2}\, dx \\ &- c_1 \int_{C(\boldsymbol{x}^1;\rho_1)} F(u_{\rho_1\rho_2})\, ds - c_2 \int_{C(\boldsymbol{x}^2;\rho_2)} F(u_{\rho_1\rho_2})\, ds, \end{aligned} \tag{4.6}$$

where $u'_{\rho_1\rho_2}$ is the shape derivative of $u_{\rho_1\rho_2}$ with respect to the change of radii of both circles; we refer to [28] for the details. By $d/d\rho$ we mean the Eulerian semiderivative of the shape functional $I_u(\rho_1, \rho_2)$ in the direction of the specific vector field $\boldsymbol{V}$ equal to the exterior unit normal field $-\boldsymbol{n}$ on $C(\boldsymbol{x}^1; \rho_1)$ and $C(\boldsymbol{x}^2; \rho_2)$. The result of the derivation conforms to the Hadamard formula [28] which says that in the case of differentiable shape functionals the shape derivatives actually depend on the normal component of the velocity field on the moving boundary.

We know that

$$u'_{\rho_1\rho_2} = \dot{u}_{\rho_1\rho_2} - \nabla u_{\rho_1\rho_2} \cdot \boldsymbol{V},$$

where $\boldsymbol{V}$ is an appropriate vector field introduced above by an extension of the normal field $-\boldsymbol{n}$ to the small neighborhoods of $C(\boldsymbol{x}^1;\rho_1)$ and $C(\boldsymbol{x}^2;\rho_2)$ and $\dot{u}$ denotes the material derivative in the direction V. Now, observe that we may take smooth $\boldsymbol{V}$ which vanishes outside Ω_δ, i.e., the only region where $\nabla u_{\rho_1\rho_2}$ actually may have singularities. Hence $u'_{\rho_1\rho_2}$ enjoys the same regularity as $\dot{u}_{\rho_1\rho_2}$, in particular, $u'_{\rho_1\rho_2} \in H^1_D(\Omega_{\rho_1\rho_2})$.

The shape derivative $u'_{\rho_1\rho_2} \in H^1_D(\Omega_{\rho_1\rho_2})$ satisfies for all $\phi \in H^1_D(\Omega_{\rho_1\rho_2})$ the following integral identity:

$$\begin{aligned}\int_{\Omega_{\rho_1\rho_2}} \nabla u'_{\rho_1\rho_2}\cdot\nabla\phi\,dx - c_1\int_{C(\boldsymbol{x}^1;\rho_1)} \nabla u'_{\rho_1\rho_2}\cdot\nabla\phi\,ds - c_2\int_{C(\boldsymbol{x}^2;\rho_2)} \nabla u'_{\rho_1\rho_2}\cdot\nabla\phi\,ds \\ = c_1\int_{C(\boldsymbol{x}^1;\rho_1)} f\phi\,ds + c_2\int_{C(\boldsymbol{x}^2;\rho_2)} f\phi\,ds.\end{aligned} \tag{4.7}$$

Now we introduce the adjoint variables w_0 and $w_{\rho_1\rho_2}$ defined by the following variational identities:

(1) Find $w_0 \in H^1_D(\Omega)$ such that

$$-\int_\Omega \nabla w_0\cdot\nabla\phi, dx = \int_\Omega F_u(u_0)\phi\,dx \quad \text{for all} \quad \phi \in H^1_D(\Omega). \tag{4.8}$$

(2) Find $w_{\rho_1\rho_2} \in H^1_D(\Omega_{\rho_1\rho_2})$ such that

$$-\int_{\Omega_{\rho_1\rho_2}} \nabla w_{\rho_1\rho_2}\cdot\nabla\phi, dx = \int_{\Omega_{\rho_1\rho_2}} F_u(u_{\rho_1\rho_2})\phi\,dx \quad \text{for all} \quad \phi \in H^1_D(\Omega_{\rho_1\rho_2}). \tag{4.9}$$

Taking into account the regularity of $u'_{\rho_1\rho_2}$, we make cross-substitutions in (4.7) and (4.9) which lead to

$$\begin{aligned}\int_{\Omega_{\rho_1\rho_2}} F_u(u_{\rho_1\rho_2})u'_{\rho_1\rho_2}\,dx = &- c_1\int_{C(\boldsymbol{x}^1;\rho_1)} [fw_{\rho_1\rho_2} + \nabla u_{\rho_1\rho_2}\cdot\nabla w_{\rho_1\rho_2}]\,ds \\ &- c_2\int_{C(\boldsymbol{x}^2;\rho_2)} [fw_{\rho_1\rho_2} + \nabla u_{\rho_1\rho_2}\cdot\nabla w_{\rho_1\rho_2}]\,ds.\end{aligned} \tag{4.10}$$

As a result, from (4.6) we obtain

$$\begin{aligned}\frac{d}{d\rho}I_u(\rho_1,\rho_2) = &- c_1\int_{C(\boldsymbol{x}^1;\rho_1)} [F(u_{\rho_1\rho_2}) + fw_{\rho_1\rho_2} + \nabla u_{\rho_1\rho_2}\cdot\nabla w_{\rho_1\rho_2}]\,ds \\ &- c_2\int_{C(\boldsymbol{x}^2;\rho_2)} [F(u_{\rho_1\rho_2}) + fw_{\rho_1\rho_2} + \nabla u_{\rho_1\rho_2}\cdot\nabla w_{\rho_1\rho_2}]\,ds.\end{aligned} \tag{4.11}$$

In the next step we observe that the derivative with respect to ρ^2 can be expressed by the derivative with respect to ρ in the following way:

$$\frac{dI_u(\rho_1,\rho_2)}{d(\pi\rho^2)} = \frac{1}{2\pi\rho}\cdot\frac{d}{d\rho}I_u(\rho_1,\rho_2). \tag{4.12}$$

Such an observation is very useful for applications, since we avoid any evaluation of the second order derivatives of shape functionals for the specific 2D problem.

In addition, it can be easily proved in the same way as it has been done in [29] by an application of the results given by [24] for the case of single hole, that $w_{\rho_1\rho_2}$ has the same kind of expansion as $u_{\rho_1\rho_2}$,

$$w_{\rho_1\rho_2} = w_0 + s_{\rho_1}(w_0) + s_{\rho_2}(w_0) + z_{\rho_1\rho_2}(w_0),$$

with $z_{\rho_1\rho_2}(w_0)$ satisfying the same type of estimates; namely on the set $C(\boldsymbol{x}^1;\rho_1)\cup C(\boldsymbol{x}^2;\rho_2)$

$$\begin{aligned} |z_{\rho_1\rho_2}(w_0)| &\le \Lambda\left(\rho_1^2|\log\rho_1| + \rho_2^2|\log\rho_2|\right),\\ |\nabla z_{\rho_1\rho_2}(w_0)| &\le \Lambda(\rho_1+\rho_2), \end{aligned}$$

and on the set $P(\boldsymbol{x}^1;R_0,R)\cap P(\boldsymbol{x}^2;R_0,R)\cap\Omega$

$$|z_{\rho_1\rho_2}(w_0)| \le \Lambda(\rho_1^2+\rho_2^2).$$

Taking into account both approximations for $u_{\rho_1\rho_2}$ and $w_{\rho_1\rho_2}$ we may compute the limit

$$\begin{aligned} \lim_{\rho\to 0+}\frac{(-1)}{2\pi\rho}\,c_1\int_{C(\boldsymbol{x}^1;\rho_1)}&\left[F(u_{\rho_1\rho_2})+fw_{\rho_1\rho_2}+\nabla u_{\rho_1\rho_2}\cdot\nabla w_{\rho_1\rho_2}\right]ds\\ &= -c_1^2\left[F(u_0)+fw_0+2\nabla u_0\cdot\nabla w_0\right]_{\boldsymbol{x}=\boldsymbol{x}^1} = c_1^2\,\mathcal{T}I_u(\boldsymbol{x}^1) \end{aligned} \tag{4.13}$$

and similarly for the integral on $C(\boldsymbol{x}^2;\rho_2)$. This gives for the increment of I_u the expression

$$dI_u = \pi\rho^2\cdot\left[c_1^2\,\mathcal{T}I_u(\boldsymbol{x}^1)+c_1^2\,\mathcal{T}I_u(\boldsymbol{x}^2)\right].$$

But $\pi\rho^2c_1^2=|B(\boldsymbol{x}^1;\rho_1)|$ and $\pi\rho^2c_2^2=|B(\boldsymbol{x}^2;\rho_2)|$; therefore the expansion (4.3) for I_u follows. The expansion for the functional I_g can be obtained in the same way. □

Remark 4.2. It is shown in [25] that the above procedure gives exactly the same result as the direct application of the asymptotic analysis to shape functionals under consideration. It means that the TDs can be obtained in a different way, but the form of such derivatives is always equivalent, which can be seen after the appropriate and involved transformations.

5. Dependence of solutions on boundary variations. We describe briefly the speed method in shape sensitivity analysis, referring the reader to [28] for details. The domains with moving boundaries are constructed in such a way that allows us to obtain expansions of solutions to BVPs in such domains with respect to the boundary perturbations.

Let us consider the BVP of the same form as the problem discussed in previous sections. Now, the Neumann part Γ^N of the boundary $\Gamma=\Gamma^N\cup\Gamma^D$ is divided into two parts: the fixed part, still denoted by Γ^N, with the nonhomogeneous Neumann conditions prescribed on Γ^N, and the part denoted by Γ^V, which is subjected to the boundary variations. For the sake of simplicity it is assumed that the homogeneous Neumann conditions are prescribed on Γ^V.

$$\Delta u_0 = f \qquad \text{in} \quad \Omega, \tag{5.1a}$$

$$u_0 = g_D \qquad \text{on} \quad \Gamma^D, \tag{5.1b}$$

$$\frac{\partial}{\partial n}u_0 = g_N \qquad \text{on} \quad \Gamma^N, \tag{5.1c}$$

$$\frac{\partial}{\partial n}u_0 = 0 \qquad \text{on} \quad \Gamma^V. \tag{5.1d}$$

In addition, we assume that Γ^V consists of a single smooth arc and the parameter m describing the regularity of the data in Theorem 2.2 now satisfies the stronger condition $m\ge 2$.

The evolution of Γ^V with respect to the parameter τ is governed by the transformation $T(\tau,\cdot) : \Omega \mapsto \Omega_\tau$ (see [28]), defined by the vector field $\boldsymbol{\Theta}$, of the form

$$T(\tau,\boldsymbol{x}) = \boldsymbol{x} + \tau\boldsymbol{\Theta}(\boldsymbol{x}), \qquad \boldsymbol{x} \in B(R) \supset \Omega. \tag{5.2}$$

We assume that the vector field $\boldsymbol{\Theta}$ satisfies the following assumption.

(H3) $\boldsymbol{\Theta} \in [C^4(\overline{B(R)})]^2$, and

$$\|\boldsymbol{\Theta}\|_{[C^4(\overline{B(R)})]^2} \le C_\theta.$$

Moreover, there exists $\delta > 0$ such that the field $\boldsymbol{\Theta}(\boldsymbol{x})$ vanishes in the region $U = \Omega \setminus (\Gamma^N \cup \Gamma^D + B(\delta))$ (see Figure 2) i.e., in an open neighborhood of the fixed part of the boundary.

The above assumptions imply several properties of the transformation $T(\tau,\cdot)$. Namely, there exists a constant $\tau_0 > 0$ such that for $0 \le \tau < \tau_0$ we have the following:

1. $T(\tau,\cdot)$ is invertible in $B(R)$;
2. $T(\tau,\boldsymbol{x}) = \boldsymbol{x}$ on U, and $T(\tau, B(R)\setminus U) = B(R)\setminus U$;
3. $T^{-1}(\tau,\boldsymbol{y}) = \boldsymbol{y} - \tau\boldsymbol{K}_1(\boldsymbol{y},\tau)$, where

$$\|\boldsymbol{K}_1(\boldsymbol{y},\tau)\|_{[C^3(\overline{B(R)})]^2} \le \Lambda C_\theta.$$

The third property requires short justification. Let $\boldsymbol{y} = T(\tau,\boldsymbol{x})$. Then

$$\begin{aligned}\boldsymbol{y} - \tau\boldsymbol{\Theta}(\boldsymbol{y}) &= \boldsymbol{x} + \tau\boldsymbol{\Theta}(\boldsymbol{x}) - \tau\boldsymbol{\Theta}(\boldsymbol{x} + \tau\boldsymbol{\Theta}(\boldsymbol{x})) \\ &= \boldsymbol{x} + \tau\boldsymbol{\Theta}(\boldsymbol{x}) - \tau\boldsymbol{\Theta}(\boldsymbol{x}) - \tau^2 D\boldsymbol{\Theta}(\boldsymbol{x} - \eta\tau\boldsymbol{\Theta}(\boldsymbol{x}))\cdot\boldsymbol{\Theta}(\boldsymbol{x}),\end{aligned}$$

where $\boldsymbol{x}^* := \boldsymbol{x} - \eta\tau\boldsymbol{\Theta}(\boldsymbol{x}) \in B(R)$ and $0 < \eta < 1$. Hence we have the representation $\boldsymbol{x} = \boldsymbol{y} - \tau\boldsymbol{K}_1(\boldsymbol{y},\tau)$, where $\boldsymbol{K}_1 = \boldsymbol{\Theta}(\boldsymbol{y}) - \tau D\boldsymbol{\Theta}(\boldsymbol{x}^*)\cdot\boldsymbol{\Theta}(\boldsymbol{x})$ as the function of y with $\boldsymbol{x} = T^{-1}(\tau,\boldsymbol{y})$ and $\boldsymbol{x}^* = T^{-1}(\tau,\boldsymbol{y}) - \eta\tau\boldsymbol{\Theta}(T^{-1}(\tau,\boldsymbol{y}))$ is bounded in the C^3-norm.

Let us now fix $\delta > 0$, and let $\boldsymbol{\Theta}(\boldsymbol{x})$ be a vector field satisfying all conditions listed above. We consider in the domain $\Omega_\tau = T(\tau,\Omega)$ the following BVP parametrized by $\tau \ge 0$:

$$\Delta u_\tau = f \qquad \text{in} \quad \Omega_\tau, \tag{5.3a}$$

$$u_\tau = g_D \qquad \text{on} \quad \Gamma^D, \tag{5.3b}$$

$$\frac{\partial}{\partial n}u_\tau = g_N \qquad \text{on} \quad \Gamma^N, \tag{5.3c}$$

$$\frac{\partial}{\partial n}u_\tau = 0 \qquad \text{on} \quad \Gamma^V_\tau = T(\tau,\Gamma^V). \tag{5.3d}$$

In the above system all the differentiations are performed with respect to the variable $\boldsymbol{y} \in \Omega_\tau$.

LEMMA 5.1. *Assume that the conditions* (H1)–(H3) *are satisfied and* $m \ge 2$. *Then, for* $\tau \ge 0$, τ *small enough, the solution of* (5.3) *can be expressed as a function of the parameter* τ *as follows:*

$$u_\tau = u_0 + \tau z,$$

where

$$\|z\|_{H^3(\Omega_\delta)} \le \Lambda(\Omega, C_\theta),$$

and we denote $C_\theta = \|\boldsymbol{\Theta}\|_{C^3(\overline{B(R)})}$.

Proof. After the change of variables defined by the transformation $\boldsymbol{y} = T(\tau, \boldsymbol{x})$, BVP (5.3) is transported in a standard way [28] to the fixed domain Ω, and the resulting BVP takes the form

$$\text{(5.4a)} \qquad \operatorname{div}\left[J(T)(D_y T^{-1} D_y^{-T})\nabla u_\tau\right] = J(T) f(T(\tau, \boldsymbol{x})) \qquad \text{in} \quad \Omega,$$

$$\text{(5.4b)} \qquad u_\tau = g_D \qquad \text{on} \quad \Gamma^D,$$

$$\text{(5.4c)} \qquad \frac{\partial}{\partial n} u_\tau = g_N \qquad \text{on} \quad \Gamma^N,$$

$$\text{(5.4d)} \qquad \boldsymbol{n}^\top (D_y T^{-1} D_y^{-T})\nabla u_\tau = 0 \qquad \text{on} \quad \Gamma^V.$$

The coefficients appearing in the above system may be expressed in the form suitable for further computations. Namely

$$\begin{aligned} D_y T^{-1}(\tau, \boldsymbol{y}) &= I - \tau D_y K_1(\boldsymbol{y}, \tau), \\ J(T) &= \det\left(I + \tau D\boldsymbol{\Theta}(\boldsymbol{x})\right) = 1 + \tau k_1(\tau, \boldsymbol{x}), \end{aligned}$$

where

$$\|k_1(\tau, \cdot)\|_{C^3(\overline{B(R)})} \leq \Lambda C_\theta.$$

Therefore, we have the representations

$$\begin{aligned} D_y T^{-1} D_y^{-T} &= (I - \tau D_y K_1)(I - \tau D_y K_1^\top) = I + \tau K_2(\tau, \boldsymbol{x}), \\ J(T)(D_y T^{-1} D_y^{-T}) &= I + \tau K_3(\tau, \boldsymbol{x}), \\ J(T) f(T(\tau, \boldsymbol{x})) &= f(\boldsymbol{x}) + \tau f_1(\boldsymbol{x}, \tau), \end{aligned}$$

with the estimates

$$\begin{aligned} &\|K_2(\tau, \cdot)\|_{\left[C^2(\overline{B(R)})\right]^4} \leq \Lambda C_\theta \ , \qquad \|K_3(\tau, \cdot)\|_{\left[C^2(\overline{B(R)})\right]^4} \leq \Lambda C_\theta, \\ &\quad \text{and} \quad \|f_1(\tau, \cdot)\|_{H^1(\Omega)} \leq \Lambda C_\theta. \end{aligned}$$

As a result the transformed BVP (5.4) takes on the form

$$\text{(5.5a)} \qquad \operatorname{div}\left[(I + \tau K_3)\nabla u_\tau\right] = f + \tau f_1 \qquad \text{in} \quad \Omega,$$

$$\text{(5.5b)} \qquad u_\tau = g_D \qquad \text{on} \quad \Gamma^D,$$

$$\text{(5.5c)} \qquad \frac{\partial}{\partial n} u_\tau = g_N \qquad \text{on} \quad \Gamma^N,$$

$$\text{(5.5d)} \qquad \boldsymbol{n}^\top (I + \tau K_2)\nabla u_\tau = 0 \qquad \text{on} \quad \Gamma^V.$$

It is important to observe that all the functions K_1, K_2, K_3, k_1 vanish on U.

Now we apply the method of matched asymptotics to the system (5.4), assuming that the solution u_τ can be expanded in the form of the series with respect to the parameter τ,

$$\text{(5.6)} \qquad u_\tau \ = \ \sum_{k=0}^{\infty} \tau^k u_k(\tau, \boldsymbol{x}).$$

Substituting (5.6) into (5.5) leads formally to the system

$$\sum_{k=0}^{\infty}\tau^k\Delta u_k+\sum_{k=0}^{\infty}\tau^{k+1}\operatorname{div}(K_3\nabla u_k)=f+\tau f_1 \qquad \text{in} \quad \Omega, \tag{5.7a}$$

$$\sum_{k=0}^{\infty}\tau^k u_k=g_D \qquad \text{on} \quad \Gamma^D, \tag{5.7b}$$

$$\sum_{k=0}^{\infty}\tau^k\frac{\partial}{\partial n}u_k=g_N \qquad \text{on} \quad \Gamma^N, \tag{5.7c}$$

$$\sum_{k=0}^{\infty}\tau^k\frac{\partial}{\partial n}u_k+\sum_{k=0}^{\infty}\tau^{k+1}\boldsymbol{n}^\top K_2\nabla u_k=0 \qquad \text{on} \quad \Gamma^V. \tag{5.7d}$$

In a standard way, comparing the terms with the powers of τ of the same order, we obtain the sequence of BVPs for subsequent functions u_k, $k=0,1,\dots$.

Thus u_0 satisfies (5.1); for $k=1$ we have

$$\begin{aligned}
\Delta u_1&=f_1-\operatorname{div}(K_3\nabla u_0) && \text{in} \quad \Omega,\\
u_1&=0 && \text{on} \quad \Gamma^D,\\
\frac{\partial}{\partial n}u_1&=0 && \text{on} \quad \Gamma^N,\\
\frac{\partial}{\partial n}u_1&=-\boldsymbol{n}^\top K_2\nabla u_0 && \text{on} \quad \Gamma^V,
\end{aligned}$$

and for $k>1$

$$\begin{aligned}
\Delta u_k&=-\operatorname{div}(K_3\nabla u_{k-1}) && \text{in} \quad \Omega,\\
u_k&=0 && \text{on} \quad \Gamma^D,\\
\frac{\partial}{\partial n}u_k&=0 && \text{on} \quad \Gamma^N,\\
\frac{\partial}{\partial n}u_k&=-\boldsymbol{n}^\top K_2\nabla u_{k-1} && \text{on} \quad \Gamma^V.
\end{aligned}$$

By the trace theorem

$$\|\operatorname{div}(K_3\nabla u_{k-1})\|_{H^1(\Omega_\delta)}+\|\boldsymbol{n}^\top K_2\nabla u_{k-1}\|_{H^{5/2}(\Gamma^V_\delta)}\le\Lambda_3\cdot C_\theta\|u_{k-1}\|_{H^3(\Omega_\delta)}.$$

Since K_3, K_2 vanish on U, we have

$$\|\operatorname{div}(K_3\nabla u_{k-1})\|_{H^1(\Omega_\delta)}=\|\operatorname{div}(K_3\nabla u_{k-1})\|_{H^1(\Omega_{\delta/2})},$$

$$\|\boldsymbol{n}^\top K_2\nabla u_{k-1}\|_{H^{5/2}(\Gamma^V_\delta)}=\|\boldsymbol{n}^\top K_2\nabla u_{k-1}\|_{H^{5/2}(\Gamma^V_{\delta/2})},$$

and therefore

$$\|\operatorname{div}(K_3\nabla u_{k-1})\|_{H^1(\Omega_{\delta/2})}+\|\boldsymbol{n}^\top K_2\nabla u_{k-1}\|_{H^{5/2}(\Gamma^V_{\delta/2})}\le\Lambda_3\cdot C_\theta\|u_{k-1}\|_{H^3(\Omega_\delta)}.$$

Thus, due to Theorem 2.2, we are able to obtain the recursive bounds for the consecutive solutions u_k:

for $k=1$ $\quad\|u_1\|_{H^3(\Omega_\delta)}\le\Lambda(\|f_1\|_{H^1(\Omega_{\delta/2})}+\Lambda C_\theta\|u_0\|_{H^3(\Omega_{\delta/2})})\le\Lambda C_\theta(1+\|u_0\|_{H^3(\Omega)})$;

for $k>1$ $\quad\|u_k\|_{H^3(\Omega_\delta)}\le\Lambda C_\theta\|u_{k-1}\|_{H^3(\Omega_\delta)}$.

As a result, there exists a constant $\tau_1 > 0$ such that for $\tau < \min[\tau_0, \tau_1]$ the series (5.6) converges in $H^3(\Omega_\delta)$.

The last step consists of verification that u_τ is indeed a solution. We take the truncated series

$$u_\tau^N = \sum_{k=0}^{N} \tau^k u_k(\tau, \boldsymbol{x})$$

and substitute it into the initial BVP (5.5), which leads to the equalities

$$\begin{aligned}
\sum_{k=0}^{N} \tau^k \Delta u_k + \sum_{k=0}^{N} \tau^{k+1} \operatorname{div}(K_3 \nabla u_k) &= f + \tau f_1 + \tau^{N+1} \operatorname{div}(K_3 \nabla u_\tau^N) \quad \text{in} \quad \Omega, \\
u_\tau^N &= g_D \quad \text{on} \quad \Gamma^D, \\
\frac{\partial}{\partial n} u_\tau^N &= g_N \quad \text{on} \quad \Gamma^N, \\
\boldsymbol{n}^\top (I + \tau K_2) \nabla u_\tau^N &= \tau^{N+1} \boldsymbol{n}^\top K_2 \nabla u_\tau^N \quad \text{on} \quad \Gamma^V.
\end{aligned}$$

Hence for the remainder of the series we have the estimate

$$\|u_\tau^N - u_\tau\|_{H^3(\Omega_\delta)} \leq \Lambda \tau^{N+1},$$

which completes the proof. □

Remark 5.2. The same result can be obtained for the nonhomogeneous Neumann condition or for the nonhomogeneous Dirichlet condition on the moving part of the boundary. The only difference in comparison to the proof of Lemma 5.1 is that the functions g and h are transported to the fixed domain and then expanded with respect to the parameter τ. The construction of the asymptotic approximations is performed in the same way as for the right-hand side of the equation in the proof of Lemma 5.1. Therefore, the method of matched asymptotic expansions with respect to τ is applicable to the problems with nonhomogeneous boundary conditions of Neumann or Dirichlet types.

6. Simultaneous topology and shape modification. In this section we shall investigate the variation of the shape functional resulting both from the nucleation of an internal hole and from the boundary variations. We assume that the volume $|\Omega|$ of the geometrical domain is preserved by such perturbations. Let us assume that all the requirements concerning the domain Ω and the field $\boldsymbol{\Theta}$ specified in (H1)–(H3) are satisfied. We assume, in addition, the following:

(H4) For a given $\delta > 0$ the support of the vector field $\boldsymbol{\Theta}$ is contained in the tubular neighborhood $\Gamma^V + B(\delta/2)$.

(H4) ensures that under our assumptions on $\boldsymbol{\Theta}(\boldsymbol{x})$ the following properties of the mapping $T(\tau, \cdot) : \mathbb{R}^2 \to \mathbb{R}^2$ are obtained for $|\tau|$ small enough:

- $T(\tau, \cdot)$ is a bijection of $\Gamma^V + B(\delta)$ onto itself;
- $T(\tau, \cdot)$ is the identity mapping on the set $\Omega \setminus (\Gamma^V + B(\delta))$.

For a given $\boldsymbol{x}^0 \in \Omega$, and $\delta > 0$ such that $\boldsymbol{x}^0 \in \Omega \setminus (\Gamma^V + B(\delta))$, we select a suitable field $\boldsymbol{\Theta}(\boldsymbol{x})$ and assume $|\tau|$ to be small enough. We denote by $\Omega_{\rho\tau}$ the domain $\Omega_\tau \setminus \overline{B(\boldsymbol{x}^0; \rho)}$.

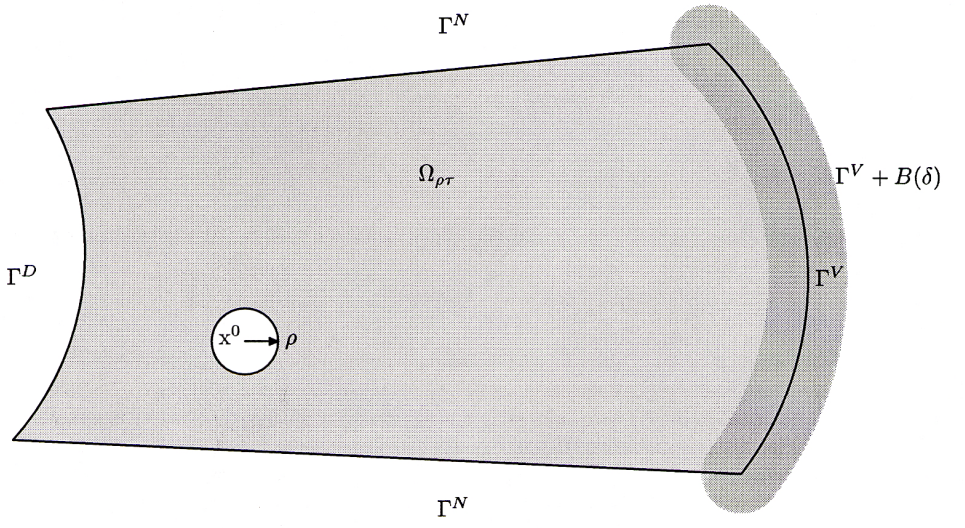


FIG. 2. *The configuration of the internal hole and variable boundary.*

Finally, $u_{\rho\tau}$ is a solution of the following BVP defined on $\Omega_{\rho\tau}$, with the regularity of the data specified in (H2) for $m \geq 2$:

$$\Delta u_{\rho\tau} = f \quad \text{in} \quad \Omega_{\rho\tau}, \tag{6.1a}$$

$$u_{\rho\tau} = g_D \quad \text{on} \quad \Gamma^D, \tag{6.1b}$$

$$\frac{\partial}{\partial n} u_{\rho\tau} = g_N \quad \text{on} \quad \Gamma^N, \tag{6.1c}$$

$$\frac{\partial}{\partial n} u_{\rho\tau} = 0 \quad \text{on} \quad C(\boldsymbol{x}^0; \rho) \cup \Gamma_\tau^V. \tag{6.1d}$$

The form of $\Omega_{\rho\tau}$ and the geometry of the problem under considerations is shown in Figure 2. Similarly, as in previous sections, we consider shape functionals of the form

$$I_u(\eta, \tau) = \int_{\Omega_{\rho\tau}} F(u_{\rho\tau})\, dx, \qquad I_g(\eta, \tau) = \int_{\Omega_{\rho\tau}} \|\nabla u_{\rho\tau}\|^{2p}\, dx\ , \tag{6.2}$$

where we denote $\eta = \pi\rho^2$.

Finally, according to (1.1), we denote by $\mathcal{S}I_u(\Omega; \boldsymbol{\Theta})$ and $\mathcal{S}I_g(\Omega; \boldsymbol{\Theta})$ the shape derivatives of the functionals $I_u(\eta, \tau)$ and $I_g(\eta, \tau)$ with respect to the variation of Γ^V taken at Ω, i.e., for $\eta = 0$, $\tau = 0$. It is well known [28] that these derivatives are given by the following formulae:

$$\mathcal{S}I_u(\Omega; \boldsymbol{\Theta}) = \int_{\Gamma^V} \left[F(u_{00}) + \nabla u_{00} \cdot \nabla w_{00}\right] (\boldsymbol{\Theta} \cdot \boldsymbol{n})\, ds, \tag{6.3a}$$

$$\mathcal{S}I_g(\Omega; \boldsymbol{\Theta}) = \int_{\Gamma^V} \left[\|\nabla u_{00}\|^{2p} + \nabla u_{00} \cdot \nabla v_{00}\right] (\boldsymbol{\Theta} \cdot \boldsymbol{n})\, ds. \tag{6.3b}$$

Furthermore, these derivatives depend only on a normal component $\Theta_n = \boldsymbol{\Theta} \cdot \boldsymbol{n}$ of $\boldsymbol{\Theta}$ on Γ^V. The adjoint variable w_{00} satisfies (4.8). The adjoint variable v_{00} is a solution

of the following adjoint equation depending on the gradient of the functional I_g: find $v_{00} \in H^1_D(\Omega)$ such that

$$-\int_\Omega \nabla v_{00} \cdot \nabla \phi, dx = \int_\Omega 2p|\nabla u_{00}|^{2p-2}(\nabla u_{00} \cdot \nabla \phi)\, dx \quad \text{for all} \quad \phi \in H^1_D(\Omega).$$

The next theorem gives the form of the variations of shape functionals resulting from modification of the geometrical domain by boundary variations as well as the nucleation of a small hole.

THEOREM 6.1. *Assume that conditions* (H1)–(H4) *are satisfied and* $m = 2$, *i.e.,* $f \in H^2(\Omega)$, *and* F *is a* C^2 *function. Then, the functionals* $I_u(\eta, \tau)$ *and* $I_g(\eta, \tau)$ *have the representation*

$$I_u(\eta, \tau) \;=\; I_u(0,0) + \eta \mathcal{T} I_u(\Omega; \boldsymbol{x}^0) + \tau \mathcal{S} I_u(\Omega; \boldsymbol{\Theta}) + r(\eta, \tau),$$

$$I_g(\eta, \tau) \;=\; I_g(0,0) + \eta \mathcal{T} I_u(\Omega; \boldsymbol{x}^0) + \tau \mathcal{S} I_g(\Omega; \boldsymbol{\Theta}) + r(\eta, \tau),$$

where the remainder is of the form $r(\eta, \tau) = o(\eta) + o(\tau) + O(\eta\tau)$ *and* $\eta = \pi\rho^2$.

Proof. The proof is given in the case of I_u, since for the second shape functional I_g the same argument applies.

Let us consider the expression

$$I_u(\eta, \tau) - I_u(0,0) = I_u(\eta, \tau) - I_u(0, \tau) + I_u(0, \tau) - I_u(0,0). \tag{6.4}$$

From the results on the shape sensitivity analysis [28] it follows that

$$I_u(0, \tau) - I_u(0,0) = \tau \mathcal{S} I_u(\Omega; \boldsymbol{\Theta}) + o(\tau).$$

It remains to analyze the first difference on the right-hand side of (6.4). By an application of Theorem 4.1 for a single hole we have

$$I_u(\eta, \tau) - I_u(0, \tau) = \eta \mathcal{T} I_u(\Omega_\tau; \boldsymbol{x}^0) + o(\eta), \tag{6.5}$$

and by Lemma 3.1 the term $o(\eta)$ is uniform with respect to τ for $|\tau|$ small enough. But we have the explicit form of the TD in Ω_τ:

$$\mathcal{T} I_u(\Omega_\tau; \boldsymbol{x}^0) = -\left[F(u_{0\tau}) + f w_{0\tau} + 2\nabla u_{0\tau} \cdot \nabla w_{0\tau}\right]_{\boldsymbol{x}=\boldsymbol{x}^0}.$$

Now using Lemma 5.1, we have

$$|u_{0\tau}(\boldsymbol{x}^0) - u_{00}(\boldsymbol{x}^0)| \le \Lambda\tau, \qquad \|\nabla u_{0\tau}(\boldsymbol{x}^0) - \nabla u_{00}(\boldsymbol{x}^0)\| \le \Lambda\tau,$$
$$|w_{0\tau}(\boldsymbol{x}^0) - w_{00}(\boldsymbol{x}^0)| \le \Lambda\tau, \qquad \|\nabla w_{0\tau}(\boldsymbol{x}^0) - \nabla w_{00}(\boldsymbol{x}^0)\| \le \Lambda\tau.$$

Substituting these estimates into (6.5) leads to

$$I_u(\eta, \tau) - I_u(0, \tau) = \eta \mathcal{T} I_u(\Omega; \boldsymbol{x}^0) + o(\eta) + O(\eta\tau),$$

and the required result for I_u follows. The case of I_g may be treated in an analogous way. □

Remark 6.2. Let us note that using the method proposed in the paper we can define the *domain differential* denoted by $\mathcal{D}J(\Omega; \boldsymbol{\Theta}, \boldsymbol{x})$ of an arbitrary shape functional $J(\Omega)$:

$$\mathcal{D}J(\Omega; \boldsymbol{\Theta}, \boldsymbol{x}^0)(\rho, \tau) = |B(\boldsymbol{x}^0; \rho)| \cdot \mathcal{T} J(\Omega; \boldsymbol{x}^0) + \tau \cdot \mathcal{S} J(\Omega; \boldsymbol{\Theta}).$$

Such a differential provides complete characterization of the variation of $J(\Omega)$ with respect to the variations of Ω, taking into account both the shape and topology changes.

From Theorem 6.1 it follows by standard arguments that an optimal domain Ω^* satisfies the condition

$$\mathcal{D}J(\Omega^*; \boldsymbol{\Theta}, \boldsymbol{x}^0)(\rho, \tau) = |B(\boldsymbol{x}^0; \rho)| \cdot \mathcal{T}J(\Omega^*; \boldsymbol{x}^0) + \tau \cdot \mathcal{S}J(\Omega^*; \boldsymbol{\Theta}) \geq 0$$

for all admissible (ρ, τ) and all admissible vector fields $\boldsymbol{\Theta}$ in an appropriate tangent set. In the case of volume constraints the above formula leads to the necessary optimality conditions of the form

$$\mathcal{T}J(\Omega^*; \boldsymbol{x}^0) \geq 0 \quad \text{in } \Omega^* \quad \text{and} \quad \mathcal{S}J(\Omega^*; \boldsymbol{\Theta}) \geq 0 \quad \text{for all } \boldsymbol{\Theta}. \tag{6.6}$$

The latter inequality follows since under the volume constraints $|\Omega_{\rho\tau}| = |\Omega|$ we have the relation [27], [28]

$$\pi\rho^2 = \tau \int_{\Gamma^V} \theta_n \, d\Gamma + o(\tau),$$

which can be neglected for the admissible tangent directions of $\boldsymbol{\Theta}$. Results on the form of tangent sets for pointwise constraints in L^∞ can be found, e.g., in [2].

7. Example. In the following example it is shown that using the method proposed in the paper we can verify that the shape which is optimal in the framework of classical theory can be improved using the topology variations. For the problem under consideration the second order sufficient optimality conditions have been established by Belov and Fujii in [3].

Let us consider the shape functional

$$J(\Omega) = \int_\Omega u^2 \, dx \quad \to \quad \max, \tag{7.1}$$

where $\Omega \subset \mathbb{R}^2$ is a ring with exterior boundary Γ and interior boundary Σ, as shown in Figure 3. The function u satisfies a BVP

$$\Delta u = -1 \qquad \text{in} \quad \Omega, \tag{7.2a}$$

$$u = 0 \qquad \text{on} \quad \Gamma, \tag{7.2b}$$

$$\frac{\partial u}{\partial n} = 0 \qquad \text{on} \quad \Sigma. \tag{7.2c}$$

We impose the volume constraints $|\Omega| = \pi$ for admissible domains. This example is motivated by optimum design problems concerning the elastic bars in torsion. First, we assume that $\Sigma = \emptyset$, i.e., only the Dirichlet boundary conditions are prescribed on the external boundary Γ. Then, it is known [3] that the unit ball $\Omega^* = B(1)$ maximizes $J(\Omega)$ over the family of simply connected domains with the Dirichlet boundary condition prescribed on the external boundary. We may easily check that the domain Ω^* is a critical point of the integral shape functional $J(\Omega)$. Let us introduce the mapping $T_\tau : \Omega \to \Omega_\tau$ associated with the vector field $\boldsymbol{V}(\cdot, \cdot)$ supported in a small neighborhood of $\Gamma^* = \partial\Omega^*$. In this section the speed vector field is denoted by $\boldsymbol{V}$ instead of $\boldsymbol{\Theta}$. If the volume of Ω_τ is to be preserved, $|\Omega_\tau| = |\Omega|$, then we have

$$\int_{\Gamma^*} (\boldsymbol{V} \cdot \boldsymbol{n}) \, ds = 0. \tag{7.3}$$

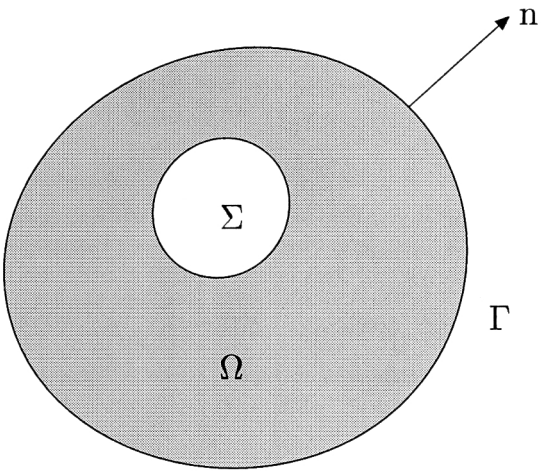


FIG. 3. *Domain* $\Omega \subset \mathbb{R}^2$.

The adjoint equation for the problem under considerations has the form

$$\Delta w = 2u \quad \text{in} \quad \Omega^*, \tag{7.4a}$$

$$w = 0 \quad \text{on} \quad \Gamma^*, \tag{7.4b}$$

$$\frac{\partial w}{\partial n} = 0 \quad \text{on} \quad \Sigma; \tag{7.4c}$$

however, the last condition can be omitted since $\Sigma = \emptyset$. For such a simple geometry of Ω^* both u and w have explicit representations in polar coordinates (r, θ):

$$u(\boldsymbol{x}) = \frac{1}{4}(1 - r^2),$$
$$w(\boldsymbol{x}) = \frac{1}{32}r^4 - \frac{1}{8}r^2 + \frac{3}{32}.$$

Therefore, the shape derivative of $J(\Omega)$ at Ω^* and in the direction $\boldsymbol{V}$ is given by

$$\begin{aligned} \mathcal{S}J(\Omega^*; \boldsymbol{V}) &= \int_{\Gamma^*} \frac{\partial w}{\partial n}\frac{\partial u}{\partial n}\,(\boldsymbol{V}\cdot\boldsymbol{n})\,ds \\ &= \frac{1}{16}\int_{\Gamma^*} (\boldsymbol{V}\cdot\boldsymbol{n})\,ds = 0 \qquad \text{due to (7.3).} \end{aligned} \tag{7.5}$$

Thus Ω^* is a critical point.

Let us consider the change of the topology by nucleation of a small circular hole $B(\rho)$ inside of the domain $\Omega^* = B(1)$, with the Neumann part of the boundary $\Sigma_\rho = \partial B(\rho)$. In order to preserve the volume we must move the exterior boundary Γ^* and expand Ω^*; thus we introduce for $\rho > 0$

$$\Omega_\rho = B(\sqrt{1+\rho^2}) \setminus \overline{B(\rho)}, \qquad \Gamma_\rho = \partial B(\sqrt{1+\rho^2})$$

and denote by u_ρ the solution to (7.2) in Ω_ρ. Again, we have the explicit expressions

$$u_\rho(\boldsymbol{x}) = u(\boldsymbol{x}) + \frac{1}{4}\rho^2 + \frac{1}{2}\rho^2 \log \frac{r}{\sqrt{1+\rho^2}},$$
$$J(\Omega_\rho) = \pi \left[\frac{1}{48} - \frac{3}{32}\rho^2 + \frac{1}{4}\rho^6 A^2 + \frac{1}{8}\rho^6 A - \frac{1}{4}\rho^4 A + \frac{1}{16}\rho^4 \right],$$

with

$$A = \log \frac{\rho}{\sqrt{1+\rho^2}}.$$

It is easy to see that both u_ρ and ∇u_ρ converge pointwise to u and ∇u in $B(1)$, with the removable singularity at $\boldsymbol{x} = 0$.

Moreover, the value $\rho = 0$, which corresponds to $\Omega_\rho = \Omega^* = B(1)$, still gives the maximum of the functional $J(\Omega_\rho)$ defined above; see Figure 4. Using explicit expression for $J(\Omega_\rho)$ we may directly compute the limit

$$\lim_{\rho\to 0+} \frac{dJ(\Omega_\rho)}{d(\pi\rho^2)} = \lim_{\rho\to 0+} \frac{1}{2\pi\rho}\frac{dJ(\Omega_\rho)}{d\rho} = -\frac{3}{32}. \tag{7.6}$$

The negative value indicates, as expected, that a hole of a sufficiently small area with the center at $\boldsymbol{x} = \mathbf{0}$ decreases the value of the functional.

Now we compute the same limit using the domain differential with the shape and topological parts. According to [29] and (4.5), the TD is given by

$$\mathcal{T}J(\Omega^*; 0) = -\left[(u(0))^2 - w(0) + 2\,\nabla u(0)\cdot\nabla w(0)\right] = -\frac{5}{32}.$$

In order to preserve the volume, we set $(\boldsymbol{V}\cdot\boldsymbol{n}) = 1$ in the shape derivative on Γ^* and select τ such that $2\pi\tau = \pi\rho^2$. Then, in view of (7.5),

$$\mathcal{S}J(\Omega^*; \boldsymbol{V}) = \frac{1}{16}\int_{\Gamma^*} (\boldsymbol{V}\cdot\boldsymbol{n})\, ds = \frac{1}{16} 2\pi.$$

Hence

$$\begin{aligned}\mathcal{D}J(\Omega^*; \boldsymbol{V}, 0)(\rho,\tau) &= \mathcal{T}J(\Omega^*;0)\cdot\pi\rho^2 + \tau\cdot\mathcal{S}J(\Omega^*;\boldsymbol{V}) \\ &= \left(-\frac{5}{32} + \frac{1}{16}\right)\pi\rho^2 = -\frac{3}{32}\pi\rho^2,\end{aligned}$$

in agreement with (7.6). This computation confirms the formula given by Theorem 6.1 for the simple case.

It is natural to ask whether without the assumption on the radial symmetry, i.e., with a small hole at $r \neq 0$, the value of the shape functional could possibly be improved. We compute the domain differential

$$\mathcal{D}J(\Omega^*; \boldsymbol{V}, \boldsymbol{x}) = \left[\frac{1}{16} - (u(r))^2 + w(r) - 2\,\frac{\partial u}{\partial r}(r)\cdot\frac{\partial w}{\partial r}(r)\right]\pi\rho^2 = f(r)\cdot\pi\rho^2.$$

The graph of the function $f(r)$ is given in Figure 5. We have $f(r) > 0$ for the location $r > r_0 \approx 0.45$, which means that the nucleation of a small hole near the edge increases $J(\Omega)$. Thus looking at the domain differential we conclude that the unit ball *is not* an optimal domain for the problem (7.1)–(7.2) with volume constraint if we change the topology and admit an arbitrarily located $\Sigma \neq \emptyset$.

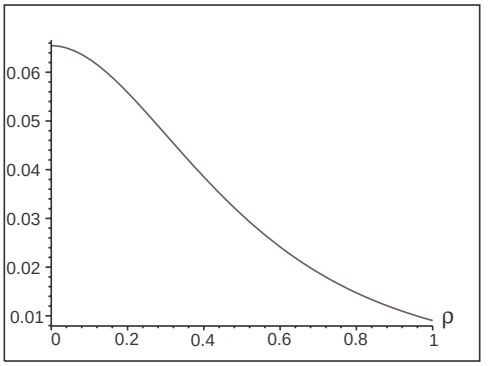


FIG. 4. *The dependence of the goal functional $J(\Omega_\rho)$ as a function of the radius of internal hole ρ.*

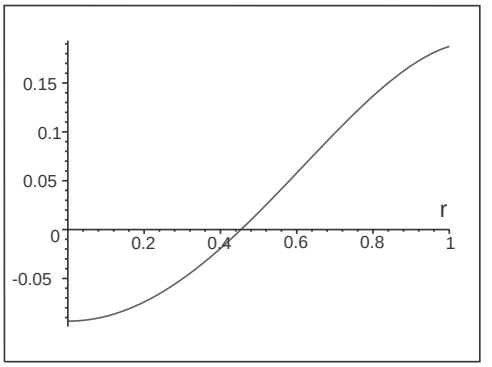


FIG. 5. *The dependence of the function $f(r)$ with respect to the location of the hole r, showing the possibility of increasing the functional.*

REFERENCES

[1] G. ALLAIRE, E. BONNETIER, G. FRANCFORT, AND F. JOUVE, *Shape optimisation by the homogenisation method*, Numer. Math., 76 (1997), pp. 27–68.

[2] E. BEDNARCZUK, M. PIERRE, E. ROUY, AND J. SOKOŁOWSKI, *Calculating tangent sets to certain sets in functional spaces*, Nonlinear Anal., 42 (2000), pp. 871–886.

[3] S. BELOV AND N. FUJII, *Symmetry and sufficient condition of optimality in a domain optimisation problem*, Control Cybernet., 26 (1997), pp. 45–56.

[4] M. PH. BENDSOE, *Optimisation of Structural Topology, Shape and Material*, Springer-Verlag, Berlin, 1995.

[5] M. C. DELFOUR AND J. P. ZOLÉSIO, *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*, Adv. Des. Control, SIAM, Philadelphia, 2001.

[6] A. V. CHERKAEV, Y. GRABOVSKY, A. B. MOVCHAN, AND S. K. SERKOV, *The cavity of the optimal shape under the shear stresses*, Internat. J. Solids Structures, 35 (1998), pp. 4391–4410.

[7] H. A. ESCHENAUER, V. V. KOBELEV, AND A. SCHUMACHER, *Bubble method for topology and shape optimisation of structures*, Struct. Optimiz., 8 (1994), pp. 42–51.

[8] S. GARREAU, P. GUILLAUME, AND M. MASMOUDI, *The topological asymptotic for PDE systems: The elasticity case*, SIAM J. Control Optim., 39 (2001), pp. 1756–1778.

[9] P. GRISVARD, *Singularities in Boundary Value Problems*, Springer-Verlag, Berlin, Masson, Paris, 1992.

[10] A. M. IL'IN, *Matching of Asymptotic Expansions of Solutions of Boundary Value Problems*, Trans. Math. Monogr. 102, AMS, Providence, RI, 1992.

[11] L. JACKOWSKA-STRUMIŁŁO, J. SOKOŁOWSKI, AND A. ŻOCHOWSKI, *The Topological Derivative Method and Artificial Neural Networks for Numerical Solution of Shape Inverse Problems*, RR-3739, INRIA-Lorraine, Nancy, France, 1999.

[12] L. JACKOWSKA-STRUMIŁŁO, J. SOKOŁOWSKI, A. ŻOCHOWSKI, AND A. HENROT, *On numerical solution of shape inverse problems*, Comput. Optim. Appl., 23 (2002), pp. 231–255.

[13] M. KACHANOV, I. TSUKROV, AND B. SHAFIRO, *Effective moduli of solids with cavities of various shapes*, Appl. Mech. Rev., 47 (1994), pp. S151–S174.

[14] I. V. KAMOTSKI AND S. A. NAZAROV, *Spectral problems in singular perturbed domains and self-adjoint extensions of differential operators*, Trudy S.-Petersburg Mat. Obshch., 6 (1998), pp. 151–212 (in Russian); Proceedings of the St. Petersburg Mathematical Society, 6 (2000), pp. 127–181, Amer. Math. Soc. Transl. Ser. 2 199, Amer. Math. Soc., Providence, RI (in English).

[15] T. LEWINSKI AND J. SOKOLOWSKI, *Optimal Shells Formed on a Sphere. The Topological Derivative Method*, RR-3495, INRIA-Lorraine, Nancy, France, 1998.

[16] T. LEWINSKI AND J. SOKOLOWSKI, *Topological derivative for nucleation of non-circular voids*, in Differential Geometric Methods in the Control of Partial Differential Equations (Boulder, CO, 1999), Contemp. Math. 268, R. Gulliver, W. Littman, and R. Triggiani, eds., AMS,

Providence, RI, 2000, pp. 341–361.

[17] T. Lewinski and J. Sokolowski, *Energy change due to appearing of cavities in elastic solids*, Internat. J. Solids Structures, 40 (2003), pp. 1765–1803.

[18] T. Lewinski, J. Sokolowski, and A. Zochowski, *Justification of the bubble method for the compliance minimization problems of plates and spherical shells*, in Third World Congress of Structural and Multidisciplinary Optimization (WCSMO-3), Buffalo/Niagara Falls, New York, 1999, CD-ROM.

[19] T. Lewinski and J. J. Telega, *Plates, Laminates and Shells*, Series on Advances in Applied Sciences, World Scientific, Singapore, 2000.

[20] W. G. Mazja, S. A. Nazarov, and B. A. Plamenevskii, *Asymptotische Theorie elliptischer Randwertaufgaben in singulär gestörten Gebieten*. 1, Akademie-Verlag, Berlin, 1991 (in German); *Asymptotic Theory of Elliptic Boundary Value Problems in Singularly Perturbed Domains*, Vol. 1, Birkhäuser-Verlag, Basel, 2000 (in English).

[21] W. G. Mazja, S. A. Nazarov, and B. A. Plamenevskii, *Asymptotische Theorie elliptischer Randwertaufgaben in singulär gestörten Gebieten*. 2, Akademie-Verlag, Berlin, 1991 (in German); *Asymptotic Theory of Elliptic Boundary Value Problems in Singularly Perturbed Domains*, Vol. 2, Birkhäuser-Verlag, Basel, 2000 (in English).

[22] S. A. Nazarov, *Asymptotic conditions at points, selfadjoint extensions of operators and the method of matched asymptotic expansions*, Trudy S.-Petersburg Mat. Obshch., 5 (1996), pp. 112–183 (in Russian); Trans. Amer. Math. Soc. Ser. 2., 193 (1999), pp. 77–126 (in English).

[23] S. A. Nazarov, *The polynomial property of self-adjoint elliptic boundary-value problems and the algebraic description of their attributes*, Uspekhi Mat. Nauk., 54 (1999), pp. 77–142 (in Russian); Russian Math. Surveys, 54 (1999), pp. 947–1014 (in English).

[24] S. A. Nazarov and B. A. Plamenevsky, *Elliptic Problems in Domains with Piecewise Smooth Boundaries*, de Gruyter Exp. Math. 13, Walter de Gruyter, Berlin, 1994.

[25] S. A. Nazarov and J. Sokołowski, *Asymptotic analysis of shape functionals*, J. Math. Pures Appl., 82 (2003), pp. 125–196.

[26] A. Shumacher, *Topologieoptimierung von Bauteilstrukturen unter Verwendung von Lochpositionierungkriterien*, Ph.D. Thesis, Universität–Gesamthochschule–Siegen, Siegen, Germany, 1995.

[27] J. Sokołowski, *Shape sensitivity analysis of thin shells*, in Optimization Methods in Partial Differential Equations, Contemp. Math. 209, S. Cox and I. Lasiecka, eds., AMS, Providence, RI, 1997, pp. 247–266.

[28] J. Sokołowski and J-P. Zolesio, *Introduction to Shape Optimization: Shape Sensitivity Analysis*, Springer-Verlag, Berlin, 1992.

[29] J. Sokołowski and A. Żochowski, *On the topological derivative in shape optimization*, SIAM J. Control Optim., 37 (1999), pp. 1251–1272.

[30] J. Sokołowski and A. Żochowski, *Topological derivative for optimal control problems*, Control Cybernet., 28 (1999), pp. 611–626.

[31] J. Sokołowski and A. Żochowski, *Topological derivatives for elliptic problems*, Inverse Problems, 15 (1999), pp. 123–134.

[32] J. Sokołowski and A. Żochowski, *Topological derivatives of shape functionals for elasticity systems*, Mech. Structures Mach., 29 (2001), pp. 333–351.

[33] J. Sokołowski and A. Żochowski, *On Topological Derivative in Shape Optimisation*, Rapport de Recherche 3170, INRIA-Lorraine, Nancy, France, 1997.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1222–1243

A CHARACTERIZATION OF BOUNDED-INPUT BOUNDED-OUTPUT STABILITY FOR LINEAR TIME-VARYING SYSTEMS WITH DISTRIBUTIONAL INPUTS*

DANIEL COBB† AND CHI-JO WANG‡

Abstract. We consider the problem of extending the concept of bounded-input bounded-output stability to linear time-varying systems with distributional inputs. In particular, the notion of impulse response is examined in a functional analytic setting. This requires that we first extend the classical notion of an integral operator to distribution space. Duality theory for several key normed spaces is then examined. Next, the adjoint operator corresponding to the given system is studied. Finally, necessary and sufficient conditions for stability are established, along with several expressions for the "gain" of the system.

Key words. bounded-input bounded-output stability, distributions, time-varying systems

AMS subject classification. 93D25

DOI. 10.1137/S0363012901384831

1. Introduction.

The concept of impulse response has traditionally played a central role in linear system theory. In spite of this fact, certain fundamental system-theoretic ideas have apparently not been developed on a mathematically rigorous level for systems with arbitrary distributional inputs and outputs. In a previous paper we addressed the problem of characterizing bounded-input bounded-output (BIBO) stability in the time-invariant case. In this paper we extend the theory to include time-varying linear systems.

To frame the problem, recall that in classical system theory a "system" is typically viewed as an integral operator

$$y(t) = \int_{-\infty}^{\infty} h(t,\tau)u(\tau)d\tau. \tag{1.1}$$

It can be shown (e.g., see [2, p. 109]) that (1.1) determines a bounded linear operator on L^∞ if and only if

$$\sup_t \int_{-\infty}^{\infty} |h(t,\tau)|\, d\tau < \infty. \tag{1.2}$$

Such a characterization is inadequate, however, for studying systems with distributional inputs u, since the integral (1.1) is not defined. In spite of this fact, the kernel $h(t,\tau)$ is often referred to as the system "impulse response." Furthermore, there are many common systems where h itself is a distribution. For example, consider the "time-varying gain"

$$y(t) = \beta(t)u(t).$$

*Received by the editors February 11, 2001; accepted for publication (in revised form) November 28, 2002; published electronically August 6, 2003.
http://www.siam.org/journals/sicon/42-4/38483.html

†Department of Electrical and Computer Engineering, University of Wisconsin, 1415 Engineering Drive, Madison, WI 53706-1691 (cobb@engr.wisc.edu).

‡Department of Electrical Engineering, Southern Taiwan University of Technology, Yung-Kang, Tainan 710, Taiwan (chijo@ms2.hinet.net).

Based on formal manipulations,

$$h(t,\tau) = \beta(t)\delta(t-\tau),$$

where δ is the unit impulse; condition (1.2) cannot be applied directly to distributions. Our goal is to develop a more general theory that characterizes stability for systems with distributional inputs and distributional impulse responses.

In section 2, we present basic analytic results that will be required in subsequent sections. In section 3, we study families of distributions in one variable satisfying certain smoothness properties in the index. These are then interpreted as distributions in two variables and used to generalize the notion of an integral operator. Section 4 applies the theory of normed-space extensions, developed in [1], to distributions in two variables. It is shown in Theorem 4.3 that the space of BIBO stable kernels (i.e., functions satisfying (1.2)) extends naturally to the space of distributions which are derivatives of functions of uniformly bounded variation $DUBV$. Section 5 contains the main results of the paper. Theorem 5.3 states that BIBO stable linear systems are precisely those with $DUBV$ kernels, Theorem 5.4 gives an expression for the adjoint system, and Theorem 5.5 establishes several equivalent representations of the system gain.

2. Preliminaries. First we present some pertinent facts concerning the theory of distributions. See [3], [4], and [5] for more detail. If $\phi : \mathbb{R}^n \to \mathbb{R}$, define the *support* of ϕ (denoted $\operatorname{supp}\phi$) to be the closure of the set $\{(t_1,\dots,t_n) \in \mathbb{R}^n | \phi(t_1,\dots,t_n) \neq 0\}$. Let K_n be the space of C^∞ functions $\phi : \mathbb{R}^n \to \mathbb{R}$ with $\operatorname{supp}\phi$ bounded. Convergence in K_n can be defined in several ways. When we assign a norm $\|\cdot\|$ to K_n we will refer to the pair $(K_n, \|\cdot\|)$. For example, $K_1 \subset L^p$, $1 \le p \le \infty$, so we may consider $(K_1, \|\cdot\|_p)$. Also,

$$\|\psi\|_{p\infty} = \left(\int_{-\infty}^{\infty} \sup_{\tau} |\psi(t,\tau)|^p \, dt \right)^{\frac{1}{p}} < \infty$$

for $\psi \in K_2$, so $(K_2, \|\cdot\|_{p\infty})$ is well defined for $1 \le p < \infty$. *Strong convergence* $\phi_k \to 0$ in K_n means that there exists $a < \infty$ such that $\operatorname{supp}\phi_k \subset [-a,a]$ for every k and $\|\phi_k\|_{C^p} \to 0$ for every integer $p \ge 0$, where

$$\|\phi\|_{C^p} = \max \left\{ \left| \frac{\partial^{i_1+\cdots+i_n}\phi(t_1,\dots,t_n)}{\partial t_1^{i_1}\cdots\partial t_n^{i_n}} \right| \mid 0 \le i_1+\cdots+i_n \le p; t_1,\dots,t_n \in \mathbb{R} \right\}.$$

A *distribution* f is an element of K_n', the dual space of K_n under strong convergence. For $f \in K_n'$, $\operatorname{supp} f$ is defined to be the complement of the largest open set $U \subset \mathbb{R}^n$ such that $\operatorname{supp}\phi \subset U$ implies $\langle f, \phi\rangle = 0$. Let K_{1+}' be the set of all $f \in K_1'$ such that there exists $a \in \mathbb{R}$ with $\operatorname{supp} f \subset [a,\infty)$. Also let K_{2+}' be the set of all $f \in K_2'$ such that there exists a function $a : \mathbb{R} \to \mathbb{R}$ with $a(\tau) \to \infty$ as $\tau \to \infty$ and $\operatorname{supp} f \subset \{t \ge a(\tau)\}$. Note that K_{2+}' is a subspace of K_2'. *Weak* convergence* $f_k \to 0$ in K_n' means that $\langle f_k, \phi\rangle \to 0$ for every $\phi \in K_n$. One basis of *weak** neighborhoods of 0 in K_n' consists of all sets of the form $\{f \mid |\langle f, \phi_i\rangle| < \varepsilon; i = 1,\dots,m\}$, where $\varepsilon > 0$ and $\phi_1,\dots,\phi_m \in K_n$ are arbitrary.

The *partial derivative* of $f \in K_n'$ with respect to t_i is defined by $\langle \frac{\partial f}{\partial t_i}, \phi\rangle = -\langle f, \frac{\partial \phi}{\partial t_i}\rangle$ for $\phi \in K_n$. It follows that the differentiation operator $f \to \frac{\partial f}{\partial t_i}$ is (*weak**) continuous. In case $n = 1$, we denote a derivative by $\frac{df}{dt}$ or by an overdot $\dot{f}$; the kth

derivative will denoted $\frac{d^k f}{dt^k}$ or $f^{(k)}$. For any $f \in K_1'$, define the t_0-*translation* $\Delta_{t_0} f$ by $\langle \Delta_{t_0} f, \phi \rangle = \langle f, \phi_{-t_0} \rangle$, where $\phi_{-t_0}(t) = \phi(t + t_0)$. By a routine calculation, $\frac{d}{dt} \Delta_{t_0} f = \Delta_{t_0} \dot{f}$. Multiplication of $f \in K_1'$ by a C^∞ function γ is defined by $\langle f\gamma, \phi \rangle = \langle f, \gamma\phi \rangle$.

A Lebesgue measurable function $f : \mathbb{R}^n \to \mathbb{R}$ is *locally integrable* if $\int_A |f| < \infty$ for every bounded interval $A \subset \mathbb{R}^n$. Every locally integrable f determines a distribution according to

$$\langle f, \phi \rangle = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(t_1, \ldots, t_n)\phi(t_1, \ldots, t_n)dt_1 \cdots dt_n.$$

Note that the *unit step function*

$$\theta(t) = \begin{cases} 1, & t \geq 0, \\ 0, & t < 0, \end{cases}$$

may be considered a distribution in K_1'; the *unit impulse* $\delta \in K_1'$ is defined by $\langle \delta, \phi \rangle = \phi(0)$. Translations of θ and δ will be denoted by θ_{t_0} and δ_{t_0}, respectively. It is easily verified that

$$\theta_{t_0}(t) = \theta(t - t_0), \qquad \langle \delta_{t_0}, \phi \rangle = \phi(t_0). \tag{2.1}$$

If f is an absolutely continuous function, f and its classical derivative $\dot{f}$ are locally integrable. In this case, the classical and distributional derivatives of f coincide, since

$$\left\langle \dot{f}, \phi \right\rangle = \int_{-\infty}^{\infty} \dot{f}\,(t)\,\phi(t)dt = -\int_{-\infty}^{\infty} f\,(t)\,\dot{\phi}(t)dt = \left\langle f, \dot{\phi} \right\rangle.$$

Consider the spaces

$$BV = \{g : \mathbb{R} \to \mathbb{R} \mid \operatorname*{var}_t g(t) < \infty\},$$
$$NBV = \{g \in BV \mid g \text{ is left-continuous and } g(\infty) = 0\}$$

with norm $\|g\|_{NBV} = \operatorname{var}_t g(t)$. (Note that we are deviating slightly from the conventional definition of NBV, as in [9, p. 171].)

In [1] we also considered the space $DBV = \{\dot{g} \mid g \in NBV\}$ with norm $\|\dot{g}\|_{DBV} = \|g\|_{NBV}$. We showed in [1, p. 989] that DBV is isometrically isomorphic to the dual space of $(K_1, \|\cdot\|_\infty)$. Furthermore, for any $g \in NBV$ and $\phi \in K_1$, $\langle \dot{g}, \phi \rangle = \int_{-\infty}^{\infty} \phi(t)dg(t)$. We need to generalize these ideas to distributions on $\mathbb{R}^2$. The appropriate construction requires a preliminary result.

LEMMA 2.1. *Let $g : \mathbb{R}^2 \to \mathbb{R}$ be Lebesgue measurable and $g(t, \cdot) \in NBV$ for a.e. t. Then the function*

$$v(t, \tau) = \operatorname*{var}_{\eta \geq \tau} g(t, \eta)$$

is Lebesgue measurable on $\mathbb{R}^2$.

Proof. Enumerate the rationals $\{r_n\}$, and consider the partition $\pi_n = (r_{k_1}, \ldots, r_{k_n})$ of $\mathbb{R}$, where $\{k_1, \ldots, k_n\} = \{1, \ldots, n\}$ and

$$r_{k_1} < r_{k_2} < \cdots < r_{k_n}.$$

Let $A = \{t \mid g(t,\cdot) \in NBV\}$. For each n , $j = 1, \ldots, n-1$, and $(t,\tau) \in A \times (r_{k_j}, r_{k_{j+1}}]$, define

$$v_n(t,\tau) = |g(t, r_{k_n})| + \sum_{i=j+1}^{n} \left|g(t, r_{k_i}) - g\left(t, r_{k_{i+1}}\right)\right| + \left|g(t,\tau) - g\left(t, r_{k_{j+1}}\right)\right|.$$

For other (t,τ), set $v_n(t,\tau) = 0$. Each v_n is Lebesgue measurable. Let $\varepsilon > 0$, $(t,\tau) \in A \times \mathbb{R}$, and $\pi = (\tau_1, \ldots, \tau_p)$ be any partition of (τ, ∞) such that

$$q(t,\tau) = |g(t,\tau_p)| + \sum_{i=1}^{p-1} |g(t,\tau_i) - g(t,\tau_{i+1})| + |g(t,\tau) - g(t,\tau_1)| > v(t,\tau) - \frac{\varepsilon}{2}.$$

Since g is left-continuous, there exists $N < \infty$ such that π_n contains rationals $r_{k_{j_i}}$ with $\tau < r_{k_{j_1}} < \cdots < r_{k_{j_p}}$ such that

$$\left|g\left(t, r_{k_{j_i}}\right) - g(t,\tau_i)\right| < \frac{\varepsilon}{4p}$$

for every i. Thus

$$\begin{aligned}
v_n(t,\tau) &\geq \left|g\left(t, r_{k_{j_p}}\right)\right| + \sum_{i=1}^{p-1} \left|g\left(t, r_{k_{j_i}}\right) - g\left(t, r_{k_{j_{i+1}}}\right)\right| + \left|g(t,\tau) - g\left(t, r_{k_{j_1}}\right)\right| \\
&\geq |g(t,\tau_p)| - \left|g\left(t, r_{k_{jp}}\right) - g(t,\tau_p)\right| \\
&\quad + \sum_{i=1}^{p-1} |g(t,\tau_i) - g(t,\tau_{i+1})| - \sum_{i=1}^{p-1} \left|g\left(t, r_{k_{j_i}}\right) - g(t,\tau_i)\right| \\
&\quad - \sum_{i=1}^{p-1} \left|g\left(t, r_{k_{j_{i+1}}}\right) - g(t,\tau_{i+1})\right| \\
&\quad + |g(t,\tau) - g(t,\tau_1)| - \left|g\left(t, r_{k_{j_1}}\right) - g(t,\tau_1)\right| \\
&\geq q(t,\tau) - \frac{\varepsilon}{2} \\
&> v(t,\tau) - \varepsilon.
\end{aligned}$$

So $v_n \to v$ a.e., and v is Lebesgue measurable. □

In particular, if g satisfies the conditions of Lemma 2.1, the map $t \to \operatorname{var}_\tau g(t,\tau)$ is Lebesgue measurable. Hence, we may define UBV to be the set of functions $g : \mathbb{R}^2 \to \mathbb{R}$ satisfying the following properties:

(UBV1) g is Lebesgue measurable.

(UBV2) $g(t,\cdot) \in NBV$ for a.e. t.

(UBV3) $\operatorname{ess\,sup}_t \operatorname{var}_\tau g(t,\tau) < \infty$.

We refer to UBV as the functions of *uniformly bounded variation*. Let $\|g\|_{UBV} = \operatorname{ess\,sup}_t \operatorname{var}_\tau g(t,\tau)$. Each $g \in UBV$ is bounded, since

$$|g(t,\tau)| \leq \operatorname*{var}_\tau g(t,\tau) \leq \|g\|_{UBV}.$$

Thus $g \in K_2'$, and we may also define the set of partial derivatives

$$DUBV = \left\{ \frac{\partial g}{\partial \tau} \mid g \in UBV \right\}.$$

It is routine to verify that UBV and $DUBV$ are linear spaces and that $\|\cdot\|_{UBV}$ and

$$\left\|\frac{\partial g}{\partial \tau}\right\|_{DUBV} = \|g\|_{UBV}$$

are norms on UBV and $DUBV$.

For $1 \leq p < \infty$, let UL^p be the space of Lebesgue measurable functions $f : \mathbb{R}^2 \to \mathbb{R}$ satisfying

$$\operatorname*{ess\,sup}_t \int_{-\infty}^{\infty} |f(t,\tau)|^p \, d\tau < \infty.$$

(Lebesgue measurability of the map $t \to \int_{-\infty}^{\infty} |f(t,\tau)|^p d\tau$ follows from [9, Theorem 7.8].) UL^p is the set of *uniformly* L^p functions and has norm

$$\|f\|_{\infty p} = \operatorname*{ess\,sup}_t \left(\int_{-\infty}^{\infty} |f(t,\tau)|^p \, d\tau \right)^{\frac{1}{p}}.$$

We may consider $UL^p \subset K_2'$, since, from Holder's inequality,

$$\begin{aligned}
\int_{-a}^{a}\int_{-a}^{a} |f(t,\tau)| \, d\tau dt &\leq \int_{-a}^{a} \left(\operatorname*{ess\,sup}_t \int_{-a}^{a} |f(t,\tau)| \, d\tau \right) dt \\
&= 2a \operatorname*{ess\,sup}_t \int_{-a}^{a} |f(t,\tau)| \, d\tau \\
&\leq (2a)^{2-\frac{1}{p}} \operatorname*{ess\,sup}_t \left(\int_{-a}^{a} |f(t,\tau)|^p \, d\tau \right)^{\frac{1}{p}}.
\end{aligned}$$

For $p = \infty$, we define UL^∞ to be the same as L^∞ on $\mathbb{R}^2$.

Support constraints may be placed on the spaces above by setting $L^p_+ = L^p \cap K'_{1+}$, $UBV_+ = UBV \cap K'_{2+}$, $DUBV_+ = DUBV \cap K'_{2+}$, and $UL^p{}_+ = UL^p \cap K'_{2+}$.

THEOREM 2.2. (1) *Let $g \in UBV$ and $g_t = g(t,\cdot)$. Then*

$$\left\langle \frac{\partial g}{\partial \tau}, \psi \right\rangle = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(t,\tau) dg_t(\tau) dt$$

for every $\psi \in K_2$.

(2) *$UL^1 \subset DUBV$ with $\|f\|_{DUBV} = \|f\|_{\infty 1}$ for every $f \in UL^1$.*

(3) *$DUBV$ is the dual of $(K_2, \|\cdot\|_{1\infty})$.*

(4) *UBV and $DUBV$ are isometrically isomorphic.*

Proof. (1) Integration by parts yields

$$\left\langle \frac{\partial g}{\partial \tau}, \psi \right\rangle = -\left\langle g, \frac{\partial \psi}{\partial \tau} \right\rangle = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_t(\tau) \frac{\partial \psi(t,\tau)}{\partial \tau} d\tau dt = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(t,\tau) dg_t(\tau) dt$$

for every ψ.

(2) If $f \in UL^1$, there exists a Lebesgue measurable $A \subset \mathbb{R}$ such that $f(t,\cdot) \in L^1$ for every $t \in A$. Let $g(t,\tau) = \int_{-\infty}^{\tau} f(t,\eta) d\eta$ for $t \in A$. Then g is Lebesgue measurable, $g(t,\cdot)$ is absolutely continuous, and $f = \frac{\partial g}{\partial \tau}$. The result follows from

$$\|f\|_{DUBV} = \operatorname*{ess\,sup}_t \operatorname*{var}_\tau g(t,\tau) = \operatorname*{ess\,sup}_t \int_{-\infty}^{\infty} |f(t,\tau)| \, d\tau = \|f\|_{\infty 1} < \infty.$$

(3) We first prove that $DUBV$ is contained in the dual of K_2. Let g and g_τ be as in (1) and $f = \frac{\partial g}{\partial \tau}$. We must show that

$$\sup_{\|\psi\|_{1\infty}=1} |\langle f, \psi\rangle| = \|f\|_{DUBV}.$$

From (1),

$$\begin{aligned}
|\langle f, \psi\rangle| &= \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \psi(t,\tau) dg_t(\tau) dt \right| \\
&\le \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\psi(t,\tau)|\, |dg_t(\tau)|\, dt \\
&\le \int_{-\infty}^{\infty} \sup_{\tau} |\psi(t,\tau)| \operatorname*{var}_{\tau} g(t,\tau) dt \\
&\le (\operatorname*{ess\,sup}_{t} \operatorname*{var}_{\tau} g(t,\tau)) \int_{-\infty}^{\infty} \sup_{\tau} |\psi(t,\tau)|\, dt \\
&= \|f\|_{DUBV}\, \|\psi\|_{1\infty},
\end{aligned}$$

so

$$\sup_{\|\psi\|_{1\infty}=1} |\langle f, \psi\rangle| \le \|f\|_{DUBV}.$$

To establish the reverse inequality, observe that, for $\phi_1, \phi_2 \in K_1$, setting $\psi(t,\tau) = \phi_1(t)\phi_2(\tau)$ yields $\psi \in K_2$ and

$$\|\psi\|_{1\infty} = \int_{-\infty}^{\infty} \sup_{\tau} |\phi_1(t)\phi_2(\tau)|\, dt = \|\phi_1\|_1\, \|\phi_2\|_\infty.$$

Thus

$$\begin{aligned}
\sup_{\|\psi\|_{1\infty}=1} |\langle f, \psi\rangle| &= \sup_{\|\psi\|_{1\infty}=1} \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \psi(t,\tau) dg_t(\tau) dt \right| \\
&\ge \sup_{\|\phi_2\|_\infty=1} \sup_{\|\phi_1\|_1=1} \left| \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \phi_2(\tau) dg_t(\tau) \right) \phi_1(t) dt \right| \\
&= \sup_{\|\phi_2\|_\infty=1} \operatorname*{ess\,sup}_{t} \left| \int_{-\infty}^{\infty} \phi_2(\tau) dg_t(\tau) \right| \\
&= \operatorname*{ess\,sup}_{t} \sup_{\|\phi_2\|_\infty=1} |\langle \dot{g}_t, \phi_2\rangle| \\
&= \operatorname*{ess\,sup}_{t} \|\dot{g}_t\|_{DBV} \\
&= \|f\|_{DUBV}.
\end{aligned}$$

Next we show that $DUBV$ contains the dual of $(K_2, \|\cdot\|_{1\infty})$. Let f be any continuous linear functional on $(K_2, \|\cdot\|_{1\infty})$. ($f$ is also a distribution, since $\psi_k \to 0$ strongly in K_2 implies $\|\psi_k\|_{1\infty} \to 0$ and $\langle f, \psi_k\rangle \to 0$.) Let $\phi_1 \in K_1$. Then each $\phi_2 \in K_1$ determines $\psi \in K_2$ by $\psi(t,\tau) = \phi_1(t)\phi_2(\tau)$ (i.e., ψ is the "direct product" $\psi = \phi_1 \times \phi_2$). The map $\phi_2 \longmapsto \psi$ is continuous from $(K_1, \|\cdot\|_\infty)$ into $(K_2, \|\cdot\|_{1\infty})$, so $\phi_2 \longmapsto \langle f, \phi_1 \times \phi_2\rangle$ is a continuous linear functional on $(K_1, \|\cdot\|_\infty)$. Since K_1 is dense in C_0, there exists $G(\phi_1;\cdot) \in NBV$ such that

$$\langle f, \phi_1 \times \phi_2\rangle = \int_{-\infty}^{\infty} \phi_2(\tau) dG(\phi_1;\tau) \tag{2.2}$$

for every ϕ_2. It is routine to show that the operator $\phi_1 \mapsto G(\phi_1;\cdot)$ is linear. Continuity also holds, since

$$\begin{aligned}\sup_{\|\phi_1\|_1=1} \|G(\phi_1;\cdot)\|_{NBV} &= \sup_{\|\phi_1\|_1=1} \sup_{\|\phi_2\|_\infty=1} \left| \int_{-\infty}^{\infty} \phi_2(\tau) dG(\phi_1;\tau) \right| \\ &= \sup_{\|\phi_1\|_1=1} \sup_{\|\phi_2\|_\infty=1} |\langle f, \phi_1 \times \phi_2 \rangle| \\ &\leq \sup_{\|\psi\|_{1\infty}=1} |\langle f, \psi \rangle| \\ &< \infty.\end{aligned}$$

Therefore, $\phi_1 \mapsto G(\phi_1;\cdot)$ is a continuous linear operator from $(K_1, \|\cdot\|_1)$ into NBV. From [7, Theorem 2.3.1] and (2.2), there exists $g \in UBV$ such that

$$\langle f, \phi_1 \times \phi_2 \rangle = \int_{-\infty}^{\infty} \phi_2(\tau) d\left(\int_{-\infty}^{\infty} g(t,\tau)\phi_1(t)dt \right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \phi_1(t)\phi_2(\tau) dg_t(\tau) dt$$

for every $\phi_1, \phi_2 \in K_1$. Let $\Pi = \{\phi_1 \times \phi_2 \mid \phi_1, \phi_2 \in K_1\}$. By linearity,

$$\langle f, \psi \rangle = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(t,\tau) dg_t(\tau) dt \tag{2.3}$$

for every $\psi \in \operatorname{span}\Pi$. From [6, p. 65], $\operatorname{span}\Pi$ is strongly dense in K_2, so it is also dense relative to $\|\cdot\|_{1\infty}$. By continuity, (2.3) holds for all $\psi \in K_2$. From part (1), $f = \frac{\partial g}{\partial \tau} \in DUBV$.

(4) Note that the map $g \to \frac{\partial g}{\partial \tau}$ from UBV into $DUBV$ is defined to be linear, onto, and norm-preserving. It remains to show that the map is one-to-one. From part (1), if $\frac{\partial g}{\partial \tau} = 0$,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(t,\tau) dg_t(\tau) dt = 0$$

for all $\psi \in K_2$. Hence

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \phi_1(t)\phi_2(\tau) dg_t(\tau) dt = 0$$

for all $\phi_1, \phi_2 \in K_1$. Since

$$\left| \int_{-\infty}^{\infty} \phi_1(t) dg_t(\tau) \right| \leq \|\phi_1\|_\infty \operatorname*{var}_\tau g(t,\tau)$$

for a.e. t, the map $t \to \int_{-\infty}^{\infty} \phi_1(t) dg_t(\tau)$ may be viewed as a distribution $T(\phi_1)$. But $\langle T(\phi_1), \phi_2 \rangle = 0$ for every ϕ_2, so $T(\phi_1) = 0$ and

$$\int_{-\infty}^{\infty} \phi_1(t) dg_t(\tau) = 0$$

a.e. for every ϕ_1. Since NBV is the dual of $(K_1, \|\cdot\|_\infty)$, $g(t,\cdot) = 0$ for a.e. t and $g = 0$ a.e. □

3. Families and integral operators. In order to generalize the concept of an integral operator as in (1.1), we must study collections of distributions indexed by a real parameter. Let $\{f_t \mid t \in \mathbb{R}\}$ be a collection of distributions in K_1'. Suppose that, for each $a < \infty$, there exist an integer $p \geq 0$ and an L^1 function $M : [-a, a] \to \mathbb{R}$ such that

$$|\langle f_t, \phi \rangle| \leq M(t) \|\phi\|_{C^p}$$

for every $\phi \in K_1$ with $\operatorname{supp} \phi \subset [-a, a]$. Then we say that $\{f_t\}$ is an L^1 *family* of distributions on $\mathbb{R}$.

THEOREM 3.1. (1) *If $\{f_t\}$ is an L^1 family, then the map $\psi \to \int_{-\infty}^{\infty} \langle f_t, \psi(t, \cdot) \rangle dt$ is a distribution in K_2'.*

(2) *If $f = \{f_t\}$ is an L^1 family, then so is $\{\dot{f}_t\}$, and $\frac{\partial f}{\partial \tau} = \{\dot{f}_t\}$.*

(3) *If $f : \mathbb{R}^2 \to \mathbb{R}$ is locally integrable, then $\{f(t, \cdot)\}$ is an L^1 family.*

(4) *If $f \in UBV$, then $\{f(t, \cdot)\}$ is an L^1 family.*

Proof. (1) Let $\psi \in K_2$. Then $\|\psi(t, \cdot)\|_{C^p} \leq \|\psi\|_{C^p}$ for every t, so

$$\int_{-a}^{a} |\langle f_t, \psi(t, \cdot) \rangle| \, dt \leq \int_{-a}^{a} M(t) \|\psi(t, \cdot)\|_{C^p} \, dt \leq \int_{-a}^{a} M(t) dt \, \|\psi\|_{C^p} .$$

Thus the map $\psi \to \int_{-\infty}^{\infty} \langle f_t, \psi(t, \cdot) \rangle dt$ is well defined, linear, and, from [4, p. 34], continuous.

(2) There exist M, p such that, for any $\phi \in K_1$ with $\operatorname{supp} \phi \subset [-a, a]$,

$$\left| \left\langle \dot{f}_t, \phi \right\rangle \right| = \left| \left\langle f_t, \dot{\phi} \right\rangle \right| \leq M(t) \left\| \dot{\phi} \right\|_{C^p} \leq M(t) \|\phi\|_{C^{p+1}} .$$

For any $\psi \in K_2$,

$$\begin{aligned} \left\langle \frac{\partial \{f_t\}}{\partial \tau}, \psi \right\rangle &= - \left\langle \{f_t\}, \frac{\partial \psi(t, \tau)}{\partial \tau} \right\rangle \\ &= - \int_{-a}^{a} \left\langle f_t, \frac{\partial \psi(t, \tau)}{\partial \tau} \right\rangle dt \\ &= \int_{-a}^{a} \left\langle \dot{f}_t, \psi(t, \cdot) \right\rangle dt \\ &= \left\langle \left\{ \dot{f}_t \right\}, \psi \right\rangle . \end{aligned}$$

(3) Let $M(t) = \int_{-a}^{a} |f(t, \tau)| d\tau$. By Fubini's theorem, M is L^1 on $[-a, a]$. For any $\phi \in K_1$,

$$|\langle f(t, \cdot), \phi \rangle| = \left| \int_{-a}^{a} f(t, \tau) \phi(\tau) d\tau \right| \leq M(t) \|\phi\|_{C^0} .$$

(4) This follows immediately from part (3) and the fact that f is bounded on $\mathbb{R}^2$. □

A slight modification of the argument used in Theorem 3.1(1) shows that each L^1 family $\{f_\tau\}$ also determines a distribution according to $\psi \to \int_{-\infty}^{\infty} \langle f_\tau, \psi(\cdot, \tau) \rangle d\tau$. We will rely on the notation $\{f_t\}$ and $\{f_\tau\}$ to distinguish these two cases.

Consider a collection of distributions $\{f_\tau\}$ in K_1' with $\tau \to \langle f_\tau, \phi \rangle$ continuous for every $\phi \in K_1$. Then we say that $\{f_\tau\}$ is a C^0 *family.*

THEOREM 3.2. *Every C^0 family is an L^1 family.*

Proof. Let $\{f_\tau\}$ be a C^0 family. According to [4, p. 34], each f_τ has finite order on each bounded interval $[-a,a]$; i.e., for every $\tau \in [-a,a]$ there exist integers M_τ, $p_\tau < \infty$ such that

$$|\langle f_\tau, \phi\rangle| \leq M_\tau \|\phi\|_{C^{p_\tau}} \tag{3.1}$$

for every $\phi \in K_1$ with $\operatorname{supp}\phi \subset [-a,a]$. Suppose each M_τ and p_τ are chosen to minimize $q_\tau = \max\{M_\tau, p_\tau\}$. If the set $\{q_\tau \mid |\tau| \leq a\}$ is unbounded, there exist $\eta_k, \eta \in [-a,a]$ such that $\eta_k \to \eta$ and $q_{\eta_k} \to \infty$. On the other hand, since $\langle f_\tau, \phi\rangle$ is continuous, $\langle f_{\eta_k}, \phi\rangle \to \langle f_\eta, \phi\rangle$ for every ϕ. From [4, p. 57], there exist M, p such that

$$\left|\left\langle f_{\eta_k}, \phi\right\rangle\right| \leq M \|\phi\|_{C^p}$$

for every k, yielding a contradiction. Hence, $\{q_\tau\}$ is bounded, and there exist M, p such that

$$|\langle f_\tau, \phi\rangle| \leq M \|\phi\|_{C^p}$$

for every τ, ϕ. □

In addition to continuity, we might also consider families $\{f_\tau\}$ which are differentiable in τ. Define the *weak* derivative* $\frac{\partial f_\tau}{\partial \tau}|_{\tau_0} \in K_2'$ of $\{f_\tau\}$ at τ_0 according to

$$\left\langle \frac{\partial f_\tau}{\partial \tau}|_{\tau_0}, \phi \right\rangle = \frac{d}{d\tau} \langle f_\tau, \phi\rangle\,|_{\tau_0} = \lim_{\tau_n \to \tau_0} \left\langle \frac{f_{\tau_n} - f_{\tau_0}}{\tau_n - \tau_0}, \phi \right\rangle, \tag{3.2}$$

whenever the limit exists for every $\phi \in K_1$. According to [3, p. 368], (3.2) determines a distribution in K_1' for each τ_0. If $\{\frac{\partial f_\tau}{\partial \tau}|_{\tau_0} \mid \tau_0 \in \mathbb{R}\}$ is a C^0 family, we denote it by $\frac{\partial f_\tau}{\partial \tau}$ and say that $\{f_\tau\}$ is a C^1 *family.* Continuing in this way, if $\{\frac{\partial f_\tau}{\partial \tau}\}$ is a C^{p-1} family, $\{f_\tau\}$ is a C^p *family.* Applying (3.2), $\{f_\tau\}$ is a C^p family if and only if $\tau \to \langle f_\tau, \phi\rangle$ is a C^p function for each ϕ. We may interpret the latter statement as a definition for $p = \infty$. Since $\langle \dot{f}_\tau, \phi\rangle = -\langle f_\tau, \dot{\phi}\rangle$, $\{\dot{f}_\tau\}$ is a C^p family whenever $\{f_\tau\}$ is a C^p family.

Next we relate two notions of differentiation for C^1 families.

THEOREM 3.3. *Suppose $g = \{g_\tau\}$ is a C^1 family. Then $\frac{\partial g}{\partial \tau} = \{\frac{\partial g_\tau}{\partial \tau}\}$.*

Proof. Suppose

$$\left\langle \frac{g_\tau - g_{\tau_0}}{\tau - \tau_0}, \psi(\cdot,\tau) \right\rangle \not\to \left\langle \frac{\partial g_\tau}{\partial \tau}|_{\tau_0}, \psi(\cdot,\tau_0) \right\rangle \tag{3.3}$$

as $\tau \to \tau_0$. Then there exist $\tau_n \to \tau_0$ and $\varepsilon > 0$ such that

$$\sup_j \left| \left\langle \frac{g_{\tau_n} - g_{\tau_0}}{\tau_n - \tau_0} - \frac{\partial g_\tau}{\partial \tau}|_{\tau_n}, \psi(\cdot,\tau_j) \right\rangle \right| \geq \left| \left\langle \frac{g_{\tau_n} - g_{\tau_0}}{\tau_n - \tau_0} - \frac{\partial g_\tau}{\partial \tau}|_{\tau_n}, \psi(\cdot,\tau_n) \right\rangle \right| > \varepsilon \tag{3.4}$$

for every n. But $\psi(\cdot,\tau_n) \to \psi(\cdot,\tau_0)$ in K_1, so, from [4, p. 31], $\{\psi(\cdot,\tau_n)\} \subset K_1$ is a bounded set. Hence, from [4, p. 56], the left side of (3.4) must converge to 0. This yields a contradiction, so we have convergence in (3.3). Thus

$$\begin{aligned}
\frac{d}{d\tau} \langle g_\tau, \psi(\cdot,\tau)\rangle\,|_{\tau_0} &= \lim_{\tau \to \tau_0} \frac{\langle g_\tau, \psi(\cdot,\tau)\rangle - \langle g_{\tau_0}, \psi(\cdot,\tau_0)\rangle}{\tau - \tau_0} \\
&= \lim_{\tau \to \tau_0} \left\langle \frac{g_\tau - g_{\tau_0}}{\tau - \tau_0}, \psi(\cdot,\tau) \right\rangle + \lim_{\tau \to \tau_0} \left\langle g_{\tau_0}, \frac{\psi(\cdot,\tau) - \psi(\cdot,\tau_0)}{\tau - \tau_0} \right\rangle \\
&= \left\langle \frac{\partial g_\tau}{\partial \tau}|_{\tau_0}, \psi(\cdot,\tau_0) \right\rangle + \left\langle g_{\tau_0}, \frac{\partial \psi}{\partial \tau}|_{\tau_0} \right\rangle,
\end{aligned}$$

and, for every $\psi \in K_2$,

$$\begin{aligned}\left\langle \frac{\partial g}{\partial \tau}, \psi \right\rangle &= -\left\langle g, \frac{\partial \psi}{\partial \tau} \right\rangle \\ &= -\int_{-\infty}^{\infty} \left\langle g_\tau, \frac{\partial \psi}{\partial \tau} \right\rangle d\tau \\ &= -\int_{-\infty}^{\infty} \left(\frac{d}{d\tau} \langle g_\tau, \psi(\cdot,\tau)\rangle - \left\langle \frac{\partial g_\tau}{\partial \tau}, \psi(\cdot,\tau) \right\rangle \right) d\tau \\ &= -\lim_{\tau\to\infty} \langle g_\tau, \psi(\cdot,\tau)\rangle + \lim_{\tau\to-\infty} \langle g_\tau, \psi(\cdot,\tau)\rangle + \int_{-\infty}^{\infty} \left\langle \frac{\partial g_\tau}{\partial \tau}, \psi(\cdot,\tau) \right\rangle d\tau \\ &= \int_{-\infty}^{\infty} \left\langle \frac{\partial g_\tau}{\partial \tau}, \psi(\cdot,\tau) \right\rangle d\tau.\end{aligned}$$

Hence $\frac{\partial g}{\partial \tau} = \{\frac{\partial g_\tau}{\partial \tau}\}$. □

Note that

$$\left\langle \frac{\partial \theta_\tau}{\partial \tau}, \phi \right\rangle = \frac{\partial}{\partial \tau} \int_{\tau}^{\infty} \phi(\eta) d\eta = -\phi(\tau) = \langle -\delta_\tau, \phi \rangle$$

for every $\phi \in K_1$, where θ_τ is a translation of the unit step as in (2.1). In view of Theorem 3.3, $\frac{\partial}{\partial \tau}\{\theta_\tau\} = \{\frac{\partial \theta_\tau}{\partial \tau}\} = -\{\delta_\tau\}$. By a similar calculation, $\frac{\partial}{\partial \tau}\{\delta_\tau^{(n-1)}\} = -\{\delta_\tau^{(n)}\}$.

Any C^∞ family $\{h_\tau\}$ belonging to K'_{2+} determines a linear operator on K'_{1+} in the following way. For each $\phi \in K_1$, let $\xi(\tau) = \langle h_\tau, \phi \rangle$. Then $\xi(\tau)$ is C^∞ with $\xi(\tau) = 0$ for large τ. Suppose $u \in K'_{1+}$ with $\operatorname{supp} u \subset [a, \infty]$, and let $\bar{\xi} \in K_1$ with $\bar{\xi}(\tau) = \xi(\tau)$ for $\tau \geq a$. It is easy to show that $y(\phi) = \langle u, \bar{\xi} \rangle$ is independent of the choice of $\bar{\xi}$. Indeed, let $\bar{\xi}_1$ and $\bar{\xi}_2$ be two such functions. Then $\operatorname{supp}(\bar{\xi}_1 - \bar{\xi}_2) \subset (-\infty, a]$, and $\langle u, \bar{\xi}_1 \rangle - \langle u, \bar{\xi}_2 \rangle = \langle u, \bar{\xi}_1 - \bar{\xi}_2 \rangle = 0$.

THEOREM 3.4. *The map $u \longmapsto y$ defines a linear operator $T : K'_{1+} \to K'_{1+}$, where $T(\delta_{t_0}) = h_{t_0}$ for every t_0.*

Proof. Linearity is obvious. Since $\{h_\tau\} \in K'_{2+}$, there exists $b \in \mathbb{R}$ such that $\operatorname{supp} \phi \subset (-\infty, b]$ implies $\operatorname{supp} \xi \subset (-\infty, a)$. For any such ϕ, $y(\phi) = 0$, so $\operatorname{supp} y \subset [b, \infty)$. We must show that $y \in K'_1$.

The topological space K'_1 is not first-countable, so we must consider nets $\{\phi_\lambda\}$ in K_1. Suppose $\phi_\lambda \to 0$. Then there exists $b < \infty$ such that $\operatorname{supp} \phi_\lambda \subset [-b, b]$ for every λ. Let $\xi_\lambda(\tau) = \langle h_\tau, \phi_\lambda \rangle$. Since $\{h_\tau\} \in K'_{1+}$, there exists $c < \infty$ such that $\operatorname{supp} \xi_\lambda(\tau) \subset (-\infty, c]$ for every λ. Arguing as in the proof of Theorem 3.2, there exist $M_n, p_n < \infty$ such that

$$\left| \xi^{(n)}(\tau) \right| \leq M_n \|\phi\|_{C^{p_n}}$$

for every $\tau \in [a, c]$ and $\phi \in K_1$ with $\operatorname{supp} \phi \in [-b, b]$. Hence, $\xi_\lambda^{(n)} \to 0$ uniformly on $[a, c]$ for every n. $\bar{\xi}_\lambda \in K_1$ can be chosen so that $\bar{\xi}_\lambda(\tau) = \xi_\lambda(\tau)$ on $[a, \infty)$ and each $\bar{\xi}_\lambda^{(n)} \to 0$ uniformly on $[a, c]$. Thus $\bar{\xi}_\lambda \to 0$ strongly in K_1, and $y(\phi_\lambda) = \langle u, \bar{\xi}_\lambda \rangle \to 0$. Finally, for $u = \delta_{t_0}$ and any $\phi \in K_1$, $\langle T(\delta_{t_0}), \phi \rangle = \bar{\xi}(t_0) = \xi(t_0) = \langle h_{t_0}, \phi \rangle$. □

Suppose $h_\tau = h(\cdot, \tau)$. Then, under mild assumptions,

$$\langle T(u), \phi \rangle = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} h(t,\tau)\phi(t)dt \right) u(\tau)d\tau = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} h(t,\tau)u(\tau)d\tau \right) \phi(t)dt,$$

so the operator determined by Theorem 3.4 is "classical." For this reason, we call every operator of the type described by Theorem 3.4 a *generalized integral operator*. Suppose h_τ is the τ-translation of a fixed $h \in K'_{1+}$. Then $\{h_\tau\}$ is a C^∞ family and $\{h_\tau\} \in K'_{2+}$, so $\{h_\tau\}$ determines a generalized integral operator according to $\langle T(u), \phi\rangle = \langle u, \xi\rangle$, where $\xi(\tau) = \langle h_\tau, \phi\rangle = \langle h, \phi_{-\tau}\rangle$. From [3, p. 103], constructing $T(u)$ in this way yields the convolution operator $u \to h * u$. This sets up the time-invariant analysis we carried out in [1].

We next address the issue of continuity of generalized integral operators.

THEOREM 3.5. *Let $T : K'_{1+} \to K'_{1+}$ be any continuous linear operator, and suppose $\{T(\delta_\tau)\} \in K'_{2+}$ is a C^∞ family. Then T is a generalized integral operator.*

Proof. Let $\phi \in K_1$, $u \in K'_{1+}$, $\operatorname{supp} u \subset [a, \infty)$, $\xi(\tau) = \langle T(\delta_\tau), \phi\rangle$, and $\bar{\xi} \in K_1$ with $\bar{\xi}(\tau) = \xi(\tau)$ for $\tau \geq a$. We must show that $\langle T(u), \phi\rangle = \langle u, \bar{\xi}\rangle$. First, set $u = \delta_{t_0}$. Then

$$\langle T(\delta_{t_0}), \phi\rangle = \xi(t_0) = \bar{\xi}(t_0) = \left\langle \delta_{t_0}, \bar{\xi}\right\rangle.$$

For arbitrary $u \in K'_{1+}$, [1, Lemma 2.2] shows that there exist $t_{ik} \geq a$, $\beta_{ik} \in \mathbb{R}$, and integers $n_k > 0$ such that

$$u_k = \sum_{i=1}^{n_k} \beta_{ik} \delta_{t_{ik}} \to u,$$

where the limit is *weak**. For every k,

$$\langle T(u_k), \phi\rangle = \sum_{i=1}^{n_k} \beta_{ik} \langle T(\delta_{t_{ik}}), \phi\rangle = \sum_{i=1}^{n_k} \beta_{ik} \left\langle \delta_{t_{ik}}, \bar{\xi}\right\rangle = \left\langle u_k, \bar{\xi}\right\rangle.$$

Taking the limit yields

$$\langle T(u), \phi\rangle = \left\langle u, \bar{\xi}\right\rangle. \qquad \square$$

The next result establishes conditions under which generalized integral operators are continuous. In particular, it shows that the converse to Theorem 3.5 is false.

THEOREM 3.6. *Let $T : K'_{1+} \to K'_{1+}$ be a generalized integral operator.*

(1) *T is continuous if and only if there exists a function $b : \mathbb{R} \to \mathbb{R}$ such that $b(\tau) \to \infty$ as $\tau \to -\infty$ and $\operatorname{supp}\{T(\delta_\tau)\} \subset \{|t| \geq b(\tau)\}$.*

(2) *T is continuous on $\{u \in K'_{1+} \mid \operatorname{supp} u \subset [a, \infty)\}$ for every $a > -\infty$.*

Proof. (1) (Sufficient) Consider any basic neighborhood $Y = \{y \in K'_{1+} \mid |\langle y, \phi_i\rangle| < \varepsilon; i = 1, \ldots, n\}$ of 0 in K'_{1+}, and let $\xi_i(\tau) = \langle T(\delta_\tau), \phi_i\rangle$. Since T is a generalized integral operator, each ξ_i is C^∞. $\{h_\tau\} \in K'_{1+}$ guarantees that $\operatorname{supp} \xi_i$ is right-bounded. The condition $\operatorname{supp}\{T(\delta_\tau)\} \subset \{|t| \geq b(\tau)\}$ guarantees that $\operatorname{supp} \xi_i$ is left-bounded. Hence, $\xi_i \in K_1$, and $\langle T(u), \phi_i\rangle = \langle u, \xi_i\rangle$. It follows that the inverse image

$$\begin{aligned} T^{-1}(Y) &= \{u \in K'_{1+} \mid |\langle T(u), \phi_i\rangle| < \varepsilon; i = 1, \ldots, n\} \\ &= \{u \in K'_{1+} \mid |\langle u, \xi_i\rangle| < \varepsilon; i = 1, \ldots, n\} \end{aligned}$$

is open. Since $\varepsilon, \phi_1, \ldots, \phi_n$ are arbitrary, T is continuous.

(Necessary) Suppose no such function b exists. Then there exists a sequence $\tau_k \to -\infty$ and $c < \infty$ such that $\operatorname{supp} T(\delta_{\tau_k}) \cap [-c, c]$ is nonempty for every k. Hence, there exist $\phi_k \in K_1$ such that $\operatorname{supp} \phi_k \subset [-c, c]$ and $\beta_k = \langle T(\delta_{\tau_k}), \phi_k\rangle \neq 0$ for each k. Let

$$\xi_k = \frac{1}{k \left\|\phi_k\right\|_{C^k}} \phi_k.$$

Then, for any integers $p \geq 0$ and $k \geq p$,

$$\|\xi_k\|_{C^p} = \frac{1}{k\,\|\phi_k\|_{C^k}}\,\|\phi_k\|_{C^p} \leq \frac{1}{k},$$

so $\xi_k \to 0$. From [4, p. 31], $\{\xi_k\}$ is a bounded set. Let

$$u_k = \frac{k\,\|\phi_k\|_{C^k}}{\beta_k}\delta_{\tau_k}.$$

Then $u_k \to 0$. On the other hand,

$$\begin{aligned}\sup_m \langle T(u_k), \xi_m\rangle &\geq \langle T(u_k), \xi_k\rangle \\ &= \left\langle \frac{k\,\|\phi_k\|_{C^k}}{\beta_k} T(\delta_{\tau_k}), \xi_k \right\rangle \\ &= \frac{1}{\beta_k}\,\langle T(\delta_{\tau_k}), k\,\|\phi_k\|_{C^k}\,\xi_k\rangle \\ &= \frac{1}{\beta_k}\,\langle T(\delta_{\tau_k}), \phi_k\rangle = 1.\end{aligned}$$

From [4, p. 56], $T(u_k) \not\to 0$, which is a contradiction.

(2) Let $u_\lambda \to 0$ be a net with $\operatorname{supp} u_\lambda \subset [a, \infty)$, $\phi \in K_1$, and $\xi(\tau) = \langle T(\delta_\tau), \phi\rangle$. Select $\bar{\xi} \in K_1$ with $\bar{\xi}(\tau) = \xi(\tau)$ for $\tau \geq a$. This gives $\langle T(u_\lambda), \phi\rangle = \langle u_\lambda, \bar{\xi}\rangle \to 0$, so $T(u_\lambda) \to 0$. □

In view of Theorem 3.6, restricting attention to continuous linear operators on K'_{1+} would be inadequate for developing a sufficiently comprehensive theory of stable linear systems. For example, even the "integrator system" $T(\delta_\tau) = \theta_\tau$ is discontinuous. On the other hand, the class of all generalized integral operators on K'_{1+} contains the full range of operators normally considered in linear system theory, so we will adopt these as our space of systems.

We end this section with a result relating impulse response and step response.

THEOREM 3.7. *Let $T : K'_{1+} \to K'_{1+}$ be a generalized integral operator. Then $\{T(\theta_\tau)\}$ is a C^∞ family and $T(\delta_\tau) = -\frac{\partial T(\theta_\tau)}{\partial \tau}$.*

Proof. From Theorem 3.6(2), T is continuous on $\{u \in K'_{1+} \mid \operatorname{supp} u \subset [\tau_0 - 1, \infty)\}$ for any τ_0. Hence, for any $\phi \in K_1$,

$$\begin{aligned}\frac{\langle T(\theta_\tau), \phi\rangle - \langle T(\theta_{\tau_0}), \phi\rangle}{\tau - \tau_0} &= \left\langle T\left(\frac{\theta_\tau - \theta_{\tau_0}}{\tau - \tau_0}\right), \phi \right\rangle \to \langle -T(\delta_{\tau_0}), \phi\rangle\,, \\ \frac{\left\langle T(\delta_\tau^{(n)}), \phi\right\rangle - \left\langle T(\delta_{\tau_0}^{(n)}), \phi\right\rangle}{\tau - \tau_0} &= \left\langle T\left(\frac{\delta_\tau^{(n)} - \delta_{\tau_0}^{(n)}}{\tau - \tau_0}\right), \phi \right\rangle \to \left\langle -T(\delta_{\tau_0}^{(n+1)}), \phi\right\rangle\end{aligned}$$

as $\tau \to \tau_0$ for any n, so $\{T(\theta_\tau)\}$ is C^∞. The first derivative is given by

$$\left\langle \frac{\partial T(\theta_\tau)}{\partial \tau}, \phi \right\rangle = \frac{\partial}{\partial \tau}\,\langle T(\theta_\tau), \phi\rangle = \langle -T(\delta_\tau), \phi\rangle\,. \qquad \square$$

4. Extension of normed linear spaces. In [1] we considered the problem of imbedding a normed linear space Y into a Hausdorff topological vector space X and extending the norm to a maximal linear subspace of X. For example, we showed that

L^1 can be imbedded in K_1' and $\|\cdot\|_1$ can be extended to all of DBV. In particular, the extended L^1 norm applied to the unit impulse evaluates to $\|\delta\|_1^e = 1$. We will again need these results to construct our time-varying theory.

Let $\mathfrak{T}$ be the topology on X, $\|\cdot\|$ the norm on $Y \subset X$, and $B(y,r) \subset Y$ the closed norm-ball about $y \in Y$ with radius r. We make the following assumptions on the 4-tuple $(X, \mathfrak{T}, Y, \|\cdot\|)$:

T1. For every nonempty $U \in \mathfrak{T}$, $U \cap Y$ is nonempty.

T2. For every $U \in \mathfrak{T}$ and every $y \in U \cap Y$, there exists $r > 0$ such that $B(y,r) \subset U$.

T3. There exists $U \in \mathfrak{T}$ such that $U \cap Y = Y - B(0,1)$.

Condition T1 says that Y is dense in X. T2 requires that the norm topology on Y is at least as strong as the topology induced on Y by X. T3 says that $B(0,1)$ in Y is closed relative to X. Thus T2 and T3 give bounds on the topology induced on Y by X. Under assumptions T1–T3, there exists a natural extension $\|\cdot\|^e$ of $\|\cdot\|$ to a subspace $Y_e \supset Y$ of X. In particular, for any $y \in Y_e$, $\|y\|^e$ is equal to the minimum value of $\lim \|y_\lambda\|$ over all $\mathfrak{T}$-approximating nets $y_\lambda \to y$, $y_\lambda \in Y$. (See [1, section 3] for details.)

As mentioned in section 1, $h \in UL^1$ is the classical condition for BIBO stability. Therefore it makes sense to examine the extension UL_e^1 of $\|\cdot\|_{\infty 1}$ in K_2' and check whether UL_e^1 actually characterizes BIBO stability for generalized integral operators. First we must establish whether K_2' and UL^1 satisfy T1–T3.

LEMMA 4.1. (1) *If Y_1 is dense in Y relative to $\|\cdot\|$, then the 4-tuple $(X, \mathfrak{T}, Y_1, \|\cdot\|)$ satisfies* T1–T3 *and $Y_{1e} = Y_e$.*

(2) *Let $Y \subset X_1 \subset X$ and $\mathfrak{T}_1$ be the relative topology on X_1 induced by $\mathfrak{T}$. Then the 4-tuple $(X_1, \mathfrak{T}_1, Y, \|\cdot\|)$ satisfies* T1–T3 *and the corresponding extension of Y is $Y_e \cap X_1$.*

Proof. (1) From T1 and T2, Y_1 is dense in X relative to $\mathfrak{T}$, so T1 holds for Y_1. If $U \in \mathfrak{T}$ and $y \in U \cap Y_1$, then $y \in U \cap Y$, so there exists $r > 0$ such that $B(y,r) \subset U$, where $B(y,r)$ is a norm-ball in Y. The corresponding ball in Y_1 is $B(y,r) \cap Y_1 \subset U$, so T2 holds. Finally, if U satisfies T3 relative to Y, then U also satisfies T3 relative to Y_1, since

$$U \cap Y_1 = (U \cap Y) \cap Y_1 = (Y - B(0,1)) \cap Y_1 = Y_1 - (B(0,1) \cap Y_1).$$

Finally, we must show that $\|\cdot\|^e$ satisfies [1, Proposition 3.1(1)–(3)] using Y_1 in place of Y. To prove (1), note that, since $\|\cdot\|$ and $\|\cdot\|^e$ coincide on Y, they must coincide on Y_1. Condition (2) holds, since it does not involve Y. To prove (3), let $x \in X$ with $\|x\|^e < \infty$, $\varepsilon > 0$, and let U be a $\mathfrak{T}$-neighborhood of x. Then (3) applied to Y guarantees that there exists $y \in U \cap Y$ such that

$$\|y\| < \|x\|^e + \frac{\varepsilon}{2}.$$

Density of Y_1 in Y relative to $\|\cdot\|$ and T2 imply that there exists $y_1 \in U \cap Y_1$ such that $\|y_1 - y\| < \frac{\varepsilon}{2}$. Then

$$\|y_1\| < \|y\| + \frac{\varepsilon}{2} < \|x\|^e + \varepsilon.$$

(2) Restricting $\mathfrak{T}$ to X_1, T1–T3 are obvious. Suppose Y_e is the extension of Y using X, and $\|\cdot\|^e$ is the corresponding norm. Let $\|\cdot\|^f$ be the restriction of $\|x\|^e$ to X_1. We must show that $\|\cdot\|^f$ satisfies [1, Proposition 3.1(1)–(3)], using X_1. To prove (1) and (2), note that $\|x\|^e$ and $\|x\|^f$ coincide on X_1; hence, $\|y\|^f = \|y\|$ for

$y \in Y$ and $\|\cdot\|^f$ is lower semicontinuous. To establish (3), let $U \in \mathfrak{T}$ $x \in U \cap X_1$ with $\|x\|^f < \infty$, and $\varepsilon > 0$. Then $\|x\|^e < \infty$ and there exists $y \in U \cap Y$ such that

$$\|y\| < \|x\|^e + \varepsilon = \|x\|^f + \varepsilon.$$

But $Y \subset X_1$, so $y \in (U \cap X_1) \cap Y$. □

THEOREM 4.2. *Let $X = K_2'$ and $Y = UL^1$. Then* T1–T3 *are satisfied.*

Proof. First we note that $K_2 \subset UL^1$. From [4, p. 118], K_2 is dense in K_2', and T1 follows. If T2 holds for $y = 0$, then it holds for all y, since $B(0,r) \subset U$ implies $B(y,r) \subset y+U$. Thus it suffices to prove that, for every $n, \varepsilon > 0$, and $\psi_1, \ldots, \psi_n \in K_2$, there exists $r > 0$ such that $B(0,r) \subset U$, where

$$U = \{f \in K_2' \mid |\langle f, \psi_i\rangle| < \varepsilon; i = 1, \ldots, n\}.$$

Let

$$r < \varepsilon \min\left\{\frac{1}{\|\psi_i\|_{1\infty}}\right\}$$

and $f \in B(0,r)$. From Theorem 2.2,

$$|\langle f, \psi_i\rangle| \le \|f\|_{DUBV} \|\psi_i\|_{1\infty} = \|f\|_{\infty 1} \|\psi_i\|_{1\infty} \le r \max\{\|\psi_j\|_{1\infty}\} < \varepsilon$$

for every i, so $f \in U$.

To prove T3, let $y \in UL^1$ with $\|y\|_{\infty 1} > 1$, and choose $\varepsilon < \frac{1}{2}(\|y\|_{\infty 1} - 1)$. From Theorem 2.2, we may select $\psi \in K_2$ with $\|\psi\|_{1\infty} = 1$ such that $|\langle y, \psi\rangle| > \|y\|_{\infty 1} - \varepsilon$. Let

$$U = \{x \in K_2' \mid |\langle x - y, \psi\rangle| < \varepsilon\}.$$

Then $y \in U \in \mathfrak{T}$, and $f \in U \cap Y$ implies

$$\|f\|_{\infty 1} \ge |\langle f, \psi\rangle| \ge |\langle y, \psi\rangle| - |\langle f - y, \psi\rangle| > \|y\|_{\infty 1} - 2\varepsilon > 1,$$

so $U \cap Y \subset Y - B(0,1)$. □

We are now in a position to characterize the extension of UL^1 into K_2'.

THEOREM 4.3. $UL_e^1 = DUBV$.

Proof. Let

$$\|f\|^e = \begin{cases} \|f\|_{DUBV}, & f \in DUBV, \\ \infty, & f \in K_2' - DUBV. \end{cases}$$

We must verify that $\|\cdot\|^e$ satisfies [1, Lemma 3.1(1)–(3)] relative to $\|\cdot\|_{\infty 1}$. Condition (1) says that $\|\cdot\|_{\infty 1}$ and $\|\cdot\|_{DUBV}$ coincide on UL^1. This was established in Theorem 2.2.

Condition (2) requires that $\|\cdot\|^e$ be lower semicontinuous on K_2'. Equivalently, we must show that the set $\Sigma_M = \{f \in K_2' \mid \|f\|^e > M\}$ is open for each M. Suppose $\|f\|^e > M$. From Theorem 2.2(3), there exists $\psi \in K_2$ such that $\|\psi\|_{1\infty} = 1$ and $|\langle f, \psi\rangle| > M$. Let $U = \{g \in K_2' \mid |\langle g, \psi\rangle| < |\langle f, \psi\rangle| - M\}$. $f + U$ is open in K_2', and

$$\|f + g\|^e \ge |\langle f + g, \psi\rangle| \ge |\langle f, \psi\rangle| - |\langle g, \psi\rangle| > M$$

for every $g \in U$. Hence, Σ_M is open.

Finally, condition (3) says that, for any $f \in DUBV$, $\varepsilon > 0$, and neighborhood U of f, there exists $y \in U \cap UL^1$ such that $\|y\|_{\infty 1} < \|f\|^e + \varepsilon$. We accomplish this by constructing a sequence $f_n \to f$ with $f_n \in UL^1$ and $\|f_n\|_{\infty 1} \leq \|f\|_{DUBV}$. Then, for large n, $f_n \in U$, and $y = f_n$ satisfies the conditions.

Our construction of f_n proceeds as follows. Let $\phi_1, \phi_2 \in K_1$ with $\phi_1(t) \geq 0$ for all t, $\int_{-\infty}^{\infty} \phi_1(t)dt = 1$, $\phi_2(0) = 1$, and $\|\phi_2\|_\infty = 1$. Set $\psi_n(t,\tau) = n\phi_1(n(t-\tau))\phi_2(\frac{\tau}{n})$. Then $\psi_n \in C^\infty$. Suppose $\operatorname{supp}\phi_1, \operatorname{supp}\phi_2 \subset [-a,a]$. If $|\tau| \geq na$, then $|\frac{\tau}{n}| \geq a$, so $\phi_2(\frac{\tau}{n}) = 0$ and $\psi(t,\tau) = 0$. If $|\tau| < na$ and $|t| \geq (n + \frac{1}{n})a$, then

$$n\,|t-\tau| \geq n\,(|t| - |\tau|) > n\left(n + \frac{1}{n}\right)a - n^2 a = a,$$

so $\phi_1(n(t-\tau)) = 0$ and $\psi_n(t,\tau) = 0$. Hence, $\psi_n \in K_2$.

Let

$$f_n(t,\tau) = \int_{-\infty}^{\infty} \psi_n(\tau,\eta)dg_t(\eta)$$

and $f = \frac{\partial g}{\partial \tau}$. Then

$$\begin{aligned}
\int_{-\infty}^{\infty} |f_n(t,\tau)|\,d\tau &\leq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\psi_n(\tau,\eta)|\,|dg_t(\eta)|\,d\tau \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\psi_n(\tau,\eta)|\,d\tau\,|dg_t(\eta)| \\
&= n\int_{-\infty}^{\infty} \left|\phi_2\left(\frac{\eta}{n}\right)\right| \left(\int_{-\infty}^{\infty} \phi_1(n(\tau-\eta))d\tau\right) |dg_t(\eta)| \\
&= \int_{-\infty}^{\infty} \left|\phi_2\left(\frac{\eta}{n}\right)\right| \left(\int_{-\infty}^{\infty} \phi_1(x)dx\right) |dg_t(\eta)| \\
&\leq \operatorname*{var}_{\eta} g(t,\eta),
\end{aligned}$$

$$\|f_n\|_{\infty 1} \leq \operatorname*{ess\,sup}_t \operatorname*{var}_\tau g(t,\tau) = \|f\|_{DUBV}.$$

To show $\langle f_n, \psi\rangle \to \langle f, \psi\rangle$, note that, for every $t, \eta, x \in \mathbb{R}$, $\psi(t, \frac{x}{n} + \eta) \to \psi(t,\eta)$ and

$$\int_{-\infty}^{\infty} \left|\phi_1(x)\psi\left(t, \frac{x}{n} + \eta\right)\right| dx \leq \|\psi\|_\infty.$$

By the dominated convergence theorem,

$$\begin{aligned}
\int_{-\infty}^{\infty} \psi_n(\tau,\eta)\psi(t,\tau)d\tau &= n\phi_2\left(\frac{\eta}{n}\right)\int_{-\infty}^{\infty} \phi_1(n(\tau-\eta))\psi(t,\tau)d\tau \\
&= \phi_2\left(\frac{\eta}{n}\right)\int_{-\infty}^{\infty} \phi_1(x)\psi\left(t, \frac{x}{n} + \eta\right)dx \\
&\to \phi_2(0)\int_{-\infty}^{\infty} \phi_1(x)\psi(t,\eta)dx \\
&= \psi(t,\eta)
\end{aligned}$$

for every t, η. Furthermore,

$$\begin{aligned}\left|\int_{-\infty}^{\infty}\psi_n(\tau,\eta)\psi(t,\tau)d\tau\right| &\le n\left|\phi_2\left(\frac{\eta}{n}\right)\right|\int_{-\infty}^{\infty}\phi_1(n(\tau-\eta))\,|\psi(t,\tau)|\,d\tau\\ &= \left|\phi_2\left(\frac{\eta}{n}\right)\right|\int_{-\infty}^{\infty}\phi_1(x)\left|\psi\left(t,\frac{x}{n}+\eta\right)\right|dx\\ &\le \|\psi(t,\cdot)\|_\infty,\end{aligned}$$

so, if $\operatorname{supp}\psi\subset[-a,a]^2$,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\|\psi(t,\cdot)\|_\infty\,|dg_t(\eta)|\,dt \le \int_{-a}^{a}\|\psi(t,\cdot)\|_\infty \operatorname*{var}_\eta g(t,\eta)dt \le 2a\,\|\psi\|_\infty\,\|g\|_{UBV}.$$

Again, by the dominated convergence theorem,

$$\begin{aligned}\langle f_n,\psi\rangle &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\psi_n(\tau,\eta)\psi(t,\tau)d\tau dg_t(\eta)dt\\ &\to \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\psi(\tau,\eta)dg_t(\eta)dt = \langle f,\psi\rangle. \qquad \square\end{aligned}$$

COROLLARY 4.4. *Let $X = K_{2+}'$ and $Y = UL_+^1$. Then X and Y satisfy* T1–T3 *and $Y_e = DUBV_+$.*

Proof. We have $K_2 \subset UL_+^1 \subset UL^1 \subset K_2'$ with K_2 dense in K_2' (see [4, p. 118]). Hence, UL_+^1 is dense in UL^1. From Lemma 4.1(1) and Theorems 4.2 and 4.3, $(K_2', \mathfrak{T}, UL_+^1, \|\cdot\|_{\infty 1})$ satisfies T1–T3 and $UL_{+e}^1 = UL_e^1 = DUBV$. Furthermore, $UL_+^1 \subset K_{2+}' \subset K_2'$, so Lemma 4.1(2) implies that $(K_{2+}', \mathfrak{T}_1, UL_+^1, \|\cdot\|_{\infty 1})$ satisfies T1–T3 and $UL_{+e}^1 = DUBV \cap K_{2+}' = DUBV_+$. $\square$

5. BIBO stability. In this section, we consider stability of linear operators $T : K_{1+}' \to K_{1+}'$. (We must restrict our attention to K_{1+}' in order to have a consistent definition of generalized integral operators.) As in [2, p. 109], we define a *BIBO stable* linear operator T to be one such that

(S1) $T(L_+^\infty) = L_+^\infty$,

(S2) T is continuous on L_+^∞ relative to $\|\cdot\|_\infty$.

We have shown in [1] that, for time-invariant (i.e., convolution) operators, (S2) follows automatically from (S1). Unfortunately, this result does not extend to the time-varying setting. For example, let $h_\tau = e^{-\tau}\delta_\tau$. For any $u \in L_+^\infty$, $\phi \in K_1$,

$$\begin{aligned}\xi(\tau) = \langle h_\tau,\phi\rangle &= e^{-\tau}\langle\delta_\tau,\phi\rangle = e^{-\tau}\phi(\tau),\\ \langle y,\phi\rangle = \langle u,\xi\rangle &= \int_{-\infty}^{\infty}u(\tau)e^{-\tau}\phi(\tau)d\tau.\end{aligned}$$

Thus $y(t) = e^{-t}u(t)$ and $y \in L_+^\infty$. However, the map $u \mapsto y$ is not continuous on L_+^∞. Indeed, let $u_k(t) = \frac{1}{k}\theta(t+k)$; then $\|u_k\|_\infty = \frac{1}{k} \to 0$. But $y_k(t) = \frac{1}{k}e^{-t}\theta(t+k)$ and $\|y_k\|_\infty = \frac{e^k}{k} \to \infty$. Thus, as in [2], we adopt (S2) as an independent assumption.

Since classical integral operators satisfying $T(\delta_\tau) \in UL^1$ are known to be BIBO stable, a natural conjecture is that a generalized integral operator is BIBO stable if and only if $\{T(\delta_\tau)\} \in UL_{e+}^1$ ($= DUBV_+$ by Corollary 4.4). The following example lends support to this idea. Let $T(\delta_\tau) = \delta_\tau^{(n)}$. Then $\langle T(u),\phi\rangle = \langle u,\xi\rangle$, where $\xi(\tau) = \langle\delta_\tau^{(n)},\phi\rangle = (-1)^n\phi^{(n)}(\tau)$. Hence $\langle T(u),\phi\rangle = \langle u,(-1)^n\phi^{(n)}\rangle = \langle u^{(n)},\phi\rangle$, and $T(u) =$

$u^{(n)}$. In view of (S1), the n-times differentiator is BIBO stable if and only if $n = 0$, since $T(\theta) = \delta^{(n-1)} \notin L^\infty$ for $n > 0$. On the other hand, $\{\delta_\tau\} = -\frac{\partial}{\partial \tau}\{\theta_\tau\}$, and $\{\theta_\tau\} \in UBV_+$, so $\{\delta_\tau\} \in UL^1_{e+} = DUBV_+$. But $\{\delta^{(n)}_\tau\} = -\frac{\partial}{\partial \tau}\{\delta^{(n-1)}_\tau\}$ for $n > 0$, and $\delta^{(n-1)}_\tau \notin UBV$, so $\{\delta^{(n)}_\tau\} \notin UL^1_{e+}$.

Corresponding to each generalized integral operator T we may associate an operator $\tilde{T} : K_1 \to C^\infty$ defined by

$$\tilde{T}(\phi)(\tau) = \langle T(\delta_\tau), \phi \rangle .$$

Let $\phi \in K_1$, $u \in L^\infty_+$, $\xi = \tilde{T}(\phi)$. Then

$$\langle T(u), \phi \rangle = \left\langle u, \bar{\xi} \right\rangle = \int_{-\infty}^{\infty} u(\tau)\tilde{T}(\phi)(\tau)d\tau.$$

This suggests an adjoint relationship between T and $\tilde{T}$, which we will explore further in Theorem 5.4. First we need a result which shows that stability of T can be characterized in terms of $\tilde{T}$.

LEMMA 5.1. *T is BIBO stable if and only if*

$$\sup_{\substack{\phi \in K_1 \\ \|\phi\|_1 = 1}} \int_{-\infty}^{\infty} \left| \tilde{T}(\phi)(\tau) \right| d\tau < \infty.$$

Proof. The proof is identical to the proof of [1, Lemma 4.1(2)], replacing the phrase "convolution operator" by "generalized integral operator." ∎

If T is BIBO stable, Lemma 5.1 indicates that $\tilde{T}$ is a continuous linear operator from $(K_1, \|\cdot\|_1)$ into $(L_1, \|\cdot\|_1)$. Since $(K_1, \|\cdot\|_1)$ is dense in L^1, $\tilde{T}$ extends uniquely to a continuous linear operator $\tilde{T}_e : L^1 \to L^1$.

LEMMA 5.2. *Let $T : K'_{1+} \to K'_{1+}$ be a BIBO stable generalized integral operator, $s(\cdot, \tau) = T(\theta_\tau)$, $s_\tau(t) = \widehat{s}_t(\tau) = s(t, \tau)$, $\phi \in L^1$, and $u \in L^\infty$. Then*

(1) *$\tau \to \int_{-\infty}^{\infty} s(t, \tau)\phi(t)dt$ is absolutely continuous,*

(2) *$\tilde{T}_e(\phi)(\tau) = -\frac{d}{d\tau}\int_{-\infty}^{\infty} s(t, \tau)\phi(t)dt$ for every $\tau \in \mathbb{R}$,*

(3) *$\int_{-\infty}^{\infty} u(\tau)d\int_{-\infty}^{\infty} s(t, \tau)\phi(t)dt = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \phi(t)u(\tau)d\widehat{s}_t(\tau)dt$.*

Proof. From [7, Theorem 2.3.9], there exists $g : \mathbb{R}^2 \to \mathbb{R}$ such that $g(\cdot, \cdot) - g(\cdot, \infty)$ $\in UBV$, the map $\tau \to \int_{-\infty}^{\infty} g(t, \tau)\phi(t)dt$ is absolutely continuous, and

$$\tilde{T}_e(\phi)(t) = \frac{d}{d\tau}\int_{-\infty}^{\infty} g(t, \tau)\phi(t)dt$$

for each $\phi \in L^1$. Let $c = \int_{-\infty}^{\infty} g(t, \infty)\phi(t)dt$. Then

$$\begin{aligned}
\int_{-\infty}^{\infty} s\left(t, t_0\right)\phi\left(t\right)dt &= \int_{-\infty}^{\infty} \theta_{t_0}\left(\tau\right)\tilde{T}\left(\phi\right)\left(\tau\right)d\tau \\
&= \int_{t_0}^{\infty} \tilde{T}\left(\phi\right)\left(\tau\right)d\tau \\
&= \int_{t_0}^{\infty} \frac{d}{d\tau}\left(\int_{-\infty}^{\infty} g(t, \tau)\phi(t)dt\right)d\tau \\
&= \int_{-\infty}^{\infty} \left(g\left(t, \infty\right) - g\left(t, t_0\right)\right)\phi\left(t\right)dt \\
&= c - \int_{-\infty}^{\infty} g(t, t_0)\phi(t)dt,
\end{aligned}$$

from which (1) and (2) follow.

To prove (3), note that, for any $\phi \in L^1$, $|\phi(t)| < \infty$ a.e. and

$$\mu_t(-\infty, \tau] = \phi(t) s(t, \tau)$$

determines a finite signed Borel measure on $\mathbb{R}$ for a.e. t as does

$$\mu(-\infty, \tau] = g(\tau) = \int_{-\infty}^{\infty} \mu_t(-\infty, \tau] dt. \tag{5.1}$$

Consider the family $\mathcal{L}$ of sets $A \subset \mathbb{R}$ such that the map $t \to \mu_t(A)$ is Borel measurable for a.e. t. Since

$$\mu_t(\mathbb{R}) = \operatorname*{var}_{\tau} s(t, \tau),$$

$\mathbb{R} \in \mathcal{L}$. If $A, B \in \mathcal{L}$ with $A \subset B$, then

$$\mu_t(B - A) = \mu_t(B) - \mu_t(A),$$

so $B - A \in \mathcal{L}$. If $A_n \in \mathcal{L}$ with $A_n \uparrow A$, then

$$\mu_t(A) = \mu_t(A_1) + \sum_n \mu_t(A_{n+1} - A_n),$$

so $A \in \mathcal{L}$. From the $\pi - \lambda$ theorem (see [8, Theorem 4.2]), every Borel set in $\mathbb{R}$ belongs to $\mathcal{L}$. Hence, $t \to \mu_t(A)$ is Borel measurable for any Borel set A and

$$\int_{-\infty}^{\infty} |\mu_t(A)|\, dt = \int_{-\infty}^{\infty} \left| \int_A \phi(t) d\widehat{s}_t(\tau) \right| dt \le \int_{-\infty}^{\infty} \int_A |\phi(t)|\, |d\widehat{s}_t(\tau)|\, dt \le \|s\|_{UBV}\, \|\phi\|_1 .$$

Also, if the A_n are pairwise disjoint,

$$\int_{-\infty}^{\infty} \mu_t \left(\underset{n}{\cup} A_n \right) dt = \int_{-\infty}^{\infty} \sum_n \mu_t(A_n) dt = \sum_n \int_{-\infty}^{\infty} \mu_t(A_n) dt,$$

so the map

$$A \to \int_{-\infty}^{\infty} \mu_t(A) dt \tag{5.2}$$

is a finite signed Borel measure. Since μ and (5.2) have the same distribution function (5.1),

$$\mu(A) = \int_{-\infty}^{\infty} \mu_t(A) dt$$

for each A.

Let I_A be the indicator function on A. Then

$$\int_{-\infty}^{\infty} I_A(\tau) d\mu = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I_A(\tau) d\mu_t dt,$$

$$\int_{-\infty}^{\infty} I_A(\tau) d \int_{-\infty}^{\infty} s(t, \tau) \phi(t) dt = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(t) I_A(\tau) d\hat{s}_t(\tau) dt. \tag{5.3}$$

Both sides of (3) are bounded by $\|u\|_\infty \|s\|_{UBV} \|\phi\|_1$, so, as functions of u, they both represent continuous linear functionals on L^∞. Since the span of the indicators I_A is dense in L^∞, 5.3 implies (3). ∎

We are now in a position to prove our main result.

THEOREM 5.3. *Let $T : K'_{1+} \to K'_{1+}$ be a generalized integral operator. The following statements are equivalent:*

(1) *T is BIBO stable.*
(2) *$\{T(\delta_\tau)\} \in UL^1_{+e}$.*
(3) *$\{T(\theta_\tau)\} \in UBV_+$.*

Proof. From Theorem 3.7, $\{T(\theta_\tau)\} \in UBV_+$ if and only if $\{T(\delta_\tau)\} \in DUBV_+$. From Corollary 4.4, $DUBV_+ = UL^1_{+e}$. Thus (2) and (3) are equivalent.

To prove that (3) implies (1), let $u \in L^\infty_+$, $\phi \in K_1$, $s_\tau = T(\theta_\tau)$, $s(t,\tau) = s_\tau(t)$, and note that

$$\operatorname*{var}_\tau \langle s_\tau, \phi \rangle = \operatorname*{var}_\tau \int_{-\infty}^{\infty} s(t,\tau)\phi(t)dt \leq \int_{-\infty}^{\infty} \left(\operatorname*{var}_\tau s(t,\tau) \right) |\phi(t)|\, dt \leq \|s\|_{UBV} \|\phi\|_1 .$$

From Theorem 3.7 and Lemma 5.2,

$$\begin{aligned} \langle T(u), \phi \rangle &= \int_{-\infty}^{\infty} u(\tau) \langle T(\delta_\tau), \phi \rangle \, d\tau \\ &= -\int_{-\infty}^{\infty} u(\tau) \left\langle \frac{\partial s_\tau}{\partial \tau}, \phi \right\rangle d\tau \\ &= -\int_{-\infty}^{\infty} u(\tau) \frac{\partial}{\partial \tau} \langle s_\tau, \phi \rangle \, d\tau \\ &= -\int_{-\infty}^{\infty} u(\tau) d \langle s_\tau, \phi \rangle , \end{aligned}$$

$$|\langle T(u), \phi \rangle| \leq \|u\|_\infty \operatorname*{var}_\tau \langle s_t, \phi \rangle \leq \|u\|_\infty \|s\|_{UBV} \|\phi\|_1 .$$

Therefore,

$$\|T(u)\|_\infty = \sup_{\substack{\phi \in K_1 \\ \|\phi\|_1 = 1}} |\langle T(u), \phi \rangle| \leq \|u\|_\infty \|s\|_{UBV} .$$

Finally, we prove that (1) implies (2). Let $\phi_1, \phi_2 \in K_1$, and let $\psi \in K_2$ be given by $\psi(t,\tau) = \phi_1(t)\phi_2(\tau)$. From Lemma 5.2,

$$\begin{aligned} \langle \{T(\delta_\tau)\}, \psi \rangle &= \int_{-\infty}^{\infty} \langle T(\delta_\tau), \phi_1 \rangle \, \phi_2(\tau) d\tau \\ &= \int_{-\infty}^{\infty} \tilde{T}(\phi_1)(\tau)\phi_2(\tau) d\tau \\ &= -\int_{-\infty}^{\infty} \left(\frac{d}{d\tau} \int_{-\infty}^{\infty} s(t,\tau)\phi_1(t)dt \right) \phi_2(\tau) d\tau \\ &= -\int_{-\infty}^{\infty} \phi_2(\tau) d \left(\int_{-\infty}^{\infty} s(t,\tau)\phi_1(t)dt \right) \\ &= -\int_{-\infty}^{\infty} \phi_1(t) \left(\int_{-\infty}^{\infty} \phi_2(\tau) ds_t(\tau) \right) dt \\ &= -\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \psi(t,\tau) ds_t(\tau) dt. \end{aligned} \tag{5.4}$$

From linearity and continuity of the functionals in (5.4) and from [6, p. 65], (5.4) holds for any $\psi \in K_2$. Thus Theorem 2.2(1) implies

$$\{T(\delta_\tau)\} = -\frac{\partial s}{\partial \tau} \in DUBV. \qquad \square$$

We now examine a certain extension of the operator T and its relation to $\tilde{T}_e$. Consider the closure L_0^∞ of $L_+^\infty \subset L^\infty$. It is easy to show that L_0^∞ is a closed proper subspace of L^∞ and

$$L_0^\infty = \left\{ f \in L^\infty \mid \operatorname*{ess\,sup}_{t\in(-\infty,-n]} |f(t)| \to 0 \text{ as } n \to \infty \right\}.$$

If T is stable, T extends uniquely to a continuous linear operator $T_0 : L_0^\infty \to L_0^\infty$. This extension can be taken further.

THEOREM 5.4. *Suppose $T : K_{1+}' \to K_{1+}'$ is a BIBO stable generalized integral operator, $s(\cdot,\tau) = T(\theta_\tau)$, and $\hat{s}_t(\tau) = s(t,\tau)$. Let $T_e : L^\infty \to L^\infty$ be the continuous linear operator defined by*

$$T_e(u)(t) = -\int_{-\infty}^{\infty} u(\tau) d\hat{s}_t(\tau).$$

Then

(1) *$T_e(u) = T_0(u)$ for all $u \in L_0^\infty$,*

(2) *T_e is the adjoint of $\tilde{T}_e$.*

Proof. (1) Let $u \in L_+^\infty$, $\phi \in K_1$. From Lemma 5.2,

$$\begin{aligned}
\langle T(u), \phi \rangle &= \int_{-\infty}^{\infty} u(\tau) \tilde{T}(\phi)(\tau) d\tau \\
&= \int_{-\infty}^{\infty} u(\tau) \langle T(\delta_\tau), \phi \rangle d\tau \\
&= -\int_{-\infty}^{\infty} u(\tau) \left\langle \frac{\partial s_\tau}{\partial \tau}, \phi \right\rangle d\tau \\
&= -\int_{-\infty}^{\infty} u(\tau) \frac{\partial}{\partial \tau} \langle s_\tau, \phi \rangle d\tau \\
&= -\int_{-\infty}^{\infty} u(\tau) d \langle s_\tau, \phi \rangle \\
&= -\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(\tau) \phi(t) d\hat{s}_t(\tau) dt.
\end{aligned}$$

Since ϕ is arbitrary,

$$T(u)(t) = -\int_{-\infty}^{\infty} u(\tau) d\hat{s}_t(\tau)$$

a.e. Since L_+^∞ is dense in L_0^∞, the result follows.

(2) Let $\phi \in L^1$ and $u \in L^\infty$. From Lemma 5.2,

$$\begin{aligned}\langle T_e(u), \phi\rangle &= \int_{-\infty}^{\infty} T_e(u)(t)\phi(t)dt \\ &= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(\tau)\phi(t)d\hat{s}_t(\tau)dt \\ &= -\int_{-\infty}^{\infty} u(\tau)d\int_{-\infty}^{\infty} s(t,\tau)\phi(t)dt \\ &= -\int_{-\infty}^{\infty} u(\tau)\left(\frac{d}{d\tau}\int_{-\infty}^{\infty} s(t,\tau)\phi(t)dt\right)d\tau \\ &= \int_{-\infty}^{\infty} u(\tau)\tilde{T}_e(\phi)(\tau)d\tau \\ &= \left\langle u, \tilde{T}_e(\phi)\right\rangle. \qquad \square\end{aligned}$$

To conclude, we give several equivalent expressions for the "gain" of a BIBO stable linear operator.

THEOREM 5.5. *For any BIBO stable generalized integral operator* $T : K'_{1+} \to K'_{1+}$,

$$\begin{aligned}\sup_{\substack{u\in L^\infty_+\\ \|u\|_\infty=1}} \|T(u)\|_\infty &= \sup_{\substack{u\in L^\infty\\ \|u\|_\infty=1}} \|T_e(u)\|_\infty \\ &= \sup_{\substack{\phi\in L^1\\ \|\phi\|_1=1}} \left\|\tilde{T}_e(\phi)\right\|_1 = \|\{T(\theta_\tau)\}\|_{UBV} = \|\{T(\delta_\tau)\}\|^e_{UL^1}.\end{aligned} \tag{5.5}$$

Proof. Let $v \in L^\infty, \|v\|_\infty = 1, v_n(\tau) = v(\tau)\theta(\tau+n)$, and $\varepsilon > 0$. From Theorem 5.4, for a.e. t there exists $N < \infty$ such that $n > N$ implies

$$\begin{aligned}|T_e(v)(t)| - |T(v_n)(t)| &\le |T_e(v)(t) - T(v_n)(t)| \\ &= \left|\int_{-\infty}^{-n} v(\tau)d\hat{s}_t(\tau)\right| \le \int_{-\infty}^{-n} |v(\tau)|\,|d\hat{s}_t(\tau)| \le \operatorname*{var}_{\tau\le -n} s(t,\tau) < \varepsilon.\end{aligned}$$

Hence,

$$\sup_{\substack{u\in L^\infty_+\\ \|u\|_\infty=1}} \|T(u)\|_\infty \ge \|T(v_n)\|_\infty \ge |T(v_n)(t)| > |T_e(v)(t)| - \varepsilon.$$

Since v, t, ε are arbitrary,

$$\sup_{\substack{u\in L^\infty_+\\ \|u\|_\infty=1}} \|T(u)\|_\infty \ge \sup_{\substack{\varepsilon>0\\ u\in L^\infty\\ \|u\|_\infty=1}} \operatorname*{ess\,sup}_t(|T_e(u)(t)| - \varepsilon) = \sup_{\substack{u\in L^\infty\\ \|u\|_\infty=1}} \|T_e(u)\|_\infty.$$

The second equality in (5.5) follows from Theorem 5.4(2). For the third equality, set

$$s(t,\tau) = T(\theta_\tau)(t).$$

From Lemma 5.2,

$$\tilde{T}_e(\phi)(\tau) = -\frac{d}{d\tau}\int_{-\infty}^{\infty} s(t,\tau)\phi(t)dt$$

for every τ, and

$$\left\|\tilde{T}_e(\phi)\right\|_1 = \int_{-\infty}^{\infty}\left|\frac{d}{d\tau}\int_{-\infty}^{\infty} s(t,\tau)\phi(t)dt\right| d\tau = \operatorname*{var}_{\tau}\int_{-\infty}^{\infty} s(t,\tau)\phi(t)dt.$$

From [7, Theorem 2.3.9],

$$\sup_{\substack{\phi\in L^1\\ \|\phi\|_1=1}}\left\|\tilde{T}_e(\phi)\right\|_1 = \|s\|_{UBV}\,.$$

The last equality in (5.5) follows from Theorems 3.7 and 4.3. □

REFERENCES

[1] C.-J. Wang and J. D. Cobb, *A characterization of bounded-input bounded-output stability for linear time-invariant systems with distributional inputs*, SIAM J. Control Optim., 34 (1996), pp. 987–1000.

[2] C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.

[3] I. M. Gelfand and G. E. Shilov, *Generalized Functions*, Vol. 1, Academic Press, New York, 1968.

[4] I. M. Gelfand and G. E. Shilov, *Generalized Functions*, Vol. 2, Academic Press, New York, 1968.

[5] J. Barros-Neto, *An Introduction to the Theory of Distributions*, Marcel Dekker, New York, 1973.

[6] K. Yosida, *Functional Analysis*, Springer-Verlag, Berlin, 1971.

[7] N. Dunford and B. J. Pettis, *Linear operations on summable functions,* Trans. Amer. Math. Soc., 47 (1940), pp. 323–392.

[8] P. Billingsley, *Probability and Measure*, 3rd ed., John Wiley, New York, 1995.

[9] W. Rudin, *Real and Complex Analysis*, McGraw–Hill, New York, 1974.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1244–1265

WELL-POSEDNESS OF BOUNDARY CONTROL SYSTEMS*

ADA CHENG† AND KIRSTEN MORRIS‡

Abstract. Continuity of the input/output map for boundary control systems is shown through the system transfer function. Our approach transforms the question of continuity of the input/output map of a boundary control system to uniform boundedness of the solution to a related elliptic problem. This is shown for a class of boundary control systems with Dirichlet, Neumann, or Robin boundary control.

Key words. infinite-dimensional systems, boundary control, input/output stability

AMS subject classifications. 35B30, 35B37, 93C20

DOI. 10.1137/S0363012902384916

1. Introduction.

Boundary control systems are an important class of infinite-dimensional control systems. Some important applications are control of annealing processes, control of structural vibrations, and active noise control.

Key questions are whether the mappings from input to state, input to output, initial state to input, and initial state to final state are well defined and bounded. When all four mappings are well defined and bounded, the system is said to be well-posed [19]. Salamon [20] showed that boundedness of the input/output map implies well-posedness of the control system with respect to some state space. (An alternative proof in [14] uses frequency domain analysis.) Since boundedness is equivalent to continuity for linear systems, ill-posedness of the input/output map indicates that the measured outputs are not continuously dependent on the inputs. This would lead to difficulties in the practical implementation of any such control system. Often, however, ill-posedness of the control system indicates modelling errors. An example illustrating this point is given in this paper. Thus, showing boundedness of the input/output map of a boundary control system is important. This problem is the focus of this paper.

Boundedness of the initial to final state map is equivalent to showing existence of a semigroup and is fairly well understood. A number of authors have obtained results on boundedness of the state/output map and input/state map. For more details see, e.g., [3, 8, 9, 10, 11, 12, 15, 16]

The literature on showing boundedness of the input/output map is less extensive. One technique for determining well-posedness is to use spectral expansion of the underlying semigroup. This technique is applicable to showing boundedness of the input/state and state/output maps as well as the input/output map. For example, in [7] it was shown that several examples of boundary control systems with one space dimension were well-posed. In [6], it was shown that the one-dimensional heat equation with Dirichlet boundary control and point observation is well-posed under a suitable choice of state space. In [18], well-posedness of an accelerometer control system was shown. The spectral expansion method requires the availability of the

*Received by the editors February 20, 2002; accepted for publication (in revised form) February 16, 2003; published electronically August 6, 2003. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.
http://www.siam.org/journals/sicon/42-4/38491.html
†Department of Mathematics, Kettering University, Flint, MI 48504 (acheng@math.waterloo.ca).
‡Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 (kmorris@riccati.uwaterloo.ca).

eigenvalues and eigenvectors of the system (or at least estimates). Also, the eigenvectors must form a Riesz basis. For many multidimensional problems it is difficult to calculate the eigenfunctions and eigenvalues of the underlying semigroup. Hence there are difficulties in extending this method to more general problems.

Well-posedness of a structural acoustic control system has been considered in [1, 2]. The authors use a state-space formulation of the control system. Partial differential equation results lead to estimates of the regularity of the resolvant and hence of the transfer function in state-space form.

Another method to determine boundedness of the input/output map uses the system transfer function. The concept of the transfer function for finite-dimensional systems extends to general well-posed systems. This is discussed in detail in the next section. Curtain and Weiss [6] showed that the input/output map is bounded if and only if the transfer function is uniformly bounded in a right half-plane. In several papers [6, 18, e.g.] well-posedness is established by showing that the system transfer function is bounded in some right-half-plane. The difficulty with this approach is that the transfer function has been rigorously obtained for only a few systems.

In [3], boundedness of the input/output map was shown for a class of structural control systems with point measurement of acceleration by showing that the system transfer function is proper. However, unlike the examples given above, justification for the transfer function was not computed directly. Instead, it was shown that the fact that the infinitesimal generator generates an analytic semigroup implies properness of the system transfer function.

In the next section systems theory for boundary control systems is discussed. The nature of the input/output map and the transfer function for these systems is explained. We give a representation for the system transfer function purely in terms of the boundary control formulation.

In section 3 we present our approach. The question of boundedness of the input/output map of a boundary control system is transformed to uniform boundedness (in a sense defined later) of solutions to a related elliptic boundary value problem. We use this approach to obtain well-posedness of several large classes of boundary control systems. Section 4 contains some background on elliptic boundary value problems. In sections 5 and 6 we show boundedness of the input/output map for a several large classes of problems with Dirichlet, Neumann, or Robin boundary control.

Our approach has several advantages. It is not necessary to compute a state-space realization. Also, the analysis of an elliptic problem is simpler than that of the original problem, and the extensive literature available on boundary value problems may be used. Our method is particularly useful for multidimensional systems with variable coefficients where the state-space realization is tedious to obtain and the system transfer function is even more difficult to obtain from the realization.

2. Transfer functions for boundary control systems. We will use the following formal definition of a *boundary control system*:

$$\left.\begin{array}{rll} \frac{d}{dt}z(t) & = Lz, & z(0) = z_0, \\ \Gamma z(t) & = u(t), & \\ y(t) & = Kz(t). & \end{array}\right\} \tag{2.1}$$

The operators $L \in \mathcal{L}(\mathcal{Z}, \mathcal{H})$, $\Gamma \in \mathcal{L}(\mathcal{Z}, \mathcal{U})$, and $K \in \mathcal{L}(\mathcal{Z}, \mathcal{Y})$. The spaces $(\mathcal{Z}, \|\cdot\|_{\mathcal{Z}}), (\mathcal{H}, \|\cdot\|_{\mathcal{H}})$, $(\mathcal{U}, \|\cdot\|_{\mathcal{U}}), (\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$ are all Hilbert spaces, and $\mathcal{Z}$ is a dense subspace of $\mathcal{H}$ with continuous, injective embedding $\iota_{\mathcal{Z}}$. The triple (L, Γ, K) refers to a boundary control system with output operator K. We shall often refer to a boundary

control system by the double (L, Γ). (The operator K is in this case understood to be the identity operator.) We will assume throughout this paper that a boundary control system (2.1) satisfies the following assumptions:

(A1) The operator Γ is onto, $\ker \Gamma$ is dense in $\mathcal{H}$, and there exists $\mu \in R$ such that $\ker(\mu I - L) \cap \ker\Gamma = 0$ and $\mu I - L$ is onto $\mathcal{H}$.

(A2) For any $z_0 \in \mathcal{Z}$ with $\Gamma z_0 = 0$, $u(\cdot) = 0$, there exists a unique solution of (Γ, L) in $C^1[0, T; \mathcal{H}] \cap C[0, T; \mathcal{Z}]$ depending continuously on z_0.

In this paper, we are solely interested in the boundedness of the input/output map from $u \in L^2(0, T; \mathcal{U})$ to $y \in L^2(0, T; \mathcal{Y})$.

DEFINITION 2.1. *The input/output map is bounded if for all times $T > 0$ and $u \in H^2(0, T; \mathcal{U})$, $z(0) = 0$, the output y is well defined and there is a constant c_T such that $\|y\|_{L^2(0,T;\mathcal{Y})} \le c_T \|u\|_{L^2(0,T;\mathcal{U})}$.*

This implies that the input/output map $u \to y$ can be extended to a bounded map on all of $L^2(0, T; \mathcal{U})$. Alternatively, one can describe the relationship between the inputs and the outputs using the Laplace transform.

DEFINITION 2.2. *Let $\hat{y}(s)$ indicate the Laplace transform of the output of a system and indicate similarly the transform of the input by $\hat{u}(s)$. The system transfer function is the operator $G(s)$ such that*

$$\hat{y}(s) = G(s)\hat{u}(s)$$

for all s, Re $s > \sigma$ *for some real σ.*

Implicit in this definition is that the input/output map is well defined and that the output is Laplace transformable. Boundedness of the input/output map can be determined using the system transfer function.

THEOREM 2.3 (see [6]). *Let (L, Γ, K) define a boundary control system. The input/output map of the system is bounded if and only if there exists a real number σ such that the transfer function $G(s)$ associated with (L, Γ, K) satisfies*

$$\sup_{\text{Re } s > \sigma} \| G(s) \|_{\mathcal{L}(\mathcal{U}, \mathcal{Y})} < \infty.$$

The function $G(s)$ is said to be proper *if the above inequality holds.*

We now consider the definition of a transfer function for a boundary control system in detail. First, consider a control system in state-space form:

$$\dot{z}(t) = Az(t) + Bu(t), \qquad z(0) = z_0, \tag{2.2}$$

$$y(t) = Cz(t), \tag{2.3}$$

where A is an infinitesimal generator of a C_0-semigroup $T(t)$ on state space $\mathcal{H}$. Also, B and C are bounded operators: $B \in \mathcal{L}(\mathcal{U}, \mathcal{H})$, $C \in \mathcal{L}(\mathcal{H}, \mathcal{Y})$. The input/output map is

$$y(t) = C \int_0^t T(t - \sigma) B u(\sigma)\, d\sigma. \tag{2.4}$$

Defining $g(t) = CT(t)B$, the output is simply the convolution of $g(t)$ and $u(t)$. Taking the Laplace transform on both sides of (2.4) gives

$$\hat{y}(s) = G(s)\hat{u}(s). \tag{2.5}$$

Here $G(s) = C(sI - A)^{-1}B$ is the system transfer function. Note that it is the Laplace transform of the function $g(t)$ that defines the input/output map. This is a direct generalization of the theory for finite-dimensional systems.

Any boundary control system can be written in state-space form (A, B, C) [19]. The operator A that generates the semigroup $T(t)$ in the state-space formulation is defined from the boundary control system as follows. Define

$$\mathcal{W} = \{\, z \in \mathcal{Z} \mid \Gamma z = 0 \,\}, \tag{2.6}$$

and let ι denote the canonical injection from $\mathcal{W}$ to $\mathcal{Z}$. Then $A = L\iota$ and $\mathcal{W} = [D(A)]$, the completion of $D(A)$ in the graph norm of A. Assumptions (A1) and (A2) imply that A generates a C_0-semigroup on $\mathcal{H}$. Techniques to define B and C also exist [19], but the input and output operators are generally unbounded on the state space. The linear operator $C \in \mathcal{L}(\mathcal{W}, \mathcal{Y})$ is defined by $C = K\iota$ and $C \in \mathcal{L}(\mathcal{W}, \mathcal{Y})$. The definition of B is more complicated and not needed here, but $B \in \mathcal{L}(\mathcal{U}, \mathcal{V})$ where $\mathcal{V} = [D(A^*)]'$, the dual space of $[D(A^*)]$. The operator A extends to an operator that generates a C_0-semigroup on $\mathcal{V}$ with domain $\mathcal{H}$. However, (2.3) is no longer well defined since $z(t)$ may not be in the domain of C.

In the following theorem we show that the output of a boundary control system is well defined, and that this output can be defined via the convolution of a Laplace-transformable distribution with the input. The following results will be required.

LEMMA 2.4 (see [19, Cor. 2.9]). *Let* (A1) *and* (A2) *be satisfied. Then for every $z_o \in \mathcal{Z}$ and every $u \in H^2(0,T;\mathcal{U})$, with $\Gamma z_o = u(0)$, there is a unique solution $z(\cdot) \in C(0,T;\mathcal{Z}) \bigcap C^1(0,T;\mathcal{H})$ of* (2.1).

THEOREM 2.5 (see [23, Theorem 6.5-1]). *Necessary and sufficient conditions for a function $G(s) \in \mathcal{L}(\mathcal{U}, \mathcal{Y})$ to be the Laplace transform of a distribution whose support is bounded on the left at $t = 0$ are that* (1) *there exists some half-plane* $\operatorname{Re} s > \sigma$ *on which $G(s)$ is analytic and* (2) *that there is a polynomial P such that for* $\operatorname{Re} s > \sigma$

$$\|G(s)\|_{\mathcal{L}(\mathcal{U},\mathcal{Y})} \le P(|s|).$$

THEOREM 2.6. *The input/output map of any boundary control system* (2.1) *is well defined for all inputs $u \in H^2(0,T;\mathcal{U})$, $u(0) = 0$. This output can be written as*

$$y(t) = g(t) * u(t),$$

where $g(t)$ is a distribution with Laplace transform $G(s)$. Let $A = L\iota$ with domain as in (2.6). *The operator $G(s) \in \mathcal{L}(\mathcal{U}, \mathcal{Y})$ for each $s \in \rho(A)$ and $G(s)$ is the system transfer function.*

Proof. First, as mentioned above, construct the state-space realization (A, B, C) using the procedure in [19]. Equation (2.2) is valid if we consider it as a differential equation on $\mathcal{V} = [D(A^*)]'$. Rewriting, we obtain, for any $\mu \in \rho(A)$,

$$\begin{aligned} z(t) &= (\mu I - A)^{-1}(\mu I - A)z(t) \\ &= (\mu I - A)^{-1}(\mu z(t) - \dot{z}(t)) + (\mu I - A)^{-1}Bu(t). \end{aligned} \tag{2.7}$$

For all initial conditions $z(0) = 0$ and smooth controls $u \in H^2(0,T;\mathcal{U})$ with $u(0) = 0$, the first term in (2.7) is in $\mathcal{W} \subset \mathcal{Z}$ for each time t (Lemma 2.4). Regarding A as a generator on $\mathcal{V}$ with domain $\mathcal{H}$, we obtain that $(\mu I - A)^{-1}B \in \mathcal{L}(\mathcal{U}, \mathcal{H})$. Furthermore, for any $\mu \in \rho(A)$, $\operatorname{Range}(\mu I - A)^{-1}B \subset \mathcal{Z}$ and so $(\mu I - A)^{-1}B \in \mathcal{L}(\mathcal{U}, \mathcal{Z})$ [19, Prop. 2.8]. Thus we may apply the operator K to the solution $z(t)$ to obtain the output $y(t)$:

$$y(t) = K(\mu I - A)^{-1}(\mu z(t) - \dot{z}(t)) + K(\mu I - A)^{-1}Bu(t). \tag{2.8}$$

Since $\mathcal{W} \subset \mathcal{Z}$, $K(\mu I - A)^{-1} \in \mathcal{L}(\mathcal{H}, \mathcal{Y})$ and $K(\mu I - A)^{-1}B \in \mathcal{L}(\mathcal{U}, \mathcal{Y})$. Since both u and z are Laplace transformable, we now take the Laplace transform of both sides of (2.8) to obtain

$$\hat{y}(s) = K(\mu I - A)^{-1}(\mu - s)(sI - A)^{-1}B\hat{u}(s) + K(\mu I - A)^{-1}B\hat{u}(s).$$

The system transfer function is thus

$$G(s) = K(\mu I - A)^{-1}(\mu - s)(sI - A)^{-1}B + K(\mu I - A)^{-1}B.$$

Setting $\mu = s$, we obtain that

$$G(s) = K(sI - A)^{-1}B \tag{2.9}$$

for any $s \in \rho(A)$. (This is formula (2.18) for the *generalized transfer function* in [19].)

For $s \in \rho(A)$, $G(s)$ is analytic and so condition (1) in Theorem 2.5 is satisfied. Since the norm on $\mathcal{H}$ is equivalent to the graph norm of A (as a generator on $\mathcal{V}$) on $\mathcal{V}$, $\|(s - A)^{-1}B\|_{\mathcal{L}(\mathcal{U},\mathcal{H})} \leq M$ for some constant M and all Re $s > \sigma$ for some σ. Thus, there is a polynomial $P(s)$ such that G satisfies condition (2) in Theorem 2.5. It follows from Theorem 2.5 that $G(s)$ is the Laplace transform of a distribution $g(t)$; hence the output $y(t)$ is the convolution of this distribution and the input. □

This representation of the input/output map is valid for any boundary control system and for $u \in H^2(0, T; \mathcal{U})$ with $u(0) = 0$. In order to extend the input/output map to all $u \in L^2(0, T; \mathcal{U})$ we need to show that the map is bounded or, equivalently, that the transfer function is proper.

We now obtain a representation of the transfer function of a boundary control system. This representation is based entirely on the boundary control description (2.1) and does not require construction of a state-space realization. The transfer function is defined in terms of an elliptic problem associated with the boundary control system.

DEFINITION 2.7. *The* abstract elliptic problem $(L, \Gamma)_e$ *corresponding to the boundary control system* (L, Γ), *as defined in* (2.1), *is*

$$\left.\begin{array}{ll} Lz & = sz, \quad s \in \mathbb{C}, \\ \Gamma z & = u. \end{array}\right\} \tag{2.10}$$

We denote the solution $z \in \mathcal{Z}$ *by* $z(s)$.

DEFINITION 2.8. *Let* $\mathcal{T}(t)$ *be a* C_0*-semigroup on* $\mathcal{H}$. *The constant* α *defined by*

$$\alpha = \inf_{t>0} \frac{1}{t} \log \|\mathcal{T}(t)\|$$

is called the growth bound *of the semigroup* $\mathcal{T}(t)$.

Let α indicate the growth bound of the semigroup associated with (L, Γ). The elliptic problem (2.13) has a unique solution $z(s)$ for all u and Re $s > \alpha$. The system transfer function may be described through the solutions to the abstract elliptic problem (2.13).

THEOREM 2.9. *Let* (L, Γ, K) *define a boundary control system. Define* $\mathcal{W}, A,$ *and* $D(A)$ *be as above. Then there exists an* $\alpha \in \Re$ *such that the transfer function,* $G(s)$, *of the boundary control system* (L, Γ, K) *is given by*

$$G(s)u = Kz(s) \quad \textit{for all } s \in \mathbb{C}, \ \textit{with } \text{Re } s > \alpha, \tag{2.11}$$

where $z(s)$ *is the solution to the abstract elliptic problem* (2.10) *with input* u.

Proof. Let α denote the growth bound of the C_0-semigroup generated by A. Then for all $s \in \mathbb{C}$ with Re $s > \alpha$, $s \in \rho(A)$.

The transfer function $G(s)$ is given by (2.9). However, $(sI-A)^{-1}B$ is the solution operator of abstract elliptic problem (2.10) with input u [19, Prop. 2.8, eqn. 2.20].

Alternatively, for any given $u \in \mathcal{U}$, choose $z \in \mathcal{Z}$ so that $\Gamma z = u$. Then $G \in \mathcal{L}(\mathcal{U}, \mathcal{Y})$ is defined by [19, Rem. 2.7]

$$G(s)\Gamma z = Kz - C(sI - A)^{-1}(sz - Lz). \tag{2.12}$$

For any $u \in \mathcal{U}$ and any $s \in \mathbb{C}$, with Re $s > \alpha$, let z solve the associated elliptic problem. From (2.12) we have

$$G(s)u = Kz(s).$$

This is precisely (2.11). □

Thus, the solution to (2.11) gives a representation of the transfer function of a boundary control system. The representation of $G(s)$ obtained above is not as surprising as the abstract elliptic problem (2.10) is the formal Laplace transform (with respect to t) of the boundary control system. Theorem 2.9 is a justification of such a process. Thus the abstract elliptic problem $(L, \Gamma)_e$ corresponding to the boundary control system (L, Γ) can be written as

$$\left.\begin{array}{rl} L\hat{z} &= s\hat{z}, \quad s \in \mathbb{C}, \\ \Gamma\hat{z} &= \hat{u}. \end{array}\right\} \tag{2.13}$$

As a simple example, we compute the transfer function for a heat transfer problem on a unit interval using (2.11).

Example 2.10 (one-dimensional heat equation with Neumann boundary control). One of the simplest examples of a well-posed boundary control system is the problem of temperature control in a one-dimensional rod of length 1 with a controlled heat flow at one end. The output is the temperature measured at x_1, $0 \le x_1 \le 1$. The system equations are

$$\left.\begin{array}{rcll} \frac{\partial z}{\partial t} &=& \frac{\partial^2 z}{\partial x^2}, & x \in [0,1], \\ z(x,0) &=& 0, & x \in [0,1], \\ \frac{\partial z}{\partial x}(0,t) &=& 0, & t > 0, \\ \frac{\partial z}{\partial x}(1,t) &=& u(t), & t > 0, \\ y(t) &=& z(x_1,t). & \end{array}\right\} \tag{2.14}$$

In this example,

$$\mathcal{Z} = \{z \in H^2(0,1); z'(0) = 0\},$$

with the norm inherited from $H^2(0,1)$, $\mathcal{U} = \mathcal{Y} = \Re$, and $\mathcal{H} = L^2(0,1)$. It is easy to verify that (A1) and (A2) are satisfied. The elliptic problem corresponding to (2.14) is

$$\left.\begin{array}{rcl} \frac{d^2\hat{z}}{dx^2} &=& s\hat{z}, \\ \hat{z}'(0) &=& 0, \\ \hat{z}'(1) &=& \hat{u}, \end{array}\right\} \tag{2.15}$$

with output equation

$$\hat{y} = \hat{z}(x_1).$$

The solution to the abstract elliptic problem is

$$\hat{z}(x,s) = \frac{\hat{u}\ \cosh(\sqrt{s}\ x)}{\sqrt{s}\ \sinh\sqrt{s}}.$$

For this problem, the growth bound $\alpha = 0$. By Theorem 2.9 we have for all $s \in \mathbb{C}$ with $\operatorname{Re} s > 0$ that the transfer function of the system is given by

$$\begin{aligned} G(s)\hat{u} &= K\left(\frac{\hat{u}\ \cosh(\sqrt{s}\ x)}{\sqrt{s}\ \sinh\sqrt{s}}\right) \\ &= \frac{\hat{u}\ \cosh(\sqrt{s}\ x_1)}{\sqrt{s}\ \sinh\sqrt{s}}. \end{aligned}$$

This is exactly the transfer function one would obtain by formally taking the Laplace transform of (2.14). Moreover, the transfer function is proper; hence the input/output map is bounded.

The following example shows that if the boundary condition is not chosen correctly, it leads to an improper system transfer function. Hence examining the nature of the input/output map is useful in determining whether the mathematical model of the system is sensible. Some choices of sensing or control operations also lead to improper transfer functions.

Example 2.11 (Euler–Bernoulli beam with Kelvin–Voigt damping). Consider the Euler–Bernoulli beam with Kelvin–Voigt damping. The beam is assumed to be fixed at $x = 0$ and free at $x = 1$. Then the equation governing the motion of the transverse displacement is

$$\left.\begin{array}{rcll} \frac{\partial^2 w}{\partial t^2} + \frac{\partial^2}{\partial x^2}\left[EI\frac{\partial^2 w}{\partial x^2} + c_d I\frac{\partial^3 w}{\partial x^2 t}\right] &=& 0, & x \in (0,1), \\ w(0,t) &=& 0, & t \geq 0, \\ \frac{\partial w}{\partial x}(0,t) &=& 0, & t \geq 0, \\ \frac{\partial^2 w}{\partial x^2}(1,t) &=& 0, & t \geq 0, \\ \frac{\partial^3 w}{\partial x^3}(1,t) &=& u(t), & t \geq 0, \\ y(t) &=& \frac{\partial w}{\partial t}(1,t), & \end{array}\right\} \tag{2.16}$$

where E, I, and c_d are positive constants. We shall compute the system transfer function via Theorem 2.9. First, we will rewrite the problem in the standard form (2.1). Define

$$z(x,t) = \begin{bmatrix} z_1(x,t) \\ z_2(x,t) \end{bmatrix} = \begin{bmatrix} w(x,t) \\ \frac{dw(x,t)}{dt} \end{bmatrix},$$

$$\begin{aligned} \frac{d}{dt}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} &= \begin{bmatrix} 0 & I \\ -EI\frac{d^4}{dx^4} & -c_d I\frac{d^4}{dx^4} \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \\ z_1'''(1,t) &= u(t), \\ y(t) &= z_2(1,t). \end{aligned}$$

For this problem,

$$\mathcal{Z} = \{(z_1, z_2) \in H^4(0,1) \times H^4(0,1); z_1(0) = z_1'(0) = z_1''(1) = 0\}.$$

The space $\mathcal{H} = \bar{H}^2(0,1) \times L^2(0,1)$, where

$$\bar{H}^2(0,1) = \{H^2(0,1); z(0) = z_1'(0) = 0\}$$

and $\mathcal{U} = \mathcal{Y} = R$. It can be verified that assumptions (A1) and (A2) are satisfied. The elliptic problem associated with (2.16) is, writing $w = z_1$ and noting that $\hat{z}_2 = s\hat{z}_1$,

$$\left.\begin{array}{rcl} (EI + sc_dI)\frac{\partial^4 \hat{w}}{\partial x^4} & = & -s^2\hat{w}, \\ \hat{w}'''(1) & = & \hat{u}. \end{array}\right\} \tag{2.17}$$

This is to be solved for $(\hat{w}, s\hat{w}) \in \mathcal{Z}$. The output equation $\hat{y} = s\hat{w}(1,s)$. The solution to the abstract elliptic problem is

$$\hat{w}(s,x) = A(s)\cosh(m(s)x) + B(s)\sinh(m(s)x) - A(s)\cos(m(s)x) - B(s)\sin(m(s)x),$$

where, letting $i = \sqrt{-1}$,

$$\begin{aligned} m(s) &= \sqrt{i}\left(\frac{s^2}{EI + sc_dI}\right)^{\frac{1}{4}}, \\ A(s) &= \frac{-\hat{u}(\sinh(m(s)) + \sin(m(s)))}{2m^3(s)(1 + \cosh(m(s))\cos(m(s)))}, \\ B(s) &= \frac{-A(s)(\cosh(m(s)) + \cos(m(s)))}{\sinh(m(s)) + \sin(m(s))}. \end{aligned}$$

Thus the system transfer function is

$$G(s) = \frac{s(\sinh(m(s))\cos(m(s)) - \cosh(m(s))\sin(m(s)))}{m^3(s)(1 + \cosh(m(s))\cos(m(s)))}.$$

One can show that for Re $s > 0$,

$$\lim_{|s|\to\infty} 4\exp\left(\sqrt{\frac{2}{i}}m(s)\right)(\sinh(m(s))\cos(m(s)) - \cosh(m(s))\sin(m(s))) = 1 - i,$$

$$\lim_{|s|\to\infty} 4\exp\left(\sqrt{\frac{2}{i}}m(s)\right)(1 + \cosh(m(s))\cos(m(s))) = 1.$$

Thus, for Re $s > 0$,

$$\lim_{|s|\to\infty} \frac{(\sinh(m(s))\cos(m(s)) - \cosh(m(s))\sin(m(s)))}{(1 + \cosh(m(s))\cos(m(s)))} = 1 - i.$$

Thus $G(s)$ is improper since $|\frac{s}{m^3(s)}|$ is unbounded as $|s| \to \infty$.

The appropriate boundary conditions should be on the bending moments and shear forces in the beam:

$$\begin{aligned} EI\frac{\partial^2 w}{\partial x^2} + c_dI\frac{\partial^3 w}{\partial x^2 t}(1,t) &= 0, \qquad t \geq 0, \\ EI\frac{\partial^3 w}{\partial x^3} + c_dI\frac{\partial^4 w}{\partial x^3 t}(1,t) &= u(t), \qquad t \geq 0. \end{aligned}$$

The original set of boundary conditions is incorrect since the moment M is equal to $\frac{\partial w^2}{\partial x^2}$ only when there is no damping in the system. With these boundary conditions, the resulting transfer function is

$$G(s) = \frac{s\Big(\sinh(m(s))\cos(m(s)) - \cosh(m(s))\sin(m(s))\Big)}{m^3(s)\Big(EI + sc_dI\Big)\Big(1 + \cosh(m(s))\cos(m(s))\Big)}.$$

Now $G(s)$ is proper since

$$\lim_{|s|\to\infty} \frac{s}{m^3(s)(EI + sc_dI)} = 0.$$

3. Boundedness of the input/output map. Theorem 2.3 implies that the boundedness of the input/output map of a boundary control system can be determined from the properness of the system transfer function. For a given observation operator K, the properness of the transfer function depends entirely on the behavior of the solution to $(L,\Gamma)_e$ as the parameter s varies.

Since we will henceforth be working entirely with the Laplace transform, we shall drop the " ˆ "notation in the interest of clarity. The following theorem provides a sufficient condition for the properness of the transfer function of a boundary control system.

DEFINITION 3.1. *Let $(\mathcal{V}, \|\cdot\|_{\mathcal{V}})$ be a normed linear space with $\mathcal{V} \subset \mathcal{H}$. We say that the solution, $\hat{z}(s)$, to the abstract elliptic problem* (2.13) *is* uniformly bounded with respect to the $\mathcal{V}$ norm *if there exist constants $\mu_1 \in \Re$ and $M \in \Re^+$ such that*

$$\|z(s)\|_{\mathcal{V}} \le M\|u\|_{\mathcal{U}} \tag{3.1}$$

for all $u \in \mathcal{U}$ and for all $s \in \mathbb{C}$ with Re $s > \mu_1$.

The following sufficient condition for properness of the system transfer function is now immediate.

THEOREM 3.2. *Let (L,Γ,K) define a boundary control system. Let $\mathcal{V}$ be a normed linear space satisfying $\mathcal{Z} \subset \mathcal{V} \subset \mathcal{H}$. If the solution to $(L,\Gamma)_e$ is uniformly bounded with respect to the $\mathcal{V}$ norm, then for all observation operators $K \in \mathcal{L}(\mathcal{V},\mathcal{Y})$, the transfer function associated with the boundary control system (L,Γ,K) is proper.*

Proof. By assumption there exist constants μ_1 and M such that inequality (3.1) holds. Let A be as defined in Theorem 2.9 with growth bound ω_0. Choose $\mu = \max\{\mu_1, \omega_0\}$ and the result follows. □

Thus, continuity of the input/output map of a boundary control system can be established by determining uniform boundedness of the solution $z(s)$ to a family of elliptic problems. Continuity of the input/output map can be established without an explicit representation of the transfer function. Also, Theorem 3.2 states that uniform boundedness of the solution to the elliptic problem $(L,\Gamma)_e$ in the $\mathcal{V}$ norm implies boundedness of the input/output map for the class of boundary control systems $\{(L,\Gamma,K) \mid K \in \mathcal{L}(\mathcal{V},\mathcal{Y})\}$. This is advantageous since there exist a large literature of results on solutions to elliptic partial differential equations, although not on uniform boundedness of the solution. A major advantage of this approach is that it is not required to compute the linear operators (A,B,C) of a state-space realization.

Example 3.3 (one-dimensional heat equation with Neumann boundary control continued). The solution to the corresponding elliptic problem is

$$z(x,s) = \frac{u\ \cosh(\sqrt{s}\ x)}{\sqrt{s}\ \sinh\sqrt{s}}.$$

Let $\mathcal{V} = H^1(0,1)$, $\mathcal{U} = \Re$, and $\mu_1 = 1$. Then for all $s \in \mathbb{C}$ with Re $s > 1$ we have

$$\|z\|^2_{L^2(0,1)} \leq \frac{|u|^2 \cosh 2}{16 \sinh 2} + \frac{|u|^2}{8 \sinh^2 2},$$
$$\left\|\frac{dz}{dx}\right\|^2_{L^2(0,1)} \leq \frac{|u|^2 \cosh 2}{2 \sinh 2} + \frac{|u|^2}{2 \sinh^2 2.}$$

Hence $\|z\|_{H^1(0,1)} \leq \sqrt{\frac{2 \cosh 2}{\sinh 2}}|u|$. Thus by Theorem 3.2, the input/output map is bounded for all $\mathcal{K} \in \mathcal{L}(\mathcal{H}^1(0,1), \Re)$. In particular, this holds for $Kz = z(x_1, t)$.

We now provide some conditions for uniform boundedness of the solution to $(L,\Gamma)_e$ with respect to $\mathcal{V}$ by rewriting $(L,\Gamma)_e$ as two subproblems.

PROPOSITION 3.4. *Let (L,Γ) define a boundary control system as in* (2.1) *and let $\mathcal{V}$ be a normed linear space satisfying $\mathcal{Z} \subset \mathcal{V} \subset \mathcal{H}$. Let $\mu \in \Re^+$ and $\mu \notin \sigma(L)$ (spectrum of L), and define the problems $(L,\Gamma)_{e1}$ and $(L,\Gamma)_{e2}$ by*

$$(L,\Gamma)_{e1} := \begin{cases} Lf = \mu f, \\ \Gamma f = u. \end{cases} \tag{3.2}$$

$$(L,\Gamma)_{e2} := \begin{cases} Lw = sw + (s-\mu)f, & s \in \mathbb{C}, \\ \Gamma w = 0. \end{cases} \tag{3.3}$$

The solution to $(L,\Gamma)_e$ is uniformly bounded with respect to the $\mathcal{V}$ norm if the following two conditions hold:

1. *There exists $f \in \mathcal{Z}$ such that f solves $(L,\Gamma)_{e1}$ and*

$$\|f\|_{\mathcal{V}} \leq C_1 \|u\|_{\mathcal{U}} \tag{3.4}$$

   for some positive constant C_1.
2. *Let $f \in \mathcal{Z}$ denote the solution to $(L,\Gamma)_{e1}$. There exists $w \in \mathcal{Z}$ such that w solves $(L,\Gamma)_{e2}$ and*

$$\|w\|_{\mathcal{V}} \leq C_2 \|f\|_{\mathcal{V}} \tag{3.5}$$

   for some positive constant C_2, independent of s.

Proof. The result is immediate by noting that $w + f$ solves the original elliptic problem $(L,\Gamma)_e$. ∎

4. Uniformly elliptic boundary value problems. In the remaining sections, we shall look at boundedness of solutions to uniformly elliptic boundary value problems. We concentrate on linear second order differential operators. Unfortunately, the traditional estimates on solutions to elliptic problems of the form (2.10) are dependent on the argument s. Our focus lies in obtaining estimates that are independent of s. We begin with some background theory and then show that under certain standard assumptions, solutions to uniformly elliptic boundary value problems of order 2 with either Dirichlet, Neumann, or Robin boundary control are uniformly bounded. The results generalize to higher order uniformly elliptic operators [5].

Let Ω be an open set in $\Re^n$. A linear second order differential operator in Ω is defined by

$$L(x,D) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(x) D_{ij} + \sum_{j=1}^{n} c_j(x) D_j + d(x). \tag{4.1}$$

We assume that the coefficients are sufficiently smooth and that the operator L is uniformly elliptic in Ω. More precisely,

[H1a] (smoothness condition 1) The coefficients $a_{ij}(x)$ are bounded and absolutely continuous in $\bar{\Omega}$, and the remaining coefficients are bounded and measurable in Ω.

[H1b] (uniform ellipticity) Define the principal part of L by

$$L^0(x,D) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(x)D_{ij} = D'A(x)D,$$

where $A(x)$ is an $n \times n$ positive definite matrix with components $a_{ij}(x)$. We assume that L is *uniformly elliptic in* Ω. That is, there exists a positive constant c_L such that for all $x \in \Omega$, $\xi \in \Re^n$,

$$L^0(x,\xi) \geq c_L|\xi|^2.$$

Our analysis is based on the boundary control system formulation. We shall no longer refer to the state-space realization. The boundary operator Γ is defined by

$$\Gamma(x,D) = b_0(x) + \sum_{i=1}^{n} b_{1i}(x)D_i = b_0(x) + B_1'(x)D, \tag{4.2}$$

where $B_1'(x) = \big(b_{11}(x),\ldots,b_{1n}(x)\big)$ and $D' = (D_1,\ldots,D_n)$. So $B_1'(x) = 0$ for Dirichlet boundary control and $b_0(x) = 0$ for Neumann boundary control. We impose the following condition on the operator Γ:

[H2] (smoothness condition 2) The coefficients of Γ are real. Also, $b_0(x) \in C^2(\partial\Omega)$ and $b_{1i}(x) \in C^1(\partial\Omega)$ for $i = 1,\ldots,n$.

Estimates of the solution to a uniformly elliptic boundary value problem depend on regularity of the region Ω.

DEFINITION 4.1 (see [4]). *Let Ω be an open set in $\Re^n$ with boundary $\partial\Omega$. Then Ω is said to be* uniformly regular of class C^m *if there exists a family of open sets $\{O_i\}$ of $\Re^n$ and of homeomorphisms $\{\Phi_i\}$ of O_i onto the unit ball $\{y : \|y\| < 1\}$ in $\Re^n$ and an integer N such that the following conditions are satisfied:*

[UR1] *For each i,*

$$\Phi_i(O_i \cap \Omega) = \{y \ : \ \|y\| < 1, y_1 > 0\},$$
$$\Phi_i(O_i \cap \partial\Omega) = \{y \ : \ \|y\| < 1, y_1 = 0\}.$$

[UR2] *Let $O_i' = \Phi_i^{-1}(\{y \in \Re^n : \|y\| < 1/2\})$. Then $\bigcup_{i=1}^{\infty} O_i'$ contains the $1/N$ neighborhood of $\partial\Omega$.*

[UR3] *Any $(N+1)$ distinct sets of $\{O_i\}$ have an empty intersection.*

[UR4] *Let $\Psi_i = \Phi_i^{-1}$. Then Ψ_i, Φ_i are mappings of class C^m. Let Φ_{ik}, Ψ_{ik} be the kth components of Φ_i, Ψ_i, respectively. Then*

$$|D^\alpha \Phi_{ik}(x)| \leq M, \quad |D^\alpha \Psi_{ik}(y)| \leq M, \quad |\Phi_{i1}(x)| \leq M\mathrm{dist}(x,\partial\Omega)$$

for $|\alpha| \leq m$, $x \in O_i$, $\|y\| < 1, k = 1,\ldots,n$, and $i = 1,2,\ldots$.

In general, it is nontrivial to show that a region is uniformly regular of class C^m. For our work, we are concerned only with bounded sets Ω in $\Re^n$ and cylinders of the form $\Omega \times \Re$ in $\Re^{n+1}$. It was stated without details in [22, p. 237] that for bounded sets with sufficiently smooth boundary, there exist mappings $\{\Phi_i\}$ such that [UR2] holds. We give a more complete discussion of this point. If Ω is bounded, then there

is a finite open cover for the boundary. If the boundary is sufficiently smooth, then it is possible to choose a covering such that [UR1] and [UR2] hold. Conditions [UR3] and [UR4] then hold trivially since the covering is finite. Thus we have the following result.

THEOREM 4.2. *If Ω is bounded with sufficiently smooth boundary, then $\Omega \times \Re$ is also uniformly regular.*

In addition to [H1a], [H1b], and [H2], we assume, unless stated otherwise, that Ω, L, and Γ also satisfy the following:

[H3] Ω is bounded and uniformly regular of class C^2.

[H4] (root condition) Let $L^0(x, D)$ denote the principal part of $L(x, D)$. For every pair of linearly independent real vectors ξ and η, the polynomial $L^0(x, \xi+\tau\eta)$ in τ has an equal number of roots with positive and negative imaginary parts.

[H5] (complementing condition) Let $B^0(x,D)$ denote the principal part of $\Gamma(x,D)$. Let x be an arbitrary point on $\partial\Omega$ and n be the outward normal unit vector to $\partial\Omega$ at x. For each tangential vector $\xi \neq 0$ to $\partial\Omega$ at x, let $\hat{\tau}$ be the root of the polynomial $L^0(x, \xi + \tau n)$ with positive imaginary part. Then $\hat{\tau}$ is not a root of $B^0(x, \xi + \tau n)$.

If $n \geq 3$, then the root condition is satisfied for all uniformly elliptic operators [21, p. 130]. If the coefficients of L are real, then the root condition is also satisfied when $n = 2$.

5. Uniformly elliptic operators with Dirichlet boundary control. It is well known that the one-dimensional heat equation on a unit interval with Dirichlet boundary control and point observations is not well-posed with respect to the usual choice of state space $L^2(0, 1)$ [6]. Thus, showing well-posedness of more general Dirichlet control problems with state-space methods is hampered by the difficulty of first obtaining an appropriate state space.

In this section we will show that a class of control problems with Dirichlet boundary control do have a bounded input/output map by showing that the associated elliptic problem is uniformly bounded and hence the transfer function is proper.

Let $\Omega \subset \Re^n$, $n = 1, 2, 3$, let L be a second order differential operator as defined in (4.1) with $d(x) \leq 0$, and define the boundary operator to be

$$\Gamma(x, D) = b_0(x), \quad b_0(x) \neq 0 \qquad \text{for all } x.$$

We shall show that if Ω, L, Γ satisfy hypotheses [H1]–[H5] and Ω satisfies an additional assumption, then the solution to the abstract elliptic problem

$$\left.\begin{array}{rcll} Lz & = & sz & \text{in } \Omega, \\ \Gamma z & = & u & \text{on } \partial\Omega \end{array}\right\} \tag{5.1}$$

is uniformly bounded with respect to the $\sup_{x\in\Omega} |\cdot|$ norm. This will imply boundedness of the input/output map for the corresponding boundary control system. The following definition is due to Browder [4].

DEFINITION 5.1. *Let Ω be an open set in $\Re^n$. If for any $a \in \partial\Omega$ the part of Ω, $\partial\Omega$ in some neighborhood of a is expressed as*

$$x_i > \psi(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n), \qquad x_i = \psi(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n),$$

respectively, for some $i = 1, \ldots, n$ and a C^{2m} function ψ, then Ω is called locally regular of class C^{2m}.

In addition to uniformly regularity of class C^2, we further assume the following:

[H6] Ω is locally regular of class C^4.

We will use Proposition 3.4 to show that the solution to the elliptic Dirichlet problem is uniformly bounded in the $C(\Omega)$-norm. The following result will be used to show that the solution to the subproblem $(L,\Gamma)_{e2}$ satisfies the second condition in Proposition 3.4.

THEOREM 5.2 (see [21, p. 216]). *Let $F \in C(\bar{\Omega})$ and consider*

$$\begin{aligned} Lw &= sw + F && \text{in } \Omega, \\ \Gamma w &= 0 && \text{on } \partial\Omega. \end{aligned}$$

The solution w exists, and $w \in C(\Omega)$. Furthermore, we have

$$\sup_{x\in\Omega} |w(x)| \le \frac{C}{|s|} \sup_{x\in\Omega} |F(x)|. \tag{5.2}$$

To prove boundedness of the input/output map we also require the maximum principle and existence of a solution to $Lf = 0$ with a Dirichlet boundary condition. This will be used to show that the first condition in Proposition 3.4 holds.

The following theorem is an immediate consequence of Theorems 8.6 and 8.12 in [13]. (The assumptions imposed on L and Ω in [13] are weaker than [H1]–[H6].)

THEOREM 5.3. *Let L and Ω satisfy assumptions* [H1]–[H6], *$\mu \in \Re^+$, $\mu \notin \sigma(L)$ be fixed, and $u \in H^2(\Omega)$; then there exists a unique $f \in H^2(\Omega)$ that solves*

$$\begin{aligned} Lf &= \mu f && \text{in } \Omega, \\ f &= u && \text{on } \partial\Omega. \end{aligned} \tag{5.3}$$

Proof. For $u \in H^1(\Omega)$, Theorem 8.6 in [13] guarantees that (5.3) has a unique (weak) solution $f \in H^1(\Omega)$. Since [H1]–[H3] hold and $u \in H^2(\Omega)$, by Theorem 8.12 in [13] the solution is in $H^2(\Omega)$. ∎

The norm $[\cdot]_{q-1/2,\partial\Omega}$ is defined by

$$[u]_{q-1/2,\partial\Omega} = \inf\{\|z\|_{H^q(\Omega)}; z \in H^q(\Omega), z = u \text{ on } \partial\Omega\}. \tag{5.4}$$

The space $H^{q-\frac{1}{2}}(\partial\Omega)$ is the space of functions defined on $\partial\Omega$ such that this norm is finite. For $u \in H^{q-\frac{1}{2}}(\partial\Omega)$, u may be extended to $\tilde{u} \in H^q(\Omega)$ such that $\tilde{u}|_{\partial\Omega} = u$ and $\|\tilde{u}\|_{H^q(\Omega)} = [u]_{q-1/2,\partial\Omega}$.

COROLLARY 5.4. *Let L and Ω satisfy assumptions* [H1]–[H6]. *For any $\mu \in \Re^+$, $\mu \notin \sigma(L)$, and $u \in H^{\frac{3}{2}}(\partial\Omega)$, there exists a unique $f \in H^2(\Omega)$ that solves*

$$\begin{aligned} Lf &= \mu f && \text{in } \Omega, \\ b_0(x) f &= u && \text{on } \partial\Omega. \end{aligned} \tag{5.5}$$

Proof. Since $b_0(x) \in C^2(\partial\Omega)$ and $b_0(x) \neq 0$ for all $x \in \partial\Omega$, we have $\tilde{u} = \frac{u}{b_0} \in H^{\frac{3}{2}}(\partial\Omega)$. Thus it can be extended to an element in $H^2(\Omega)$ which we shall denote by the same symbol. By Theorem 5.3 there exists a unique $f \in H^2(\Omega)$ that solves

$$\begin{aligned} Lf &= \mu f && \text{in } \Omega, \\ f &= \tilde{u} && \text{on } \partial\Omega. \end{aligned}$$ ∎

The following maximum principle is required. The stated assumptions are stronger than those given in [13].

THEOREM 5.5 (see, e.g., [13, Thm. 8.1]). *Let $f \in H^2(\Omega)$ satisfy $Lf - \mu f = 0$ in Ω. Then*

$$\sup_{x\in\Omega} f(x) \le \sup_{x\in\partial\Omega} \max\{f(x), 0\}.$$

We can now state our main theorem for this section. It implies in particular that Dirichlet boundary control with point observation is a well-posed control system.

THEOREM 5.6. *Consider the pair L, Γ with Dirichlet control $\Gamma = b_0(x)$. Assume that assumptions* [H1]–[H6] *are satisfied on the region Ω. The operators L, Γ define a boundary control system with $\mathcal{U} = H^{\frac{3}{2}}(\partial\Omega)$, $\mathcal{Z} = H^2(\Omega)$, and $\mathcal{H} = L^2(\Omega)$. The input/output map of the boundary control system* (5.1) *is bounded for all observation operators $K \in \mathcal{L}(C(\Omega), \mathcal{Y})$.*

Proof. Let $\mu \in \Re^+$ and $\mu \notin \sigma(L)$, and write (L, Γ) as $(L, \Gamma)_{e1}$ and $(L, \Gamma)_{e2}$ as in Proposition 3.4. We will use $\mathcal{V} = C(\Omega)$. Since $\Omega \subset R^n$, $n \le 3$, Sobolev's imbedding theorem, e.g., [21, Thm. 3.20], implies that $\mathcal{V} \subset \mathcal{Z}$. Define $C_1 = \sup_{x\in\partial\Omega} \frac{1}{|b_0(x)|}$. Using Theorem 5.5, we have

$$\begin{aligned}\sup_{x\in\Omega} |f(x)| &\le \sup_{x\in\partial\Omega} |f(x)| \\ &\le C_1 \sup_{x\in\partial\Omega} |u(x)| \\ &\le C_1 \|u\|_{H^{\frac{3}{2}}}.\end{aligned}$$

The latter inequality also follows from Sobolev's imbedding theorem. Thus, the solution to the subproblem $(L, \Gamma)_{e1}$ satisfies inequality (3.4). Inequality (3.5) then follows from inequality (5.2). Therefore, by Theorem 3.2 the system transfer function associated with (L, Γ, K) is proper for all observation operators $K \in \mathcal{L}(C(\Omega), \mathcal{Y})$. That is, the input/output map of the boundary control system (L, Γ, K) is bounded. ∎

6. Uniformly elliptic operators with Neumann or Robin boundary control. In this section we will show that a class of control problems with Neumann/Robin boundary control have a bounded input/output map. In the interests of clarity and brevity we will give only the proofs for second order elliptic operators. The generalization to higher order operators is straightforward. Details are in [5].

In special cases results have been obtained to these problems by transforming the boundary control system to state-space form and then using the analyticity of the underlying semigroup to show well-posedness of the input/output map. The transformation to state-space form is not necessary. As for Dirichlet problems, well-posedness for general Neumann problems is shown by direct analysis of the boundary control formulation.

Let L and Γ be defined as in (4.1) and (4.2). In this section we will assume $B_1'(x) \neq 0$. Hence Γ represents a Neumann boundary control when $b_0(x) = 0$ and a Robin boundary control otherwise. We shall show that if Ω, L, and Γ satisfy hypotheses [H1]–[H5], then the solution to the abstract elliptic problem is uniformly bounded with respect to the $H^1(\Omega)$ norm. This implies boundedness of the input/output map for the corresponding boundary control system.

It is not enough to use regularity of the solution to elliptic problems. We must show that the solution is uniformly bounded in the parameter s. We first state two theorems concerning estimates of solutions to elliptic problems. These theorems are key to showing uniform boundedness of solutions to Neumann/Robin boundary control problems.

THEOREM 6.1 (see [21, Thm. 4.10]). *Let Ω be uniformly regular of class C^2 and $L(x,D)$, $B(x,D)$ be defined as in* (4.1) *and* (4.2). Assume that $L(x,D)$ and $\Gamma(x,D)$ satisfy assumptions [H2]–[H5]. Then there exists a positive constant m_1 such that for all $z \in H^2(\Omega)$ the following inequality holds:

$$\|z\|_{H^2(\Omega)} \leq m_1 \left[\|Lz\|_{L^2(\Omega)} + [\Gamma z]_{1/2,\partial\Omega} + \|z\|_{L^2(\Omega)} \right]. \tag{6.1}$$

THEOREM 6.2 (see [21, Lem. 5.7]). *Let L, Γ and Ω be as defined in* (4.1) *and* (4.2)*, and assume that they satisfy assumptions* [H1]–[H5]. *Let $\theta \in [-\pi, \pi)$ be fixed but arbitrary and t be a new real variable. Set*

$$Q = \Omega \times \Re,$$
$$\mathcal{L}_\theta(x,D) = \mathcal{L}_\theta(x,D_x,D_t) = L(x,D_x) + \exp(i\theta) D_t^2,$$

and define $\mathcal{B}(x,D_x)$ to be the extension of $\Gamma(x,D_x)$ to $\partial Q = \partial\Omega \times \Re$. If $\mathcal{L}_\theta, \mathcal{B}, Q$ also satisfy [H1]–[H5]*, then there exists a constant M_θ such that for any $z \in H^2(\Omega)$, $u \in H^{2-m_j}(\Omega)$[1] satisfying $\Gamma z = u$ on $\partial\Omega$ and any s satisfying* arg $s = \theta$, $|s| > M_\theta$, *the following inequality holds:*

$$|s|^{1/2} \|z\|_{H^1(\Omega)} + \|z\|_{H^2(\Omega)} \leq M_\theta \left[\|(L-s)z\|_{L^2(\Omega)} + |s|^{1-m_j/2} \|u\|_{L^2(\Omega)} + \|u\|_{H^{2-m_j}(\Omega)} \right]. \tag{6.2}$$

The outline of the proof is as follows: For any $\theta \in [-\pi, \pi)$, define $Q, \mathcal{L}_\theta$, and $\mathcal{B}$ by

$$\left.\begin{array}{rcl} Q & := & \Omega \times \Re, \\ \mathcal{L}_\theta(x,D) = \mathcal{L}_\theta(x,D_x,D_t) & := & L(x,D_x) + \exp(i\theta) D_t^2, \text{ and} \\ \mathcal{B}(x,D_x) & := & \text{the extension of } \Gamma(x,D_x) \text{ to } \partial Q = \partial\Omega \times \Re. \end{array}\right\} \tag{6.3}$$

From Theorem 6.2 we know that if $\{L, \Gamma, \Omega\}$ and $\{\mathcal{L}_\theta, \mathcal{B}, Q\}$ both satisfy [H1]–[H5], then there exists a constant M_θ such that the following a priori estimate holds for any $z \in H^2(\Omega)$, $u \in H^1(\Omega)$ satisfying $\Gamma z = u$ on $\partial\Omega$ and any s satisfying arg $s = \theta$, $|s| > M_\theta$, $\theta \in [-\pi, \pi)$:

$$|s|^{1/2} \|z\|_{H^1(\Omega)} + \|z\|_{H^2(\Omega)} \leq M_\theta \left[\|(L-s)z\|_{L^2(\Omega)} + |s|^{1/2} \|u\|_{L^2(\Omega)} + \|u\|_{H^1(\Omega)} \right].$$

If z solves $Lz = sz$, then

$$\|z\|_{H^1(\Omega)} \leq M_\theta \left(\|u\|_{L^2(\Omega)} + \frac{1}{|s|^{1/2}} \|u\|_{H^1(\Omega)} \right).$$

If in addition $|s| > 1$, then

$$\|z\|_{H^1(\Omega)} \leq 2M_\theta \|u\|_{H^1(\Omega)}.$$

We will show that for $\theta \in [-\pi/2, \pi/2]$, M_θ can be chosen independently of θ. This will imply that the solution to the elliptic problem is uniformly bounded with respect to the H^1-norm and thus the input/output map is bounded for any observation operator $K \in \mathcal{L}(\mathcal{H}^1(\Omega), \mathcal{Y})$.

[1] m_j=0 if Γ is the Dirichlet boundary condition, and $m_j = 1$ if Γ is a Neumann or Robin boundary condition.

First we show that Q is uniformly regular of class C^2 and for each $\theta \in [-\pi/2, \pi/2]$, $\mathcal{L}_\theta, \mathcal{B}, Q$ satisfy assumptions [H1], [H2], [H4], and [H5]. This ensures the existence of M_θ.

LEMMA 6.3. *Let $L(x, D_x), \Gamma(x, D_x)$, and Ω satisfy assumptions* [H1]–[H5]. *For any $\theta \in [-\pi/2, \pi/2]$, define $\mathcal{L}_\theta, \mathcal{B}$, and Q be as in* (6.3). *Then Q is uniformly regular of class C^2 and $\{\mathcal{L}_\theta, \mathcal{B}\}$ satisfy assumptions* [H1], [H2], [H4], *and* [H5] *in Q.*

Proof. Since Ω satisfies [H3], Q is uniformly regular. Next we show that $\mathcal{L}_\theta$ is uniformly elliptic. That is, there exists a positive constant c_1 such that for all $(\xi, \eta) \in \Re^n \times \Re$ and $x \in \Omega$ the following inequality holds:

$$|\mathcal{L}_\theta^0(x, \xi, \eta)| \geq c_1 \left(|\xi|^2 + \eta^2\right).$$

By assumption, there exists a positive constant c_L such that for all $x \in \Omega$, $\xi \in \Re^n$

$$|L^0(x, \xi)| \geq c_L |\xi|^2.$$

Since the matrix A associated with L^0 is positive definite, this means $L^0(x, \xi) \geq 0$ for all $x \in \Omega$ and $\xi \in \Re^n$. Let $c = \min\{c_L^2, 1\}$. Then for any $(x, t) \in \Omega \times \Re$, $(\xi, \eta) \in \Re^n \times \Re$, and $\theta \in [-\pi/2, \pi/2]$, we have

$$\begin{aligned}
\left|\mathcal{L}_\theta^0\big((x, t), (\xi, \eta)\big)\right|^2 &= |L^0(x, \xi) + \exp(i\theta)\eta^2|^2 \\
&= |L^0(x, \xi)|^2 + 2\cos(\theta) L^0(x, \xi)\eta^2 + \eta^4 \\
&\geq c_L^2 |\xi|^4 + \eta^4 \\
&\geq c\left(|\xi|^4 + \eta^4\right) \\
&\geq \frac{c}{2}\left(|\xi|^4 + 2|\xi|^2\eta^2 + \eta^4\right) \\
&= \frac{c}{2}\left(|\xi|^2 + \eta^2\right)^2.
\end{aligned}$$

This implies the inequality

$$|\mathcal{L}_\theta^0(x, \xi, \eta)| \geq \sqrt{\frac{c}{2}}\left(|\xi|^2 + \eta^2\right),$$

which proves that $\mathcal{L}$ is uniformly elliptic in Q. Clearly [H2] holds. Also since $n \geq 2$, $n + 1 \geq 3$, the root condition holds. It remains to show that [H5] is satisfied. Let (x, t) be an arbitrary point on ∂Q, n_1 be the unit outward normal vector to $\partial\Omega$ at x, and ξ_1 be any nonzero tangential vector to $\partial\Omega$ at x. The outward normal unit vector to ∂Q at (x, t) is then $n = (n_1', 0)$ and any nonzero tangential vector has the form $\xi = (\xi_1', 0)$. Let $\hat{\tau}$ be a root of $\bar{B}^0(x, \xi + \tau n)$. Then $\hat{\tau}$ is a root of $B^0(x, \xi_1 + \tau n_1)$, which by assumption is not a root of $L^0(x, \xi_1 + \tau n_1)$. This implies that

$$\mathcal{L}(x, \xi + \hat{\tau} n) = L(x, \xi_1 + \hat{\tau} n_1) + \exp(i\theta)(\xi_2 + \hat{\tau} n_2)^2 = L(x, \xi_1 + \hat{\tau} n_1) \neq 0.$$

Hence $\hat{\tau}$ is not a root of $\mathcal{L}(x, \xi + \hat{\tau} n)$. So $\{\mathcal{L}, \mathcal{B}\}$ satisfies [H5]. ∎

For each $\theta \in [-\pi/2, \pi/2]$, $\mathcal{L}, \mathcal{B}, Q$ satisfy [H1], [H2], [H4], and [H5]; thus the hypotheses of Theorem 6.2 have been justified. It remains to show that M_θ may be chosen independent of θ in this range. The following lemma is needed to prove this claim.

LEMMA 6.4. *Let $\mathcal{L}_\theta(x, D)$ be defined as in* (4.1). *Then $\mathcal{L}_\theta$ is continuous in θ. That is, for any $\epsilon > 0$, there exists $\delta > 0$ such that whenever $|\theta_1 - \theta_2| < \delta$, $\theta_1, \theta_2 \in [-\pi/2, \pi/2]$ we have*

$$\|\mathcal{L}_{\theta_1} v - \mathcal{L}_{\theta_2} v\|_{L^2(Q)} < \epsilon \|v\|_{H^2(Q)} \qquad \text{for all } v \in H^2(Q).$$

Proof. For any $0 < \epsilon < \sqrt{2}$, choose $\delta = \arccos(1 - \frac{\epsilon^2}{2})$, where arccos denotes the principal branch; then if $|\theta_1 - \theta_2| < \delta$ and $\theta_1, \theta_2 \in [-\pi/2, \pi/2]$), we have

$$\begin{aligned}\|\mathcal{L}_{\theta_1} v - \mathcal{L}_{\theta_2} v\|_{L^2(Q)} &\leq |\exp(i\theta_1) - \exp(i\theta_2)| \|v\|_{H^2(Q)} \\ &= \sqrt{(2 - 2\cos(\theta_1 - \theta_2))} \|v\|_{H^2(Q)} \\ &= \sqrt{(2 - 2\cos(|\theta_1 - \theta_2|))} \|v\|_{H^2(Q)}.\end{aligned}$$

Since $\epsilon < \sqrt{2}$, $\delta < \pi/2$; hence the function $f(x) = 2 - 2\cos(x)$ is nonnegative and monotone increasing on the interval $[0, \delta]$. Thus

$$\begin{aligned}\|\mathcal{L}_{\theta_1} v - \mathcal{L}_{\theta_2} v\|_{L^2(Q)} &< \sqrt{(2 - 2\cos(\delta))} \|v\|_{H^2(Q)} \\ &= \epsilon \|v\|_{H^2(Q)}.\end{aligned}$$

For any $\epsilon \geq \sqrt{2}$, choose $\delta = \pi/2$; then if $|\theta_1 - \theta_2| < \pi/2$ and $\theta_1, \theta_2 \in [-\pi/2, \pi/2]$) we have

$$\begin{aligned}\|\mathcal{L}_{\theta_1} v - \mathcal{L}_{\theta_2} v\|_{L^2(Q)} &\leq \sqrt{(2 - 2\cos(|\theta_1 - \theta_2|))} \|v\|_{H^2(Q)} \\ &< \sqrt{2} \|v\|_{H^2(Q)} \\ &< \epsilon \|v\|_{H^2(Q)}. \qquad \square\end{aligned}$$

Due to Theorem 6.1, for each $\theta \in [-\pi/2, \pi/2]$, there exists a constant m_θ such that for any $v \in H^2(Q)$,

$$\|v\|_{H^2(Q)} \leq m_\theta \left(\|\mathcal{L}_\theta v\|_{L^2(Q)} + [\mathcal{B}v]_{0,\partial Q} + \|v\|_{L^2(Q)} \right). \tag{6.4}$$

For each θ, define $m(\theta) = \inf\{m_\theta : \text{inequality (6.4) holds}\}$. The infimum exists since clearly 1 is a lower bound for m_θ. The next theorem proves that $m(\theta)$ is bounded above.

THEOREM 6.5. *Let $m(\theta)$ be as defined above. Then $\{m(\theta); -\pi/2 \leq \theta \leq \pi/2\}$ is bounded above. Hence there exists a positive constant $\bar{m}$ such that the following inequality holds for all $\theta \in [-\pi/2, \pi/2]$:*

$$\|v\|_{H^2(Q)} \leq \bar{m} \left(\|\mathcal{L}_\theta v\|_{L^2(Q)} + [\mathcal{B}v]_{1/2,\partial Q} + \|v\|_{L^2(Q)} \right). \tag{6.5}$$

Proof. Suppose not. Then for each n, there exists $\theta_n \in [-\pi/2, \pi/2]$ such that $m(\theta_n) > n$. The sequence $\{\theta_n\}$ is bounded; thus it contains a convergent subsequence $\{\theta_{k_n}\}$ which converges to $\bar{\theta} \in [-\pi/2, \pi/2]$. Theorem 6.1 ensures that $m(\bar{\theta})$ is positive and finite; thus there exists some n such that $m(\bar{\theta}) < n$. Let $\epsilon = \frac{1}{m(\bar{\theta})} - \frac{1}{n} > 0$. By Lemma 6.4, there exists $N > n$ such that for all $k_n > N$ (k_n are the indices of the convergent subsequence),

$$\|\mathcal{L}_{\bar{\theta}} v - \mathcal{L}_{\theta_{k_n}} v\|_{L^2(Q)} < \epsilon \|v\|_{H^2(Q)} \qquad \text{for all } v \in H^2(Q).$$

Pick a k_n such that $m(\theta_{k_n}) - 1 > n$. By definition, $m(\theta_{k_n})$ is the smallest constant such that for all $v \in H^2(Q)$, inequality (6.4) holds. Thus there exists some $v_0 \in H^2(Q)$ such that

$$\|v_0\|_{H^2(Q)} > (m_{\theta_{k_n}} - 1) \left(\|\mathcal{L}_{\theta_{k_n}} v_0\|_{L^2(Q)} + [\mathcal{B}v_0]_{1/2,\partial Q} + \|v_0\|_{L^2(Q)} \right).$$

But then

$$\begin{aligned}\epsilon\|v_0\|_{H^2(Q)} &= \left(\frac{1}{m(\bar\theta)} - \frac{1}{n}\right)\|v_0\|_{H^2(Q)}\\ &< \left(\frac{1}{m(\bar\theta)} - \frac{1}{m(\theta_{k_n})-1}\right)\|v_0\|_{H^2(Q)}\\ &< \left(\|\mathcal{L}_{\bar\theta}v_0\|_{L^2(Q)} + [\mathcal{B}v_0]_{1/2,\partial Q} + \|v_0\|_{L^2(Q)}\right)\\ &\quad - \left(\|\mathcal{L}_{\theta_{k_n}}v_0\|_{L^2(Q)} + [\mathcal{B}v_0]_{1/2,\partial Q} + \|v_0\|_{L^2(Q)}\right)\\ &\le \|\mathcal{L}_{\bar\theta}v_0 - \mathcal{L}_{\theta_{k_n}}v_0\|_{L^2(Q)}\\ &< \epsilon\|v_0\|_{H^2(Q)},\end{aligned}$$

a contradiction. Thus $m(\theta)$ is bounded above. Let $\bar m = \sup\{m(\theta), -\pi/2 \le \theta \le \pi/2\}$. Then for any $\theta \in [-\pi/2, \pi/2]$ and $v \in H^2(Q)$, inequality (6.5) holds. ∎

We now state a modification of Theorem 6.2.

THEOREM 6.6. *Let Ω, L, Γ, (4.1), (4.2) define a boundary control system with $\mathcal{H} = L^2(\Omega)$ and $\mathcal{U} = H^{\frac{1}{2}}(\partial\Omega)$. Assume that* [H1]–[H5] *are satisfied. Then there exists a positive constant R such that for any $z \in H^2(\Omega)$, $u \in \mathcal{U}$ satisfying $\Gamma z = u$ on $\partial\Omega$ and any complex number s on the open right half-plane $\mathbb{C}_{R^2} := \{s : \operatorname{Re} s > R^2\}$, the following inequality holds:*

$$|s|^{1/2}\|z\|_{H^1(\Omega)} + \|z\|_{H^2(\Omega)} \le m\left[\|(L-s)z\|_{L^2(\Omega)} + |s|^{1/2}\|u\|_{L^2(\Omega)} + \|u\|_{H^1(\Omega)}\right], \tag{6.6}$$

where m is a positive constant dependent only on L and Ω.

Proof. The proof is along the lines given in [21] except that we show that the constant is independent of θ. Let ζ be a function in $C^\infty(-\infty,\infty)$ such that $\zeta(t) = 0$ for $|t| > 1$, $\zeta(t) = 1$ for $|t| < 1/2$. Let m_1 be a constant chosen such that $\|\zeta\|_{H^2(\Re)} \le m_1$. Let $\bar m = \max\{m(\theta), -\pi/2 \le \theta \le \pi/2\}$ and $m_2 = \max\{\bar m, m_1\}$. Define

$$R := \text{largest root of the quadratic } r^2 - 6m_2^2 r - 6m_2^2.$$

We note that R is necessarily positive and real. In fact $R = \frac{6m_2^2 + m_2\sqrt{36m_2^2+24}}{2}$. Moreover, since $m(\theta)$ is bounded below by 1, $\bar m$ and hence m_2 is always greater than 1. Thus $R > 6$. For any $z \in H^2(\Omega)$ and any $s \in \mathbb{C}_{R^2}$, set $\theta = \arg s$, $r = |s|^{1/2}$, and $v(x,t) = \zeta(t)\exp(irt)z(x)$. Clearly $v \in H^2(Q)$; hence (6.5) implies

$$\begin{aligned}\|v\|_{H^2(Q)} &\le \bar m\left(\|\mathcal{L}_\theta v\|_{L^2(Q)} + [\mathcal{B}v]_{1/2,\partial Q} + \|v\|_{L^2(Q)}\right)\\ &\le m_2\left(\|\mathcal{L}_\theta v\|_{L^2(Q)} + [\mathcal{B}v]_{1/2,\partial Q} + \|v\|_{L^2(Q)}\right).\end{aligned} \tag{6.7}$$

Now a lower bound for $\|v\|_{H^2(Q)}$, an upper bound for $[\mathcal{B}v]_{1/2,\partial Q}$, and an upper bound for $\|\mathcal{L}_\theta v\|_{L^2(Q)}$ need to be computed. The final inequality is then obtained via simple algebra. First we compute a lower bound for $\|v\|_{H^2(Q)}$. By definition of $\|\cdot\|_{H^2(Q)}$ we have

$$\begin{aligned}\|v\|^2_{H^2(Q)} &= \sum_{|\alpha|+k\le 2}\int_{-\infty}^{\infty}\int_\Omega \left|D_x^\alpha D_t^k v(x,t)\right|^2 dxdt\\ &\ge \sum_{|\alpha|+k\le 2}\int_{-\frac{1}{2}}^{\frac{1}{2}}\int_\Omega \left|D_x^\alpha D_t^k \exp(irt)z(x)\right|^2 dxdt\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=0}^{2} (r)^{2k} \sum_{|\alpha|+k\leq 2} \int_{\Omega} \left| D_x^{\alpha} z(x) \right|^2 dx \\
&= \sum_{k=0}^{2} (r)^{2k} \left\| z \right\|^2_{H^{2-k}(\Omega)} \\
&\geq (r)^{2k} \left\| z \right\|^2_{H^{2-k}(\Omega)}
\end{aligned}
$$

for any $k = 0, 1, 2$. Hence

$$
\left\| v \right\|_{H^2(Q)} \geq (r)^k \left\| z \right\|_{H^{2-k}(\Omega)}
$$

for any $k = 0, 1, 2$. Thus

$$
3 \left\| v \right\|_{H^2(Q)} \geq \sum_{k=0}^{2} (r)^k \left\| z \right\|_{H^{2-k}(\Omega)} . \tag{6.8}
$$

Next we compute an upper bound for $[\mathcal{B}v]_{1/2,\partial Q}$. By definition of $[\cdot]_{1/2,\partial\Omega}$ we have for $\Gamma z \in H^2(\Omega)$ such that $z = u$ on $\partial\Omega$, and

$$
\begin{aligned}
[\mathcal{B}v]^2_{1/2,\partial Q} &= [\zeta(t) \exp(irt) B z(x)]^2_{1/2,\partial Q} \\
&= [\zeta(t) \exp(irt) u]^2_{1/2,\partial Q} \\
&\leq \left\| \zeta(t) \exp(irt) u \right\|^2_{H^1(Q)} \\
&= \sum_{|\alpha|+k\leq 1} \int_{-\infty}^{\infty} \int_{\Omega} \left| D_x^{\alpha} D_t^k \zeta(t) \exp(irt) u \right|^2 dx dt \\
&= \int_{-\infty}^{\infty} \int_{\Omega} |\zeta(t) \exp(irt) u|^2 \, dx dt + \int_{-\infty}^{\infty} \int_{\Omega} |\zeta(t) \exp(irt) Du|^2 \, dx dt \\
&\quad + \int_{-\infty}^{\infty} \int_{\Omega} |\zeta'(t) \exp(irt) u + ir\zeta(t) \exp(irt) u|^2 \, dx dt \\
&\leq m_1^2 \left\| u \right\|^2_{L^2(\Omega)} + m_1^2 \left\| Du \right\|^2_{L^2(\Omega)} + m_1^2 \left\| u \right\|^2_{L^2(\Omega)} + 2rm_1^2 \left\| u \right\|^2_{L^2(\Omega)} \\
&\quad + r^2 m_1^2 \left\| u \right\|^2_{L^2(\Omega)} \\
&= 2m_1^2 \left\| u \right\|^2_{L^2(\Omega)} + m_1^2 \left\| Du \right\|^2_{L^2(\Omega)} + (2r + r^2) m_1^2 \left\| u \right\|^2_{L^2(\Omega)} .
\end{aligned}
$$

Since $r = |s|^{1/2} > R > 6$, $2r < r^2$. Hence

$$
\begin{aligned}
[\mathcal{B}v]^2_{1/2,\partial Q} &\leq 2m_1^2 \left(\left\| u \right\|^2_{H^1(\Omega)} + r^2 \left\| u \right\|^2_{L^2(\Omega)} \right) \\
&\leq 2m_1^2 \left(r \left\| u \right\|_{L^2(\Omega)} + \left\| u \right\|_{H^1(\Omega)} \right)^2 \\
&\leq 2m_2^2 \left(r \left\| u \right\|_{L^2(\Omega)} + \left\| u \right\|_{H^1(\Omega)} \right)^2 .
\end{aligned}
$$

Thus

$$
[\mathcal{B}v]_{1/2,\partial Q} \leq \sqrt{2} m_2 \left(r \left\| u \right\|_{L^2(\Omega)} + \left\| u \right\|_{H^1(\Omega)} \right) . \tag{6.9}
$$

This is the upper bound on $[\mathcal{B}v]_{1/2,\partial Q}$. Now we calculate an upper bound on $\mathcal{L}_\theta v$. Substituting the expression for $v(x,t)$ into $\mathcal{L}_\theta v$, we find

$$
\mathcal{L}_\theta v = \zeta(t) \exp(irt)(L - r^2 \exp(i\theta)) z + 2ir \exp(i\theta) \zeta'(t) \exp(irt) z + \exp(i\theta) \zeta''(t) \exp(irt) z .
$$

Therefore

$$\begin{aligned}\|\mathcal{L}_\theta v\|_{L^2(Q)} &\le \left\|\zeta(t)\exp(irt)(L-r^2\exp(i\theta))z\right\|_{L^2(Q)} + 2\left\|r\exp(i\theta)\zeta'(t)\exp(irt)z\right\|_{L^2(Q)} \\ &\quad + \left\|\exp(i\theta)\zeta''(t)\exp(irt)z\right\|_{L^2(Q)} \\ &\le m_1\left(\left\|(L-r^2\exp(i\theta))z\right\|_{L^2(\Omega)} + 2r\|z\|_{L^2(\Omega)} + \|z\|_{L^2(\Omega)}\right) \\ &\le m_2\left(\left\|(L-r^2\exp(i\theta))z\right\|_{L^2(\Omega)} + 2r\|z\|_{L^2(\Omega)} + \|z\|_{L^2(\Omega)}\right). \qquad (6.10)\end{aligned}$$

Also,

$$\|v\|_{L^2(Q)} \le m_2\|z\|_{L^2(\Omega)}. \qquad (6.11)$$

Substituting inequality (6.8) into (6.7), we obtain

$$r^2\|z\|_{L^2(\Omega)} + r\|z\|_{H^1(\Omega)} + \|z\|_{H^2(\Omega)} \le 3m_2\left(\|\mathcal{L}_\theta v\|_{L^2(Q)} + [\mathcal{B}v]_{1/2,\partial Q} + \|v\|_{L^2(Q)}\right). \qquad (6.12)$$

Next, substitute inequalities (6.9), (6.10), and (6.11) into inequality (6.12) to obtain

$$\begin{aligned}r^2\|z\|_{L^2(\Omega)} &+ r\|z\|_{H^1(\Omega)} + \|z\|_{H^2(\Omega)} \\ &\le 3m_2^2\left(\left\|(L-r^2\exp(i\theta))z\right\|_{L^2(\Omega)} + 2r\|z\|_{L^2(\Omega)} + \|z\|_{L^2(\Omega)}\right. \\ &\qquad \left. + \sqrt{2}r\|u\|_{L^2(\Omega)} + \sqrt{2}\|u\|_{H^1(\Omega)} + \|z\|_{L^2(\Omega)}\right). \qquad (6.13)\end{aligned}$$

After rearrangement we obtain

$$\begin{aligned}(r^2 - 6m_2^2 r - 6m_2^2)\|z\|_{L^2(\Omega)} &+ r\|z\|_{H^1(\Omega)} + \|z\|_{H^2(\Omega)} \\ &\le 3\sqrt{2}m_2^2\left(\left\|(L-r^2\exp(i\theta))z\right\|_{L^2(\Omega)} + r\|u\|_{L^2(\Omega)} + \|u\|_{H^1(\Omega)}\right). \qquad (6.14)\end{aligned}$$

By definition of R we have $r^2 - 6m_2^2 r - 6m_2^2 \ge 0$. Hence (6.14) implies

$$r\|z\|_{H^1(\Omega)} + \|z\|_{H^2(\Omega)} \le 3\sqrt{2}m_2^2\left(\left\|(L-r^2\exp(i\theta))z\right\|_{L^2(\Omega)} + r\|u\|_{L^2(\Omega)} + \|u\|_{H^1(\Omega)}\right). \qquad (6.15)$$

Substituting back $s = r^2\exp(i\theta)$ above and defining $m = 3\sqrt{2}m_2^2$, we have the desired result. ∎

The boundedness of the input/output map for Neumann boundary control with observation now follows.

COROLLARY 6.7. *The input/output map of the boundary control system is bounded for all observation operators $K \in \mathcal{L}(H^1(\Omega), \mathcal{Y})$.*

Proof. By Theorem 6.6, the solution to the abstract elliptic problem (L, Γ) is uniformly bounded with respect to the $H^1(\Omega)$ norm. Hence by Theorem 3.2, the system transfer function associated with (L, Γ, K) is proper for all observation operators $K \in \mathcal{L}(H^1(\Omega), \mathcal{Y})$. Thus by Theorem 2.3 the input/output map is bounded to the boundary control system (L, Γ, K). ∎

Remark 6.8. The main result above is stated for a control space $\mathcal{U} = H^{\frac{1}{2}}(\partial\Omega)$. This space can be regarded as the traces of functions in $H^1(\Omega)$ (5.4). Consider the following characterization of these functions.

THEOREM 6.9 (see, e.g., [17, sect. 1.1.3]). *If a function u defined on Ω is absolutely continuous on almost all straight lines that are parallel to coordinate axes and the first classical derivatives of u belong to $L_2(\Omega)$, then $u \in H^1(\Omega)$.*

Thus, $H^{\frac{1}{2}}(\Omega)$ includes piecewise continuous functions, provided that Ω is such that we can extend u into the interior so that it satisfies the above theorem. The singularities on the boundary of Ω remain.

Remark 6.10. If Γ is Dirichlet boundary control, then $m_j = 0$ in Theorem 6.2. Using the same technique as Theorem 6.6 we can show that there exists a positive constant R such that for any $z \in H^2(\Omega)$, $u \in H^2(\Omega)$ satisfying $\Gamma z = u$ on $\partial\Omega$, and any complex number s on the open right half-plane $\mathbb{C}_{R^2} := \{s : \text{Re } s > R^2\}$, the following inequality holds:

$$|s|^{1/2} \|z\|_{H^1(\Omega)} + \|z\|_{H^2(\Omega)} \leq m \left[\|(L-s)z\|_{L^2(\Omega)} + |s| \|u\|_{L^2(\Omega)} + \|u\|_{H^2(\Omega)} \right],$$

where m is a positive constant dependent only on L and Ω. Unfortunately this implies the solution to $Lz = sz$ in Ω and $\Gamma z = u$ on $\partial\Omega$ satisfies only

$$\|z\|_{H^1(\Omega)} \leq m|s|^{1/2} \|u\|_{H^2(\Omega)}.$$

So we cannot conclude that the solution is uniformly bounded in the H^1-norm. In the case of Dirichlet boundary control on a one-dimensional rod, it can easily be shown that the solution to the elliptic problem is not uniformly bounded in the H^1-norm.

7. Conclusions. The input/output map and the transfer function are well defined for abstract boundary control systems. We showed that the question of continuity of the input/output map can be transformed to boundedness of solutions to a related elliptic problem. It is not necessary to construct a state-space realization.

This approach enabled us to show boundedness of the input/output map for general classes of boundary control systems involving uniformly elliptic operators with Dirichlet, Neumann, or Robin boundary control.

We are currently working on extending our approach to problems that are second order in time.

REFERENCES

[1] G. Avalos, I. Lasiecka, and R. Rebarber, *Lack of time-delay robustness for stabilization of a structural acoustics model*, SIAM J. Control Optim., 37 (1999), pp. 1394–1418.

[2] G. Avalos, I. Lasiecka, and R. Rebarber, *Well-posedness of a structural acoustics control model with observation of the pressure*, J. Differential Equations, 173 (2000), pp. 1057–1072.

[3] H.T. Banks and K.A. Morris, *Input/output stability for accelerometer control systems*, Control Theory Adv. Tech., 10 (1994), pp. 1–17.

[4] F.E. Browder, *On the spectral theory of elliptic differential operators* I, Math. Anal., 142 (1961), pp. 22–130.

[5] A. Cheng, *Well-Posedness of Boundary Control Systems*, Ph.D. Thesis, University of Waterloo, Waterloo, Ontario, Canada, 2000.

[6] R.F. Curtain and G. Weiss, *Well-posedness of triples of operators (in the sense of linear systems theory)*, Internat. Ser. Numer. Math., 91 (1989), pp. 41–59.

[7] R.F. Curtain, *Well-Posedness of Infinite-Dimensional Linear Systems in Time and Frequency Domain*, Report TW-287, Mathematics Institute, University of Groningen, Groningen, The Netherlands, 1988.

[8] P. Grabowski and F.M. Callier, *Admissible observation operators. Duality of observation and control using factorizations*, Dynam. Contin. Discrete Impuls. Systems, 6 (1999), pp. 81–119.

[9] P. Grabowski and F.M. Callier, *Boundary control systems in factor form: Transfer functions and input-output maps*, Integral Equations Operator Theory, 41 (2001), pp. 1–37.

[10] P. Grabowski, *On the spectral-Lyapunov approach to parametric optimization of distributed parameter systems*, IMA J. Math. Control Inform., 7 (1990), pp. 317–338.

[11] P. Grabowski, *Admissibility of observation functionals*, Internat. J. Control, 62 (1995), pp. 1163–1173.
[12] P. Grabowski and F.M. Callier, *Admissibility of observation operators: Semigroup criteria for admissibility*, Integral Equations Operator Theory, 25 (1996), pp. 183–196.
[13] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1977.
[14] B. Jacob and H.J. Zwart, *Properties of the realization of inner functions*, Math. Control. Signal Systems, 15 (2002), pp. 356–379.
[15] I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations: Continuous and Approximation Theories*. I., Cambridge University Press, Cambridge, UK, 2000.
[16] I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations*: *Continuous and Approximation Theories*. II., Cambridge University Press, Cambridge, UK, 2000.
[17] V.G. Maz'ja, *Sobolev Spaces*, Springer-Verlag, Berlin, 1985.
[18] K.A. Morris, *The well-posedness of accelerometer control systems*, in Analysis and Optimization of Systems: State and Frequency Domain Approaches for Infinite-Dimensional Systems, R.F. Curtain, ed., Springer-Verlag, Berlin, 1992, pp. 378–387.
[19] D. Salamon, *Infinite-dimensional linear systems with unbounded control and observation: A functional analytic approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.
[20] D. Salamon, *Realization theory in Hilbert space*, Math. Systems Theory, 21 (1989), pp. 147–164.
[21] H. Tanabe, *Functional Analytic Methods for Partial Differential Equations*, Marcel Dekker, New York, 1997.
[22] F. Treves, *Basic Linear Partial Differential Equations*, Academic Press, New York, 1975.
[23] A. H. Zemanian, *Realizability Theory for Continuous Linear Systems*, Dover Publications, New York, 1972.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1266–1292

BARRIER OPERATORS AND ASSOCIATED GRADIENT-LIKE DYNAMICAL SYSTEMS FOR CONSTRAINED MINIMIZATION PROBLEMS*

JÉRÔME BOLTE† AND MARC TEBOULLE‡

Abstract. We study some continuous dynamical systems associated with constrained optimization problems. For that purpose, we introduce the concept of elliptic barrier operators and develop a unified framework to derive and analyze the associated class of gradient-like dynamical systems, called A-driven descent method (*A-DM*). Prominent methods belonging to this class include several continuous descent methods studied earlier in the literature such as steepest descent method, continuous gradient projection methods and Newton-type methods as well as continuous interior descent methods such as Lotka–Volterra-type differential equations and Riemannian gradient methods. Related discrete iterative methods such as proximal interior point algorithms based on Bregman functions and second order homogeneous kernels can also be recovered within our framework and allow for deriving some new and interesting dynamics. We prove global existence and strong viability results of the corresponding trajectories of (*A-DM*) for a smooth objective function. When the objective function is convex, we analyze the asymptotic behavior at infinity of the trajectory produced by the proposed class of dynamical systems (*A-DM*). In particular, we derive a general criterion ensuring the global convergence of the trajectory of (*A-DM*) to a minimizer of a convex function over a closed convex set. This result is then applied to several dynamics built upon specific elliptic barrier operators. Throughout the paper, our results are illustrated with many examples.

Key words. dynamical systems, continuous gradient-like systems, elliptic barrier operators, Lotka–Volterra differential equations, asymptotic analysis, viability, Lyapunov functionals, explicit and implicit discrete schemes, interior proximal algorithms, global convergence, constrained convex minimization, Riemannian gradient methods

AMS subject classifications. 34G20, 34D05, 37C10, 37N40, 49K15, 52A41, 53B21, 90C21, 90C25

DOI. 10.1137/S0363012902410861

1. Introduction.

This paper proposes to study some continuous dynamical systems in relation with the constrained optimization problem

$$(\mathcal{P}) \qquad \inf\{f(x) : \ x \in \overline{C}\},$$

where C is a nonempty *open* convex subset of $\mathbb{R}^n$, $n \geq 1$, $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a convex function, and $\overline{C}$ denotes the closure of C.

Our first aim is to give a unified framework to smooth continuous interior descent methods studied earlier in the literature: the steepest descent method, Lotka–Volterra-type equations, the continuous Newton method, and the continuous gradient projection method. Another goal of this study is to enlighten the local geometric aspects of some discrete *implicit* dynamics related to $(\mathcal{P})$ (particularly proximal-type algorithms) by associating them to some adequate vector fields. More precisely, we will also show that one of our continuous models can be cast as a specific Riemannian gradient method.

*Received by the editors July 8, 2002; accepted for publication (in revised form) March 26, 2003; published electronically September 9, 2003.
http://www.siam.org/journals/sicon/42-4/41086.html
†Département de Mathématiques, Case 51, ACSIOM-CNRS FRE 2311, Université Montpellier II, 34095 Montpellier, Cedex 5, France (bolte@math.univ-montp2.fr).
‡School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel (teboulle@post.tau.ac.il).

This has led us to introduce the class of gradient-like dynamical systems

$$(A\text{-}DM) \qquad \begin{cases} \dot{x}(t) + A_{x(t)}\nabla f(x(t)) = 0 \quad \forall t \geq 0, \\ x(0) \in C, \end{cases}$$

with

$$A : \begin{cases} C \times \mathbb{R}^n & \mapsto & \mathbb{R}^n, \\ (x, v) & \mapsto & A_x v. \end{cases} \tag{1.1}$$

The notation (*A-DM*) stands for *A*-driven descent method. To make (*A-DM*) an interior descent method, we introduce a class of mappings of the type (1.1) called *elliptic barrier operators*. This is an alternative approach to the classical barrier methods (see, for instance, Auslender, Cominetti, and Haddou [9]), since the penalization does not act on the objective function f but on its gradient. Roughly speaking, this implies two major requirements on the map A:

• the mapping $x \in C \to A_x \nabla f(x)$ must preserve the local optimality information given by $\nabla f(\cdot)$, and

• the operator A has to vanish on $\{(x, -\nu)\,, x \in \overline{C},\ \nu \in N_{\overline{C}}(x)\}$, where $N_{\overline{C}}(x)$ is the normal cone to $\overline{C}$ at $x \in \overline{C}$.

In the next section, a formal definition and the basic properties of elliptic barrier operators are given. The relevance of this notion is first illustrated by the general properties of (*A-DM*) systems. We prove existence and viability results. If ∇f is locally Lipschitz continuous, then the trajectories of (*A-DM*) are defined for all $t \geq 0$ and remain in C. Let us emphasize the fact that, unlike in Nagumo-type theorems used in viability theory (Aubin and Cellina [7]), the trajectories never encounter the boundary of C and thus make (*A-DM*) an *interior* method.

In section 3, we propose a general and unifying framework to generate in a systematic way elliptic barrier operators. This is achieved by developing an abstract setting with the help of proximal-like maps involving appropriately defined distance-like functions. Given a convenient distance-like function $d : \mathbb{R}^n \times C \mapsto \mathbb{R} \cup \{+\infty\}$, closed, proper, and convex with respect of its first variable, we introduce the class of mappings

$$A^d_x v = x - \arg\min\left\{\langle u, v\rangle + d(u, x) \mid u \in \mathbb{R}^n\right\},\ (x, v) \in C \times \mathbb{R}^n, \tag{1.2}$$

where $\langle \cdot, \cdot \rangle$ stands for the Euclidean inner product of $\mathbb{R}^n$. Aside from the fact that slight assumptions on d allow us to make A^d an elliptic barrier operator, the associated A^d-driven descent method (*A-DM*) can be seen as another step toward a unified approach to both continuous and discrete gradient-like dynamics. Indeed, one of the main facts underlying the introduction of the d operator is that (A^d-*DM*) systems can be reformulated as the differential inclusion

$$\partial_1 d(\dot{x}(t) + x(t), x(t)) + \nabla f(x(t)) \ni 0,\ t \geq 0, \tag{1.3}$$

where, for each $t \geq 0$, $\partial_1 d(\cdot, x(t))$ denotes the subdifferential of $d(\cdot, x(t))$.

This structure is at the heart of the so-called proximal-like methods (see the examples below)

$$\partial_1 d(x^{k+1}, x^k) + \nabla f(x^{k+1}) \ni 0,\ x^0 \in C,\ k \geq 0. \tag{1.4}$$

For instance, with $d(u, x) = 2^{-1}|u - x|^2$, where $|\ \cdot\ |$ denotes the Euclidean norm, the inclusion (1.4) reduces to the proximal minimization algorithm; see, e.g., Martinet

[33], Lemaire [30], and references therein. Then, according to the classical idea that consists in interpreting an iterative scheme as some discretization of a continuous dynamical system, the differential inclusion (1.3), i.e., A^d-*DM*, can be proposed as a continuous model for the proximal method (1.4). This opens new perspectives on crossed investigations, and from that viewpoint it is important to realize that the interplay between discrete and continuous dynamical systems goes far beyond the fruitful finite-time approximation aspects. For instance, in Alvarez and Attouch [2] and Antipin [4] crucial features of the asymptotic analysis appear also as closely related matters.

To give the reader a concrete idea on the type of operators A that will emerge in this study, we outline below some specific models.

(a) *The gradient projection operator.* The first natural example is given by

$$A^P : \left\{ \begin{array}{lcl} C \times \mathbb{R}^n & \mapsto & \mathbb{R}^n, \\ (x, v) & \mapsto & x - P_{\overline{C}}(x - v), \end{array} \right. \tag{1.5}$$

where $P_{\overline{C}}$ is the orthogonal projection on $\overline{C}$. A^P-*DM* is the continuous gradient projection method as introduced in [4],

$$\dot{x}(t) + x(t) - P_{\overline{C}}[x(t) - \nabla f(x(t))] = 0,\ x(0) \in C \quad \forall t \geq 0. \tag{1.6}$$

The operator A^P ruling (1.6) can be recovered thanks to (1.2) with a distance-like function of the type $d : \mathbb{R}^n \times C \ni (u, x) \mapsto \frac{1}{2}|u - x|^2 + \delta_{\overline{C}}(u)$, where $\delta_{\overline{C}}$ is the indicator function of $\overline{C}$. Let us emphasize the fact that the trajectory of the continuous system (1.6) is interior, which is not the case for the well-known explicit discretization

$$x^{k+1} = P_{\overline{C}}[x^k - \mu_k \nabla f(x^k)],\ x^0 \in C,\ \mu_k > 0;$$

see, e.g., [31], [19].

(b) *The Bregman operators.* The Bregman proximal method (BPM) is obtained by replacing the quadratic kernel in the proximal minimization algorithm by a distance-like function based on a Bregman function $h : \overline{C} \to \mathbb{R}$. Defining

$$\forall (x, y) \in \overline{C} \times C,\ D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle \tag{1.7}$$

leads to the scheme

$$(BPM) \qquad x^{k+1} \in \arg\min \left\{ f(x) + c_k D_h(x, x_k) | x \in \overline{C} \right\},\ c_k > 0,\ x^0 \in \overline{C}.$$

(BPM) has been studied and generalized from many viewpoints; see, for instance, Censor and Zenios [17], Chen and Teboulle [18], Eckstein [20], Kiwiel [27], and Teboulle [37]. One of the corresponding continuous models that is proposed here is given by barrier operators A^{q_h} of the type

$$A^{q_h} : \left\{ \begin{array}{lcl} C \times \mathbb{R}^n & \mapsto & \mathbb{R}^n, \\ (x, v) & \mapsto & \nabla^2 h(x)^{-1} v, \end{array} \right.$$

where $\nabla^2 h(x)$ is the Hessian of some convenient Bregman function with zone C and with $q_h(u, x) = \langle \nabla^2 h(x)(u - x), u - x \rangle$, $(u, x) \in \mathbb{R}^n \times C$. The A^{q_h}-driven descent method (A^{q_h}-*DM*), actually a Riemannian gradient method, is then given by

$$(A^{q_h}\text{-}DM) \qquad \dot{x}(t) + \nabla^2 h(x(t))^{-1} \nabla f(x(t)) = 0,\ x(0) \in C.$$

Besides its links with (BPM) developed in section 4, the latter system allows us to recover several dynamics. With $h_1(x) = \frac{\alpha}{2}|x|^2 + \beta \sum_{i=1\dots N} x_i \log x_i$, α, $\beta > 0$ on $C = \mathbb{R}^n_{++} := \{x \in \mathbb{R}^n,\ x_i > 0\}$, we obtain the regularized Lotka–Volterra equation recently proposed, from a completely different viewpoint, in Attouch and Teboulle [6],

(1.8)
$$(A^{q_{h_1}}\text{-}DM) \qquad \dot{x}_i(t) + \frac{x_i(t)}{\beta + \alpha x_i(t)} \frac{\partial f}{\partial x_i}(x(t)) = 0 \quad \forall i = 1, \dots, n,\ x(0) \in \mathbb{R}^n_{++},$$

where f is to be optimized on $\mathbb{R}^n_+$.

If $h(x) = \frac{\alpha}{2}|x|^2$ and $C = \mathbb{R}^n$, $(A^{q_h}\text{-}DM)$ is the classical continuous steepest descent method $\dot{x}(t) + \nabla f(x(t)) = 0$, $t \geq 0$; see Brézis [14].

For $h(x) = f(x)$ and $C = \mathbb{R}^n$, we obtain the continuous Newton descent method studied in Alvarez and Pérez [3] (see also [7])

$$(A^{q_f}\text{-}DM) \qquad \dot{x}(t) + \nabla^2 f(x(t))^{-1} \nabla f(x(t)) = 0. \tag{1.9}$$

Another surprising fact of this dynamics is to be physically meaningful in infinite-dimensional spaces. Naturally those problems are out of the scope of the present paper, but the reader interested by thermodynamical evolution equations of the form $A^{q_h}\text{-}DM$ is referred to Kenmochi and Pawlow [26] and references therein.

(c) *Barrier operators based on interior methods for the positive orthant.* Another line of research pursued by Auslender, Teboulle, and Ben-Tiba [8] concerning proximal interior methods is based on the distance-like function

$$\forall (x, y) \in (\mathbb{R}^n_{++})^2 \ \ d_\varphi(x, y) = \sum_{i=1}^{n} y_i^2 \varphi\left(\frac{x_i}{y_i}\right), \tag{1.10}$$

where $\varphi : \mathbb{R}_{++} \to \mathbb{R}$ is some relevant convex function.

The associated iterative proximal interior method is given by

$$(RIPM) \qquad x^{k+1} \in \arg\min \left\{ f(x) + c_k d_\varphi(x, x_k) | x \in \mathbb{R}^n_+ \right\},\ c_k > 0,\ x^0 \in \mathbb{R}^n_{++},$$

where $(RIPM)$ stands for regularized interior proximal method. Like (BPM) this algorithm can be applied to a minimize a general closed convex function. However, it enjoys stronger convergence properties, particularly when applied to a dual problem of a convex program; see [8] for further details and results.

Our continuous approach to $(RIPM)$ is obtained by considering barrier operators of the form

$$A^{d_\varphi} : \left\{ \begin{array}{lll} \mathbb{R}^n_{++} \times \mathbb{R}^n & \mapsto & \mathbb{R}^n, \\ (x, v) & \mapsto & \left(x_i - x_i(\varphi^*)'(x_i^{-1} v_i)\right)_{i=1,\dots,n}, \end{array} \right.$$

where φ^* is the Legendre–Fenchel conjuguate of the function φ used in $(RIPM)$.

All these continuous models are derived and analyzed in section 4. Section 5 is devoted to the asymptotic analysis of $(A\text{-}DM)$ in the convex case. We derive a general criterion ensuring the global convergence of the trajectories of $(A\text{-}DM)$ to a minimizer of f over $\overline{C}$. We then apply this general result to the dynamics built upon A^P, A^{q_h}, and A^{d_φ}. The proof relies on the existence of Lyapunov functionals measuring a sort of distance between the state variable and the set of equilibria. This approach is inspired at the same time by Opial's lemma [35] and the techniques used in

monotone optimization algorithms. We also prove a general localization result for the limit point of the trajectories produced by (*A-DM*), which extends results of the same type obtained recently in [6] and in [30] for the classical continuous gradient descent scheme. Throughout this paper we give many examples exhibiting some explicit and new systems of the type (*A-DM*). For instance, with $C = \mathbb{R}^n_{++}$, one obtains the systems

$$(A^{q_h} - DM) \qquad \dot{x}_i(t) + \frac{2x_i(t)^{3/2}}{x_i(t)^{3/2}+1}\frac{\partial f}{\partial x_i}(x(t)) = 0,\ x_i(0) > 0 \quad \forall i \in \{1,\dots,n\}$$

or

$$(A^{d_\varphi} - DM) \qquad \dot{x}_i(t) + x_i(t) + \frac{1}{2}\frac{\partial f}{\partial x_i}(x(t)) - \sqrt{\frac{1}{4}\frac{\partial f}{\partial x_i}(x(t))^2 + x_i(t)^2} = 0,$$

with $i \in \{1,\dots,n\}$, $t \geq 0$, and $x(0) \in \mathbb{R}^n_{++}$. The first equation is given by the Bregman function $h(s) = s^2/4 - 2\sqrt{s}$, $s \geq 0$, while the second one corresponds to a continuous model of the logarithmic-quadratic method [8] obtained with the choice $\varphi(s) = 1/2(s-1)^2 - \log s + s - 1$, $s > 0$.

Notation. Our notation is fairly standard. The Euclidean space $\mathbb{R}^n$ is equipped with the scalar product $\langle \cdot, \cdot \rangle$; the related norm is denoted $|\cdot|$. The boundary of C is denoted $\mathrm{bd}C$. $N_{\overline{C}}(x)$ and $T_{\overline{C}}(x)$ denote, respectively, the normal cone and the tangent cone of $\overline{C}$ at $x \in \overline{C}$. We recall that $N_{\overline{C}}(x) = \{v \in \mathbb{R}^n \ \langle v, z - x\rangle \leq 0 \text{ for all } z \in \overline{C}\} = \{v \in \mathbb{R}^n | \text{ for all } u \in T_{\overline{C}}(x),\ \langle v, u\rangle \leq 0\}$. If $\phi : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$, $p \geq 1$, is a closed proper convex function, its domain is defined by $\mathrm{dom}\ \phi = \{x \in \mathbb{R}^p | \phi(x) < +\infty\}$, and its Legendre–Fenchel conjuguate, $y \in \mathbb{R}^p \to \sup\{\langle y, x\rangle - \phi(x) | x \in \mathbb{R}^p\}$, is denoted ϕ^*. If S is a closed convex subset of $\mathbb{R}^n$, the set of minimizers of ϕ on S is denoted $\arg\min_S \phi$. The indicator function of $\overline{C}$ is denoted by $\delta_{\overline{C}}$. Other notation and definitions not explicitly stated here can be found in the classical book of Rockafellar [36].

2. Elliptic barrier operators and viability results. In this section, the definition and the first properties of elliptic barrier operators are introduced. Then, in view of constrained minimization, we study the corresponding A-driven descent methods, proving in particular that the obtained trajectories $\{x(t)\}$ are *interior* and defined for any $t \in [0, +\infty)$.

2.1. Elliptic barrier operators: Definition and properties.

DEFINITION 2.1. *$A : C \times \mathbb{R}^n \to \mathbb{R}^n$ is an elliptic barrier operator on C if it satisfies the following:*

($r1$) A is Lipschitz continuous on every compact subset of $C \times \mathbb{R}^n$.

($r2$) There exists $\alpha > 0$, such that for every $(x, v) \in C \times \mathbb{R}^n$, $\langle A_x v, v\rangle \geq \alpha |A_x v|^2$.

($r3$) For all $x \in C$, $A_x v = 0$ implies $v = 0$.

(v) For all $b \in \mathrm{bd}\ C$, for all $\nu \in N_{\overline{C}}(b)$, for all $M > 0$, $\exists \epsilon, K > 0$ such that $|x - b| < \epsilon$, $x \in C$, $|v| \leq M$ implies

$$\langle -A_x v, \nu\rangle \leq K\langle b - x, \nu\rangle. \tag{2.1}$$

This definition is motivated by the study of (*A-DM*) systems. The *regularity* assumption ($r1$) naturally meets the conditions of the Cauchy–Lipschitz theorem. The *ellipticity* condition ($r2$) and the *nondegeneracy* assumption ($r3$) allow us to obtain a proper descent method. An important consequence of ($r2$) is that the term

$1/\alpha$ can be seen as an upper bound for the gradient stepsize in (*A-DM*). Indeed, it follows readily from $(r2)$ that $|A_x v| \leq \alpha^{-1}|v|$, and therefore a trajectory $x(\cdot)$ of (*A-DM*) satisfies

$$|\dot{x}(t)| \leq \alpha^{-1}|\nabla f(x(t))|$$

whenever $x(t)$ is defined and belongs to C.

The *normal boundary* property (v) is required to control the outward normal impulses near the boundary of C, making the trajectories of (*A-DM*) strongly viable; i.e., $x(t) \in C$, $t \geq 0$. The choice of the term $\langle b - x, \nu\rangle$ in (2.1) also has a regularizing effect. Indeed, as it will be proved in Theorem 2.4 (see also Remark 2.1 (b)), it contributes to the fact that the trajectories of (*A-DM*) are defined on $[0, +\infty)$.

Remark 2.1. (a) A natural extension of Definition 2.1 can be obtained by replacing assumptions $(r2)$ and $(r3)$, respectively, by

$(r2)'$ For every $(x, v) \in C \times \mathbb{R}^n$, $v \neq 0$ $\langle A_x v, v\rangle > 0$.

$(r3)'$ For all $x \in C$, $v = 0$ implies $A_x v = 0$.

Observing that $(r2)'$ and $(r3)'$ imply $(r3)$, it follows that an elliptic barrier operator satisfies this new definition. This widened concept opens new perspectives but also raises some difficulties in the study of (*A-DM*): finite-time solutions, loss of regularity (see Theorem 2.4 in the elliptic case), no upper bound for the gradient step-sizes, etc. The study of such a class of mappings will not be carried out in the present paper but appears as an interesting matter for future research.

(b) If the left term in (2.1) is replaced, for instance, by $\langle b-x, \nu\rangle^{1-\theta}$, $\theta \in (0, 1)$, the well-posedness of (*A-DM*) may fail: take, for instance, $A : (x, v) \in \mathbb{R}_+ \times \mathbb{R} \to x^{1-\theta}.v$, $\theta \in (0, 1)$, $f(x) = x + 1$, and observe that the maximal solutions of (*A-DM*) are not defined on $[0, +\infty)$.

In what follows, it is of interest to strengthen $(r1)$ by assuming the additional hypothesis

$(r4)$ A is continuous on $\overline{C} \times \mathbb{R}^n$.

The following result shows that an elliptic barrier operator on C can be continuously extended to

$$C \times \mathbb{R}^n \cup \big\{(x, v) | x \in \operatorname{bd} C,\ v \in -N_{\overline{C}}(x)\big\}$$

by setting $A_x v = 0$, if $x \in \operatorname{bd} C$, $v \in -N_{\overline{C}}(x)$.

PROPOSITION 2.2. *Let $A : C \times \mathbb{R}^n \to \mathbb{R}^n$ be an elliptic barrier operator. Assume that $(x^k, v^k), k \in N$, is a sequence in $C \times \mathbb{R}^n$ such that $x^k \to x \in \overline{C}$ and $v^k \to v \in -N_{\overline{C}}(x)$ as $k \to +\infty$. Then*

(i) *$A_{x^k} v^k \to 0$ as $k \to +\infty$.*

(ii) *In addition, if A satisfies $(r4)$, then for all $x \in \overline{C}$ one has*

$$A_x^{-1}(\{0\}) \supset -N_{\overline{C}}(x). \tag{2.2}$$

Proof. If $x \in C$, the conclusion follows from $(r1)$ and $(r3)$. Else $x \in \operatorname{bd} C$. $(r2)$ and the Cauchy–Schwarz inequality yield $|A_{x^k} v^k|.|v^k| \geq \alpha |A_{x^k} v^k|^2$ for all $k \in N$ and some $\alpha > 0$. Since the sequence $v^k, k \in N$, is bounded, so is $A_{x^k} v^k, k \in N$. From (v) it follows that for k large enough $\langle -A_{x^k} v^k, -v\rangle \leq K\langle x - x^k, -v\rangle$, and therefore

$$\limsup_{k \to +\infty} \langle A_{x^k} v^k, v\rangle \leq 0. \tag{2.3}$$

On the other hand, we have

$$\langle A_{x^k} v^k, v\rangle = \langle A_{x^k} v^k, v - v^k\rangle + \langle A_{x^k} v^k, v^k\rangle \quad \forall k \in N,$$

and since $A_{x^k}v^k, k \in N$, is bounded, we obtain

$$\liminf_{k\to+\infty}\langle A_{x^k}v^k, v\rangle = \liminf_{k\to+\infty}\langle A_{x^k}v^k, v^k\rangle \geq 0. \tag{2.4}$$

From (2.3) and (2.4), we deduce that $\lim_{k\to+\infty}\langle A_{x^k}v^k, v\rangle = \liminf_{k\to+\infty}\langle A_{x^k}v^k, v^k\rangle = 0$, and thus by $(r2)$, $\lim_{k\to+\infty}|A_{x^k}v^k|^2 = 0$. □

Remark 2.2. For simplicity, assume that f is convex, with $\arg\min_{\overline{C}} f \neq \emptyset$, and that A satisfies $(r4)$. Subdifferential calculus (see, e.g., [36]) allows us to associate to $(\mathcal{P})$ the following variational characterization:

$$x^* \text{ solves } (\mathcal{P}) \text{ iff } \nabla f(x^*) + N_{\overline{C}}(x^*) = 0.$$

Using (2.2), we know that the solutions of $(\mathcal{P})$ are contained in the set of zeros of the gradient-like map $x \in \overline{C} \to A_x\nabla f(x)$. This is only a necessary condition for optimality, and it can be written as

$$\text{if } x^* \text{ solves } (\mathcal{P}), \text{ then } A_{x^*}\,\nabla f(x^*) = 0. \tag{2.5}$$

The important point here is to realize that our approach to optimization is given *throughout (A-DM) dynamics*, and thus x^* is obtained as a limit point of some descent method. Indeed, as we shall see, most of the systems and examples of section 4 satisfy (2.2) with a *strict* inclusion, yet their orbits converge to a minimizer of f on $\overline{C}$; see section 5.

We conclude these introductory notions by stating a useful criterion implying assumption (v) of Definition 2.1.

LEMMA 2.3. *Let $A : C \times \mathbb{R}^n \to \mathbb{R}^n$, $m > 0$, and $k : C \times \mathbb{R}^n \to [m, +\infty)$ be such that*

$$x - k(x,v)A_x v \in \overline{C} \quad \forall (x,v) \in C \times \mathbb{R}^n.$$

Then A satisfies (v).

Proof. The proof relies on the fact that $x - k(x,v)A_x v - b \in T_{\overline{C}}(b)$ for every (x,v) in $C \times \mathbb{R}^n$ and for every $b \in \overline{C}$. By definition we have, for all $\nu \in N_{\overline{C}}(b)$, $\langle x - k(x,v)A_x v - b, \nu\rangle \leq 0$, and therefore

$$\langle -A_x v, \nu\rangle \leq \frac{1}{k(x,v)}\langle b - x, \nu\rangle \leq \frac{1}{m}\langle b - x, \nu\rangle. \qquad \square$$

2.2. Global existence and viability results. From now on, the function $f : \mathbb{R}^n \to \mathbb{R}$ is C^1 and satisfies

$(\mathcal{H}_1)$ ∇f is Lipschitz continuous on bounded sets;

$(\mathcal{H}_2)$ $\inf_{\overline{C}} f > -\infty$.

Observe that for the moment the function f is not supposed to be convex.

THEOREM 2.4. *Let A be an elliptic barrier operator. Then, the following hold:*

(i) *The system (A-DM) admits a unique C^1 solution x defined on $[0, +\infty)$.*

Moreover,

(ii) *for all $t \geq 0$, $x(t) \in C$.*

(iii) *The function $t \in [0, +\infty) \to f(x(t))$ is nonincreasing and has a limit as $t \to +\infty$.*

(iv) $\dot{x} \in L^2(0, +\infty; \mathbb{R}^n)$.

(v) *If A satisfies $(r4)$ and $x(\cdot)$ is bounded, then $\dot{x}(t) \to 0$ as $t \to 0$, and all limit points x^* of $x(\cdot)$ satisfy the weak optimality condition*

$$A_{x^*} \nabla f(x^*) = 0.$$

Proof. Fix $T > 0$ and consider the assertion $E(T)$: "There exists a solution of $(A\text{-}DM)$ defined on $[0, T]$, and such that $x(t) \in C$ for all $t \in [0, T]$."

Set $T_{max} := \sup\{T \mid E(T) \text{ is satisfied}\}$. From $(r1)$, $(\mathcal{H}_1)$, and the fact that $x(0) \in C$, it follows by Cauchy–Lipschitz theorem that $T_{max} > 0$ and that the solution of $(A\text{-}DM)$ defined on $[0, T_{max})$ is unique.

Let us derive some a priori estimates. Let $T \in (0, T_{max})$; by the $(A\text{-}DM)$ system we have for all $t \in [0, T]$

$$\langle \dot{x}(t), \nabla f(x(t)) \rangle + \langle A_{x(t)} \nabla f(x(t)), \nabla f(x(t)) \rangle = 0,$$

and thus by $(r2)$ and $(A\text{-}DM)$ again,

$$\frac{d}{dt} f(x(t)) + \alpha |\dot{x}(t)|^2 \leq 0. \tag{2.6}$$

Integrating over some interval $(0, t)$ with $t \leq T$ gives

$$f(x(t)) - f(x(0)) + \alpha \int_0^t |\dot{x}|^2 \leq 0. \tag{2.7}$$

Note that if $T_{max} = +\infty$, (iii) and (iv) follow from (2.6), (2.7), and $(\mathcal{H}_2)$. Let us argue by contradiction and assume that $T_{max} < +\infty$.

Using the Cauchy–Schwarz inequality and the fact that $\dot{x} \in L^2(0, T_{max}; \mathbb{R}^n)$, we obtain that x is a Cauchy net at T_{max}. Therefore, x can be continuously extended by an application still denoted by x. Set $x(T_{max}) := b \in \overline{C}$.

By definition of T_{max}, b necessarily belongs to $\operatorname{bd} C$. The function $t \in [0, T_{max}] \to \nabla f(x(t))$ is bounded by a positive constant M. Owing to the continuity of x and (v), there exist $t_0 \in (0, T_{max})$, $\epsilon > 0$, $K > 0$, and $\nu \in N_{\overline{C}}(b)$, $\nu \neq 0$, such that for all $t \in (t_0, T_{max})$

$$\langle -A_{x(t)} \nabla f(x(t)), \nu \rangle \leq K \langle b - x(t), \nu \rangle. \tag{2.8}$$

Let us project $(A\text{-}DM)$ on $\mathbb{R}\nu := \{\tau\nu \mid \tau \in \mathbb{R},\ 0 \neq \nu \in N_{\overline{C}}(b)\}$; this gives for all $t \in (t_0, T_{max})$

$$\frac{d}{dt} \langle x(t), -\nu \rangle + \langle A_{x(t)} \nabla f(x(t)), -\nu \rangle = 0,$$

and using (2.8) we obtain

$$\frac{d}{dt} \langle b - x(t), \nu \rangle + K \langle b - x(t), \nu \rangle \geq 0.$$

Multiplying the above inequality by $\exp Kt$ and integrating over (t_0, T_{max}), it follows that

$$\langle b - x(T_{max}), \nu \rangle \geq \exp[-K(T_{max} - t_0)] \langle b - x(t_0), \nu \rangle.$$

Observe that by definition, $b = x(T_{max})$; hence to draw a contradiction from the latter we just have to prove that the second term of the inequality is positive. Indeed,

$x(t_0) \in C$, which is open convex, and $0 \neq \nu \in N_{\overline{C}}(b)$; thus there exists $\eta > 0$ such that $x(t_0)+\eta\nu \in C$, and a fortiori $x(t_0)+\eta\nu-b \in T_{\overline{C}}(b)$. This implies $\langle x(t_0)+\eta\nu-b, \nu\rangle \leq 0$ or, equivalently, $\langle b - x(t_0), \nu\rangle \geq \eta|\nu|^2 > 0$, and (i) is proved.

Let us prove the last statement (v). From the boundedness property of x, along with $(r4)$ and $(\mathcal{H}_1)$, it follows that $\dot{x}$ is bounded, and therefore x is a Lipschitz continuous map. The properties $(r4)$, $(\mathcal{H}_1)$ imply that $t \geq 0 \to A_{x(t)}\nabla f(x(t))$ is uniformly continuous, and therefore so is $\dot{x}(\cdot)$. Combining this fact with (iv), it follows by a classical argument that $\dot{x}(t) \to 0$ as $t \to +\infty$. Using $(r4)$, it follows that a cluster point x^* of x satisfies $A_{x^*}\nabla f(x^*) = 0$. □

3. A general abstract framework for dynamical systems with elliptic barrier operators. In this section, we propose with the help of proximal maps a systematic and unifying way to generate elliptic barrier operators. We start with an informal motivation. Given a convenient distance-like function $d : \mathbb{R}^n \times C \mapsto \mathbb{R} \cup \{+\infty\}$, the idea is to realize the descent direction $-A_x\nabla f(x)$, $x \in C$, as a vector based on x and pointing on some proximal point $u^d(x, \nabla f(x))$.

Indeed, assume that d is convex with respect to its first variable, and for $x \in C$ define formally

$$u^d(x, \nabla f(x)) \in \arg\min \{\langle u, \nabla f(x)\rangle + d(u, x) | u \in \mathbb{R}^n\}. \tag{3.1}$$

In this definition, the objective function has been replaced by its first order approximation at the point x, the constraints are supposed to be naturally taken into account by $d(\cdot,\cdot)$, and the descent direction obtained is $-A^d_x\nabla f(x) := u^d(x, \nabla f(x)) - x$. It is of interest to note that this approach is akin to the following well-known fixed point reformulation of the optimization problem $(\mathcal{P})$:

$$x^* \text{ solves } (\mathcal{P}) \text{ iff } x^* \in \arg\min \{\langle u, \nabla f(x^*)\rangle \mid u \in \overline{C}\} \tag{3.2}$$

whenever f is convex. From that viewpoint, the formal definition (3.1) may appear as a proximal regularization of some possibly ill-posed problem. On the other hand, the corresponding A^d-DM can be written as a fixed-point-like dynamics

$$\dot{x}(t) + x(t) = u^d[x(t), \nabla f(x(t))],\ x(0) \in C \quad \forall t \geq 0. \tag{3.3}$$

The solution of (3.3) is then expected to asymptotically provide a solution of $x^* = u^d(x^*, \nabla f(x^*))$, and when it makes sense, this last problem corresponds to another formulation of (3.2).

As a first example, consider $d(u, x) = 1/2|u - x|^2 + \delta_{\overline{C}}(u)$, $(u, x) \in \mathbb{R}^n \times C$. The definition of u^d writes

$$\nabla f(x) + u^d(x, \nabla f(x)) - x + N_{\overline{C}}[u^d(x, \nabla f(x))] \ni 0,$$

which in turn is equivalent to

$$u^d(x, \nabla f(x)) \in (I + N_{\overline{C}})^{-1}(x - \nabla f(x)).$$

Recalling that $(I + N_{\overline{C}})^{-1} = P_{\overline{C}}$, the proximal point is thus given by $u^d(x, \nabla f(x)) = P_{\overline{C}}(x-\nabla f(x))$. This gives rise to the descent direction $-A^d_x\nabla f(x) = P_{\overline{C}}(x-\nabla f(x))-x$, and the projected gradient dynamics (1.6) is recovered. As mentioned in the above discussion, note that the reformulation of (3.2) throughout $d(\cdot,\cdot)$, that is, $x^* = u^d(x^*, \nabla f(x^*))$, leads to the fixed point problem $x^* = P_{\overline{C}}(x^* - \nabla f(x^*))$.

Let us now develop an abstract setting that shall be illustrated in the next section with various useful kernels $d(\cdot,\cdot)$.

Let $d_0 : \mathbb{R}^n \times C \mapsto \mathbb{R}_+ \cup \{+\infty\}$ be such that the following hold:

$(P1)$ d_0 is C^1 on $C \times C$.
$(P2)$ $\nabla_1 d_0(u,u) = 0$ for all $u \in C$.
$(P3)$ For every $x \in C$, the mapping $u \in \mathbb{R}^n \mapsto d_0(u,x)$ is a closed convex function.

In $(P1)$, $\nabla_1 d_0(\cdot, u)$ is the gradient of $d(\cdot, u)$, (more generally its subdifferential is denoted by $\partial_1 d_0(\cdot, u)$). Note that, since C is nonempty, $(P1)$ ensures that $u \in \mathbb{R}^n \mapsto d_0(u,x)$ is also proper.

Denote by $\mathcal{D}$ the set of mappings $d : \mathbb{R}^n \times C \mapsto \mathbb{R}_+ \cup \{+\infty\}$ that can be written

$$d(u,x) = \frac{\alpha}{2}|u-x|^2 + d_0(u,x), \tag{3.4}$$

with $\alpha > 0$ and with d_0 satisfying $(P1)$, $(P2)$, and $(P3)$.

DEFINITION 3.1. *Let d be in $\mathcal{D}$. For all $(x,v) \in C \times \mathbb{R}^n$, set*

$$u^d(x,v) \in \arg\min \{\langle u, v\rangle + d(u,x) | u \in \mathbb{R}^n\}, \tag{3.5}$$

and define A^d by

$$A^d_x v = x - u^d(x,v). \tag{3.6}$$

The following proposition justifies the second part (3.6) of this definition (u^d could be multivalued) and describes some of the properties of the operator A^d.

PROPOSITION 3.2. *Let $d \in \mathcal{D}$.*

(i) *For each $x \in C$, the map $v \in \mathbb{R}^n \mapsto u^d(x,v)$ is a single valued α^{-1}–Lipschitz continuous map.*
(ii) *A^d satisfies $(r2)$ and $(r3)$, and for each $x \in C$, $v \in \mathbb{R}^n \mapsto A^d_x v$ is Lipschitz continuous.*
(iii) *Moreover, if d satisfies the property*

$$(p) \quad \forall x \in C, \ \text{dom } d(\cdot\,, x) \subset \overline{C},$$

then A^d satisfies (v) *of Definition* 2.1.

Proof. Let $(x,v) \in C \times \mathbb{R}^n$. From $(P3)$ and the fact that $\alpha > 0$, it follows that $u \in \mathbb{R}^n \mapsto \langle u, v\rangle + d(u,x)$ is strongly convex and has a nonempty bounded lower level set. This implies that $u^d(x,v)$ exists and is unique. Using $(P1)$ and $(P3)$ allows us to write the optimality condition in (3.5) as

$$v + \partial_1 d(\cdot, x)(u^d(x,v)) \ni 0,$$

and therefore by uniqueness of $u^d(x,v)$ (recalling (cf. [36]) that for any closed proper convex function F, one has $(\partial F)^{-1} = \partial F^*$), it follows that

$$u^d(x,v) = \partial_1 d^*(\cdot, x)(-v). \tag{3.7}$$

Denoting by I the identity map of $\mathbb{R}^n$, we observe using the definition of $d \in \mathcal{D}$ that $\partial_1 d^*(\cdot, x)$ can also be written

$$(\alpha I + \partial_1 d_0(\cdot, x) - \alpha x)^{-1}$$

or, equivalently, as the composition

$$(I + \alpha^{-1}\partial_1 d_0(\cdot\,, x) - x)^{-1} \circ \alpha^{-1} I.$$

By $(P3)$, the operator $\alpha^{-1}\partial_1 d_0(\cdot\,,x) - x$ is maximal monotone, and therefore by [14, Proposition 2.2], $(I + \frac{1}{\alpha}\partial_1 d_0(\cdot\,,x) - x)^{-1}$ is a contraction defined on $\mathbb{R}^n$. Recalling that $u^d(x,v) = (I + \alpha^{-1}\partial_1 d_0(\cdot\,,x) - x)^{-1} \circ \alpha^{-1} I$ and $A^d_x v = x - u^d(x,v)$, the above arguments prove (i) and the second part of statement (ii).

Assume that d complies with the property (p). By the definition of u^d, this implies that $u^d(x,v) = x - A^d_x v \in \overline{C}$, and therefore (iii) is a consequence of Lemma 2.3. It remains to prove the first two assertions of (ii). Let us prove that A^d satisfies $(r3)$. Let $(x,v) \in C \times \mathbb{R}^n$ be such that $A_x v = 0$. Then by (3.7), $x = \partial_1 d^*(\cdot\,,x)(-v)$, which implies that $\partial_1 d(x,x) = \nabla_1 d(x,x) = -v$. Therefore, by $(P2)$ one has $v = 0$. Now to prove that $(r2)$ is also satisfied, we use the following lemma.

LEMMA 3.3 (Baillon–Haddad [10]). *Let H, $\langle\,,\,\rangle$ be a Hilbert space whose norm is denoted $|\,.\,|$, $\phi : H \mapsto \mathbb{R}$ a C^1 convex function, and $L > 0$. The following statements are equivalent:*

(i) *For all $(x,y) \in H^2$, $|\nabla\phi(x) - \nabla\phi(y)| \le L|x-y|$.*

(ii) *For all $(x,y) \in H^2$, $\langle \nabla\phi(x) - \nabla\phi(y), x - y\rangle \ge \frac{1}{L}|\nabla\phi(x) - \nabla\phi(y)|^2$.*

In view of (3.7) and (i), this result can be applied to $\phi := d^*(\cdot\,,x)$. Hence, for x fixed in C and for all $(v_1,v_2) \in \mathbb{R}^n \times \mathbb{R}^n$, it gives

$$\langle \partial_1 d^*(\cdot\,,x)(v_1) - \partial_1 d^*(\cdot\,,x)(v_2), v_1 - v_2\rangle \ge \alpha|\partial_1 d^*(\cdot\,,x)(v_1) - \partial_1 d^*(\cdot\,,x)(v_2)|^2.$$

Now, letting $v_1 = 0$ and $v_2 = -v$ in the latter yields

$$\langle x - u^d(x,v), v\rangle \ge \alpha|x - u^d(x,v)|^2,$$

which, according to (3.6), is exactly $(r2)$. ∎

4. Elliptic barrier operators and continuous models for proximal algorithms: Examples and properties. In this section, we show that for various minimization algorithms one can derive an elliptic barrier operator and construct the associated $(A\text{-}DM)$-dynamical system. It is worth mentioning that many of the examples to follow will generate convergent trajectories to the minimizer of a convex function f over the closed convex set $\overline{C}$. From now on α will always denote the positive parameter involved in the definition of the class $\mathcal{D}$; cf. (3.4).

4.1. Projection-like methods. Let $h_0 : \mathbb{R}^n \mapsto \mathbb{R}$ be a C^1 convex function whose gradient is Lipschitz continuous on bounded sets, and set

$$\tilde{D}_h : \begin{cases} \mathbb{R}^n \times C & \to & \mathbb{R}_+ \cup \{+\infty\}, \\ (u,x) & \mapsto & D_h(u,x) + \delta_{\overline{C}}(u), \end{cases}$$

with $h(u) = \frac{\alpha}{2}|u|^2 + h_0(u)$, $u \in \mathbb{R}^n$, and where D_h is given by (cf. (1.7))

$$\forall (x,y) \in \mathbb{R}^n \times C,\ D_h(x,y) = h(x) - h(y) - \langle \nabla h(y), x - y\rangle. \tag{4.1}$$

PROPOSITION 4.1. *Let $\tilde{D}_h$ be as defined above. Then $A^{\tilde{D}_h}$ is an elliptic barrier operator that satisfies $(r4)$. Moreover, we have for all $(x,v) \in C \times \mathbb{R}^n$*

$$A^{\tilde{D}_h}_x v = x - (\nabla h + N_{\overline{C}})^{-1}(\nabla h(x) - v). \tag{4.2}$$

Proof. An easy computation gives $\tilde{D}_h(u,x) = \frac{\alpha}{2}|u-x|^2 + D_{h_0}(u,x) + \delta_{\overline{C}}(u)$. Letting $d_0(u,x) = D_{h_0}(u,x) + \delta_{\overline{C}}(u)$, we obtain that d_0 satisfies $(P1)$ and $(P3)$. For $(u,x) \in C \times C$, we have $\nabla_1 d_0(u,x) = \nabla h_0(u) - \nabla h_0(x)$, and as a consequence $(P2)$

is satisfied as well. Therefore $\tilde{D}_h$ is in $\mathcal{D}$ and clearly verifies (p). Now applying Proposition 3.2, it follows that $A^{\tilde{D}_h}$ satisfies $(r2)$, $(r3)$, and (v). The explicit formula of $A^{\tilde{D}_h}$ follows from (3.7). To obtain $(r1)$ and $(r4)$, we just have to observe that $(\nabla h + N_{\overline{C}})^{-1}$ and ∇h_0 are locally Lipschitz continuous on $\mathbb{R}^n$. □

The terminology of projection relies on the fact that (4.2) can be seen as some twisted projection in the Bregman sense. Indeed, defining the projection of $z \in \mathbb{R}^n$ on $\overline{C}$ by

$$P^h_{\overline{C}}(z) := \arg\min \{D_h(u, z) \mid u \in \overline{C}\},$$

we obtain that $P^h_{\overline{C}}(z) = (\nabla h + N_{\overline{C}})^{-1}(\nabla h(z))$ (recall that $\alpha > 0$), and therefore since $\nabla h^* = (\nabla h)^{-1}$, one can write

$$A^{\tilde{D}_h}_x v = x - P^h_{\overline{C}}(\nabla h^*(\nabla h(x) - v)) \quad \forall (x, v) \in C \times \mathbb{R}^n.$$

It is worth noting that in the framework of convex minimization, the gradient-like map $x \mapsto A^{\tilde{D}_h}_x \nabla f(x)$ enjoys remarkable properties. As a matter of fact, assume that the objective function f is convex, and observe that the following characterization holds:

$$x^* \text{ solves } (\mathcal{P}) \text{ iff } A^{\tilde{D}_h}_{x^*} \nabla f(x^*) = 0.$$

The associated $A^{\tilde{D}_h}$-driven descent method ($A^{\tilde{D}_h}$-DM) leads to the following differential equation:

$$\dot{x}(t) + x(t) - P^h_{\overline{C}}(\nabla h^*[\nabla h(x(t)) - \nabla f(x(t))]) = 0,\ x(0) \in C,\ \forall t \geq 0. \tag{4.3}$$

Note that with $h_0 = 0$ and $\alpha = 1$, the corresponding dynamical system ($A^{\tilde{D}_h}$-DM) (with corresponding operator A^P) is nothing else but the continuous gradient projection method (1.6), that is,

$$\dot{x}(t) + x(t) - P_{\overline{C}}[x(t) - \nabla f(x(t))] = 0,\ x(0) \in C,\ \forall t \geq 0.$$

We remark that if $x(0) \notin C$, we still obtain convergent trajectories (with f convex) (see [4] or Bolte [13]) but the dynamical system is neither a descent nor an interior method.

4.2. Continuous models for Bregman proximal minimization algorithms. In this section, we give two quite different continuous models associated with proximal methods based on Bregman distances.

Continuous model I: A Riemannian gradient method. Our model appears as a particular case of Riemannian gradient methods on the smooth manifold C. Let us specify the setting. Denote by $S^{++}_n(\mathbb{R})$ the cone of real definite positive symmetric matrices and let $\mathcal{T}_x C$ be the tangent space to C at $x \in C$. In what follows, we make the usual identification $\mathcal{T}_x C \simeq \mathbb{R}^n$ for all $x \in C$. If g is some differentiable metric on C, there exists a unique differentiable application $\lambda : C \to S^{++}_n(\mathbb{R})$ such that, for all $(x, u, v) \in C \times \mathbb{R}^n \times \mathbb{R}^n$,

$$g_x(u, v) = \langle \lambda(x)u, v \rangle.$$

The gradient of a smooth function ϕ with respect to the metric g is then given by the formula $\nabla_g \phi(x) = \lambda(x)^{-1} \nabla \phi(x)$ for all $x \in C$, and the corresponding gradient method is

$$\begin{cases} \dot{x}(t) + \nabla_g \phi(x(t)) = 0, \\ x(0) \in C. \end{cases} \tag{4.4}$$

For $C = \mathbb{R}^n$, ϕ real analytic, and g differentiable, a deep result of Lojasiewicz [32] allows us to prove that all bounded trajectories are converging to a critical point of ϕ.

Readers interested in the use of geometric tools in optimization are referred to Bayer and Lagarias [11], [12] in the context of linear programming and for more general results to the recent monograph of Helmke and Moore [22] and references therein.

Remark 4.1. Although our primary concerns in this paper are far removed from the complexity analysis of optimization algorithms, let us mention that there exists an intimate relation between Riemannian geometry and the complexity analysis of interior point optimization methods (see, e.g., the work of Karmarkar [25] in the context of linear programming). More generally, in the context of convex programming, Nesterov and Nemirovskii [34] introduced the fundamental concept of self-concordant barrier functions for a constraint set C, which plays a central role in the design and analysis of interior methods with polynomial complexity. Thus an interesting topic which is left for future research would be to study a Riemannian metric defined on C, based, for example, on the Hessian of a self-concordant barrier, and which could lead to further insights on the performance or complexity of barrier methods.

We focus here on the special choice of the application $\lambda : C \to S_n^{++}(\mathbb{R})$ defined by $\lambda = \nabla^2 h$, where h is some C^3 Bregman function with zone C; see Definition 4.2. The idea is to penalize the Euclidean scalar product, rather than the objective function, and to study the corresponding Riemannian gradient method

$$\dot{x}(t) + \nabla^2 h(x(t))^{-1} \nabla f(x(t)) = 0 \tag{4.5}$$

or, *equivalently*,

$$\frac{d}{dt} \nabla h(x(t)) + \nabla f(x(t)) = 0. \tag{4.6}$$

When the objective function is linear, this differential equation has been considered in Iusem, Svaiter, and Da Cruz [24]; however, their approach to the asymptotic behavior strongly relies on the linear properties of f; see Remark 5.3 (b) for an insight. Observe that this dynamics has, in its first form (4.5), the structure of (*A-DMs*). We shall see actually that most of classical Bregman functions can generate a barrier operator. Moreover, as shown below, the general framework developed in section 3 allows us to recover those methods by considering families of quadratic forms.

For the moment, let us compare (4.6) with (BPM) as given in the introduction. By an Euler implicit discretization we formally obtain

$$\frac{1}{\Delta t_k} [\nabla h(x^{k+1}) - \nabla h(x^k)] + \nabla f(x^{k+1}) = 0,\ \Delta t_k > 0. \tag{4.7}$$

Now observe that (BPM) has exactly the form of (4.7), provided that the iterates remain in C [17, 18, 20].

Before going further, we need to recall some of the basic facts concerning Bregman functions. Their definition relies mainly on their D function, as specified in (1.7).

DEFINITION 4.2. *A function $h : \overline{C} \to \mathbb{R}$ is called a Bregman function with zone C if it satisfies the following:*

(i) *h is C^1 on C.*

(ii) *h is continuous and strictly convex on $\overline{C}$.*

(iii) *For every* $r \in \mathbb{R}$, *the partial level subset* $L_h(x_0, r) = \{y \in C | D_h(x_0, y) \leq r\}$ *is bounded for every* $x_0 \in \overline{C}$.

(iv) *Let* $(y^k, k \in N)$ *be a sequence in* C *and* $x \in \overline{C}$. *If* $y^k \to x$ *as* $k \to +\infty$, *then* $D_h(x, y^k) \to 0$ *as* $k \to +\infty$.

This definition weakens the usual definition of the Bregman function proposed by Censor and Lent in [16] and is actually inspired by the more general notion of the B-functions introduced by Kiwiel in [28]. Because of (iv) and the smoothness property of h, we have kept the terminology of the Bregman function.

For the asymptotic analysis of (4.6) which will be developed in section 5, we record here the following useful lemma due to Kiwiel [28, Lemma 2.16].

LEMMA 4.3. *Let* h *be a Bregman function with zone* C *and* $x \in \overline{C}$. *If* y^k, $k \in N$, *is a bounded sequence in* C *such that* $D_h(x, y^k) \to 0$ *as* $k \to +\infty$, *then* $y^k \to x$ *as* $k \to +\infty$.

In relation to the barrier operators to follow, let us now define a subclass of Bregman functions with zone C.

For $h : \overline{C} \to \mathbb{R}$, we consider the following assumptions:

(r_h) There exist $\alpha > 0$ and a C^3 Bregman function with zone C denoted by h_0 such that for all $x \in \overline{C}$,

$$h(x) = \frac{\alpha}{2}|x|^2 + h_0(x).$$

(v_h) For every $b \in \operatorname{bd} C$ and every $\nu \in N_{\overline{C}}(b)$, there exists $K, \epsilon > 0$ such that for every $x \in C$, $|x - b| < \epsilon$,

$$|\nabla^2 h(x)^{-1}\nu| \leq K\langle b - x, \nu\rangle.$$

The set of such functions is denoted by $\mathcal{B}_C$, and for each $h \in \mathcal{B}_C$ we define a family of quadratic forms by

$$q_h : \begin{cases} \mathbb{R}^n \times C & \to & \mathbb{R}^n, \\ (u, x) & \mapsto & \langle \nabla^2 h(x)(u - x), u - x\rangle. \end{cases}$$

PROPOSITION 4.4. *For every* $h \in \mathcal{B}_C$, A^{q_h} *is an elliptic barrier operator on* C. *Moreover, for all* $(x, v) \in C \times \mathbb{R}^n$, *the following formula holds:*

$$A_x^{q_h} v = \nabla^2 h(x)^{-1} v. \tag{4.8}$$

Proof. To prove that $q_h \in \mathcal{D}$, it suffices to note that by (r_h),

$$q_h(u, x) = \alpha/2|u - x|^2 + \langle \nabla^2 h_0(x)(u - x), u - x\rangle,$$

where $\langle \nabla^2 h_0(x)(u - x), u - x\rangle$ satisfies $(P1)$, $(P2)$, $(P3)$. This implies by Proposition 3.2 that the operator A^{q_h} satisfies $(r2)$, $(r3)$. Note that q_h never satisfies the property (p), which precludes the use of Proposition 3.2 (iii).

Applying Definition 3.1, formula (4.8) can be derived easily from

$$\nabla^2 h(x)[u^{q_h}(x, v) - x] + v = 0 \quad \forall (x, v) \in C \times \mathbb{R}^n.$$

Since the mapping $M \in S_n^{++}(\mathbb{R}) \to M^{-1}$ is C^∞, we obtain by (r_h) that A^{q_h} satisfies $(r1)$.

Let us prove that A^{q_h} complies with (v) of Definition 2.1. Take $b \in \operatorname{bd} C$ and ν in $N_{\overline{C}}(b)$, and let us apply (v_h). There exist $K, \epsilon > 0$ such that for every $v \in \mathbb{R}^n$, $x \in C$, $|x - b| < \epsilon$,

$$\langle -A_x^h v, \nu \rangle = -\langle \nabla^2 h(x)^{-1} v, \nu \rangle = -\langle v, \nabla^2 h(x)^{-1} \nu \rangle \leq K|v| \langle b - x, \nu \rangle.$$

Therefore, if v is bounded, the latter amounts exactly to (v). ∎

The next lemma gives a practical means to prove that a Bregman function is in the class $\mathcal{B}_{\mathcal{C}}$.

For $a < b$ in $\overline{\mathbb{R}}$, $\varphi : (a, b) \to \mathbb{R}$, a C^2 Bregman function with zone (a, b), consider the following assumptions.

(v_l) If a is finite, there exist a neighborhood U of a in $\mathbb{R}$ and a positive constant K_l such that

$$\forall u \in U \cap (a, b) \quad \varphi''(u) \geq K_l/(u - a).$$

(v_r) If b is finite, there exist a neighborhood V of b in $\mathbb{R}$ and a positive constant K_r such that

$$\forall u \in V \cap (a, b) \quad \varphi''(u) \geq K_r/(b - u).$$

LEMMA 4.5. *Let $\varphi_1, \ldots, \varphi_n$ be some C^3 Bregman functions on $\mathbb{R}$ with zones $(a_1, c_1), \ldots, (a_n, c_n)$, $a_i < c_i$, $a_i, c_i \in \overline{\mathbb{R}}$, for all $i \in \{1, \ldots, n\}$. Assume that $\varphi_1, \ldots, \varphi_n$ satisfy (v_l), (v_r) on their respective zones, and for $\alpha > 0$ set*

$$h(x) = \frac{\alpha}{2}|x|^2 + \sum_{i=1}^{n} \varphi_i(x_i).$$

Then h belongs to $\mathcal{B}_K$, where $K = \prod_{i=1}^n (a_i, c_i)$, and A^{q_h} is an elliptic barrier operator that satisfies $(r4)$.

Proof. The fact that h is a C^3 Bregman function with zone K follows from [28, Lemma 2.8(d)], and therefore (r_h) is satisfied.

To simplify the notation, let us assume that for all $i \in \{1, \ldots, n\}$, $a_i = 0$ and $c_i = +\infty$ (which implies $K = \mathbb{R}^n_+$). For $b = (b_1, \ldots, b_n) \in \operatorname{bd} \mathbb{R}^n_+$, set $I(b) = \{i \in \{1, \ldots, n\} | b_i = 0\} \neq \emptyset$ and $J(b) = \{i \in \{1, \ldots, n\} | b_i \neq 0\}$. For each $i \in I(b)$, (v_l) yields the existence of a neighborhood U_i of 0 in $\mathbb{R}$ and $K_i > 0$ such that

$$\forall u \in U_i \cap (0, +\infty) \quad \varphi''(u) \geq K_i/u. \tag{4.9}$$

Set $U_i = \mathbb{R}^n$ for each $i \in J(b)$, and $U = \mathbb{R}^n_{++} \cap \prod_{i=1\ldots n} U_i$. Let $\nu \in N_{\overline{K}}(b)$, and observe that $\nu_i = 0$ for all $i \in J(b)$ and that $\nu_i < 0$ for all $i \in I(b)$. Therefore, for $x \in \mathbb{R}^n$, an easy computation gives

$$|\nabla^2 h(x)^{-1} \nu| \leq \sum_{i \in I(b)} \frac{-v_i}{|\alpha + \varphi_i''(x_i)|}.$$

Now if $x \in U$, (4.9) implies that

$$\begin{aligned} |\nabla^2 h(x)^{-1} \nu| &\leq \sum_{i \in I(b)} -\frac{1}{K_i} \nu_i . x_i \\ &\leq \sup_{i \in I(b)} \frac{1}{K_i} \langle b - x, \nu \rangle. \end{aligned}$$

A direct computation gives for all $x \in K$, $i, j \in \{1, \dots, n\}$,

$$\left(\nabla^2 h(x)^{-1}\right)_{i,j} = \frac{\delta_{ij}}{\alpha + \varphi_i''(x_i)},$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. Applying again (v_l), we see that A^{q_h} can be continuously extended on $\overline{K}$. Hence A^{q_h} satisfies $(r4)$. □

Example 4.1 (Bregman-based Barrier operators and their dynamics). The list of examples below shows thanks to Lemma 4.5 that many classical Bregman functions can be used to provide an elliptic barrier operator. In what follows, α is the positive regularizing term as defined in (v_h), and β is a positive parameter. For a Bregman function h with zone $I \subset \mathbb{R}$, set $h_n(x) = \sum_{i=1}^n h(x_i)$ for all $x \in I^n$.

(a) For $\theta \in (0,1)$, consider $h(s) = \frac{\alpha}{2}s^2 - \beta\frac{s^\theta}{\theta}$, $s \in \mathbb{R}_+$. Then $h \in \mathcal{B}_{\mathbb{R}_{++}}$, $h_n \in \mathcal{B}_{\mathbb{R}^n_{++}}$, and the corresponding $(A^{q_{h_n}}\text{-}DM)$ system is

$$\dot{x}_i(t) + \frac{x_i(t)^{2-\theta}}{\alpha x_i(t)^{2-\theta} + \beta(1-\theta)} \frac{\partial f}{\partial x_i}(x(t)) = 0, \quad x_i(0) > 0, \quad \forall i \in \{1, \dots, n\}. \tag{4.10}$$

(b) $h(s) = \frac{\alpha}{2}s^2 + \beta s \log s$ on $\mathbb{R}_+$ is in $\mathcal{B}_{\mathbb{R}_{++}}$, $h_n \in \mathcal{B}_{\mathbb{R}^n_{++}}$, and the associated system is

$$\dot{x}_i(t) + \frac{x_i(t)}{\alpha x_i(t) + \beta} \frac{\partial f}{\partial x_i}(x(t)) = 0, \quad x_i(0) > 0, \quad \forall i \in \{1, \dots, n\}.$$

This system is exactly the regularized Lotka–Volterra equation (1.8) recently proposed in [6]. However, it is worth noting that (1.8) was introduced there as a continuous model based not on (BPM) but on the proximal-like method

$$x^{k+1} \in \arg\min \left\{ f(x) + c_k d_\varphi(x, x^k) | x \in \mathbb{R}^n_+ \right\}, \; c_k > 0,$$

where $\varphi(s) = s - \log s - 1$ and $d_\varphi(x,y) = \frac{\alpha}{2}|x-y|^2 + \beta \sum_{i=1}^n y_i \varphi(y_i^{-1}x_i)$ for all x, y in $\mathbb{R}^n_{++}$. For more results and applications on classical Lotka–Volterra systems, see, e.g., Hofbauer and Sigmund [23].

(c) $h(s) = \frac{\alpha}{2}s^2 - \beta\sqrt{1-s^2}$ on $[-1,1]$ is in $\mathcal{B}_{(-1,1)}$, $h_n \in \mathcal{B}_{(-1,1)^n}$, and the corresponding system is

$$\dot{x}_i(t) + \frac{\left(1 - x_i(t)^2\right)^{3/2}}{\alpha\left(1 - x_i(t)^2\right)^{3/2} + \beta} \frac{\partial f}{\partial x_i}(x(t)) = 0, \quad x_i(0) \in (-1,1), \quad \forall i \in \{1, \dots, n\}.$$

(d) $h(s) = \frac{\alpha}{2}s^2 - \beta\sqrt{s(1-s)}$ on $[0,1]$ is in $\mathcal{B}_{(0,1)}$, $h_n \in \mathcal{B}_{(0,1)^n}$, and the corresponding system is

$$\dot{x}_i(t) + \frac{4x_i(t)^{3/2}(1-x_i(t))^{3/2}}{4\alpha x_i(t)^{3/2}(1-x_i(t))^{3/2} + \beta} \frac{\partial f}{\partial x_i}(x(t)) = 0, \quad x_i(0) \in (0,1), \quad \forall i \in \{1, \dots, n\}.$$

Remark 4.2. For $\epsilon, \gamma \geq 0$, and $f \in C^3(\mathbb{R}^n, \mathbb{R})$, set $h_{\epsilon,\gamma}(x) = \frac{\epsilon}{2}|x|^2 + \gamma f(x)$ for all $x \in \mathbb{R}^n$. Then we have $h_{\epsilon,\gamma} \in \mathcal{B}_{\mathbb{R}^n}$, under one of the following assumptions:

($\star$) f is strongly convex, i.e., $\nabla^2 f - \lambda I$ is positive semidefinite, with $\lambda > 0$.

($\star$) f is convex and $\epsilon > 0$.

($\star$) $\gamma = 0$, $\epsilon > 0$.

Letting $\epsilon = 0$, $\gamma = 1$ in the first case yields the continuous Newton descent method (1.9). The second version can be seen, for ϵ small, as a regularized Newton method

$$(A^{q_{h_\epsilon,\gamma}} - DM) \qquad \dot{x}(t) + [\epsilon Id + \gamma \nabla^2 f(x(t))]^{-1} \nabla f(x(t)) = 0.$$

The last point with $\gamma = 0$, $\epsilon > 0$ gives rise to the classical steepest descent method.

In the examples just described, the $A^{q_{h_\epsilon,\gamma}}$ are elliptic barrier operators *on* $\mathbb{R}^n$ so that the feasible set $\overline{C}$ is the whole space $\mathbb{R}^n$, and $(v)_h$ holds vacuously. It actually raises another interesting aspect of barrier operators: they can also be used as a geometrical means to improve convergence rate as well as well-posedness properties. This suggests, for instance, to go further in the study of the Newton–Barrier methods

$$\dot{x}(t) + [\lambda \nabla^2 h(x(t)) + \mu \nabla^2 f(x(t))]^{-1} \nabla f(x(t)) = 0,\ t \geq 0,$$

with $\lambda, \mu > 0$ and where h is a C^3 Bregman function.

Continuous model II. The Bregman distances appearing in the definition of projection methods (section 4.1) can be used in a quite different way in order to provide some other continuous model of (BPM). Indeed, replacing the kernel h_0 defined on the whole space $\mathbb{R}^n$ by some essentially smooth convex function (see definition below) allows us to get rid of the normal cone and to reformulate (4.3) as

$$\nabla h(x(t) + \dot{x}(t)) - \nabla h(x(t)) + \nabla f(x(t)) = 0 \quad \forall t \geq 0.$$

This can be discretized as

$$\nabla h(x_{k+1}) - \nabla h(x_k) + \nabla f(x_{k+1}) = 0 \quad \forall k \in N,$$

and (BPM) is recovered with a sequence of step-sizes satisfying $c_k = 1$ for all $k \in N$.

This model will be derived from our general framework developed in section 3. First, we recall the definition of essentially smooth convex functions; see [36].

DEFINITION 4.6. *A proper convex function $\phi : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ is essentially smooth if it satisfies the following:*

(i) *The interior of* dom ϕ *is nonempty; i.e.,* int dom $\phi \neq \emptyset$.
(ii) ϕ *is differentiable on* int dom ϕ.
(iii) *For all b in the boundary of* int dom ϕ *and all sequence x_k, $k \in N$, in* int dom ϕ *such that $x_k \to b$ as $k \to +\infty$, we have $|\nabla \phi(x_k)| \to +\infty$ as $k \to +\infty$.*

As in subsection 4.1, we now study operators of the form A^{D_h} (cf. (4.1)) for some relevant kernels h. Let $h_0 : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ be a closed proper convex function such that

$(\mathrm{i})_{h_0}$ h_0 is essentially smooth with in addition int dom $h_0 = C$, and
$(\mathrm{ii})_{h_0}$ ∇h_0 is Lipschitz continuous on compact subsets of C.

For such a function h_0, we set $h(u) = \alpha/2|u|^2 + h_0(u)$ for all $u \in \mathbb{R}^n$. In the following proposition, it is important to recall that D_h is an extended real function defined on the whole of $\mathbb{R}^n \times C$.

PROPOSITION 4.7. *Let h be as above. Then A^{D_h} is an elliptic barrier operator on C, and for all $(x, v) \in C \times \mathbb{R}^n$ we have*

$$A_x^{D_h} v = x - \nabla h^*(\nabla h(x) - v). \tag{4.11}$$

Proof. From $(\mathrm{i})_{h_0}$ it follows that $D_h \in \mathcal{D}$. Using the fact that h is essentially smooth with int dom $h = C$, we deduce that (p) is satisfied. By Proposition 3.2, we

see that A^{D_h} verifies $(r2)$, $(r3)$, and (v). The formula (4.11) follows from (3.6), and $(r1)$ follows from $(\mathrm{ii})_{h_0}$. □

The associated A^{D_h}-DM is thus given by

$$\dot{x}(t) + x(t) - \nabla h^*[\nabla h(x(t)) - \nabla f(x(t))] = 0,\ x(0) \in C, \quad \forall t \geq 0, \tag{4.12}$$

or using $\nabla h^* = (\nabla h)^{-1}$ equivalently as

$$\nabla h(x(t) + \dot{x}(t)) - \nabla h(x(t)) + \nabla f(x(t)) = 0,\ x(0) \in C, \quad \forall t \geq 0.$$

Example 4.2. Consider the regularized Burg's entropy obtained with $g(s) = (\alpha/2)s^2 - \beta \log s,\ s > 0$, where β is a positive parameter. For $x \in \mathbb{R}^n_{++}$ set $h(x) = \sum_{i=1}^n g(x_i)$. The function h satisfies the requirements of Proposition 4.7. A direct computation shows that

$$(g^*)'(u) = \frac{u + \sqrt{u^2 + 4\alpha\beta}}{2\alpha} \quad \forall u \in \mathbb{R}.$$

Substituting in (4.12), the following descent method is derived: For all $i = 1, \ldots, n$,

$$\dot{x}_i(t) + x_i(t)/2 + (2\alpha)^{-1}\Bigg(\beta/x_i(t) + \frac{\partial f}{\partial x_i}(x(t)) - \sqrt{[\alpha x_i(t) - \beta/x_i(t) - \frac{\partial f}{\partial x_i}(x(t))]^2 + 4\alpha\beta}\Bigg) = 0$$

for all $t \geq 0$ and with $x_i(0) > 0$ for all $i \in \{1, \ldots, n\}$.

It is interesting to note that as $\alpha \to 0$ we do not recover here the Lotka–Volterra system; compare this with the system given in Example 4.1 (b).

4.3. A continuous model for proximal algorithms with second order kernels. The class of operators A^{d_φ} defined in this section are built upon the kernels φ which are used to realize the $(RIPM)$ method introduced in [8], and which we now recall. Let $\varphi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be a closed proper function whose domain dom φ is a subset of $[0, +\infty)$. Consider the following assumptions on φ:

- $(\mathrm{i})_\varphi$ φ is finite and C^2 on $(0, +\infty)$.
- $(\mathrm{ii})_\varphi$ φ is strictly convex on $(0, +\infty)$.
- $(\mathrm{iii})_\varphi$ $\lim_{s>0,\ s\to 0} \varphi'(s) = -\infty$.
- $(\mathrm{iv})_\varphi$ $\varphi(1) = \varphi'(1) = 0$ and $\varphi''(1) > 0$.
- $(\mathrm{v})_\varphi$ For all $s > 0$, $\varphi''(1)(1 - \frac{1}{s}) \leq \varphi'(s) \leq \varphi''(1)(s - 1)$.

Now for $\alpha, \beta > 0$, set

$$\varphi(s) = \frac{\alpha}{2}(s-1)^2 + \beta\varphi_0(s), \tag{4.13}$$

where φ_0 satisfies $(i)_\varphi - (v)_\varphi$, and denote by Φ the class of such functions. For $\varphi \in \Phi$, set

$$\forall (u, x) \in \mathbb{R}^n \times \mathbb{R}^n_{++} \quad d_\varphi(u, x) = \sum_{i=1}^n x_i^2 \varphi(x_i^{-1} u_i). \tag{4.14}$$

It is proved in [8] that the associated proximal method,

$$(RIPM) \qquad x^{k+1} \in \arg\min \left\{ f(x) + c_k d_\varphi(x, x_k) | x \in \mathbb{R}^n_+ \right\},\ c_k > 0,$$

generates a positive sequence $\{x^k\}$ provided that $x^0 \in \mathbb{R}^n_{++}$. As a consequence, an equivalent formulation of $(RIPM)$ is

$$c_k \partial_1 d_\varphi(x^{k+1}, x^k) + \nabla f(x^{k+1}) = 0 \quad \forall k \geq 1. \tag{4.15}$$

Under the additional assumptions that $\arg\min_{\mathbb{R}^n_+} f \neq \emptyset$, $\sum_{k=1}^{+\infty} c_k = \infty$, and

$$\alpha \geq \beta \varphi_0''(1), \tag{4.16}$$

it is proved in [8] that the sequence x^k, $k \in N$, converges to a minimizer of f.

Following the general framework developed in section 3, we generate the elliptic barrier operator and dynamical system associated with $(RIPM)$.

PROPOSITION 4.8. *Let $\varphi \in \Phi$. Then A^{d_φ} is an elliptic barrier operator, and one has for all $(x, v) \in \mathbb{R}^n_{++} \times \mathbb{R}^n$*

$$\left(A^{d_\varphi}_x v\right)_i = x_i - x_i(\varphi^*)'(-x_i^{-1} v_i) \quad \forall i = 1, \ldots, n. \tag{4.17}$$

Proof. For all $(u, x) \in \mathbb{R}^n_+ \times \mathbb{R}^n_{++}$ we have $d_\varphi(u, x) = \alpha/2|u - x|^2 + \beta d_{\varphi_0}(u, x)$, and therefore to prove that $d_\varphi \in \mathcal{D}$, we need to show that βd_{φ_0} satisfies $(P1)$, $(P2)$, and $(P3)$. $(P1)$ follows from $(\mathrm{i})_\varphi$, while $(P3)$ is a consequence of the definition of φ_0. Using $(\mathrm{iv})_\varphi$, we see by a direct computation that $(P2)$ is satisfied and thus that $d_\varphi \in \mathcal{D}$.

Using Definition 3.1 with $d := d_\varphi \in \mathcal{D}$, the optimality conditions for (3.5) yield

$$v_i + x_i \varphi'(u_i x_i^{-1}) = 0 \quad \forall i = 1, \ldots, n,$$

from which formula (4.17) follows easily using $(\varphi^*)' = (\varphi')^{-1}$. Since dom $\varphi \subset \mathbb{R}_+$, we have for all $x \in \mathbb{R}^n_{++}$, dom $d_\varphi(., x) \subset \mathbb{R}^n_+$, and therefore by Proposition 3.2 A^{d_φ} satisfies $(r2)$, $(r3)$, and (v).

It remains to prove that $(r1)$ holds. Using formula (4.17), and since $x \in \mathbb{R}^n_{++}$, it thus suffices to show that $(\varphi^*)'$ is Lipschitz continuous. However, since here φ is a smooth α-strongly convex function, one has

$$(t - s)(\varphi'(t) - \varphi'(s)) \geq \alpha(t - s)^2 \quad \forall t, s > 0,$$

and thus recalling that $(\varphi^*)' = (\varphi')^{-1}$, one easily deduces the required Lipschitz property for $(\varphi^*)'$ and $(r1)$ follows. ∎

Remark 4.3. (a) Requirement $(\mathrm{v})_\varphi$ allows acute controls on d_φ in the asymptotic analysis of $(RIPM)$ and $(A^{d_\varphi}\text{-}DM)$ (see section 5, Theorem 5.4) and is actually not needed for the above result. Technically those controls are the reason why our operator is based on φ and not on φ^*.

(b) The assumption $(\mathrm{iii})_\varphi$ reduces the computation of $(\varphi^*)'$ to the inversion of $\varphi'_{|(0,+\infty)}$.

(c) Note also that A^{d_φ} does not satisfy $(r4)$ in general, but as we shall see in the next section it has no consequence on the asymptotic study of $(A^{d_\varphi}\text{-}DM)$ when f is convex.

(d) One could also develop a similar construction with "regularized φ-divergence" distance-like functions, that is,

$$d(u, x) = \frac{\alpha}{2}|u - x|^2 + \sum_{i=1}^{n} x_i \varphi(u_i x_i^{-1}),\ u, x \in \mathbb{R}^n_+,$$

where $\varphi : \mathbb{R} \mapsto \mathbb{R} \cup \{+\infty\}$ is an essentially smooth convex function such that $(0, +\infty) \subset \operatorname{dom} \varphi \subset [0, +\infty)$. Unfortunately the parameter α forbids the computation of the Legendre conjugates of $\partial d(\cdot, x)$, $x \in C := \mathbb{R}^n_+$, and leads to purely theoretical conclusions. This gives new motivation to study barrier operators for which $\alpha = 0$ (see Remark 2.1 (a)).

The $(A^{d_\varphi}\text{-}DM)$ system is given by

$$\dot{x}_i(t) + x_i(t) - x_i(t)(\varphi^*)' \left(-x_i(t)^{-1} \frac{\partial f}{\partial x_i}(x(t)) \right) = 0 \quad \forall t \geq 0,$$

or, equivalently, as

$$x_i(t)\varphi' \left(\frac{\dot{x}_i(t) + x_i(t)}{x_i(t)} \right) + \frac{\partial f}{\partial x_i}(x(t)) = 0, \ t \geq 0.$$

To recover $(RIPM)$ by some discretization of $(A^{d_\varphi}\text{-}DM)$, the latter can be reformulated in the following way:

$$\partial_1 d_\varphi \left(x(t) + \dot{x}(t), x(t) \right) + \nabla f(x(t)) = 0, \ x(0) \in \mathbb{R}^n_{++}, \quad \forall t \geq 0. \tag{4.18}$$

Now, if we perform an implicit discretization of (4.18), it yields

$$\partial_1 d_\varphi \left(x^{k+1}, x^k \right) + \nabla f(x^{k+1}) = 0, x^0 = x(0), \ k \in N.$$

which is exactly (4.15), with $c_k = 1$.

Example 4.3. It is a delicate matter to build a function in Φ whose Fenchel conjuguate is easily computable. As in [8] we focus on the important special choice given by a logarithmic-quadratic kernel,

$$\varphi(s) = \frac{\alpha}{2}(s-1)^2 + \beta(-\log s + s - 1), \ s > 0,$$

which admits (see [8, p. 665]) an explicit conjugate $\varphi^* \in C^\infty(\mathbb{R})$, and with

$$(\varphi^*)'(s) = \frac{1}{2\alpha}[\alpha - \beta + s + \sqrt{(\alpha - \beta + s)^2 + 4\alpha\beta}] \quad \forall s \in \mathbb{R}.$$

The corresponding $(A^\varphi\text{-}DM)$ system is then given by

$$\dot{x}_i(t) + \frac{\alpha + \beta}{2\alpha} x_i(t) + \frac{1}{2\alpha} \frac{\partial f(x(t))}{\partial x_i} - \sqrt{\frac{1}{4\alpha^2} \left[(\alpha - \beta) x_i(t) + \frac{\partial f(x(t))}{\partial x_i} \right]^2 + 4\alpha\beta x_i(t)^2} = 0, \tag{4.19}$$

with $i \in \{1, \ldots, n\}$, $t \geq 0$, and $x(0) \in \mathbb{R}^n_{++}$. An interesting fact to note is that (4.19) has a sense for any $x(0) \in \mathbb{R}^n$; this suggests like in [13] a study of its properties for *nonfeasible* initial data.

5. Asymptotic analysis for a convex objective function. In what follows, f satisfies the additional assumptions

$$(\mathcal{H}') : \begin{cases} f \text{ is convex,} \\ \arg\min_{\overline{C}} f \neq \emptyset. \end{cases}$$

This section proposes a criterion concerning elliptic barrier operators to obtain the convergence of the trajectories of (*A-DM*). It is based on Lyapunov functionals and to their (theoretical) decreasing rate. This natural approach is inspired by the classical result of Bruck [15] on the generalized steepest descent method, and by the notions of Fejer or quasi-Fejer sequences which go back to the work of Ermoliev [21] and arise in monotone and generalized gradient optimization algorithms. Such techniques have also been applied successfully to second order in time systems by Alvarez [1] and Alvarez and Attouch [2]. Before stating the main result of this section, let us describe the typical properties of those Lyapunov functionals, sometimes called *relative entropy*, when working on systems in the nonnegative orthant; see, e.g., [23]. In what follows, S should be understood as the set of equilibria of some convex function.

We suggest the following general definition for viable Lyapunov functionals.

DEFINITION 5.1. *Let $S \subset \overline{C}$ be a nonempty set. A family of functions $\{e_a, a \in S\}$ is Lyapunov viable if it satisfies*

- $(\mathrm{i})_e$ *For all $a \in S$, $e_a : C \to \mathbb{R}$ is C^1.*
- $(\mathrm{ii})_e$ *The functions e_a are nonnegative for all $a \in S$.*
- $(\mathrm{iii})_e$ *For all $a \in S$, e_a is* inf *bounded. That is, for every $r \in \mathbb{R}$, the set $\{y \in C | e_a(y) \leq r\}$ is bounded.*
- $(\mathrm{iv})_e$ *Let $x^k, k \in N$ be a sequence in C. Then for all $a \in S$,*

$$e_a(x^k) \to 0 \text{ as } k \to +\infty \iff x^k \to a \text{ as } k \to +\infty.$$

The next result is a key lemma that can be used to establish convergence of trajectories of (*A-DM*). First, we recall the following classical result (see, e.g., [1, Lemma 2.2]) which will be useful to us.

LEMMA 5.2. *Let $h : \mathbb{R} \to \mathbb{R}^+$ be a C^1 function. If $(h')^+ := \max(0, h')$ is in $L^1(0, +\infty; \mathbb{R})$, then $\lim_{t \to +\infty} h(t)$ exists.*

Let us set $S := \arg\min_{\overline{C}} f$.

LEMMA 5.3. *Let A be an elliptic barrier operator on C and f a function satisfying $(\mathcal{H}_1)$, $(\mathcal{H}_2)$, $(\mathcal{H}')$. Assume that there exist $\lambda > 0$, $\mu \in \mathbb{R}$, and a family of functions $\{e_a, a \in S\}$ that is Lyapunov viable (i.e., satisfying $(\mathrm{i})_e - (\mathrm{iv})_e$). Suppose, in addition, that for all $x \in C$,*

$$\langle -A_x \nabla f(x), \nabla e_a(x) \rangle + \lambda \langle \nabla f(x), x - a \rangle \leq \mu |A_x \nabla f(x)|^2. \tag{5.1}$$

If $x(t)$ is the solution of (A-DM), then the following hold:

(i) *$f(x(t)) \to \inf_{\overline{C}} f$ as $t \to +\infty$, with the estimation*

$$f(x(t)) - \inf_{\overline{C}} f \leq Mt^{-1} \text{ for some } M > 0.$$

(ii) *$\dot{x}(t) \to 0$ as $t \to +\infty$.*

(iii) *There exists $x^* \in S$ such that $x(t) \to x^*$ as $t \to +\infty$.*

Proof. Let $a \in S$; by (5.1) and (*A-DM*) we obtain

$$\frac{d}{dt} e_a(x(t)) + \lambda \langle \nabla f(x(t)), x(t) - a \rangle \leq \mu |\dot{x}(t)|^2, \ t \geq 0. \tag{5.2}$$

From the convex inequality it follows that for all $y \in \overline{C}$,

$$0 \geq f(a) - f(y) \geq \langle \nabla f(y), a - y \rangle. \tag{5.3}$$

Combining (ii) of Theorem 2.4, (5.3), and (5.2) yields $[\frac{d}{dt}e_a(x(t))]^+ \leq \mu|\dot{x}(t)|^2$, $t \geq 0$. From $(\text{ii})_e$ and Lemma 5.2, we deduce that $e_a(x(t))$ converges as $t \to +\infty$. Hence, by $(\text{iii})_e$, $x(\cdot)$ is bounded.

Coming back to (5.2), we obtain for all $T \geq 0$

$$\lambda \int_0^T \langle \nabla f(x(t)), x(t) - a\rangle dt \leq \int_0^T |\dot{x}(t)|^2 dt + e_a(x(0)) - e_a(x(T)),$$

and since $\lambda > 0$,

$$\langle \nabla f(x(\cdot)), x(\cdot) - a\rangle \in L^1(0, \infty; \mathbb{R}). \tag{5.4}$$

From (5.4), $(\mathcal{H}_1)$, and the boundedness property of x, we obtain that there exist $x^* \in \overline{C}$ and a nondecreasing sequence $t_k, k \in N$ such that $\langle \nabla f(x(t_k)), x(t_k) - a\rangle \to 0$ and $x(t_k) \to x^*$ as $k \to +\infty$. Using (5.3), it follows that $f(x^*) \leq f(a)$ and thus $x^* \in S$.

By Theorem 2.4 (iii) and the continuity of f, we see that the latter argument implies $f(x(t)) \to \inf_{\overline{C}} f$ as $t \to +\infty$ and that all limit points of x are in S.

To prove the second part of (i), we first deduce from (5.2) and (5.3) that

$$\frac{d}{dt}e_a(x(t)) + \lambda(f(x(t)) - f(a)) \leq \mu|\dot{x}(t)|^2,\ t \geq 0.$$

By integration it follows from Theorem 2.4 (iii) that for $t \geq 0$, $t\lambda[f(x(t)) - \inf_{\overline{C}} f] \leq e_a(x(0)) - e_a(x(t)) + \mu \int_0^t |\dot{x}|^2$. Using $(\text{iii})_e$, we obtain for all $t > 0$

$$\lambda \left[f(x(t)) - \inf_{\overline{C}} f \right] \leq \frac{1}{t} \left[e_a(x(0)) + \mu \int_0^t |\dot{x}|^2 \right]. \tag{5.5}$$

The estimate announced in (i) is then a consequence of Theorem 2.4 (iv).

Let x_1^* and x_2^* be two cluster points of $x(\cdot)$ and $t_k, \tau_k, k \in N$, increasing sequences in $\mathbb{R}^+$, such that $x(t_k) \to x_1^*$, $x(\tau_k) \to x_2^*$ as $k \to +\infty$. From $(\text{iv})_e$, we deduce that $e_{x_1^*}(x(t_k)) \to 0$ as $k \to +\infty$. However, since the function $e_{x_1^*}(x(\cdot))$ has a limit as $t \to +\infty$, we also have $e_{x_1^*}(x(\tau_k)) \to 0$ as $k \to +\infty$, and by applying $(\text{iv})_e$ again, we obtain $x_1^* = x_2^*$.

Let x^* be the limit point of $x(\cdot)$, it verifies the classical relation $\nabla f(x^*) \in -N_{\overline{C}}(x^*)$, and therefore $(\mathcal{H}_1)$ implies that $(x(t), \nabla f(x(t)))$ has its limit point in $\{x^*\} \times -N_{\overline{C}}(x^*)$. Applying Proposition 2.2, it follows that $\dot{x}(t) \to 0$ as $t \to +\infty$. □

Remark 5.1. (a) If $\mu \leq 0$, we have by (5.5)

$$f(x(t)) - \inf_{\overline{C}} f \leq \frac{1}{\lambda t} e_a(x(0)) \quad \forall t > 0.$$

(b) Note that Lemma 5.3 allows us to handle the case $\mu > 0$ in (5.1), which corresponds to quasi-Fejer convergence.

(c) The property $(r3)$ has not been used, but it is implicitly contained in (5.1).

(d) Note also that the above result holds for an elliptic barrier operator which is possibly undefined on $\operatorname{bd} C \times \mathbb{R}^n$.

Let us apply this result to some of the operators defined in section 4. In what follows, *it is implicitly assumed that* $C = \mathbb{R}^n_{++}$ *when dealing with operators of the type* A^{d_φ}, $\varphi \in \Phi$, while A^P is the gradient projection operator (cf. subsection 4.1).

THEOREM 5.4. *Let $\varphi \in \Phi$ such that $\alpha \geq \beta\varphi_0''(1)$, $h \in \mathcal{B}_C$, and assume that f satisfies $(\mathcal{H}_1)$, $(\mathcal{H}_2)$, $(\mathcal{H}')$. Then the trajectories of $(A^P$-$DM)$, $(A^{q_h}$-$DM)$, and $(A^{d_\varphi}$-$DM)$ converge to some minimizer of f on $\overline{C}$. Moreover, for all trajectories x, the following properties hold:*

(i) $f(x(t)) \to \inf_{\overline{C}} f$ *as* $t \to +\infty$, *with the estimation*

$$f(x(t)) - \inf_{\overline{C}} f \leq Mt^{-1}, \quad \text{where } M > 0.$$

(ii) $\dot{x}(t) \to 0$ *as* $t \to +\infty$.

Proof. By Propositions 4.1, 4.4, and 4.8, we know that A^P, A^{q_h}, and $A^{d\varphi}$ are elliptic barrier operators. For every $a \in S$ and for all $x \in C$, set

$$\begin{aligned} e_a^P(x) &= f(x) - f(a) + \tfrac{1}{2}|x-a|^2, \\ e_a^h(x) &= D_h(a,x) = \tfrac{\alpha}{2}|x-a|^2 + D_{h_1}(a,x), \\ e_a^\varphi(x) &= f(x) - f(a) + \theta|x-a|^2, \end{aligned}$$

where $\theta = (\alpha + \varphi_0''(1))/2$. Naturally the idea is to apply Lemma 5.3 to the operators A^P, A^{q_h}, and A^{d_φ}. Let $a \in S$. The functions e_a^φ, e_a^h, and e_a^P clearly satisfy $(\mathrm{i})_e$, $(\mathrm{ii})_e$. To obtain $(\mathrm{iii})_e$, just notice that in the three cases, the structure of the functions has the form

$$\xi_a(x) = k|x-a|^2 + \rho_a(x) \quad \forall x \in C,$$

with $\rho_a \geq 0$, $k > 0$. By definition of a Bregman function and by Lemma 4.3, e_a^h verifies $(\mathrm{iv})_e$. To prove that e_a^P and e_a^φ satisfy $(\mathrm{iv})_e$, we just have to combine $(\mathcal{H})$ and the fact that a is a minimizer of f on $\overline{C}$. Let us prove that the property (5.1) holds for the couples (e_a^P, A^P), (e_a^h, A^{q_h}), and (e_a^P, A^{d_φ}).

• The continuous gradient projection method has already been studied from different viewpoints in [13], but for the sake of completeness we recall the argument. Let $x \in C$ and $a \in S$. The optimality property of the orthogonal projection operator gives, for all $\xi \in \overline{C}$, $\langle x - \nabla f(x) - P_{\overline{C}}(x - \nabla f(x)), \xi - P_{\overline{C}}(x - \nabla f(x))\rangle \leq 0$. Therefore, if $\xi = a$, we obtain

$$\langle -\nabla f(x) + A_x^P \nabla f(x), a - x + A_x^P \nabla f(x)\rangle \leq 0$$

or, equivalently, $\langle -A_x^P \nabla f(x), x - a + \nabla f(x)\rangle + |A_x^P \nabla f(x)|^2 + \langle \nabla f(x), x - a\rangle \leq 0$, which is (5.1) with $\mu = -1$.

• Now, let us consider A^{q_h}, where h is Bregman function that belongs to $\mathcal{B}_C$. Let us compute the gradient of e_a^h for all $a \in S$. For all $x \in C$, we have

$$\begin{aligned} \nabla e_a^h(x) &= \nabla[h(a) - h(\cdot) - \langle \nabla h(\cdot), a - \cdot\rangle](x) \\ &= \nabla^2 h(x)(x-a). \end{aligned}$$

And therefore, $\langle -A_x^{q_h} \nabla f(x), \nabla e_a^h(x)\rangle = -\langle \nabla^2 h(x)^{-1} \nabla f(x), \nabla^2 h(x)(x-a)\rangle = -\langle \nabla f(x), x - a\rangle$, which verifies (5.1) with $\mu = 0$ and $\lambda = 1$.

• Finally, let us deal with e_a^P, A^{d_φ}. Our approach relies on the following key lemma proven in [8, Lemma 3.4].

LEMMA 5.5. *For every $y_1 \in \mathbb{R}_+^n$ and for every $(y_1, y_2) \in \mathbb{R}_{++}^n \times \mathbb{R}_{++}^n$, we have*

$$\langle y_1 - y_2, \partial_1 d_\varphi(y_2, y_3)\rangle \leq \theta\left(|y_1 - y_3|^2 - |y_1 - y_2|^2\right).$$

Note that it is here that the property $(\mathrm{v})_\varphi$ is needed. Indeed, the proof of this lemma is based on that assumption, together with the condition $\alpha \geq \beta\varphi_0''(1)$.

For all $i \in \{1 \dots, n\}$ and all $x \in \mathbb{R}^n_{++}$, set $(v_x)_i = -(A^{d_\varphi}_x \nabla f(x))_i$. The A^{d_φ}-*DM* can be rewritten as

$$\partial_1 d_\varphi \left(x(t) + v_x, x\right) + \nabla f(x) = 0 \quad \forall x \in \mathbb{R}^n_{++}. \tag{5.6}$$

Observe that $x \in \mathbb{R}^n_{++}$ implies $x + v_x \in \mathbb{R}^n_{++}$. Now for $a \in \arg\min_{\mathbb{R}^n_+} f$ and for all $x \in \mathbb{R}^n_{++}$, let us multiply (5.6) by $a - x - v_x$; this gives

$$\langle a - (v_x + x), \partial_1 d_\varphi(x + v_x, x) \rangle + \langle \nabla f(x), a - x - v_x \rangle = 0,$$

and therefore by Lemma 5.5

$$\theta \left(|a - x|^2 - |a - x - v_x|^2 \right) + \langle \nabla f(x), a - x - v_x \rangle \geq 0.$$

After direct algebra, this reduces to

$$\langle v_x, 2\theta(x - a) + \nabla f(x) \rangle + \langle \nabla f(x), x - a \rangle + |v_x|^2 \leq 0 \quad \forall x \in \mathbb{R}^n_{++}.$$

Recalling that $v_x = -A^{d_\varphi}_x \nabla f(x)$, we easily see that (5.1) is satisfied. □

Remark 5.2. The convergence of the orbits generated by the other operators proposed in section 4 remains an open question.

Localization of the limit point. Let A be an elliptic barrier operator, and let e_a be a family of viable Lyapunov functionals satisfying (5.1) with $\mu \leq 0$. We assume, moreover, that for all a in $S \subset \overline{C}$, there exist a nonnegative convex function $\rho_a : C \mapsto \mathbb{R}$ and $k > 0$ such that

$$e_a(x) = k|x - a|^2 + \rho_a(x) \quad \forall x \in C. \tag{5.7}$$

As in Lemaire [29], and inspired by the recent non Euclidean extension given in [6], the limit point of the trajectory produced by (*A-DM*) can be localized.

PROPOSITION 5.6. *Let A be an elliptic barrier operator on C, and let $\{e_a, a \in S\}$ be as defined in* (5.7). *Then the trajectory of (A-DM), with $x(0) \in C$, converges to a minimizer x_∞ of f on $\overline{C}$, with the following estimation:*

$$|x_\infty - x(0)|^2 \leq \inf \left\{ 4|x(0) - a|^2 + \frac{2}{k} \rho_a(x(0)) \mid a \in S \right\}.$$

Proof. The convergence result of the trajectory $x(t)$ to $x_\infty \in S = \arg\min_{\overline{C}} f$ is a direct consequence of Lemma 5.3. To prove the estimation, let us come back to the inequality (5.2) proven in Lemma 5.3:

$$\frac{d}{dt} e_a(x(t)) + \lambda \langle \nabla f(x(t)), x(t) - a \rangle \leq \mu |\dot{x}(t)|^2, \; t \geq 0.$$

The convexity property of f and the fact that $\mu \leq 0$ imply that $\mathbb{R}_+ \ni t \mapsto e_a(x(t))$ is nonincreasing. Therefore, for all $a \in S$, we have $e_a(x(t)) \leq e_a(x(0))$, where $t \geq 0$. Since $\rho_a \geq 0$, by letting $t \to +\infty$, (5.7) yields

$$k|x_\infty - a|^2 \leq k|x(0) - a|^2 + \rho_a(x(0)). \tag{5.8}$$

Now for all $a \in S$, we have

$$\begin{aligned} |x_\infty - x(0)|^2 &\leq [|x_\infty - a| + |a - x(0)|]^2 \\ &\leq 2|x_\infty - a|^2 + 2|a - x(0)|^2 \\ &\leq 4|x(0) - a|^2 + \tfrac{2}{k} \rho_a(x(0)), \end{aligned}$$

where the third inequality is a consequence of (5.8). The desired result is then obtained by taking the infimum over all $a \in S$. □

As a consequence, we then have the following corollary.

COROLLARY 5.7. *Under the assumptions of Theorem* 5.4*, we have*

$$|x_\infty - x(0)|^2 \leq 4 \inf \left\{ |x(0) - a|^2 + \frac{1}{\alpha} D_{h_1}(a, x(0)) \mid a \in S \right\}$$

if $A = A^{q_h}, h(\cdot) = \alpha/2|\cdot|^2 + h_1(\cdot)$.

Defining $s : \mathbb{R}^n \mapsto \mathbb{R}$ *as* $s(y) := \inf\{|y - a|^2 \mid a \in S\}$*, we then also have*

$$|x_\infty - x(0)|^2 \leq 4 \left(s(x(0)) + f(x(0)) - \inf_{\overline{C}} f \right)$$

if $A = A^P$*, and*

$$|x_\infty - x(0)|^2 \leq 4s(x(0)) + \frac{2}{\theta} \left(f(x(0)) - \inf_{\overline{C}} f \right)$$

if $A = A^{d_\varphi}$.

Proof. The families $\{e_a^h, e_a^P, e_a^\varphi, a \in S\}$ introduced in the beginning of the proof of Theorem 5.4 satisfy the assumptions of Proposition 5.6, and thus the claimed results follow easily. □

Remark 5.3. (a) The estimations given in Corollary 5.7 for $A = A^{q_h}$ allow us to recover the results obtained in [6, 29].

(b) Assume that f is a linear function, that is, $f(x) = \langle c, x \rangle$ for all $x \in \mathbb{R}^n$, where $c \in \mathbb{R}^n$. Take h as in Theorem 5.4. A straightforward integration of (A^{q_h}-DM) in its form given in (4.6) yields

$$\nabla h(x(t)) - \nabla h(x(0)) + tc = 0 \quad \forall t \geq 0. \tag{5.9}$$

As already noted in [24], the trajectory of (A^{q_h}-DM) can be viewed as an optimal path relatively to the barrier function D_h. Indeed, since for all $(y, z) \in C \times C$, $\nabla_1 D_h(y, z) = \nabla h(y) - \nabla h(z)$, (5.9) can be reformulated as

$$x(t) \in \arg\min \left\{ \langle c, u \rangle + \frac{1}{t} D_h(u, x(0)) \mid u \in \mathbb{R}^n \right\}, \ t > 0.$$

The convergence techniques developed in [24] but also the viscosity methods studied in Attouch [5] allow us to fully characterize the limit point as

$$x_\infty \in \arg\min \{D_h(a, x(0)) \mid a \in S\}.$$

REFERENCES

[1] F. ALVAREZ, *On the minimizing property of a second order dissipative system in Hilbert spaces*, SIAM J. Control Optim., 38 (2000), pp. 1102–1119.

[2] F. ALVAREZ AND H. ATTOUCH, *An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping*, Set-Valued Anal., 9 (2001), pp. 3–11.

[3] F. ALVAREZ AND J. M. PÉREZ, *A dynamical system associated with Newton's method for parametric approximations of convex minimization problems*, Appl. Math. Optim., 38 (1998), pp. 193–217.

[4] A. S. Antipin, *Minimization of convex functions on convex sets by means of differential equations*, Differential Equations, 30 (1994), pp. 1365–1375.
[5] H. Attouch, *Viscosity solutions of minimization problems*, SIAM J. Optim., 6 (1996), pp. 769–806.
[6] H. Attouch and M. Teboulle, *A regularized Lotka Volterra dynamical system as a continuous proximal-like method in optimization*, J. Optim. Theory Appl., to appear.
[7] J. P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, New York, 1984.
[8] A. Auslender, M. Teboulle, and S. Ben-Tiba, *Interior proximal and multiplier method based on second order homogeneous kernels*, Math. Oper. Res., 24 (1999), pp. 645–668.
[9] A. Auslender, R. Cominetti, and M. Haddou, *Asymptotic analysis for penalty and barrier methods in convex and linear programming*, Math. Oper. Res., 22 (1997), pp. 43–62.
[10] J. B. Baillon and G. Haddad, *Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones*, Israel J. Math., 26 (1977), pp. 137–150.
[11] D. A. Bayer and J. C. Lagarias, *The nonlinear geometry of linear programming.* I. *Affine and projective scaling trajectories*, Trans. Amer. Math. Soc., 314 (1989), pp. 499–526.
[12] D. A. Bayer and J. C. Lagarias, *The nonlinear geometry of linear programming.* II. *Legendre transform coordinates and central trajectories*, Trans. Amer. Math. Soc., 314 (1989), pp. 527–581.
[13] J. Bolte, *Continuous gradient projection method in Hilbert spaces*, J. Optim. Theory Appl., to appear.
[14] H. Brézis, *Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dars les Espaces Hilbert*, North-Holland Math. Stud. 5, North-Holland, Amsterdam, 1973.
[15] R. E. Bruck, *Asymptotic convergence of nonlinear contraction semi-groups in Hilbert space*, J. Funct. Anal., 18 (1974), pp. 15–26.
[16] Y. Censor and A. Lent, *An iterative row-action method for interval convex programming*, J. Optim. Theory Appl., 34 (1981), pp. 321–353.
[17] Y. Censor and S. Zenios, *The proximal minimization algorithm with D-functions*, J. Optim. Theory Appl., 73 (1992), pp. 451–464.
[18] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM J. Optim., 3 (1993), pp. 538–543.
[19] J. C Dunn, *Global and asymptotic convergence rate estimates for a class of projected gradient processes*, SIAM J. Control Optim., 19 (1981), pp. 368–400.
[20] J. Eckstein, *Nonlinear proximal point algorithms using Bregman functions, with application to convex programming*, Math. Oper. Res., 18 (1993), pp. 202–226.
[21] Y. M. Ermoliev, *On the method of generalized stochastic gradients and quasi-Fejer sequences*, Cybernetics, 5 (1969), pp. 208–220.
[22] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems*, Springer-Verlag, London, 1994.
[23] J. Hofbauer and K. Sigmund, *The Theory of Evolution and Dynamical Systems*, Cambridge University Press, Cambridge, UK, 1988.
[24] A. N. Iusem, B. F. Svaiter, and J. X. Da Cruz Neto, *Central paths, generalized proximal point methods, and Cauchy trajectories in Riemannian manifolds*, SIAM J. Control Optim., 37 (1999), pp. 566–588.
[25] N. K. Karmarkar, *Riemannian geometry underlying interior point methods for linear programming*, in Mathematical Developments Arising from Linear Programming, Contemp. Math. 114, J. C. Lagarias and M. Todd, eds., AMS, Providence, RI, 1990, pp. 51–75.
[26] N. Kenmochi and I. Pawlow, *A class of doubly nonlinear elliptic-parabolic equations with time dependent constraints*, Nonlinear Anal., 10 (1986), pp. 1181–1202.
[27] K. C. Kiwiel, *Proximal mimization methods with generalized Bregman functions*, SIAM J. Control Optim., 35 (1997), pp. 1142–1168.
[28] K. C. Kiwiel, *Free-steering relaxation methods for problems with strictly convex costs and linear constraints*, Math. Oper. Res., 22 (1997), pp. 326–349.
[29] B. Lemaire, *An asymptotical variational principle associated with the steepest descent method for a convex function*, J. Convex Anal., 3 (1996), pp. 63–70.
[30] B. Lemaire, *The proximal point algorithm*, in General Inequalities, Internat. Ser. Numer. Math. 80, J. P. Penot, ed., Birkhäuser, Basel, 1987, pp. 43–87.
[31] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, USSR Comp. Math. Phys., 6 (1965), pp. 787–823.
[32] S. Lojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, in Les équations aux dérivées partielles, Editions du centre National de la Recherche Scientifique, Paris, 1962, pp. 87–89.

[33] B. Martinet, *Perturbation des méthodes d'optimisation: Applications*, RAIRO Anal. Numér., 12 (1978), pp. 153–171.
[34] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, Philadelphia, PA, 1994.
[35] Z. Opial, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.
[36] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[37] M. Teboulle, *Entropic proximal mappings with applications to nonlinear programming*, Math. Oper. Res., 17 (1992), pp. 670–690.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1293–1309

H_∞ MODEL REDUCTION IN THE STOCHASTIC FRAMEWORK*

SHENGYUAN XU† AND TONGWEN CHEN†

Abstract. This paper investigates the problems of H_∞ model reduction for both continuous and discrete stochastic systems. In terms of certain linear matrix inequalities (LMIs) and a coupling nonconvex rank constraint, necessary and sufficient conditions for the existence of solutions to such problems are obtained. An explicit parametrization of all reduced-order models corresponding to a feasible solution is also proposed. In particular, when a zeroth-order H_∞ approximation is desired, conditions are obtained using LMIs only without any rank constraints, and a parametrization of all solutions is also presented. Finally, an illustrative example is provided to demonstrate the effectiveness of the proposed approach.

Key words. H_∞ model reduction, linear matrix inequality, stochastic systems

AMS subject classifications. 37H99, 37M99, 41A29

DOI. 10.1137/S0363012902403109

1. Introduction.

The problem of model reduction has received considerable attention in the past decades due to the fact that mathematical modeling of physical systems often involves high-order models, and such high-order models will consequently result in high-order controllers, which may make simulation and physical implementation difficult. The purpose of model reduction is to obtain a lower-order system which approximates a high-order system according to some given criterion. Many results on the model reduction problem have been reported, and various approaches, such as the aggregation method, balanced truncation approach, optimal Hankel norm approximation method, to name just a few, have been proposed in the literature; see, e.g., [2, 8, 19, 21, 24, 25, 26] and the references therein.

Recently, the H_∞ model reduction problem has been studied by many researchers [9, 14, 15, 18, 20]. The essence of this problem is to find a desired lower-order system such that the H_∞ norm of the difference between the original system and the desired lower-order one satisfies a prescribed H_∞ norm bound constraint. By converting this problem into a Hankel norm model reduction through an imbedding process, [18] proposed a characterization of the optimal solutions to the H_∞ model reduction problem. Based on this, a suboptimal computational procedure for the general multivariable continuous optimal H_∞ model reduction problem was developed in [16]. The zeroth-order H_∞ model reduction problem was investigated in [14], [15], and [17], respectively, and solutions to this problem were presented. More recently, the H_∞ model reduction problem was dealt with in [9, 20], where a linear matrix inequality (LMI) approach was proposed and necessary and sufficient conditions for the existence of solutions to both the continuous-time and discrete-time cases were provided. When time-delay and parameter uncertainty appear in a system model, the H_∞ model reduction results for continuous systems in [9] were extended in [23]. It is worth noting that all these results were obtained in the context of deterministic systems.

*Received by the editors February 22, 2002; accepted for publication (in revised form) March 7, 2003; published electronically September 9, 2003. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

http://www.siam.org/journals/sicon/42-4/40310.html

†Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada T6G 2V4 (syxu@mail.com, tchen@ee.ualberta.ca). The first author is now with the Department of Automation, Nanjing University of Science and Technology, Nanjing 210094, People's Republic of China. The second author served as Corresponding Author.

On the other hand, stochastic systems have received much attention since stochastic modeling has come to play an important role in many branches of science and engineering. The problem of model reduction for stochastic systems was studied [1, 4, 5]. For example, based on the theory of stochastic realization, an algorithm for obtaining reduced-order models was proposed in [1], while in [5] a new and direct approach to stochastic model reduction problem was proposed by establishing an equivalence between canonical correlation analysis and solutions to algebraic Riccati equations, which was developed by using the balanced stochastic realization. For stochastic systems, however, as far as the H_∞ model reduction problem is concerned, it seems that no results on this topic are available in the literature.

In this paper, we deal with the H_∞ model reduction problem for stochastic systems. For a given stochastic system with mean-square stability, the purpose is to find a lower-order mean-square stable stochastic system such that the norm of the perturbation operator of the error system satisfies a prescribed H_∞ bound. To solve this problem, an algebraic method similar to the deterministic case [9] is adopted. The main contributions of the paper are as follows: necessary and sufficient conditions for the solvability of the mentioned problem are obtained for both continuous and discrete stochastic systems. These conditions are given in terms of certain LMIs and a coupling nonconvex rank constraint set. It is worth pointing out that although the solutions are expressed in some nonconvex inequalities, fortunately these can be solved by either an efficient numerical algorithm based on alternating projections given in [9, 10, 11, 20] or some other algorithms proposed in [3]. When these conditions are feasible, an explicit parametrization of all reduced-order models is derived. In particular, sole LMI conditions without any rank constraints are obtained for the zeroth-order H_∞ approximation problem, and a parametrization of all solutions is also presented. When a stochastic system reduces to a deterministic one, it is shown that the results in this paper coincide with those on H_∞ model reduction for deterministic systems in [9]; therefore, our results can be regarded as extensions of existing results on H_∞ model reduction from deterministic systems to stochastic systems.

2. H_∞ model reduction: Continuous time. Consider a continuous-time stochastic system described by

$$(\Sigma_c): \qquad dx(t) = [Ax(t) + Bu(t)]dt + [A_0x(t) + B_0u(t)]d\omega(t), \tag{2.1}$$

$$y(t)dt = [Cx(t) + Du(t)]dt + [A_1x(t) + B_1u(t)]d\omega(t), \tag{2.2}$$

where $x(t) \in \mathbb{R}^n$ is the state, $y(t) \in \mathbb{R}^p$ is the output, $u(t) \in \mathbb{R}^q$ is the control input, A, A_0, A_1, B, B_0, B_1, C, and D are known real constant matrices, $\omega(t)$ is a zero-mean real scalar Wiener process on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ relative to an increasing family $(\mathcal{F}_t)_{t>0}$ of σ-algebras $\mathcal{F}_t \subset \mathcal{F}$, where Ω is the sample space, $\mathcal{F}$ is the σ-algebra of subsets of the sample space, and $\mathcal{P}$ is the probability measure on $\mathcal{F}$. We assume

$$\mathcal{E}\{d\omega(t)\} = 0, \quad \mathcal{E}\{d\omega(t)^2\} = dt, \tag{2.3}$$

where $\mathcal{E}\{\cdot\}$ is the expectation operator. We denote by $L_2[\Omega, \mathbb{R}^k)$ the space of square-integrable $\mathbb{R}^k$-valued vector functions on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$; we also denote by $L_{E_2}([0,\infty); \mathbb{R}^k)$ the space of nonanticipatory square-integrable stochastic processes $f(\cdot) = (f(t))_{t\in[0,\infty)}$ in $\mathbb{R}^k$ with respect to $(\mathcal{F}_t)_{t\in[0,\infty)}$ satisfying

$$\|f\|_{E_2}^2 = \mathcal{E}\left\{\int_0^\infty |f(t)|^2\, dt\right\} = \int_0^\infty \mathcal{E}\left\{|f(t)|^2\right\} dt < \infty,$$

where $|\cdot|$ is the standard Euclidean vector norm. In this paper we assume $u(t)$ belongs to $L_{E_2}([0,\infty);\mathbb{R}^q)$ and is $(\mathcal{F}_s)_{s<t}$ measurable [12].

DEFINITION 2.1 (see [12, 22]). *The stochastic system (Σ_c) is said to be mean-square stable if all initial states $x(0)$, subject to $u(t)=0$, yield*

$$\lim_{t\to\infty}\mathcal{E}\,|x(t)|^2=0.$$

DEFINITION 2.2 (see [12]). *The system (Σ_c) is said to be externally stable if, for every $u(t)\in L_{E_2}([0,\infty);\mathbb{R}^q)$,*

$$y(t)\in L_{E_2}([0,\infty);\mathbb{R}^p),$$

and there exists a scalar $\mu>0$ such that

$$\|y\|_{E_2}\le\mu\,\|u\|_{E_2},\quad u(t)\in L_{E_2}([0,\infty);\mathbb{R}^q). \tag{2.4}$$

DEFINITION 2.3 (see [12]). *Suppose that the system (Σ_c) is externally stable. Under the zero initial condition for the system (Σ_c), the operator*

$$\mathbb{L}_c: L_{E_2}([0,\infty);\mathbb{R}^q)\to L_{E_2}([0,\infty);\mathbb{R}^p),$$

defined by

$$(\mathbb{L}_c u)(t)=y(t), \tag{2.5}$$

is called the perturbation operator of the system (Σ_c). Its norm $\|\mathbb{L}_c\|$ is defined as the minimum $\mu\ge 0$ such that (2.4) holds.

Assume the system (Σ_c) is mean-square stable; then the H_∞ model reduction problem addressed in this section is as follows: given a scalar $\gamma>0$, find a mean-square stable system

$$(\hat\Sigma_c):\qquad d\hat x(t)=[\hat A\hat x(t)+\hat B u(t)]dt+[\hat A_0\hat x(t)+\hat B_0 u(t)]d\omega(t), \tag{2.6}$$

$$\hat y(t)dt=[\hat C\hat x(t)+\hat D u(t)]dt+[\hat A_1\hat x(t)+\hat B_1 u(t)]d\omega(t), \tag{2.7}$$

where $\hat x(t)\in\mathbb{R}^{\hat n}$, $\hat y(t)\in\mathbb{R}^p$, and $\hat n<n$ such that

$$\|\mathbb{L}_{\tilde c}\|<\gamma, \tag{2.8}$$

where $\mathbb{L}_{\tilde c}$ is the perturbation operator of the resulting error system from (Σ_c) and $(\hat\Sigma_c)$, which is defined as

$$(\mathbb{L}_{\tilde c}u)(t)=y(t)-\hat y(t). \tag{2.9}$$

If $\hat n=0$, then the reduced-order system in (2.6) and (2.7) becomes

$$\hat y(t)dt=\hat D u(t)dt+\hat B_1 u(t)d\omega(t). \tag{2.10}$$

In this case, the model reduction problem reduces to the zeroth-order H_∞ approximation problem. It should be pointed out that in the deterministic case, the zeroth-order H_∞ approximation problem involves only finding a constant matrix $\hat D$ [14, 15].

Before proceeding further, we give the following lemmas which will be used in the proof of our main results.

LEMMA 2.4. *The stochastic system* (Σ_c) *is mean-square stable and* $\|\mathbb{L}_c\| < \gamma$ *if and only if there exists a matrix* $P > 0$ *such that*

$$\begin{bmatrix} PA + A^TP + C^TC & C^TD + PB & A_1^T & A_0^TP \\ D^TC + B^TP & D^TD - \gamma^2 I & B_1^T & B_0^TP \\ A_1 & B_1 & -I & 0 \\ PA_0 & PB_0 & 0 & -P \end{bmatrix} < 0. \tag{2.11}$$

Proof. The proof can be carried out using the same argument as in the proof of Theorem 2.8 in [12], and thus is omitted. □

LEMMA 2.5 (see [7, 13]). *Given a symmetric matrix* Ξ *and two matrices* Γ *and* Π*, consider the problem of finding some matrix* Θ *such that*

$$\Xi + \Gamma\Theta\Pi + (\Gamma\Theta\Pi)^T < 0. \tag{2.12}$$

Then (2.12) *is solvable for* Θ *if and only if*

$$\Gamma^{\perp}\Xi\Gamma^{\perp T} < 0,\ \Pi^{T\perp}\Xi\Pi^{T\perp T} < 0.$$

Here, if $\Gamma \in \mathbb{R}^{n\times m}$ *and* $\operatorname{rank}\Gamma = r$*, the orthogonal complement* $\Gamma^{\perp}$ *is defined as a (possibly nonunique)* $(n-r)\times n$ *matrix with rank* $n-r$ *such that* $\Gamma^{\perp}\Gamma = 0$.

Now we are in a position to give the condition for the solvability of the H_∞ model reduction problem for continuous-time stochastic systems.

THEOREM 2.6. *There exists a stochastic system with* $\hat{n}$*th order in the form of* (2.6) *and* (2.7) *such that the* H_∞ *model reduction problem for the continuous stochastic system* (Σ_c) *is solvable if and only if there exist matrices* $X > 0$ *and* $Y > 0$ *satisfying*

$$\begin{bmatrix} XA + A^TX + C^TC & A_1^T & A_0^TX \\ A_1 & -I & 0 \\ XA_0 & 0 & -X \end{bmatrix} < 0, \tag{2.13}$$

$$\begin{bmatrix} AY + YA^T & B & YA_0^T \\ B^T & -\gamma^2 I & B_0^T \\ A_0Y & B_0 & -Y \end{bmatrix} < 0, \tag{2.14}$$

$$\begin{bmatrix} X & I \\ I & Y \end{bmatrix} \geq 0, \tag{2.15}$$

and

$$\operatorname{rank}\begin{bmatrix} X & I \\ I & Y \end{bmatrix} \leq n + \hat{n}. \tag{2.16}$$

In this case, all desired $\hat{n}$*th-order reduced systems corresponding to a feasible solution* (X, Y) *to* (2.13)–(2.16) *are given by*

$$\begin{bmatrix} \hat{D} & \hat{C} \\ \hat{B} & \hat{A} \\ \hat{B}_1 & \hat{A}_1 \\ \hat{B}_0 & \hat{A}_0 \end{bmatrix} = -W^{-1}\Psi^T\Lambda\Phi^T(\Phi\Lambda\Phi^T)^{-1} + W^{-1}S^{\frac{1}{2}}L(\Phi\Lambda\Phi^T)^{-\frac{1}{2}}, \tag{2.17}$$

where

$$S = W - \Psi^T\left[\Lambda - \Lambda\Phi^T\left(\Phi\Lambda\Phi^T\right)^{-1}\Phi\Lambda\right]\Psi,$$

$$\Lambda = (\Psi W^{-1}\Psi^T - \Omega)^{-1},$$

$$\Omega = \begin{bmatrix} A^TX + XA & A^TS & XB & C^T & A_1^T & A_0^TX & A_0^TS \\ S^TA & 0 & S^TB & 0 & 0 & 0 & 0 \\ B^TX & B^TS & -\gamma^2 I & D^T & B_1^T & B_0^TX & B_0^TS \\ C & 0 & D & -I & 0 & 0 & 0 \\ A_1 & 0 & B_1 & 0 & -I & 0 & 0 \\ XA_0 & 0 & XB_0 & 0 & 0 & -X & -S \\ S^TA_0 & 0 & S^TB_0 & 0 & 0 & -S^T & -U \end{bmatrix},$$

$$\Psi = \begin{bmatrix} 0 & S & 0 & 0 \\ 0 & U & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -I & 0 & 0 & 0 \\ 0 & 0 & -I & 0 \\ 0 & 0 & 0 & S \\ 0 & 0 & 0 & U \end{bmatrix},$$

$$\Phi = \begin{bmatrix} 0 & 0 & I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

and L is any matrix satisfying $\bar{\sigma}(L) < 1$, where $\bar{\sigma}(\cdot)$ is the maximum singular value of a matrix; moreover, $S \in \mathbb{R}^{n\times\hat{n}}$, $U \in \mathbb{R}^{\hat{n}\times\hat{n}}$, $W > 0$, and $U > 0$ satisfy

$$\Lambda > 0,\ X - Y^{-1} = SU^{-1}S^T \geq 0.$$

Proof. From (2.1), (2.2), (2.6), and (2.7), the error system resulting from the system (Σ_c) and $(\hat{\Sigma}_c)$ can be written as the augmented system

$$(\tilde{\Sigma}_c): \qquad d\tilde{x}(t) = [\tilde{A}\tilde{x}(t) + \tilde{B}u(t)]dt + [\tilde{A}_0\tilde{x}(t) + \tilde{B}_0u(t)]d\omega(t), \tag{2.18}$$

$$\tilde{y}(t)dt = [\tilde{C}\tilde{x}(t) + \tilde{D}u(t)]dt + [\tilde{A}_1\tilde{x}(t) + \tilde{B}_1u(t)]d\omega(t), \tag{2.19}$$

where

$$\tilde{x}(t) = \begin{bmatrix} x(t)^T & \hat{x}(t)^T \end{bmatrix}^T,\ \tilde{y}(t) = y(t) - \hat{y}(t), \tag{2.20}$$

$$\tilde{A} = \bar{A} + \bar{F}\bar{G}_c\bar{H},\ \tilde{B} = \bar{B} + \bar{F}\bar{G}_c\bar{N},\ \tilde{A}_0 = \bar{A}_0 + \bar{M}\bar{G}_c\bar{H},\ \tilde{B}_0 = \bar{B}_0 + \bar{M}\bar{G}_c\bar{N}, \tag{2.21}$$

$$\tilde{C} = \bar{C} + \bar{S}\bar{G}_c\bar{H},\ \tilde{D} = \bar{D} + \bar{S}\bar{G}_c\bar{N},\ \tilde{A}_1 = \bar{A}_1 + \bar{K}\bar{G}_c\bar{H},\ \tilde{B}_1 = \bar{B}_1 + \bar{K}\bar{G}_c\bar{N}, \tag{2.22}$$

$$\bar{A} = \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix},\ \bar{B} = \begin{bmatrix} B \\ 0 \end{bmatrix},\ \bar{A}_0 = \begin{bmatrix} A_0 & 0 \\ 0 & 0 \end{bmatrix},\ \bar{B}_0 = \begin{bmatrix} B_0 \\ 0 \end{bmatrix}, \tag{2.23}$$

$$\bar{C} = \begin{bmatrix} C & 0 \end{bmatrix},\ \bar{A}_1 = \begin{bmatrix} A_1 & 0 \end{bmatrix},\ \bar{D} = D,\ \bar{B}_1 = B, \tag{2.24}$$

$$\bar{F} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 \end{bmatrix},\ \bar{H} = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix},\ \bar{M} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I \end{bmatrix},\ \bar{N} = \begin{bmatrix} I \\ 0 \end{bmatrix}, \tag{2.25}$$

$$\bar{K} = \begin{bmatrix} 0 & 0 & -I & 0 \end{bmatrix},\ \bar{S} = \begin{bmatrix} -I & 0 & 0 & 0 \end{bmatrix},\ \bar{G}_c = \begin{bmatrix} \hat{D} & \hat{C} \\ \hat{B} & \hat{A} \\ \hat{B}_1 & \hat{A}_1 \\ \hat{B}_0 & \hat{A}_0 \end{bmatrix}. \tag{2.26}$$

By Lemma 2.4, it is easy to see that the error system $(\tilde{\Sigma}_c)$ is mean-square stable and satisfies $\|\mathbb{L}_{\tilde{e}}\| < \gamma$ if and only if there exists a matrix $P > 0$ such that

$$\begin{bmatrix} P\tilde{A} + \tilde{A}^T P + \tilde{C}^T \tilde{C} & \tilde{C}^T \tilde{D} + P\tilde{B} & \tilde{A}_1^T & \tilde{A}_0^T P \\ \tilde{D}^T \tilde{C} + \tilde{B}^T P & \tilde{D}^T \tilde{D} - \gamma^2 I & \tilde{B}_1^T & \tilde{B}_0^T P \\ \tilde{A}_1 & \tilde{B}_1 & -I & 0 \\ P\tilde{A}_0 & P\tilde{B}_0 & 0 & -P \end{bmatrix} < 0. \tag{2.27}$$

Using the Schur complement lemma and noting the expressions in (2.21) and (2.22), it can be shown that the inequality (2.27) is equivalent to

$$\begin{bmatrix} P\left(\bar{A} + \bar{F}\bar{G}_c\bar{H}\right) + \left(\bar{A} + \bar{F}\bar{G}_c\bar{H}\right)^T P & P\left(\bar{B} + \bar{F}\bar{G}_c\bar{N}\right) \\ \left(\bar{B} + \bar{F}\bar{G}_c\bar{N}\right)^T P & -\gamma^2 I \\ \bar{C} + \bar{S}\bar{G}_c\bar{H} & \bar{D} + \bar{S}\bar{G}_c\bar{N} \\ \bar{A}_1 + \bar{K}\bar{G}_c\bar{H} & \bar{B}_1 + \bar{K}\bar{G}_c\bar{N} \\ P\left(\bar{A}_0 + \bar{M}\bar{G}_c\bar{H}\right) & P\left(\bar{B}_0 + \bar{M}\bar{G}_c\bar{N}\right) \end{bmatrix}$$

$$\begin{bmatrix} \left(\bar{C} + \bar{S}\bar{G}_c\bar{H}\right)^T & \left(\bar{A}_1 + \bar{K}\bar{G}_c\bar{H}\right)^T & \left(\bar{A}_0 + \bar{M}\bar{G}_c\bar{H}\right)^T P \\ \left(\bar{D} + \bar{S}\bar{G}_c\bar{N}\right)^T & \left(\bar{B}_1 + \bar{K}\bar{G}_c\bar{N}\right)^T & \left(\bar{B}_0 + \bar{M}\bar{G}_c\bar{N}\right)^T P \\ -I & 0 & 0 \\ 0 & -I & 0 \\ 0 & 0 & -P \end{bmatrix} < 0. \tag{2.28}$$

This can be rewritten as

$$\Omega_c + \Psi_c \bar{G}_c \Phi_c + (\Psi_c \bar{G}_c \Phi_c)^T < 0, \tag{2.29}$$

where

$$\Omega_c = \begin{bmatrix} P\bar{A} + \bar{A}^T P & P\bar{B} & \bar{C}^T & \bar{A}_1^T & \bar{A}_0^T P \\ \bar{B}^T P & -\gamma^2 I & \bar{D}^T & \bar{B}_1^T & \bar{B}_0^T P \\ \bar{C} & \bar{D} & -I & 0 & 0 \\ \bar{A}_1 & \bar{B}_1 & 0 & -I & 0 \\ P\bar{A}_0 & P\bar{B}_0 & 0 & 0 & -P \end{bmatrix},$$

$$\Psi_c = \begin{bmatrix} P\bar{F} \\ 0 \\ \bar{S} \\ \bar{K} \\ P\bar{M} \end{bmatrix},$$

$$\Phi_c = \begin{bmatrix} \bar{H} & \bar{N} & 0 & 0 & 0 \end{bmatrix}.$$

It is noted that the parameters of the unknown reduced-order model are included in the matrix $\bar{G}_c$. From Lemma 2.5 it is easy to see that a necessary and sufficient condition for the LMI in (2.29) to have a solution $\bar{G}_c$ is

$$\Psi_c^{\perp} \Omega_c \Psi_c^{\perp T} < 0, \ \Phi_c^{T\perp} \Omega_c \Phi_c^{T\perp T} < 0. \tag{2.30}$$

Now, by some calculations, it can be verified that

$$\Psi_c^{\perp} = \begin{bmatrix} [I & 0] & 0 & 0 & 0 & [0 & 0] \\ [0 & 0] & I & 0 & 0 & [0 & 0] \\ [0 & 0] & 0 & 0 & 0 & [I & 0] \end{bmatrix} \begin{bmatrix} P^{-1} & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & P^{-1} \end{bmatrix},$$

$$\Phi_c^{T\perp} = \begin{bmatrix} [I \;\; 0] & 0 & 0 & 0 & [0 \;\; 0] \\ [0 \;\; 0] & 0 & I & 0 & [0 \;\; 0] \\ [0 \;\; 0] & 0 & 0 & I & [0 \;\; 0] \\ [0 \;\; 0] & 0 & 0 & 0 & [I \;\; 0] \\ [0 \;\; 0] & 0 & 0 & 0 & [0 \;\; I] \end{bmatrix}.$$

Partition P and P^{-1} as

$$P = \begin{bmatrix} X & X_{12} \\ X_{12}^T & X_{22} \end{bmatrix},$$

$$P^{-1} = \begin{bmatrix} Z & Z_{12} \\ Z_{12}^T & Z_{22} \end{bmatrix}, \tag{2.31}$$

where

$$X \in \mathbb{R}^{n\times n},\ X_{12} \in \mathbb{R}^{n\times\hat{n}},\ Z_{12} \in \mathbb{R}^{n\times\hat{n}},\ X_{22} \in \mathbb{R}^{\hat{n}\times\hat{n}},\ Z_{22} \in \mathbb{R}^{\hat{n}\times\hat{n}}.$$

Then we can show that

$$\Psi_c^{\perp}\Omega_c\Psi_c^{\perp T} = \begin{bmatrix} AZ + ZA^T & B & ZA_0^T \\ B^T & -\gamma^2 I & B_0^T \\ A_0 Z & B_0 & -Z \end{bmatrix}, \tag{2.32}$$

$$\Phi_c^{T\perp}\Omega_c\Phi_c^{T\perp T} = \begin{bmatrix} XA + A^TX & C^T & A_1^T & A_0^TX & A_0^TX_{12} \\ C & -I & 0 & 0 & 0 \\ A_1 & 0 & -I & 0 & 0 \\ XA_0 & 0 & 0 & -X & -X_{12} \\ X_{12}^TA_0 & 0 & 0 & -X_{12}^T & -X_{22} \end{bmatrix}. \tag{2.33}$$

Let $Y = Z$; then the inequality $\Psi_c^{\perp}\Omega_c\Psi_c^{\perp T} < 0$ gives (2.14). It can also be shown that using the Schur complement formula to $\Phi_c^{T\perp}\Omega_c\Phi_c^{T\perp T} < 0$ results in (2.13). On the other hand, noting the (1,1) block of P^{-1} in (2.31), we have

$$X - Y^{-1} = X_{12}X_{22}^{-1}X_{12}^T \geq 0. \tag{2.34}$$

This, by the Schur complement formula, is equivalent to (2.15). Also, (2.34) implies

$$\operatorname{rank}(X - Y^{-1}) \leq \hat{n}, \tag{2.35}$$

since $X_{22} \in \mathbb{R}^{\hat{n}\times\hat{n}}$ is a symmetric positive definite matrix. It then follows that (2.35) provides the condition (2.16). In addition, when (2.13)–(2.16) are satisfied, the parametrization (2.17) of all reduced models corresponding to a feasible solution can be obtained by using the results in [7] and [13]. This completes the proof. □

Remark 2.7. Theorem 2.6 presents a necessary and sufficient condition for the solvability of the H_∞ model reduction problem for continuous-time stochastic systems. It is noted that the inequalities in (2.13)–(2.16) are nonconvex though the constraints (2.13)–(2.15) are convex. For such nonconvex inequalities, either an efficient numerical algorithm based on alternating projections in [9, 10, 11, 20] or some other algorithms in [3] can be employed. Moreover, as in [9], a bisection approach can be used to seek the minimum H_∞ performance level γ in order to solve the optimal H_∞ model reduction problem for continuous-time stochastic systems.

Next, we shall establish a solvability condition based on LMIs only for the zeroth-order H_∞ approximation problem, which is given in the following theorem.

THEOREM 2.8. *There exist two matrices $\hat{B}_1$ and $\hat{D}$ solving the zeroth-order H_∞ approximation problem for the continuous-time stochastic system (Σ_c) if and only if there exists a matrix $X > 0$ satisfying*

$$\begin{bmatrix} XA + A^TX + C^TC & A_1^T & A_0^TX \\ A_1 & -I & 0 \\ XA_0 & 0 & -X \end{bmatrix} < 0, \tag{2.36}$$

$$\begin{bmatrix} XA + A^TX & XB & A_0^TX \\ B^TX & -\gamma^2 I & B_0^TX \\ XA_0 & XB_0 & -X \end{bmatrix} < 0. \tag{2.37}$$

In this case, all the solutions $\hat{B}_1$ and $\hat{D}$ to the zeroth-order H_∞ approximation problem corresponding to a feasible solution X to (2.36) *and* (2.37) *are given by*

$$\begin{bmatrix} \hat{D} \\ \hat{B}_1 \end{bmatrix} = G_1 + G_2LG_3, \tag{2.38}$$

$$G_1 = -Q_{12}Q_{22}^{-1}M^T(MQ_{22}^{-1}M^T)^{-1}, \tag{2.39}$$

$$G_2 = (Q_{12}Q_{22}^{-1}Q_{12}^T - Q_{11} - G_1G_3^{-2}G_1^T)^{\frac{1}{2}}, \tag{2.40}$$

$$G_3 = (-MQ_{22}^{-1}M^T)^{-\frac{1}{2}}, \tag{2.41}$$

$$Q_{11} = \begin{bmatrix} -I & 0 \\ 0 & -I \end{bmatrix},\ Q_{12} = \begin{bmatrix} C & D & 0 \\ A_1 & B_1 & 0 \end{bmatrix}, \tag{2.42}$$

$$Q_{22} = \begin{bmatrix} XA + A^TX & XB & A_0^TX \\ B^TX & -\gamma^2 I & B_0^TX \\ XA_0 & XB_0 & -X \end{bmatrix}, \tag{2.43}$$

$$M = \begin{bmatrix} 0 & -I & 0 \end{bmatrix}, \tag{2.44}$$

where L is any matrix satisfying $\bar{\sigma}(L) < 1$.

Proof. From (2.1), (2.2), and (2.10), we obtain the error system as

$$(\tilde{\Sigma}_{cd}):\quad dx(t) = [Ax(t) + Bu(t)]dt + [A_0x(t) + B_0u(t)]d\omega(t), \tag{2.45}$$

$$\tilde{y}(t)dt = [Cx(t) + (D - \hat{D})u(t)]dt + [A_1x(t) + (B_1 - \hat{B}_1)u(t)]d\omega(t), \tag{2.46}$$

where $\tilde{y}(t)$ is defined in (2.20). By Lemma 2.4, it is easy to see that this system is mean-square stable and satisfies a prescribed H_∞ performance level $\gamma > 0$ if and only if there exists a matrix $X > 0$ such that

$$\begin{bmatrix} XA + A^TX & XB & C^T & A_1^T & A_0^TX \\ B^TX & -\gamma^2 I & (D - \hat{D})^T & (B_1 - \hat{B}_1)^T & B_0^TX \\ C & D - \hat{D} & -I & 0 & 0 \\ A_1 & B_1 - \hat{B}_1 & 0 & -I & 0 \\ XA_0 & XB_0 & 0 & 0 & -X \end{bmatrix} < 0.$$

This can be rewritten as

$$\check{\Omega}_c + \check{\Psi}_c\check{G}_c\check{\Phi}_c + (\check{\Psi}_c\check{G}_c\check{\Phi}_c)^T < 0, \tag{2.47}$$

where

$$\check{G}_c = \begin{bmatrix} XA + A^TX & XB & C^T & A_1^T & A_0^TX \\ B^TX & -\gamma^2 I & D^T & B_1^T & B_0^TX \\ C & D & -I & 0 & 0 \\ A_1 & B_1 & 0 & -I & 0 \\ XA_0 & XB_0 & 0 & 0 & -X \end{bmatrix},$$

$$\check{\Psi}_c = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ I & 0 \\ 0 & I \\ 0 & 0 \end{bmatrix},$$
$$\check{\Phi}_c = \begin{bmatrix} 0 & -I & 0 & 0 & 0 \end{bmatrix}.$$

Therefore, by Lemma 2.5, we have that the necessary and sufficient condition for the above LMI to have a solution $\hat{B}_1$ and $\hat{D}$ is

$$\check{\Psi}_c^{\perp}\check{G}_c\check{\Psi}_c^{\perp T} < 0, \quad \check{\Phi}_c^{T\perp}\check{G}_c\check{\Phi}_c^{T\perp T} < 0. \tag{2.48}$$

By some algebraic manipulations, the LMIs in (2.36) and (2.37) can be established by using (2.48). Furthermore, note that

$$\check{\Psi}_c^T\check{\Psi}_c > 0, \quad \check{\Phi}_c\check{\Psi}_c^T\check{\Psi}_c\check{\Phi}_c^T > 0.$$

Then the parametrization in (2.38)–(2.44) of all $\hat{B}_1$ and $\hat{D}$ that satisfy the LMI in (2.47) can be obtained by using the results in [7] and [13] when (2.13)–(2.16) are satisfied. □

Remark 2.9. In the case when $A_0 = 0$, $A_1 = 0$, $B_0 = 0$, and $B_1 = 0$, that is, when the stochastic system (Σ_c) reduces to a deterministic system, it can be easily shown that Theorems 2.6 and 2.8 coincide with Theorems 1 and 2 in [9], respectively. Therefore, Theorems 2.6 and 2.8 can be viewed as extensions of existing results on H_∞ model reduction from deterministic systems to stochastic systems.

3. H_∞ model reduction: Discrete time. In this section, we consider the H_∞ model reduction problem for discrete-time stochastic systems. The system we consider is described by the following model:

$$(\Sigma_d): \qquad x(k+1) = Ax(k) + Bu(k) + [A_0x(k) + B_0u(k)]\omega(k), \tag{3.1}$$
$$y(k) = Cx(k) + Du(k) + [A_1x(k) + B_1u(k)]\omega(k), \tag{3.2}$$

where $x(k) \in \mathbb{R}^n$ is the state, $y(k) \in \mathbb{R}^p$ is the output, $u(k) \in \mathbb{R}^q$ is the control input, A, A_0, A_1, B, B_0, B_1, C, and D are known real constant matrices, and $\omega(k)$ is a zero-mean real scalar process on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ relative to an increasing family $(\mathcal{F}_k)_{k\in\mathbb{N}}$ of σ-algebras $\mathcal{F}_k \subset \mathcal{F}$. We assume

$$\mathcal{E}\{\omega(k)\} = 0, \quad \mathcal{E}\{\omega(k)^2\} = 1, \tag{3.3}$$

and $\omega(0)$, $\omega(1)$, ..., are independent. We denote by $l_2[\Omega, \mathbb{R}^k)$ the space of square-summable $\mathbb{R}^k$-valued vector functions on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$, and we also denote by $l_{e_2}([0,\infty); \mathbb{R}^k)$ the space of k-dimensional nonanticipatory square-summable stochastic processes $f(\cdot) = (f(k))_{k\in\mathbb{N}}$ on $\mathbb{N}$ with respect to $(\mathcal{F}_k)_{k\in\mathbb{N}}$ satisfying

$$\|f\|_{e_2}^2 = \mathcal{E}\left\{\sum_{k\in\mathbb{N}} |f(k)|^2\right\} = \sum_{k\in\mathbb{N}} \mathcal{E}\left\{|f(k)|^2\right\} < \infty.$$

In this section we assume $u(k)$ belongs to $l_{e_2}([0,\infty); \mathbb{R}^q)$ and is $\mathcal{F}_{k-1}$ measurable for all $k \in \mathbb{N}$ [6].

DEFINITION 3.1 (see [6]). *The stochastic system* (Σ_d) *is said to be mean-square stable if all initial states* $x(0)$*, subject to* $u(k) = 0$*, yield*

$$\lim_{k\to\infty} \mathcal{E}\,|x(k)|^2 = 0.$$

DEFINITION 3.2 (see [6]). *The system* (Σ_d) *is said to be externally stable if, for every* $u(k) \in l_{e_2}([0,\infty);\mathbb{R}^q)$*,*

$$y(k) \in l_{e_2}([0,\infty);\mathbb{R}^p),$$

and there exists a scalar $\mu > 0$ *such that*

$$\|y\|_{e_2} \le \mu \|u\|_{e_2}, \quad u(k) \in l_{e_2}([0,\infty);\mathbb{R}^q). \tag{3.4}$$

DEFINITION 3.3 (see [6]). *Suppose that the system* (Σ_d) *is externally stable. Under the zero initial condition for the system* (Σ_d)*, the operator*

$$\mathbb{L}_d : l_{e_2}([0,\infty);\mathbb{R}^q) \to l_{e_2}([0,\infty);\mathbb{R}^p),$$

defined by

$$(\mathbb{L}_d u)(k) = y(k), \tag{3.5}$$

is called the perturbation operator of the system (Σ_d)*. Its norm* $\|\mathbb{L}_d\|$ *is defined as the minimum* $\mu \ge 0$ *such that* (3.4) *holds.*

Assume the system (Σ_d) is mean-square stable; then the H_∞ model reduction problem addressed in this section is as follows: given a scalar $\gamma > 0$, find a mean-square stable system

$$(\hat{\Sigma}_d): \quad \hat{x}(k) = \hat{A}\hat{x}(k) + \hat{B}u(k) + [\hat{A}_0\hat{x}(k) + \hat{B}_0u(k)]d\omega(k), \tag{3.6}$$

$$\hat{y}(k) = \hat{C}\hat{x}(k) + \hat{D}u(k) + [\hat{A}_1\hat{x}(k) + \hat{B}_1u(k)]d\omega(k), \tag{3.7}$$

where $\hat{x}(t) \in \mathbb{R}^{\hat{n}}$, $\hat{y}(t) \in \mathbb{R}^p$, and $\hat{n} < n$ such that

$$\left\|\mathbb{L}_{\tilde{d}}\right\| < \gamma, \tag{3.8}$$

where $\mathbb{L}_{\tilde{d}}$ is the perturbation operator of the resulting error system from (Σ_d) and $(\hat{\Sigma}_d)$ defined as

$$(\mathbb{L}_{\tilde{d}}u)(k) = y(k) - \hat{y}(k). \tag{3.9}$$

Similar to the continuous-time case, if $\hat{n} = 0$, then the reduced-order system (3.6) and (3.7) becomes

$$\hat{y}(k) = \hat{D}u(k) + \hat{B}_1u(k)\omega(k). \tag{3.10}$$

In this case, the model reduction problem reduces to the zeroth-order H_∞ approximation problem.

The following lemma is essential in the derivation of our main results in this section.

LEMMA 3.4. *The stochastic system* (Σ_d) *is mean-square stable and* $\|\mathbb{L}_d\| < \gamma$ *if and only if there exists a matrix* $P > 0$ *such that*

$$\begin{bmatrix} -P & 0 & A^TP & A_0^TP & C^T & A_1^T \\ 0 & -\gamma^2 I & B^TP & B_0^TP & D^T & B_1^T \\ PA & PB & -P & 0 & 0 & 0 \\ PA_0 & PB_0 & 0 & -P & 0 & 0 \\ C & D & 0 & 0 & -I & 0 \\ A_1 & B_1 & 0 & 0 & 0 & -I \end{bmatrix} < 0. \tag{3.11}$$

Proof. Following similar reasoning as in the proof of Theorem 2.5 in [6], the desired results can be obtained. □

Now we present the necessary and sufficient conditions for the solvability of the H_∞ model reduction problem for discrete stochastic systems in the following theorem.

THEOREM 3.5. *There exists a stochastic system with* $\hat{n}$*th order in the form of* (3.6) *and* (3.7) *such that the* H_∞ *model reduction problem for the discrete stochastic system* (Σ_d) *is solvable if and only if there exist matrices* $X > 0$ *and* $Y > 0$ *satisfying*

$$\begin{bmatrix} C^TC + A_1^TA_1 - X & A^TX & A_0^TX \\ XA & -X & 0 \\ XA_0 & 0 & -X \end{bmatrix} < 0, \tag{3.12}$$

$$\begin{bmatrix} -Y & 0 & YA^T & YA_0^T \\ 0 & -\gamma^2 I & B^T & B_0^T \\ AY & B & -Y & 0 \\ A_0Y & B_0 & 0 & -Y \end{bmatrix} < 0, \tag{3.13}$$

$$\begin{bmatrix} X & I \\ I & Y \end{bmatrix} \geq 0, \tag{3.14}$$

and

$$\operatorname{rank} \begin{bmatrix} X & I \\ I & Y \end{bmatrix} \leq n + \hat{n}. \tag{3.15}$$

In this case, all desired $\hat{n}$*th-order reduced systems corresponding to a feasible solution* (X, Y) *to* (3.12)–(3.15) *are given by*

$$\begin{bmatrix} \hat{D} & \hat{C} \\ \hat{B} & \hat{A} \\ \hat{B}_1 & \hat{A}_1 \\ \hat{B}_0 & \hat{A}_0 \end{bmatrix} = -W^{-1}\Psi^T\Lambda\Phi^T(\Phi\Lambda\Phi^T)^{-1} + W^{-1}S^{\frac{1}{2}}L(\Phi\Lambda\Phi^T)^{-\frac{1}{2}}, \tag{3.16}$$

where

$$S = W - \Psi^T\left[\Lambda - \Lambda\Phi^T\left(\Phi\Lambda\Phi^T\right)^{-1}\Phi\Lambda\right]\Psi,$$
$$\Lambda = (\Psi W^{-1}\Psi^T - \Omega)^{-1},$$

$$\Omega = \begin{bmatrix} -X & -S & 0 & A^TX & A^TS & A_0^TX & A_0^TS & C^T & A_1^T \\ -S^T & -U & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\gamma^2 I & B^TX & B^TS & B_0^TX & B_0^TS & D & B_1^T \\ XA & 0 & XB & -X & -S & 0 & 0 & 0 & 0 \\ S^TA & 0 & S^TB & -S^T & -U & 0 & 0 & 0 & 0 \\ XA_0 & 0 & XB_0 & 0 & 0 & -X & -S & 0 & 0 \\ S^TA_0 & 0 & S^TB_0 & 0 & 0 & -S^T & -U & 0 & 0 \\ C & 0 & D^T & 0 & 0 & 0 & 0 & -I & 0 \\ A_1 & 0 & B_1 & 0 & 0 & 0 & 0 & 0 & -I \end{bmatrix},$$

$$\Psi = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & S & 0 & 0 & 0 \\ 0 & U & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & S \\ 0 & 0 & 0 & 0 & U \\ -I & 0 & 0 & 0 & 0 \\ 0 & 0 & -I & 0 & 0 \end{bmatrix},$$

$$\Phi = \begin{bmatrix} 0 & 0 & I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

and L is any matrix satisfying $\bar{\sigma}(L) < 1$; $S \in \mathbb{R}^{n\times\hat{n}}$, $U \in \mathbb{R}^{\hat{n}\times\hat{n}}$, $W > 0$, and $U > 0$ satisfy

$$\Lambda > 0,\ X - Y^{-1} = SU^{-1}S^T \geq 0.$$

Proof. The error system from the systems (Σ_d) and $(\hat{\Sigma}_d)$ is given by

$$(\tilde{\Sigma}_d): \qquad \tilde{x}(k+1) = \tilde{A}\tilde{x}(k) + \tilde{B}u(k) + [\tilde{A}_0 x(k) + \tilde{B}_0 u(k)]\omega(k), \tag{3.17}$$

$$\tilde{y}(k) = \tilde{C}\tilde{x}(k) + \tilde{D}u(k) + [\tilde{A}_1 x(k) + \tilde{B}_1 u(k)]\omega(k), \tag{3.18}$$

where

$$\tilde{x}(k) = \begin{bmatrix} x(k)^T & \hat{x}(k)^T \end{bmatrix}^T,\ \tilde{y}(k) = y(k) - \hat{y}(k), \tag{3.19}$$

and $\tilde{A}$, $\tilde{A}_0$, $\tilde{A}_1$, $\tilde{B}$, $\tilde{B}_0$, $\tilde{B}_1$, $\tilde{C}$, and $\tilde{D}$ are defined in (2.21)–(2.26). It follows from Lemma 3.4 that the error system $(\tilde{\Sigma}_d)$ satisfies $\left\|\mathbb{L}_{\tilde{d}}\right\|_\infty < \gamma$ if and only if there exists a matrix $P > 0$ such that

$$\begin{bmatrix} -P & 0 & \tilde{A}^TP & \tilde{A}_0^TP & \tilde{C}^T & \tilde{A}_1^T \\ 0 & -\gamma^2 I & \tilde{B}^TP & \tilde{B}_0^TP & \tilde{D}^T & \tilde{B}_1^T \\ P\tilde{A} & P\tilde{B} & -P & 0 & 0 & 0 \\ P\tilde{A}_0 & P\tilde{B}_0 & 0 & -P & 0 & 0 \\ \tilde{C} & \tilde{D} & 0 & 0 & -I & 0 \\ \tilde{A}_1 & \tilde{B}_1 & 0 & 0 & 0 & -I \end{bmatrix} < 0.$$

Then, by using this inequality and following similar reasoning as in the proof of Theorem 2.6, the desired results can be deduced. □

Similar to the continuous-time case, where the zeroth-order H_∞ approximation problem is concerned, the conditions in (3.12)–(3.15) can be simplified and the sole LMI condition can be derived, which is provided in the following theorem.

THEOREM 3.6. *There exist two matrices $\hat{B}_1$ and $\hat{D}$ solving the zeroth-order H_∞ approximation problem for the discrete-time stochastic system (Σ_d) if and only if there exists a matrix $X > 0$ satisfying*

$$\begin{bmatrix} -X & 0 & A^TX & A_0^TX \\ 0 & -\gamma^2 I & B^TX & B_0^TX \\ XA & XB & -X & 0 \\ XA_0 & XB_0 & 0 & -X \end{bmatrix} < 0, \tag{3.20}$$

$$\begin{bmatrix} C^TC + A_1^TA_1 - X & A^TX & A_0^TX \\ XA & -X & 0 \\ XA_0 & 0 & -X \end{bmatrix} < 0. \tag{3.21}$$

In this case, all the solutions $\hat{B}_1$ and $\hat{D}$ to the zeroth-order H_∞ approximation problem corresponding to a feasible solution X to (3.20) *and* (3.21) *are given by*

$$\begin{bmatrix} \hat{D} \\ \hat{B}_1 \end{bmatrix} = G_1 + G_2LG_3, \tag{3.22}$$

$$G_1 = -Q_{12}Q_{22}^{-1}M^T(MQ_{22}^{-1}M^T)^{-1}, \tag{3.23}$$

$$G_2 = (Q_{12}Q_{22}^{-1}Q_{12}^T - Q_{11} - G_1G_3^{-2}G_1^T)^{\frac{1}{2}}, \tag{3.24}$$

$$G_3 = (-MQ_{22}^{-1}M^T)^{-\frac{1}{2}}, \tag{3.25}$$

$$Q_{11} = \begin{bmatrix} -I & 0 \\ 0 & -I \end{bmatrix}, \ Q_{12} = \begin{bmatrix} C & D & 0 & 0 \\ A_1 & B_1 & 0 & 0 \end{bmatrix}, \tag{3.26}$$

$$Q_{22} = \begin{bmatrix} -X & 0 & A^TX & A_0^TX \\ 0 & -\gamma^2 I & B^TX & B_0^TX \\ XA & XB & -X & 0 \\ XA_0 & XB_0 & 0 & -X \end{bmatrix}, \tag{3.27}$$

$$M = \begin{bmatrix} 0 & -I & 0 & 0 \end{bmatrix}, \tag{3.28}$$

where L is any matrix satisfying $\bar{\sigma}(L) < 1$.

Proof. The proof can be carried out by using Lemma 3.4 and following a similar argument as in the proof of Theorem 2.8 and thus is omitted. □

4. An illustrative example. In this section, we present an illustrative example to demonstrate the applicability of the proposed approach.

Consider a continuous-time stochastic system (Σ_c) with parameters as follows:

$$A = \begin{bmatrix} -1.5 & 1 & -1 & 1 & 0 \\ 1 & -3.5 & 1 & 0 & 0.5 \\ 1 & 1 & -2 & -1 & 1 \\ 1 & 0 & 0.5 & -2 & 0 \\ -1 & 0 & 0 & 1 & -2.6 \end{bmatrix}, \quad A_0 = \begin{bmatrix} 0 & 0.2 & -0.2 & 0 & -0.3 \\ -1 & 1 & 0 & 0 & 1 \\ -0.6 & 0.1 & 0 & -0.5 & 0 \\ 1 & 0 & -0.6 & 0 & 0 \\ 0.5 & 0.1 & -0.5 & 0.3 & -0.4 \end{bmatrix},$$

$$B = \begin{bmatrix} -0.1 & 0 & 0.3 \\ 0.5 & -1 & 0.2 \\ 0.6 & 1 & 0.5 \\ 0 & 1 & -0.2 \\ 1 & -0.2 & 0 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0 & -0.1 & 0.1 \\ 1 & 0.5 & 0 \\ 1 & -0.6 & 0.2 \\ 0.5 & -1 & 0 \\ -0.4 & -0.8 & 0.5 \end{bmatrix},$$

$$C = \begin{bmatrix} 0 & -0.1 & 0.6 & 0 & 1 \\ -0.5 & 0 & 0.5 & -0.3 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 1 & -1 \\ -1 & 1 & 0 \end{bmatrix},$$

$$A_1 = \begin{bmatrix} 0 & 0.3 & 0.1 & -0.5 & 0 \\ 0 & 0.6 & -0.8 & 0.2 & -0.2 \end{bmatrix}, \quad B_1 = \begin{bmatrix} -1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix}.$$

It can be shown that this continuous stochastic system is mean-square stable. The purpose of this example is to find a first-order mean-square stable stochastic system in the form of (2.6) and (2.7) such that the error system is mean-square stable and (2.8) is satisfied. In this example, the H_∞ performance bound γ is required to be 1.5. By solving (2.13)–(2.16), we obtain

$$X = \begin{bmatrix} 2.1126 & 0.1145 & -0.5027 & 0.4792 & -0.3174 \\ 0.1145 & 0.3060 & -0.1633 & 0.1056 & 0.1476 \\ -0.5027 & -0.1633 & 0.6472 & -0.4421 & 0.2837 \\ 0.4792 & 0.1056 & -0.4421 & 0.8905 & -0.3450 \\ -0.3174 & 0.1476 & 0.2837 & -0.3450 & 0.9093 \end{bmatrix},$$

$$Y = \begin{bmatrix} 0.8436 & -0.0300 & 0.4938 & -0.1746 & 0.0791 \\ -0.0300 & 4.9692 & 1.6507 & -0.3162 & -1.4521 \\ 0.4938 & 1.6507 & 3.2800 & 0.8601 & -0.7925 \\ -0.1746 & -0.3162 & 0.8601 & 1.8450 & 0.4221 \\ 0.0791 & -1.4521 & -0.7925 & 0.4221 & 1.7705 \end{bmatrix}.$$

Then, from Theorem 2.6, we have that the H_∞ model reduction problem is solvable. It is easy to show that

$$X - Y^{-1} = \operatorname{diag}(0.5, 0, 0, 0, 0), \ \ \operatorname{rank}(X - Y^{-1}) = 1.$$

Therefore, we can choose

$$S = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}^T, \quad U = 2.$$

In this case, we have

$$\Psi^T = \left[\begin{array}{ccccc|c|ccc|ccc|ccc|ccc|c} 0&0&0&0&0&0&0&0&0&-1&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&-1&0&0&0&0&0&0&0&0\\ \hline 1&0&0&0&0&2&0&0&0&0&0&0&0&0&0&0&0&0&0\\ \hline 0&0&0&0&0&0&0&0&0&0&0&-1&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0&-1&0&0&0&0&0&0\\ \hline 0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&2 \end{array}\right],$$

$$\Phi = \left[\begin{array}{ccccc|c|ccc|ccc|ccc|ccc|c} 0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\ \hline 0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0 \end{array}\right].$$

Furthermore, if we choose

$$L = \begin{bmatrix} 0.6 & 0 & 0 & 0 \\ 0 & 0.6 & 0 & 0 \\ 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 0.6 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

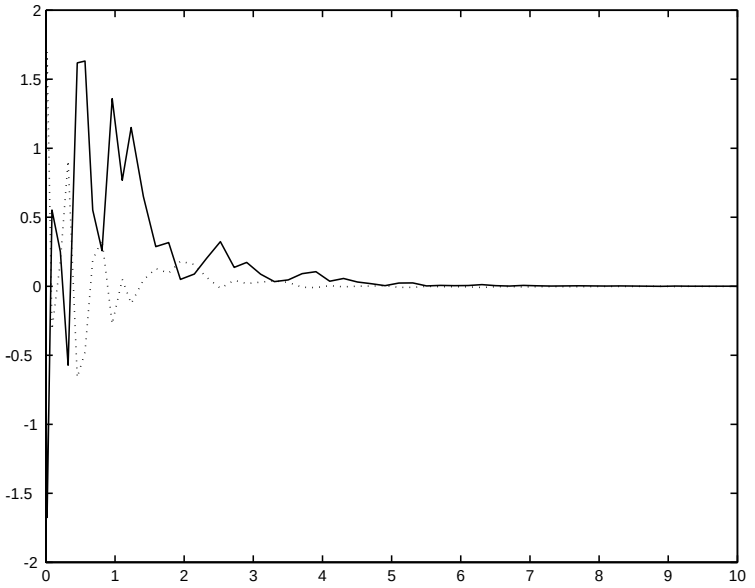


FIG. 4.1. $y_1(t)$ (—) *and* $y_2(t)$ $(\cdots)$.

$$W = \begin{bmatrix} 0.2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.2 \end{bmatrix},$$

then, according to Theorem 2.6, a desired reduced model is given by

$$\left[\begin{array}{c|c} \hat{D} & \hat{C} \\ \hline \hat{B} & \hat{A} \\ \hline \hat{B}_1 & \hat{A}_1 \\ \hline \hat{B}_0 & \hat{A}_0 \end{array}\right] = \left[\begin{array}{rrr|r} 0.6095 & 1.3235 & -0.8778 & -0.3896 \\ -0.6180 & 1.1462 & -0.0091 & 0.4345 \\ \hline 0.5173 & 0.2807 & 0.6798 & -2.9245 \\ \hline -0.8248 & -0.1522 & -0.8731 & 1.1514 \\ -0.0830 & 0.5068 & 1.0020 & -0.0939 \\ \hline 0.0033 & 0.1111 & -0.0705 & 0.0377 \end{array}\right].$$

That is,

$$\begin{aligned} d\hat{x}(t) &= \left(-2.9245\hat{x}(t) + \begin{bmatrix} 0.5173 & 0.2807 & 0.6798 \end{bmatrix} u(t)\right) dt \\ &\quad + \left(0.0377\hat{x}(t) + \begin{bmatrix} 0.0033 & 0.1111 & -0.0705 \end{bmatrix} u(t)\right) d\omega(t), \\ \hat{y}(t)dt &= \left(\begin{bmatrix} -0.3896 \\ 0.4345 \end{bmatrix} \hat{x}(t) + \begin{bmatrix} 0.6095 & 1.3235 & -0.8778 \\ -0.6180 & 1.1462 & -0.0091 \end{bmatrix} u(t)\right) dt \\ &\quad + \left(\begin{bmatrix} 1.1514 \\ -0.0939 \end{bmatrix} \hat{x}(t) + \begin{bmatrix} -0.8248 & -0.1522 & -0.8731 \\ -0.0830 & 0.5068 & 1.0020 \end{bmatrix} u(t)\right) d\omega(t). \end{aligned}$$

The simulation results of the output of the original system and the error system are given in Figures 4.1 and 4.2, respectively, where the input is specified as $\exp(-t)I$.

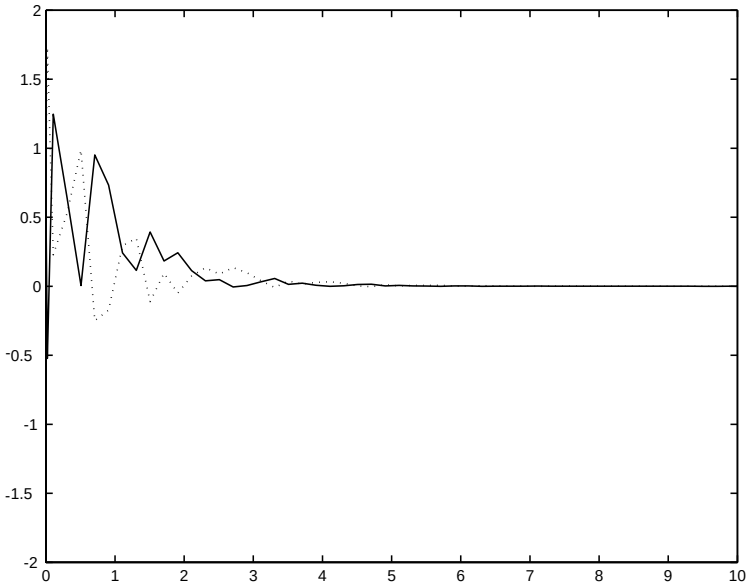


FIG. 4.2. $\hat{y}_1(t)$ (—) *and* $\hat{y}_2(t)$ $(\cdots)$.

5. Conclusions. In this paper, we have studied the problems of H_∞ model reduction for both continuous-time and discrete-time stochastic systems. Necessary and sufficient conditions for the solvability of these problems have been obtained in terms of certain LMIs and a coupling nonconvex rank constraint set. An explicit parametrization of all reduced-order models corresponding to feasible solutions has been given. Results on the zeroth-order H_∞ approximation have also been provided, which involve only LMIs without any rank constraints. The results in this paper can be viewed as extensions of existing results on H_∞ model reduction from deterministic systems to stochastic systems.

REFERENCES

[1] Y. BARAMA, *Realization and reduction of Markovian models from nonstationary data*, IEEE Trans. Automat. Control, 26 (1981), pp. 1225–1231.

[2] T. CHIU, *Model reduction by the low-frequency approximation balancing method for unstable systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 995–997.

[3] M. C. DE OLIVEIRA AND J. C. GEROMEL, *Numerical comparison of output feedback design methods*, in Proceedings of the American Control Conference, Albuquerque, NM, 1997, pp. 72–76.

[4] U. B. DESAI AND D. PAL, *A realization approach to stochastic model reduction and balanced stochastic realizations*, in Proceedings of the 21st IEEE Conference on Decision and Control, Orlando, FL, 1982, pp. 1105–1112.

[5] U. B. DESAI AND D. PAL, *A transformation approach to stochastic model reduction*, IEEE Trans. Automat. Control, 29 (1984), pp. 1097–1100.

[6] A. EL BOUHTOURI, D. HINRICHSEN, AND A. J. PRITCHARD, *H^∞-type control for discrete-time stochastic systems*, Internat. J. Robust Nonlinear Control, 9 (1999), pp. 923–948.

[7] P. GAHINET AND P. APKARIAN, *A linear matrix inequality approach to H_∞ control*, Internat. J. Robust Nonlinear Control, 4 (1994), pp. 421–448.

[8] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L_∞ error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.

[9] K. M. Grigoriadis, *Optimal H_∞ model reduction via linear matrix inequalities: Continuous- and discrete-time cases*, Systems Control Lett., 26 (1995), pp. 321–333.

[10] K. M. Grigoriadis and E. B. Beran, *Alternating projection algorithms for linear matrix inequalities problems with rank constraints*, in Advances in Linear Matrix Inequality Methods in Control, Adv. Des. Control 2, SIAM, Philadelphia, PA, 1999, pp. 251–267.

[11] K. M. Grigoriadis and R. E. Skelton, *Low-order control design for LMI problems using alternating projection methods*, Automatica J. IFAC, 32 (1996), pp. 1117–1125.

[12] D. Hinrichsen and A. J. Pritchard, *Stochastic H^∞*, SIAM J. Control Optim., 36 (1998), pp. 1504–1538.

[13] T. Iwasaki and R. E. Skelton, *All controllers for the general H_∞ control problems: LMI existence conditions and state space formulas*, Automatica J. IFAC, 30 (1994), pp. 1307–1317.

[14] D. Kavranoğlu, *H_∞ norm approximation of systems by constant matrices and related results*, in Proceedings of the 31st Conference on Decision and Control, Tucson, AZ, 1992, pp. 3271–3275.

[15] D. Kavranoğlu, *Zeroth order H_∞ norm approximation of multivariable systems*, Numer. Funct. Anal. Optim., 14 (1993), pp. 89–101.

[16] D. Kavranoğlu, *A computational scheme for H_∞ model reduction*, IEEE Trans. Automat. Control, 39 (1994), pp. 1447–1450.

[17] D. Kavranoğlu, *H_∞-norm approximation of systems by constant matrices and related results*, IEEE Trans. Automat. Control, 39 (1994), pp. 1006–1009.

[18] D. Kavranoğlu and M. Bettayeb, *Characterization of the solution to the optimal H_∞ model reduction problem*, Systems Control Lett., 20 (1993), pp. 99–107.

[19] C. P. Kwong, *Optimal chained aggregation for reduced-order modeling*, Internat. J. Control, 35 (1982), pp. 965–982.

[20] R. E. Skelton, T. Iwasaki, and K. M. Grigoriadis, *A Unified Algebraic Approach to Linear Control*, Taylor & Francis, London, 1998.

[21] V. Sreeram and P. Agathoklis, *Model reduction of linear discrete systems via weighted impulse response Gramians*, Internat. J. Control, 53 (1991), pp. 129–144.

[22] J. L. Willems and J. C. Willems, *Feedback stabilizability for stochastic systems with state and control dependent noise*, Automatica J. IFAC, 12 (1976), pp. 277–283.

[23] S. Xu, J. Lam, S. Huang, and C. Yang, *H_∞ model reduction for linear time-delay systems: Continuous-time case*, Internat. J. Control, 74 (2001), pp. 1062–1074.

[24] W. Y. Yan and J. Lam, *An approximate approach to H_2 optimal model reduction*, IEEE Trans. Automat. Control, 44 (1999), pp. 1341–1358.

[25] K. Zhou, *Frequency-weighted L_∞ norm and optimal Hankel norm model reduction*, IEEE Trans. Automat. Control, 40 (1995), pp. 1687–1699.

[26] K. Zhou, *Relative/multiplicative model reduction for unstable and non-minimum-phase systems*, Automatica J. IFAC, 31 (1995), pp. 1087–1098.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1310–1324

FRICTIONAL VERSUS VISCOELASTIC DAMPING IN A SEMILINEAR WAVE EQUATION*

MARCELO MOREIRA CAVALCANTI† AND HIGIDIO PORTILLO OQUENDO‡

Abstract. In this article we show exponential and polynomial decay rates for the partially viscoelastic nonlinear wave equation subject to a nonlinear and localized frictional damping. The equation that models this problem is given by

$$u_{tt} - \kappa_0 \Delta u + \int_0^t \operatorname{div}[a(x)g(t-s)\nabla u(s)]ds + f(u) + b(x)h(u_t) = 0 \text{ in } \Omega \times \mathbb{R}^+, \tag{0.1}$$

where a, b are nonnegative functions, $a \in C^1(\overline{\Omega})$, $b \in L^\infty(\Omega)$, satisfying the assumption

$$a(x) + b(x) \geq \delta > 0 \quad \forall x \in \Omega, \tag{0.2}$$

and f and h are power-like functions.

We observe that the assumption (0.2) gives us a wide assortment of possibilities from which to choose the functions $a(x)$ and $b(x)$, and the most interesting case occurs when one has simultaneous and complementary damping mechanisms. Taking this point of view into account, a distinctive feature of our paper is exactly to consider different and localized damping mechanisms acting in the domain but not necessarily "strategically localized dissipations" as considered in the prior literature.

Key words. stability, wave equation, frictional damping, viscoelasticity

AMS subject classifications. 35L05, 74Dxx, 93D20, 35B35, 35B40

DOI. 10.1137/S0363012902408010

1. Introduction.

Let Ω be an open bounded set of $\mathbb{R}^n$ with smooth boundary Γ. Stability for the wave equation

$$u_{tt} - \Delta u + b(x)h(u_t) = 0 \text{ in } \Omega \times \mathbb{R}^+$$

has been studied for a long time by many authors. For example, Zuazua [15] and Nakao [14] established the uniform decay of solutions provided the function b is positive in the whole domain. When the feedback term depends on the velocity in a linear way, Zuazua [16] proved that the energy related to the above equation decays exponentially if the damping region contains a neighborhood of the boundary Γ or at least contains a neighborhood ω of the particular part given by $\{x \in \Gamma : (x - x_0) \cdot \nu(x) \geq 0\}$. In the same direction, it is important to mention the result of Bardos, Lebeau, and Rauch [2], based on microlocal analysis, that ensures a necessary and sufficient condition to obtain exponential decay; namely, the damping region satisfies the well-known *geometric control condition.* The classical example of an open subset ω verifying this condition is when ω is a neighborhood of the boundary. Later, Nakao [12, 13] extended the results of Zuazua [16], treating first the case of a linear degenerate equation and then the case of a nonlinear dissipation $\rho(x, u_t)$, assuming, as usual, that the function ρ has a polynomial growth near the origin. More recently, Martinez [8] improved the previous results mentioned above in what concerns the linear wave equation subject

*Received by the editors June 10, 2002; accepted for publication (in revised form) April 24, 2003; published electronically October 2, 2003. This work was supported by CNPq (Brazil).
http://www.siam.org/journals/sicon/42-4/40801.html

†Department of Mathematics, State University of Maringá, 87020-900 Maringá - PR, Brazil (mmcavalcanti@uem.br).

‡Department of Mathematics, Federal University of Paraná, 81531-990 Curitiba - PR, Brazil and the National Laboratory for Scientific Computation, Rio de Janeiro, Brazil (higidio@mat.ufpr.br).

to a nonlinear dissipation $\rho(x, u_t)$, avoiding the polynomial growth of the function $\rho(x, s)$ in zero. His proof is based on the piecewise multiplier technique developed by Liu [7] combined with nonlinear integral inequalities to show that the energy of the system decays to zero with a precise decay rate estimate if the damping region satisfies some geometrical conditions. It is important to mention that Lasiecka and Tataru [6] studied the nonlinear wave equation subject to a nonlinear feedback acting on a part of the boundary of the system, and they were the first to prove that the energy decays to zero as fast as the solution of some associated differential equation and without assuming that the feedback has a polynomial growth in zero, although no decay rate has been shown in the general case.

On the other hand, the uniform decay of solutions for the viscoelastic wave equation

$$u_{tt} - \Delta u + \int_0^t \operatorname{div}[a(x)g(t-s)\nabla u(s)]\, ds = 0 \text{ in } \Omega \times \mathbb{R}^+,$$

was obtained by Muñoz Rivera, Barbosa Sobrinho, and Peres Salvatierra [9, 10]. Here, they also assumed that the function a is positive in the whole domain or in ω. At this point it is important to mention some papers in connection with viscoelastic effects; among them are Aassila, Cavalcanti, and Soriano [1], Cavalcanti et al. [4], Dafermos and Nohel [5], Munõz Rivera and Oquendo [11], and references therein.

The goal of the present paper is to study the wave equation with both frictional and viscoelastic dampings, where every one of these dissipations can vanish in a part of Ω and ω. Moreover, we investigate the influence of these dissipations on the rate of decay of the solutions. Our results generalize substantially the results in Cavalcanti, Domingos Cavalcanti, and Soriano [3] and complement the previous ones in the prior literature. The equation that models this problem is given by

$$(1.1)\quad u_{tt} - \kappa_0\Delta u + \int_0^t \operatorname{div}[a(x)g(t-s)\nabla u(s)]\, ds + f(u) + b(x)h(u_t) = 0 \text{ in } \Omega \times \mathbb{R}^+,$$

satisfying the Dirichlet boundary condition, i.e.,

$$(1.2)\qquad u = 0 \text{ on } \Gamma \times \mathbb{R}^+,$$

and initial data

$$(1.3)\qquad u(x,0) = u^0(x), \quad u_t(x,0) = u^1(x) \text{ in } \Omega.$$

Here, k_0 is a positive constant, the functions $a,\ b$ are nonnegative, $a \in C^1(\overline{\Omega})$, $b \in L^\infty(\Omega)$, satisfying

$$(1.4)\qquad meas\{x \in \Gamma : a(x) > 0\} > 0,$$

the relaxation function $g : [0, \infty[\to \mathbb{R}^+$ is nonincreasing and satisfies

$$(1.5)\qquad \|a\|_{L^\infty} \int_0^\infty g(s)\, ds < \kappa_0,$$

and the functions $f,\ h : \mathbb{R} \to \mathbb{R}$ satisfy

$$f(s)s \geq 0, \quad h(s)s \geq 0 \quad \forall s \in \mathbb{R}.$$

Additionally, we suppose that f is superlinear, that is,

$$(\rho+1)F(s) \le f(s)s, \quad F(z) := \int_0^z f(s)ds \quad \forall s \in \mathbb{R},$$

with the growth conditions

$$|f(x) - f(y)| \le C(1 + |x|^{\rho-1} + |y|^{\rho-1})|x-y| \quad \forall x, y \in \mathbb{R},$$

for some $C > 0$ and $\rho \ge 1$ such that $(n-2)\rho \le n$.

To make our calculations more simple, we introduce the following binary operators:

$$\begin{aligned}
(g * w)(t) &:= \int_0^t g(t-s)w(s)\,ds, \\
(g \square w)(t) &:= \int_0^t g(t-s)|w(t) - w(s)|^2\,ds, \\
(g \diamond w)(t) &:= \int_0^t g(t-s)\big(w(t) - w(s)\big)\,ds.
\end{aligned}$$

Some important relations between these operators are given by the following lemma.

LEMMA 1.1. *For any two functions $g, w \in C^1(\mathbb{R})$ and $\theta \in [0,1]$, the following inequalities hold:*

$$\begin{aligned}
2\,[g * w]\,w' &= g' \square w - g(t)|w|^2 - \frac{d}{dt}\left\{ g \square w - \left(\int_0^t g\,ds \right) |w|^2 \right\}, \\
|(g \diamond w)(t)|^2 &\le \left[\int_0^t |g(s)|^{2(1-\theta)}\,ds \right] |g|^{2\theta} \square w.
\end{aligned}$$

Proof. Differentiating the expression

$$g \square h - \left(\int_0^t g\,ds \right) |w|^2,$$

the first part of our conclusion follows. The second part is a consequence of Hölder's inequality. ∎

The existence and regularity of solutions of (1.1)–(1.3) is given by the following theorem.

THEOREM 1.2. *If $(u^0, u^1) \in [H^2(\Omega) \cap H_0^1(\Omega)] \times H_0^1(\Omega)$, then there exists a unique regular solution of* (1.1)–(1.3) *in the class*

$$u \in L^\infty_{loc}(0,\infty; H_0^1(\Omega) \cap H^2(\Omega)),\, u' \in L^\infty_{loc}(0,\infty; H_0^1(\Omega)),\, u'' \in L^\infty_{loc}(0,\infty; L^2(\Omega)).$$

The proof of the above theorem can be easily obtained, making use, for instance, of the Faedo–Galerkin method.

Now, if $\big(u^0, u^1\big) \in H_0^1(\Omega) \times L^2(\Omega)$ and considering standard arguments of density, we can prove that problem (1.1)–(1.3) has a unique solution in the class

$$u \in C^0\left([0,\infty); H_0^1(\Omega)\right) \cap C^1\left([0,\infty); L^2(\Omega)\right).$$

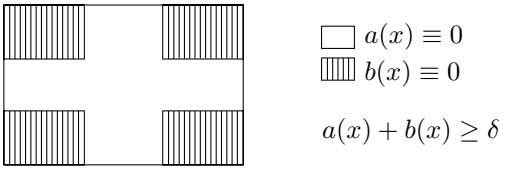


FIG. 1. *A set Ω without geometric control condition for $a(x)$ and $b(x)$.*

2. Exponential decay. In this section we shall prove that the solution of system (1.1)–(1.3) decays exponentially to zero provided the relaxation function g decays exponentially to zero and the function h is linear. In the remainder of this paper, we denote by C a positive constant independent of the initial data (u^0, u^1) which takes different values in different places. Also, we denote by C_σ a positive constant depending on the parameter σ.

The precise assumptions on the coefficients of viscoelastic and frictional dissipations are given in what follows.

Let us assume that

$$a(x) + b(x) \geq \delta \quad \forall x \in \Omega, \tag{2.1}$$

for some $\delta > 0$.

We observe that the assumption (2.1) gives us a wide assortment of possibilities from which to choose the functions $a(x)$ and $b(x)$, and the most interesting case occurs when one has simultaneous and complementary damping mechanisms. Taking this point of view into account, a distinctive feature of our paper is exactly to consider different and localized damping mechanisms acting in the domain but not necessarily "strategically localized dissipations" as considered in the prior literature (see Figure 1).

For the relaxation function g and the function h, we assume that

$$g'(t) \leq -c_1 g(t) \quad \forall t \geq 0, \tag{2.2}$$

$$c_2|w| \leq |h(w)| \leq c_3|w| \quad \forall w \in \mathbb{R}, \tag{2.3}$$

for some positive constants c_1, c_2, c_3.

Additionally, we will need the technical inequalities

$$g''(t) \leq Cg(t), \quad g'''(t) \geq Cg'(t) \quad \forall t \geq 0, \tag{2.4}$$

for some $C > 0$. We will study the asymptotic behavior of system (1.1)–(1.3) when the initial data, for an arbitrary positive λ, satisfy

$$\|u^0\|^2_{H^1} + \|u^1\|^2_{L^2} \leq \lambda. \tag{2.5}$$

Remark 1. We would like to observe that assumption (2.5) does not imply that we are considering small initial data since λ is arbitrary. Indeed, this hypothesis is required because the decay is not uniform for all arbitrary initial data. In other words, we prove that the decay is uniform for initial data taken in bounded sets.

The first order energy of system (1.1)–(1.3) is given by

$$E(t) := \frac{1}{2}\int_\Omega \left(|u_t|^2 + \kappa(x,t)|\nabla u|^2 + g\Box\nabla u\right) dx + \int_\Omega F(u)\, dx,$$

where $\kappa(x,t) := \kappa_0 - a(x)\int_0^t g(s)\,ds$. Note that, in view of (1.5), we have that

$$0 < \kappa_0 - \|a\|_{L^\infty}\int_0^\infty g(s)\,ds \le \kappa(x,t) \le \kappa_0 \quad \forall (x,t) \in \Omega \times \mathbb{R}_0^+.$$

The main result of this section is given by the following theorem.

THEOREM 2.1. *Let us consider the assumptions* (2.1)–(2.4). *If* (u^0, u^1) *satisfy* (2.5)*, then there exists a positive constant* γ *such that*

$$E(t) \le 4E(0)e^{-\frac{\gamma t}{1+\lambda\rho^{-1}}}.$$

Remark 2. The hypothesis (2.2) implies that the relaxation function g decays exponentially, that is, $g(t) \le g(0)e^{c_1 t}$. Consequently we may conjecture that the associated energy decays exponentially, where the function $a(x)$ is, in fact, effective. This kind of conjecture was first introduced by Muñoz Rivera, Barbosa Sobrinho, and Peres Salvatierra [9, 10]. In these works the authors considered the case $b(x) \equiv 0$, while $a(x)$ is effective on the whole domain Ω or at least on a strategic subset $\omega \subset \Omega$. On the other hand, the assumption (2.3) implies that $h(v)$ is similar to a linear function or, in other words, that the dissipation is *almost proportional* to the velocity u_t. We observe that, taking Zuazua's work [16] into account, where $a(x) \equiv 0$ was considered, the energy decays exponentially if $b(x)$ is effective at least on a strategic region $\omega \subset \Omega$. In the present work and keeping in mind that one has complementary dissipations given by the hypothesis (2.1), the above mentioned restrictions on Ω are unnecessary in order to obtain the exponential decay of the energy.

We shall prove Theorem 2.1 for strong solutions, that is, for solutions with initial data $(u^0, u^1) \in [H^2(\Omega) \cap H_0^1(\Omega)] \times H_0^1(\Omega)$. Our conclusion follows by a density argument. We shall apply a piecewise multiplier method to obtain appropriate inequalities for the strong solutions of system (1.1)–(1.3).

The dissipative property of the solutions of system (1.1)–(1.3) is given by the following lemma.

LEMMA 2.2. *The first order energy satisfies the following identity:*

$$\frac{d}{dt}E(t) = \frac{1}{2}\int_\Omega a(x)\big[g'\Box\nabla u - g(t)|\nabla u|^2\big]\,dx - \int_\Omega b(x)h(u_t)u_t\,dx.$$

Proof. Multiplying (1.1) by u_t, performing an integration by parts, and using Lemma 1.1, our conclusion follows. □

Let us consider a nonnegative function $\varphi \in C^1(\bar{\Omega})$, $\operatorname{supp}(\varphi) \subset \operatorname{supp}(a)$ and such that

$$\begin{aligned} &\varphi(x) \ge \delta/2 \quad \text{if} \quad x \in a^{-1}([\delta/2, \infty[), \\ &\varphi(x) = 0 \qquad \text{if} \quad x \in a^{-1}([0, \delta/4]). \end{aligned}$$

Observe that if $a(x) \le \delta/2$ for all $x \in \Omega$, this implies that $b(x) > \delta/2$ for all $x \in \Omega$, since, on the contrary, if $b(x) \le \delta/2$ for some $x \in \Omega$, then

$$a(x) + b(x) \le \delta/2 + \delta/2 = \delta \quad \text{for some } x \in \Omega,$$

which contradicts the assumption (2.1), namely,

$$a(x) + b(x) > \delta \quad \forall x \in \Omega.$$

Consequently, $a(x) \le \delta/2$ for all $x \in \Omega$ implies that $b(x) > \delta/2$ for all $x \in \Omega$. Therefore, we have the frictional damping acting on the whole domain Ω. Analogously,

$b(x) \leq \delta/2$ for all $x \in \Omega$ implies that $a(x) > \delta/2$ for all $x \in \Omega$, which shows us that the viscoelastic damping acts on the whole domain Ω. Now, when one has $a(x) > \delta/2$ for some $x \in \Omega$, and, having in mind that a is a continuous function, then, $a(x) > \delta/2$ holds for a neighborhood W of Ω (which can be considered the maximal one satisfying the property $a(x) > \delta/2$ for all $x \in W$). This means at least that $b(x) > \delta/2$ in $\Omega \backslash W$. Of course the most interesting case occurs when one has simultaneous but complementary damping effects.

Next, we will present two inequalities that will play an essential role when establishing the desired decay rates. First, in view of (2.1), we find that

$$\varphi(x) + b(x) \geq \frac{\delta}{2} \quad \forall x \in \Omega. \tag{2.6}$$

Indeed, we have two cases to consider.

(i) $x \in a^{-1}([\delta/2, +\infty))$. In this case, since $\varphi(x) \geq \delta/2$ and $b(x) \geq 0$, we obtain

$$\varphi(x) + b(x) \geq \delta/2 \quad \forall x \in \Omega.$$

(ii) $x \notin a^{-1}([\delta/2, +\infty))$. We have $0 \leq a(x) < \delta/2$, which implies that $-a(x) > -\delta/2$. From this last inequality and taking assumption (2.1) into account, we deduce

$$\varphi(x) + b(x) \geq b(x) \geq \delta - a(x) > \delta - \delta/2 = \delta/2 \quad \forall x \in \Omega,$$

which proves the inequality (2.6)

Second, from assumption (1.4) and Poincaré's inequality, we have that

$$\int_\Omega \big(\varphi(x) + |\nabla\varphi(x)|\big)|w|^2\,dx \leq C \int_\Omega a(x)|\nabla w|^2\,dx \quad \forall w \in H_0^1(\Omega), \tag{2.7}$$

for some positive constant C.

Indeed, before proving the above inequality, let us remember a useful result which is, in fact, a variant of the Poincaré inequality: *Let Ω_1, Ω_2, and Ω be subsets of $\mathbb{R}^n$ with positive measure and such that $\overline{\Omega}_1 \subset \Omega_2$, $\Omega_2 \subset \Omega$. Then, assuming that Ω is bounded and moreover that meas $(\partial\Omega_2 \cap \partial\Omega) \neq 0$, we have*

$$\int_{\Omega_1} |\omega|^2\,dx \leq C \int_{\Omega_2} |\nabla\omega|^2\,dx \quad \forall \omega \in H_0^1(\Omega),$$

where C is a positive constant.

The proof of the above inequality is immediate. Indeed, it is sufficient to observe that $\omega|_{\partial\Omega_2 \cap \partial\Omega} = 0$ and *meas* $(\partial\Omega_2 \cap \partial\Omega) > 0$.

On the other hand, from assumption (1.4) and since a is continuous, there exist $\varepsilon_0 > 0$ and $V \subset \overline{\Omega}$, a neighborhood of $\partial\Omega$ such that $meas(\partial V \cap \partial\Omega) > 0$ and $a(x) \geq \varepsilon_0$ for all $x \in V$. Setting $\Omega_1 := \text{supp}(\varphi)$, $\Omega_2 := \{x \in \Omega; a(x) > \min\{\delta/4, \epsilon_0\}\}$

and considering $\omega \in H_0^1(\Omega)$, from the above statements we deduce that

$$\begin{aligned}\int_\Omega (\varphi(x)+|\nabla\varphi(x)|)\,|\omega|^2\,dx &= \int_{\Omega_1} (\varphi(x)+|\nabla\varphi(x)|)\,|\omega(x)|^2\\ &\le \max_{x\in\overline{\Omega}} (\varphi(x)+|\nabla\varphi(x)|)\int_{\Omega_1}|\omega(x)|^2\,dx\\ &\le C_1 \max_{x\in\overline{\Omega}} (\varphi(x)+|\nabla\varphi(x)|)\int_{\Omega_2}|\nabla\omega(x)|^2\,dx\\ &\le C_2 \max_{x\in\overline{\Omega}} (\varphi(x)+|\nabla\varphi(x)|)\int_{\Omega_2} a(x)|\nabla\omega(x)|^2\,dx\\ &\le C\int_\Omega a(x)|\nabla\omega(x)|^2\,dx,\end{aligned}$$

which proves (2.7). Let us introduce the following functional:

$$R_1(t) := -\int_\Omega \varphi(x)\left\{u_t(g*u)_t + \frac{1}{2}g''\Box u - \frac{1}{2}g'(t)|u|^2 - \frac{1}{2}a(x)|g*\nabla u|^2\right\}dx.$$

The following lemma retrieves a part of the energy.

LEMMA 2.3. *Given $\epsilon>0$, there exists a positive constant C such that*

$$\begin{aligned}\frac{d}{dt}R_1(t) \le -g(0)\int_\Omega \frac{\delta}{2}|u_t|^2\,dx + \epsilon(1+\lambda^{\rho-1})\int_\Omega \kappa(x,t)|\nabla u|^2\,dx\\ + \frac{C}{\epsilon}\int_\Omega a(x)\big(g(t)|\nabla u|^2 - g'\Box\nabla u\big)\,dx + C\int_\Omega a(x)g\Box\nabla u\,dx\\ + C\int_\Omega b(x)\big(|u_t|^2+|h(u_t)|^2\big)\,dx\end{aligned}$$

for any strong solutions of (1.1)–(1.3).

Proof. Multiplying (1.1) by $\varphi(x)(g*u)_t$, integrating by parts, and using Lemma 1.1, we obtain the following identity:

$$\begin{aligned}\frac{d}{dt}R_1(t) = -\int_\Omega \varphi(x)\left\{g(0)|u_t|^2 + \frac{1}{2}g'''\Box u - \frac{1}{2}g''(t)|u|^2\right\}dx\\ + \int_\Omega (\kappa(x,t)\nabla u + a(x)g\diamond\nabla u)\cdot\nabla\varphi(x)(g*u)_t\,dx\\ + \int_\Omega \kappa_0\nabla u\cdot[\varphi(x)(g*\nabla u)_t]\,dx + \int_\Omega [f(u)+b(x)h(u_t)]\varphi(x)(g*u)_t\,dx.\end{aligned}$$

Using hypothesis (2.4), inequality (2.7), the identity $(g*u)_t = g(t)u - g'\diamond u$, Lemma 1.1, and the Young inequality, we get, for $\eta>0$,

$$\begin{aligned}\frac{d}{dt}R_1(t) \le -g(0)\int_\Omega \varphi(x)|u_t|^2\,dx + C\int_\Omega \big(a(x)g\Box\nabla u + b(x)|h(u_t)|^2\big)\,dx\\ + \eta\int_\Omega \big(\kappa(x,t)|\nabla u|^2 + |f(u)|^2\big)\,dx + \frac{C}{\eta}\int_\Omega a(x)\big(g(t)|\nabla u|^2 - g'\Box\nabla u\big)\,dx. \qquad (2.8)\end{aligned}$$

On the other hand, from the growth conditions of function f and Sobolev imbedding,

we obtain

$$\begin{aligned}\int_\Omega |f(u)|^2\,dx &\le C\left\{\int_\Omega |u|^2\,dx + \int_\Omega |u|^{2\rho}\,dx\right\}\\ &\le C\left\{\int_\Omega |\nabla u|^2\,dx + \left(\int_\Omega |\nabla u|^2\,dx\right)^\rho\right\}\\ &\le C\big(1+E(0)^{\rho-1}\big)\int_\Omega |\nabla u|^2\,dx.\end{aligned}$$

Substituting this inequality into (2.8) and considering (2.6), we arrive at

$$\begin{aligned}\frac{d}{dt}R_1(t) \le -g(0)\int_\Omega \frac{\delta}{2}|u_t|^2\,dx + \eta C\big(1+E(0)^{\rho-1}\big)\int_\Omega \kappa(x,t)|\nabla u|^2\,dx\\ + \frac{C}{\eta}\int_\Omega a(x)\big(g(t)|\nabla u|^2 - g'\Box\nabla u\big)\,dx + C\int_\Omega a(x)g\Box\nabla u\,dx\\ + C\int_\Omega b(x)\big(|u_t|^2+|h(u_t)|^2\big)\,dx.\end{aligned}$$

From (2.5) our conclusion follows. □

Let us introduce the following functional:

$$R_2(t) := \int_\Omega u_t u\,dx.$$

The following lemma retrieves the complementary part of the energy of that given in the previous lemma.

LEMMA 2.4. *There exists a positive constant C such that*

$$\begin{aligned}\frac{d}{dt}R_2(t) \le \int_\Omega |u_t|^2\,dx - \frac{1}{2}\int_\Omega \kappa(x,t)|\nabla u|^2\,dx - (\rho+1)\int_\Omega F(u)\,dx\\ + C\int_\Omega \big(a(x)g\Box\nabla u + b(x)|h(u_t)|^2\big)\,dx\end{aligned}$$

for any strong solutions of (1.1)–(1.3).

Proof. Multiplying (1.1) by u and integrating by parts, we get

$$\frac{d}{dt}R_2(t) = \int_\Omega |u_t|^2 - \kappa(x,t)|\nabla u|^2 - a(x)g\diamond\nabla u\cdot\nabla u\,dx - \int_\Omega [f(u)+b(x)h(u_t)]u\,dx.$$

Applying Young's inequality and using Lemma 1.1, we obtain

$$\begin{aligned}\frac{d}{dt}R_2(t) \le \int_\Omega |u_t|^2\,dx - \frac{1}{2}\int_\Omega \kappa(x,t)|\nabla u|^2\,dx - \int_\Omega f(u)u\,dx\\ + C\int_\Omega \big(a(x)g\Box\nabla u + b(x)|h(u_t)|^2\big)\,dx.\end{aligned}$$

From the superlinearity of the function f, our conclusion follows. □

Let us consider the following functional:

$$R(t) := R_1(t) + \frac{\delta g(0)}{4}R_2(t).$$

The following lemma summarizes the results obtained in the previous lemmas.

LEMMA 2.5. *There exist positive constants k_1 and C such that*

$$\frac{d}{dt}R(t) \le -k_1E(t) + C(1+\lambda^{\rho-1})\int_\Omega a(x)\big(g(t)|\nabla u|^2 - g'\Box\nabla u + g\Box\nabla u\big)\,dx + C\int_\Omega b(x)\big(|u_t|^2 + |h(u_t)|^2\big)\,dx$$

for any strong solutions of (1.1)–(1.3).

Proof. Let us fix ϵ_0 such that

$$-\frac{1}{2}\frac{\delta g(0)}{4} + \epsilon_0(1+\lambda^{\rho-1}) = -\frac{1}{4}\frac{\delta g(0)}{4}.$$

Taking $\epsilon = \epsilon_0$ in Lemma 2.3 and combining it with Lemma 2.4, we get

$$\begin{aligned}\frac{d}{dt}R(t) \le -\frac{\delta g(0)}{4}\int_\Omega\left(|u_t|^2 + \frac{1}{4}\kappa(x,t)|\nabla u|^2 + (\rho+1)F(u)\right)dx \\ + C(1+\lambda^{\rho-1})\int_\Omega a(x)\big(g(t)|\nabla u|^2 - g'\Box\nabla u + g\Box\nabla u\big)\,dx \\ + C\int_\Omega b(x)\big(|u_t|^2 + |h(u_t)|^2\big)\,dx,\end{aligned}$$

from which our conclusion follows. □

Proof of Theorem 2.1 Using hypotheses (2.2) and (2.3) in Lemma 2.5, we get

$$\frac{d}{dt}R(t) \le -k_1E(t) + C(1+\lambda^{\rho-1})\int_\Omega a(x)\big(g(t)|\nabla u|^2 - g'\Box\nabla u\big)\,dx \tag{2.9}$$

$$+ C\int_\Omega b(x)h(u_t)u_t\,dx. \tag{2.10}$$

Let N be a positive constant, and let us introduce the Lyapunov functional

$$F(t) := N(1+\lambda^{\rho-1})E(t) + R(t).$$

It is easy to verify that, for N large, we get

$$\frac{N}{2}(1+\lambda^{\rho-1})E(t) \le F(t) \le 2N(1+\lambda^{\rho-1})E(t) \quad \forall t \ge 0. \tag{2.11}$$

From Lemma 2.2, inequality (2.10), and taking N large, we obtain

$$\frac{d}{dt}F(t) \le -k_1E(t),$$

from which follows, in view of inequality (2.11), that

$$\frac{d}{dt}F(t) \le -\frac{k_1}{2N(1+\lambda^{\rho-1})}F(t).$$

This inequality implies that

$$F(t) \le F(0)e^{-\frac{k_1t}{2N(1+\lambda^{\rho-1})}},$$

and in view of inequality (2.11), we conclude that

$$E(t) \le 4E(0)e^{-\frac{k_1t}{2N(1+\lambda^{\rho-1})}}.$$

Hence the proof is complete. □

3. Polynomial decay. Here our attention will be focused on the uniform rate of decay when the function $g(t)$ is polynomially decreasing as $(1+t)^{-p}$ or the function $h(w)$ is nonlinear of the type $|w|^{1+\frac{1}{q}}$ on a neighborhood of zero. In this case, we will show that the solution decays polynomially.

The hypotheses we will use in this section for the functions g and h are

$$g'(t) \leq -c_1 g^{1+\frac{1}{p}}(t) \quad \forall t \geq 0, \tag{3.1}$$

$$\begin{aligned} c_2|w| \leq |h(w)| \leq c_3|w| \qquad &\text{for} \qquad |w| > 1, \\ c_4|w|^{1+\frac{1}{q}} \leq |h(w)| \leq c_5|w|^{\frac{q}{q+1}} \qquad &\text{for} \qquad |w| \leq 1, \end{aligned} \tag{3.2}$$

where $p > 2$, $q > 1/2$, and $c_1, \ldots, c_5$ are positive.

We summarize the main result of this section in the following theorem.

THEOREM 3.1. *Let us consider the assumptions* (2.1), (2.4), (3.1)–(3.2). *If* (u^0, u^1) *satisfy* (2.5), *then there exists a positive constant* $M = M(\lambda)$ *such that*

$$E(t) \leq \frac{M}{(1+t)^r}$$

for $r = \min\{p,\ 2q\}$.

Remark 3. The above theorem states that the decay rate of the energy is driven by the weakest dissipation, that is, the slowest one. When $b(x) \equiv 0$ and assuming that $a(x) \geq \delta > 0$ for all $x \in \Omega$ (that is, the viscoelastic dissipation acts on the whole domain Ω), and $g(t) = \frac{1}{(1+t)^p}$, Muñoz Rivera [9] proved that the energy decays with the same rate, namely, $E(t) \leq \frac{C}{(1+t)^p}$. On the other hand, if $a(x) \equiv 0$ and $h(w) = |w|^{\frac{1}{q}}w$, where $b(x)$ is effective at least on a strategic part $\omega \subset \Omega \subset \mathbb{R}^2$, Nakao [12] showed that the energy decays with the following rate: $E(t) \leq \frac{C}{(1+t)^{2q}}$. In the present manuscript and taking into consideration that $a(x)$ e $b(x)$ are complementary, it is expected that the energy decays at least according to the weakest (slowest) dissipation, or, in other words, $E(t) \leq \frac{C}{(1+t)^r}$ com $r = \min\{p, 2q\}$.

We start by stating some technical lemmas.

LEMMA 3.2. *Suppose that* $g \in C([0,\infty[)$, $w \in L^1_{loc}(0,\infty)$, *and* $0 \leq \theta \leq 1$; *then we have that*

$$\int_0^t |g(\tau)w(\tau)|\, d\tau \leq \left\{\int_0^t |g(\tau)|^{1-\theta}|w(\tau)|\, d\tau\right\}^{\frac{1}{\sigma+1}} \left\{\int_0^t |g(\tau)|^{1+\frac{\theta}{\sigma}}|w(\tau)|\, d\tau\right\}^{\frac{\sigma}{\sigma+1}}.$$

Proof. For any fixed t we have

$$\int_0^t |g(\tau)w(\tau)|\, d\tau = \int_0^t \underbrace{|g(\tau)|^{\frac{1-\theta}{\sigma+1}}|w(\tau)|^{\frac{1}{\sigma+1}}}_{:=w_1} \underbrace{|g(\tau)|^{1-\frac{1-\theta}{\sigma+1}}|w(\tau)|^{\frac{\sigma}{\sigma+1}}}_{:=w_2}\, d\tau.$$

Note that $w_1 \in L^s_{loc}(0,\infty)$, $w_2 \in L^{s'}_{loc}(0,\infty)$, where $s = \sigma + 1$ and $s' = \frac{\sigma+1}{\sigma}$. Using Hölder's inequality, we get

$$\int_0^t |g(\tau)w(\tau)|\, d\tau \leq \left\{\int_0^t |g(\tau)|^{1-\theta}|w(\tau)|\, d\tau\right\}^{\frac{1}{\sigma+1}} \left\{\int_0^t |g(\tau)|^{1+\frac{\theta}{\sigma}}|w(\tau)|\, d\tau\right\}^{\frac{\sigma}{\sigma+1}}.$$

This completes the proof. ∎

LEMMA 3.3. *Let us suppose that $v \in L^\infty(0,T;H^1(\Omega))$ and g is a continuous function. Then, there exists $C>0$ such that*

$$\int_\Omega a(x) g \Box \nabla v \, dx \le C \left\{ \int_0^t \|v(\tau)\|^2_{H^1} \, d\tau + t\|v(t)\|^2_{H^1} \right\}^{\frac{1}{p+1}} \left\{ \int_\Omega a(x) g^{1+\frac{1}{p}} \Box \nabla v \, dx \right\}^{\frac{p}{p+1}}.$$

Moreover, if there exists $0<\theta<1$ such that $\int_0^\infty g^{1-\theta}(s)\,ds < \infty$, then we have

$$\int_\Omega a(x) g \Box \nabla v \, dx \le C \left\{ \left(\int_0^\infty g^{1-\theta} \, d\tau \right) \|v\|^2_{L^\infty(0,T;H^1(\Omega))} \right\}^{\frac{1}{\theta p+1}} \left\{ \int_\Omega a(x) g^{1+\frac{1}{p}} \Box \nabla v \, dx \right\}^{\frac{\theta p}{\theta p+1}}.$$

Proof. From the hypothesis on v and Lemma 3.2, we get

$$\begin{aligned}
&\int_\Omega a(x) g \Box \nabla v \, dx \\
&\quad = \int_\Omega \int_0^t g(t-\tau) \underbrace{a(x)|\nabla v(t) - \nabla v(\tau)|^2}_{=w(\tau)} \, d\tau \, dx \\
&\quad \le \left\{ \int_\Omega \int_0^t g^{1-\theta}(t-\tau) w(\tau) d\tau \, dx \right\}^{\frac{1}{\theta p+1}} \left\{ \int_\Omega \int_0^t g^{1+\frac{1}{p}}(t-\tau) w(\tau) d\tau \, dx \right\}^{\frac{\theta p}{\theta p+1}} \\
&\quad \le \left\{ \int_\Omega a(x) g^{1-\theta} \Box \nabla v \, dx \right\}^{\frac{1}{\theta p+1}} \left\{ \int_\Omega a(x) g^{1+\frac{1}{p}} \Box \nabla v \, dx \right\}^{\frac{\theta p}{\theta p+1}}. \qquad (3.3)
\end{aligned}$$

Now, for $0<\theta<1$, we have

$$\begin{aligned}
\int_\Omega a(x) g^{1-\theta} \Box \nabla v \, dx &= \int_0^t g^{1-\theta}(t-\tau) \int_\Omega a(x) |\nabla v(t) - \nabla v(\tau)|^2 \, dx d\tau \\
&\le C \left(\int_0^t g^{1-\theta}(\tau) \, d\tau \right) \|v\|^2_{L^\infty(0,T;L^2(\Gamma_2))},
\end{aligned}$$

from which the second inequality of this lemma follows. When $\theta = 1$, we get

$$\begin{aligned}
\int_\Omega a(x) 1 \Box \nabla v \, dx &= \int_0^t \int_\Omega a(x) |\nabla v(t) - \nabla v(\tau)|^2 \, dx \, d\tau \\
&\le C \left\{ t \int_\Omega |\nabla v(t)|^2 \, dx + \int_0^t \int_\Omega |\nabla v(\tau)|^2 \, dx \, d\tau \right\}.
\end{aligned}$$

Substitution of this inequality into (3.3) yields the first inequality. The proof is now complete. ∎

Proof of Theorem 3.1 We shall use some estimates of the previous section which do not depend on the behavior of the functions g and h. First, we will estimate the term $\int_\Omega b(x)\big(|u_t|^2 + |h(u_t)|^2\big)\,dx$. Let us consider the following decomposition of Ω:

$$\Omega^+ := \{x \in \Omega \ : \ |u_t(x)| > 1\} \quad \text{and} \quad \Omega^- := \{x \in \Omega \ : \ |u_t(x)| \le 1\}.$$

From the first hypothesis of (3.2), we get

$$\int_{\Omega^+} b(x)\big(|u_t|^2 + |h(u_t)|^2\big)\,dx \le C \int_\Omega b(x) h(u_t) u_t \, dx \qquad (3.4)$$

for some $C > 0$. On the other hand, the second part of the hypothesis (3.2) implies that

$$|u_t|^2 \le C[h(u_t)u_t]^{\frac{2q}{2q+1}}, \quad |h(u_t)|^2 \le C[h(u_t)u_t]^{\frac{2q}{2q+1}}$$

for any $x \in \Omega^-$. Moreover, using Holder's inequality, we have that

$$\int_{\Omega^-} b(x)[h(u_t)u_t]^{\frac{2q}{2q+1}}\,dx \le C\left(\int_\Omega b(x)h(u_t)u_t\,dx\right)^{\frac{2q}{2q+1}}.$$

Therefore, these two last inequalities imply that

$$\int_{\Omega^-} b(x)\big(|u_t|^2 + |h(u_t)|^2\big)\,dx \le C\left(\int_\Omega b(x)h(u_t)u_t\,dx\right)^{\frac{2q}{2q+1}}. \tag{3.5}$$

Finally, from inequalities (3.4) and (3.5), we conclude that

$$\begin{aligned}
&\int_\Omega b(x)\big(|u_t|^2 + |h(u_t)|^2\big)\,dx \\
&\qquad = \int_{\Omega^+} b(x)\big(|u_t|^2 + |h(u_t)|^2\big)\,dx + \int_{\Omega^-} b(x)\big(|u_t|^2 + |h(u_t)|^2\big)\,dx \\
&\qquad \le C\left\{\int_\Omega b(x)h(u_t)u_t\,dx + \left(\int_\Omega b(x)h(u_t)u_t\,dx\right)^{\frac{2q}{2q+1}}\right\}. \qquad (3.6)
\end{aligned}$$

Next, we will estimate the term $\int_\Omega a(x)g\Box\nabla u\,dx$. From hypothesis (3.1), it is easy to verify that $g(t) \le C(1+t)^{-p}$ for some $C > 0$. Let us fix $\theta = 1/2$, and then $(1-\theta)p > 1$, from which follows that

$$\int_0^\infty g^{1-\theta}(s)\,ds \le C\int_0^\infty \frac{1}{(1+s)^{(1-\theta)p}}\,ds < \infty.$$

Using this estimate in the second part of Lemma 3.3, we get

$$\int_\Omega a(x)g\Box\nabla u\,dx \le CE(0)^{\frac{1}{\theta p+1}}\left(\int_\Omega a(x)g^{1+\frac{1}{p}}\Box\nabla u\,dx\right)^{\frac{\theta p}{\theta p+1}}. \tag{3.7}$$

Substituting (3.6) and (3.7) into Lemma 2.5, we arrive at

$$\begin{aligned}
\frac{d}{dt}R(t) &\le -k_1E(t) + C_\lambda\int_\Omega \big[a(x)\big(g(t)|\nabla u|^2 - g'\Box\nabla u\big) + b(x)h(u_t)u_t\big]\,dx \\
&\quad + C_\lambda\left\{\left(\int_\Omega a(x)g^{1+\frac{1}{p}}\Box\nabla u\,dx\right)^{\frac{\theta p}{\theta p+1}} + \left(\int_\Omega b(x)h(u_t)u_t\,dx\right)^{\frac{2q}{2q+1}}\right\}.
\end{aligned}$$

Let us take $r := \min\{\theta p,\ 2q\}$. Since $R(t) \le CE(t)$ for some $C > 0$, the above

inequality implies that

$$\begin{aligned}\frac{d}{dt}[E^{\frac{1}{r}}R](t) =& \frac{1}{r}R(t)E^{\frac{1}{r}-1}(t)\frac{d}{dt}E(t) + E^{\frac{1}{r}}(t)\frac{d}{dt}R(t)\\ \leq& -CE^{\frac{1}{r}}(t)\frac{d}{dt}E(t) + E^{\frac{1}{r}}(t)\frac{d}{dt}R(t)\\ \leq& -k_2\frac{d}{dt}E^{1+\frac{1}{r}}(t) - k_1E^{1+\frac{1}{r}}(t)\\ &+ C_\lambda E^{\frac{1}{r}}(0)\int_\Omega \big[a(x)\big(g(t)|\nabla u|^2 - g'\Box\nabla u\big) + b(x)h(u_t)u_t\big]\,dx\\ &+ C_\lambda E^{\frac{1}{r}}(t)\left\{\left(\int_\Omega a(x)g^{1+\frac{1}{p}}\Box\nabla u\,dx\right)^{\frac{\theta p}{\theta p+1}} + \left(\int_\Omega b(x)h(u_t)u_t\,dx\right)^{\frac{2q}{2q+1}}\right\} \end{aligned} \tag{3.8}$$

for some positive constant k_2. Now, we will estimate the two last terms of the above inequality. Since

$$E^{\frac{1}{r}}(t)\left(\int_\Omega a(x)g^{1+\frac{1}{p}}\Box\nabla u\,dx\right)^{\frac{\theta p}{\theta p+1}} \leq E^{\frac{\theta p-r}{r(\theta p+1)}}(0)E^{\frac{r+1}{r(\theta p+1)}}(t)\left(\int_\Omega a(x)g^{1+\frac{1}{p}}\Box\nabla u\,dx\right)^{\frac{\theta p}{\theta p+1}},$$

applying Young's inequality yields, for $\epsilon > 0$,

$$E^{\frac{1}{r}}(t)\left(\int_\Omega a(x)g^{1+\frac{1}{p}}\Box\nabla u\,dx\right)^{\frac{\theta p}{\theta p+1}} \leq \epsilon E^{\frac{r+1}{r}}(t) + C_\epsilon E^{\frac{\theta p-r}{r\theta p}}(0)\int_\Omega a(x)g^{1+\frac{1}{p}}\Box\nabla u\,dx. \tag{3.9}$$

Similarly, we have the following estimate:

$$E^{\frac{1}{r}}(t)\left(\int_\Omega b(x)h(u_t)u_t\,dx\right)^{\frac{2q}{2q+1}} \leq \epsilon E^{\frac{r+1}{r}}(t) + C_\epsilon E^{\frac{2q-r}{r(2q)}}(0)\int_\Omega b(x)h(u_t)u_t\,dx. \tag{3.10}$$

Substituting (3.9) and (3.10) into (3.8) and taking ϵ small, we arrive at

$$\begin{aligned}\frac{d}{dt}[E^{\frac{1}{r}}(R+k_2E)](t) \leq& -\frac{k_1}{2}E^{1+\frac{1}{r}}(t)\\ &+ C_\lambda\int_\Omega \big[a(x)\big(g(t)|\nabla u|^2 - g'\Box\nabla u\big) + b(x)h(u_t)u_t\big]\,dx.\end{aligned} \tag{3.11}$$

Let N be a positive constant, and let us introduce the Lyapunov functional

$$F(t) := NE(t) + E^{\frac{1}{r}}(t)\big(R(t) + k_2E(t)\big).$$

It is easy to verify that, for N large, we have

$$\frac{N}{2}E(t) \leq F(t) \leq 2NE(t) \quad \forall t \geq 0. \tag{3.12}$$

From Lemma 2.2 and inequality (3.11), we get, for $N = N(\lambda)$ large,

$$\frac{d}{dt}F(t) \leq -\frac{k_1}{2}E^{1+\frac{1}{r}}(t),$$

from which follows, in view of (3.12), that

$$\frac{d}{dt}F(t) \leq -k_3F^{1+\frac{1}{r}}(t) \tag{3.13}$$

for some $k_3 = k_3(\lambda) > 0$. Hence we obtain

$$F(t) \le \frac{C_\lambda}{(1+t)^r} \quad \text{and consequently} \quad E(t) \le \frac{C_\lambda}{(1+t)^r}.$$

Since $p > 2,\ \theta = 1/2$, and $q > 1/2$, we have that $r > 1$. Therefore,

$$\int_0^\infty \|u(\tau)\|_{H^1}^2\, d\tau + t\|u(t)\|_{H^1}^2 \le C\left\{\int_0^\infty E(\tau)\, d\tau + tE(t)\right\} < \infty.$$

From the first part of Lemma 3.3, we get the following estimate:

$$\int_\Omega a(x) g \Box \nabla u\, dx \le C_\lambda \left(\int_\Omega a(x) g^{1+\frac{1}{p}} \Box \nabla u\, dx\right)^{\frac{p}{p+1}}.$$

Using this inequality instead of (3.7) and repeating the same calculations and changing θp to p, we conclude that

$$E(t) \le \frac{C_\lambda}{(1+t)^r}$$

for $r := \min\{p,\ 2q\}$. This completes the proof. □

Acknowledgments. The authors are deeply grateful to the referees for their comments since they gave us a new vision of the problem and *forced us to clarify some obscure points in our previous version.* The authors would like to thank Enrique Zuazua for his kind attention during the refereeing process.

REFERENCES

[1] M. Aassila, M. M. Cavalcanti, and J. A. Soriano, *Asymptotic stability and energy decay rates for solutions of the wave equation with memory in a star-shaped domain*, SIAM J. Control Optim., 38 (2000), pp. 1581–1602.

[2] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[3] M. M. Cavalcanti, V. N. Domingos Cavalcanti, and J. A. Soriano, *Exponential decay for the solution of semilinear viscoelastic wave equations with localized damping*, Electron. J. Differential Equations, 2002 (2002), pp. 1–14.

[4] M. M. Cavalcanti, V. N. Domingos Cavalcanti, T. F. Ma, and J. A. Soriano, *Global existence and asymptotic stability for viscoelastic problems*, Differential Integral Equations, 15 (2002), pp. 731–748.

[5] C. M. Dafermos and J. A. Nohel, *Energy methods for nonlinear hyperbolic integro differential equations*, Comm. Partial Differential Equations, 4 (1979), pp. 218–278.

[6] I. Lasiecka and D. Tataru, *Uniform boundary stabilization of semilinear wave equation with nonlinear boundary damping*, Differential Integral Equations, 6 (1993), pp. 507–533.

[7] K. Liu, *Locally distributed control and damping for the conservative systems*, SIAM J. Control Optim., 35 (1997), pp. 1574–1590.

[8] P. Martinez, *A new method to obtain decay rate estimates for dissipative systems with localized damping*, Rev. Mat. Complut., 12 (1999), pp. 251–283.

[9] J. E. Muñoz Rivera and J. Barbosa Sobrinho, *Existence and uniform rates of decay for contact problems in viscoelasticity*, Appl. Anal., 67 (1997), pp. 175–199.

[10] J. E. Muñoz Rivera and A. Peres Salvatierra, *Asymptotic behaviour of the energy in partially viscoelastic materials*, Quart. Appl. Math., 59 (2001), pp. 557–578.

[11] J. E. Muñoz Rivera and H. P. Oquendo, *Exponential stability to a contact problem of partially viscoelastic materials*, J. Elasticity, 63 (2001), pp. 87–111.

[12] M. Nakao, *Decay of solutions of the wave equation with a local nonlinear dissipation*, Math. Ann., 305 (1996), pp. 403–417.

[13] M. Nakao, *Decay of solutions of the wave equation with local degenerate dissipation*, Israel J. Math., 95 (1996), pp. 25–42.
[14] M. Nakao, *A difference inequality and its applications to nonlinear evolution equations*, J. Math. Soc. Japan, 30 (1978), pp. 747–762.
[15] E. Zuazua, *Stability and decay for a class of nonlinear hyperbolic problems,* Asymptot. Anal., 1 (1988), pp. 161–185.
[16] E. Zuazua, *Exponential decay for the semilinear wave equation with locally distributed damping,* Comm. Partial Differential Equations, 15 (1990), pp. 205–235.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1325–1346

APPROXIMATION OF TIME-OPTIMAL CONTROL PROBLEMS VIA NONLINEAR POWER MOMENT PROBLEMS*

G. M. SKLYAR† AND S. YU. IGNATOVICH‡

Abstract. In this paper, we continue the construction of the time-optimal control problems classification for nonlinear affine analytic systems in a neighborhood of the equilibrium. The moment approach is the basis of this investigation. All possible asymptotics for solutions of the nonlinear time-optimal control problems are described in terms of special right ideals in the algebra of nonlinear power moments. We demonstrate the possible asymptotic behavior of the time-optimal control for a certain class of essentially nonlinear three-dimensional systems.

Key words. nonlinear time-optimal control problem, nonlinear power moment min-problem, algebra of nonlinear power moments

AMS subject classifications. 93B10, 93B25

DOI. 10.1137/S0363012901398253

1. Introduction.

One of the most powerful tools in nonlinear control theory is the representation of nonlinear control systems in the form of the series of Volterra type [4]. This approach gives the proper language for the analysis of mappability and approximation of nonlinear systems. Such problems are studied in many remarkable works. In addition, the study of the problem of realizability of the series as certain systems became a special area where control systems are treated as abstract algebraic structures.

It turns out that the series method may be useful in the analysis of specific problems of optimal control theory such as the time-optimal problem.

In this paper, we consider the problem of time-optimal control to the equilibrium for the nonlinear affine control system given in the form (2.1). Applying the series method for the system (2.1), we get the formula which expresses each initial point x^0 via the control $u(t)$, steering this point to the equilibrium in terms of *the series of nonlinear power moments* (Theorem 2.1). As a result, the steering problem is reduced to the *nonlinear power moment problem*. We note that such a reduction is generally accepted for linear systems.

One can see that, like in the linear case, in steering to the equilibrium the (nonautonomous) system corresponding to the nonlinear power moment problem is not uniquely defined.

From the technique viewpoint, the representation given in Theorem 2.1 can be obtained by certain transformations of the Fliess formula for a trajectory of the system [9]. However, from the algebraic point of view, the series of nonlinear power moments turns out to be essentially different. Thus in [25] we show that the structure of the algebra of nonlinear power moments differs from the Fliess one.

*Received by the editors November 15, 2001; accepted for publication (in revised form) February 27, 2003; published electronically October 2, 2003. This research was supported by KBN Polish grant 5 PO3A 030 21.

http://www.siam.org/journals/sicon/42-4/39825.html

†Instytut Matematyki, Uniwersytet Szczeciński, ul. Wielkopolska 15, 70-451, Szczecin, Poland (sklar@sus.univ.szczecin.pl, sklyar@univer.kharkov.ua).

‡Department of Differential Equations and Control, Kharkov National University, Svoboda sqr., 4, 61077, Kharkov, Ukraine (bob@online.kharkiv.com).

Extending the approach proposed for the linear time optimality [15], we consider the nonlinear time-optimal control problem in terms of *the Markov moment min-problem* (Definition 2.2).

In [16, 17, 18], the technique of the moment theory was developed and applied to obtain the *exact* solutions of several special linear time-optimal control problems. Further, in [23], the time-optimal control problem for arbitrary *linear* systems with real analytic coefficients was considered, and the possible asymptotic behavior of its solution in a neighborhood of the equilibrium was completely described. In analyzing the case of *affine nonlinear* systems [24], the conditions under which the nonlinear time-optimal control problem is equivalent to a linear one in the sense of asymptotic behavior of their solutions were given.

In the present paper, the asymptotic approximation of time-optimal control problems in a neighborhood of the equilibrium for *arbitrary affine nonlinear* control systems is studied (Definition 2.3).

The question of approximation of nonlinear systems has been of interest for the last three decades. In this context, the problem of approximation can be regarded as follows: for the given control system, construct a new system having the simpler structure and preserving certain properties of the initial system [19, 10]. In a number of works, the set of vector fields corresponding to a control system is approximated by the set of vector fields whose Lie algebra is nilpotent [11]. The construction is applied first to the problems of controllability [26, 2] and stabilization [12]. On the other hand, in [6], the method of lifting the system to a greater dimension space is proposed, and the approximating system is constructed as the realization of a truncated Volterra series of the system. In the present paper, we suggest another approach to the problem of approximation of the nonlinear systems—namely, an approximation in the sense of time optimality. More specifically, we say that *a certain system approximates the given one in the sense of time optimality* in a neighborhood of the equilibrium if the asymptotic behavior of the solutions of the time-optimal control problems for these two systems is the same.

The paper is organized as follows. Background and the main results are given in section 2. The first step in our consideration is the representation of the series of nonlinear power moments in a *canonical form* with homogeneous *principal part.* The main result concerning *the approximation* is given in Theorem 2.4. We prove that the asymptotic behavior of the solution of the moment min-problem is preserved when the series of nonlinear power moments is substituted by its *principal part.* The proof of the theorem is given in section 3. Theorem 2.6 is the main result *on the classification of the canonical forms.* In this theorem, feasible principal parts of series of nonlinear power moments are given, and the corresponding approximating nonlinear control systems are constructed. The method of the series transformation to the canonical form given in Theorem 2.6 is based on the study of the properties of *the algebra of nonlinear power moments* and is given in section 4. The essential point of this method is Theorem 4.5, which develops the result obtained by Ree [22] (see our Theorem 2.5). Section 5 contains the proof of Theorem 2.6.

In section 6, we construct the series of nonlinear power moments and the canonical form for the system of the Euler equations for a spacecraft. In section 7, we analyze carefully the problem of time-optimal control to the origin for the following two systems:

$$\text{(A)}\ \dot{x}_1 = u,\ \dot{x}_2 = \frac{1}{2}x_1^2,$$

$$\text{(B)}\ \dot{x}_1 = u,\ \dot{x}_2 = x_1,\ \dot{x}_3 = \frac{1}{2}x_1^2.$$

It turns out that in case (B) the optimal control is not the bang-bang. However, we show that for these two systems all conditions of Theorem 2.6 are satisfied, and, therefore, they approximate certain classes of nonlinear systems in the sense of time optimality. We also give a simulation of the time-optimal control.

2. Background and the main results. We consider the class $\mathcal{U}$ of the affine control systems of the form

$$\dot{x} = a(t,x) + ub(t,x), \qquad a(t,0) \equiv 0, \qquad |u(t)| \le 1 \quad \text{a.e.}, \tag{2.1}$$

where $a(t,x)$, $b(t,x)$ are real analytic vector functions defined on some neighborhood of the origin of $\mathbb{R}^{n+1}$. Further, we denote the system (2.1) by $\{a,b\}$.

2.1. Series of nonlinear power moments. For a sufficiently small $\theta > 0$, the system (2.1) naturally generates the mapping $S_{a,b} : (\theta, u) \to x^0$, where $|u(t)| \le 1$ a.e. for $t \in [0,\theta]$ and $x^0 = x(0)$ is the initial point transferred to the origin by the control $u(t)$ in time θ by virtue of the system (2.1). The basic point for our consideration is the following representation of $S_{a,b}$ which can be obtained by transforming the Fliess formula [9].

THEOREM 2.1. *Assume that the functions $a(t,x)$, $b(t,x)$ are analytic on some neighborhood of the origin of $\mathbb{R}^{n+1}$. Then there exists $T_0 > 0$ such that for any $\theta \in (0,T_0)$ the operator $S_{a,b}(\theta,\cdot)$ admits the representation in the form of a series of nonlinear power moments,*

$$S_{a,b}(\theta,u) = \sum_{m=1}^{\infty} \sum_{\substack{m_1+\cdots+m_k+k=m \\ k\ge 1, m_j \ge 0}} v_{m_1\dots m_k} \xi_{m_1\dots m_k}(\theta,u), \tag{2.2}$$

where $\xi_{m_1\dots m_k}(\theta,u)$ are nonlinear power moments of the function $u(t)$ of the form

$$\xi_{m_1\dots m_k}(\theta,u) = \int_0^\theta \int_0^{\tau_1} \cdots \int_0^{\tau_{k-1}} \tau_1^{m_1}\tau_2^{m_2}\cdots\tau_k^{m_k} \prod_{j=1}^{k} u(\tau_j)\, d\tau_k \cdots d\tau_2 d\tau_1, \tag{2.3}$$

and $v_{m_1\dots m_k}$ are constant vector coefficients defined as follows. Let R_a, R_b be the operators acting by the rule

$$R_a d(t,x) = d_t(t,x) + d_x(t,x)\cdot a(t,x), \qquad R_b d(t,x) = d_x(t,x)\cdot b(t,x)$$

for any vector function $d(t,x)$ analytic on some neighborhood of the origin of $\mathbb{R}^{n+1}$ and

$$\operatorname{ad}_{R_a}^{m+1} R_b = [R_a, \operatorname{ad}_{R_a}^m R_b],\ m \ge 0;\ \operatorname{ad}_{R_a}^0 R_b = R_b,$$

where $[\cdot,\cdot]$ is the operator commutator. Then

$$v_{m_1\dots m_k} = \frac{(-1)^k}{m_1!\cdots m_k!} \operatorname{ad}_{R_a}^{m_1} R_b \circ \operatorname{ad}_{R_a}^{m_2} R_b \circ \cdots \circ \operatorname{ad}_{R_a}^{m_k} R_b E(x) \Big|_{\substack{t=0 \\ x=0}}, \tag{2.4}$$

where $E(x) \equiv x$. Moreover, there exists a constant $C_0 > 0$ such that for any $\theta \in (0,T_0)$ and any $u(t)$ such that $|u(t)| \le 1$ a.e., $t \in [0,\theta]$, the following estimate holds:

$$\sum_{m=1}^{\infty} \left\| \sum_{\substack{m_1+\cdots+m_k+k=m \\ k\ge 1, m_j \ge 0}} v_{m_1\dots m_k} \xi_{m_1\dots m_k}(\theta,u) \right\| \le C_0. \tag{2.5}$$

Further, under $S_{a,b}$, we mean the series in the right-hand side of (2.2).

Throughout the paper, we refer to the number $m = m_1 + \cdots + m_k + k$ as *the order of the nonlinear power moment* $\xi_{m_1 \ldots m_k}$. Thus the method of summation of the series $S_{a,b}$ corresponds to the order of its terms.

2.2. Nonlinear Markov moment min-problem. Following the approach first proposed for the linear case in [15], we reformulate the time-optimal control problem for the system $\{a, b\}$ as *the nonlinear Markov moment min-problem* for the series $S_{a,b}$.

DEFINITION 2.2. *The nonlinear Markov moment min-problem for the series $S_{a,b}$ of the form* (2.2) *is as follows: for a given vector $s \in \mathbb{R}^n$ find (if possible) the smallest interval $[0, \theta_s]$ such that the set $U_s^{a,b}(\theta_s)$ of functions $\tilde{u}(t)$, $|\tilde{u}(t)| \le 1$, $t \in [0, \theta_s]$, satisfying for $\theta = \theta_s$ the moment equalities*

$$s = S_{a,b}(\theta, \tilde{u}) \tag{2.6}$$

is nonempty.

Further on, we identify the time-optimal control problem for the system $\{a, b\}$ and the Markov moment min-problem for the series $S_{a,b}$.

Note that for any $\{a, b\} \in \mathcal{U}$ and any analytic transformation F of a neighborhood of the origin in $\mathbb{R}^n$, $F(0) = 0$, there exists the number T_1 such that for $\theta \in (0, T_1)$ the mapping $\tilde{S}(\theta, \cdot) = F(S_{a,b}(\theta, \cdot))$ is expanded in the series of nonlinear power moments that represents the system

$$\dot{z} = \tilde{a}(t, z) + \tilde{b}(t, z)u, \qquad \text{where} \quad z = F(x), \tag{2.7}$$

i.e., $\tilde{S} = S_{\tilde{a},\tilde{b}}$. Thus the representation of the time-optimal control problem in the form of the nonlinear moment min-problem allows us to consider transformations of the series instead of changes of variables in the system.

2.3. Approximation in the sense of time optimality. Developing the approach of [23, 24], we adopt the following definition.

DEFINITION 2.3. *Consider two systems $\{a, b\}, \{a^*, b^*\} \in \mathcal{U}$, and suppose there exists such an open domain $\Omega \subset \mathbb{R}^n \backslash \{0\}$, $0 \in \overline{\Omega}$, that for any $x_0 \in \overline{\Omega}$ there exists the unique solution $(\theta^*_{x_0}, u^*_{x_0}(t))$ of the time-optimal control problem for the system $\{a^*, b^*\}$. Denote by $U_s^{a,b}(\theta)$ the set of all admissible controls which transfer the point s to the origin by virtue of the system $\{a, b\}$ in time θ, and denote by θ_s the optimal time for the system $\{a, b\}$, i.e., $\theta_s = \min\{\theta : U_s^{a,b}(\theta) \ne \emptyset\}$.*

We say that the nonlinear time-optimal control problem for the system $\{a^*, b^*\}$ approximates the time-optimal control problem for the system $\{a, b\}$ (in the domain Ω) if there exist a nonsingular transformation Φ of a neighborhood of the origin of $\mathbb{R}^n$, $\Phi(0) = 0$, and a set of pairs $(\tilde{\theta}_s, \tilde{u}_s(t))$, $s \in \Omega$, such that $\tilde{u}_s(t) \in U^{a,b}_{\Phi(s)}(\tilde{\theta}_s)$ and

$$\frac{\theta_{\Phi(s)}}{\theta^*_s} \to 1, \qquad \frac{\tilde{\theta}_s}{\theta^*_s} \to 1, \tag{2.8}$$

$$\frac{1}{\theta} \int_0^\theta |u^*_s(t) - \tilde{u}_s(t)| dt \to 0 \qquad as \qquad s \to 0, \quad s \in \Omega, \tag{2.9}$$

where $\theta = \min\{\tilde{\theta}_s, \theta^*_s\}$.

The first main result of the paper is the following theorem on approximation of time-optimal control problems.

THEOREM 2.4 (on approximation). *Let a transformation F reduce the series $S_{a,b}$ of the form* (2.2) *representing a system $\{a,b\} \in \mathcal{U}$ to the form*

$$S_{\widetilde{a,b}} = F(S_{a,b}) = \Xi + \rho, \tag{2.10}$$

such that the qth component Ξ_q of Ξ contains nonlinear power moments of order r_q only, and the qth component ρ_q of ρ contains moments of order greater than r_q, $q = 1, \dots, n$. Let $\Omega \subset \mathbb{R}^n$, $0 \in \overline{\Omega}$, be an open domain such that

(i) *the moment min-problem*

$$s = \Xi(\theta, u) \tag{2.11}$$

has the unique solution $(\theta_s^*, u_s^*(t))$ for any $s \in \Omega$;

(ii) *the function θ_s^* is continuous at any $s \in \Omega$;*

(iii) *for the set $K = \{u_s^*(t\theta_s^*),\ t \in [0,1] : s \in \Omega\}$, the following condition* (C) *holds:*

(C) *in considering K as a set in the space $L_2[0,1]$, the weak convergence of a sequence of elements from K implies the strong convergence.*

Then there exists such a set $\{\Omega(\delta)\}_{\delta>0}$ of embedded domains ($\Omega(\delta_1) \subset \Omega(\delta_2)$ for $\delta_1 > \delta_2$) that the moment min-problem (2.11) *approximates the time-optimal control problem for the system $\{a,b\}$ in any domain $\Omega(\delta)$ (with $\Phi = F^{-1}$) and $\cup_{\delta>0}\Omega(\delta) = \Omega$.*

In other words, if after *a certain* transformation of the series (what is equivalent to the change of variables in the control system) it turns out that the moment min-problem for the *principal* part Ξ (with homogeneous components) satisfies conditions (i)–(iii), then the solution of this homogeneous moment min-problem asymptotically approximates the solution of the initial time-optimal control problem.

The proof of this theorem can be found in section 3.

2.4. Algebra of nonlinear power moments. The statement of Theorem 2.4 leads us to the problem of description of *all possible* principal parts Ξ in *the canonical form* (2.10). To this end, we propose the following algebraic approach. Consider the linear span $\mathcal{A}$ of all nonlinear power moments $\xi_{m_1 \dots m_k}$ over $\mathbb{R}$ as the free algebra with the basis

$$\{\xi_{m_1 \dots m_k} :\ k \geq 1,\ m_1, \dots, m_k \geq 0\} \tag{2.12}$$

and the multiplication of the form

$$\xi_{m_1 \dots m_k} \xi_{n_1 \dots n_s} = \xi_{m_1 \dots m_k n_1 \dots n_s}.$$

Consider also *the Lie algebra* L over $\mathbb{R}$ generated by the elements $\{\xi_m\}_{m=0}^{\infty}$ with the commutation $[\ell_1, \ell_2] = \ell_1\ell_2 - \ell_2\ell_1$, $\ell_1, \ell_2 \in L$.

We introduce *the inner product* in $\mathcal{A}$ considering the basis (2.12) as *the orthonormal one*. Given a subspace $P \subset \mathcal{A}$, by $P^\perp$ we denote its orthogonal complement, $P^\perp = \{a \in \mathcal{A} :\ (a, y) = 0 \text{ for all } y \in P\}$.

Introduce further the graded structure in the algebra $\mathcal{A}$ putting the order of basis elements by the formula

$$\operatorname{ord}(\xi_{m_1 \dots m_k}) = m_1 + \dots + m_k + k.$$

We say that the element $x \in \mathcal{A}$ has an order iff x is a linear combination of basis elements of the same order. We write $\operatorname{ord}(x) = m$ iff x is a linear combination of

basis elements of order m. Obviously, the subspaces $\mathcal{A}_m = \{x \in \mathcal{A} : \operatorname{ord}(x) = m\}$ are *finite-dimensional and orthogonal to each other.*

The main tool of our further construction is the operation of the *shuffle product* [22, 1, 5, 13] given in $\mathcal{A}$ by the following recursion formula:

$$\begin{aligned} &\xi_m * \xi_n = \xi_{mn} + \xi_{nm}, \\ &\xi_{m_1 \dots m_k} * \xi_n = (\xi_{m_1 \dots m_{k-1}} * \xi_n)\xi_{m_k} + \xi_{m_1 \dots m_k}\xi_n, \\ &\xi_{m_1 \dots m_k} * \xi_{n_1 \dots n_s} = (\xi_{m_1 \dots m_{k-1}} * \xi_{n_1 \dots n_s})\xi_{m_k} + (\xi_{m_1 \dots m_k} * \xi_{n_1 \dots n_{s-1}})\xi_{n_s}. \end{aligned} \tag{2.13}$$

For a given $P \subset \mathcal{A}$, we denote $P^{\mathrm{sh}} = \operatorname{Lin}\{a_1 * \cdots * a_n : \ n \geq 2, a_1, \dots, a_n \in P\}$.

The shuffle product operation is associative and commutative; if $\xi_{m_1 \dots m_k}$ are considered as the nonlinear power moments (2.3), then the shuffle product corresponds to the "usual" product of moments (i.e., integrals) *as functionals of* u.

The connection among the Lie algebra, the shuffle product operation, and the inner product of $\mathcal{A}$ is described by the remarkable theorem given by Ree, which is reformulated under the introduced notation as follows.

THEOREM 2.5 (See Ree [22, Theorem 2.2]). *The Lie algebra L allows the representation*

$$L = (\mathcal{A} * \mathcal{A})^{\perp},$$

where $\mathcal{A} * \mathcal{A} = \operatorname{Lin}\{a_1 * a_2 : \ a_1, a_2 \in \mathcal{A}\}$.

2.5. The structure of the right ideal induced by the system in the algebra and canonical forms classification. We consider now the system $\{a, b\} \in \mathcal{U}$. The series $S_{a,b}$ (of the form (2.2)) naturally defines the linear mapping $v : \mathcal{A} \to \mathbb{R}^n$ by the rule $v(\xi_{m_1 \dots m_k}) = v_{m_1 \dots m_k}$. We note that under this definition each element of the Lie algebra L is mapped to the vector field from the Lie algebra generated by the sequence of operators $\operatorname{ad}^m_{R_a} R_b$, $m \geq 0$, applied to $E(x) \equiv x$ and evaluated at the point $t = 0$, $x = 0$. Namely, denote $R = \frac{(-1)^k}{m_1! \dots m_k!}[\operatorname{ad}^{m_1}_{R_a} R_b, \cdots [\operatorname{ad}^{m_{k-1}}_{R_a} R_b, \operatorname{ad}^{m_k}_{R_a} R_b] \cdots]$. Then $v([\xi_{m_1}, \cdots [\xi_{m_{k-1}}, \xi_{m_k}] \cdots]) = RE(x)\big|_{\substack{t=0 \\ x=0}}$. In addition, $Rd(t, x) = d_x(t, x) \cdot R\,E(x)$ for any analytic vector function $d(t, x)$. This results for any $\ell \in L$ in the condition

$$v(\ell \xi_{m_1 \dots m_k}) = \frac{(-1)^k}{m_1! \dots m_k!} (\operatorname{ad}^{m_1}_{R_a} R_b \circ \cdots \circ \operatorname{ad}^{m_k}_{R_a} R_b E(x))'_x \Big|_{\substack{t=0 \\ x=0}} v(\ell), \tag{2.14}$$

which gives the certain relations between coefficients $v_{m_1 \dots m_k}$ of the series $S_{a,b}$. So, if $v(\ell) = 0$, then $v(\ell z) = 0$ for any element $z \in \mathcal{A}$. This observation leads us to the following concept of *the right ideal* $J_{a,b}$ generated by the system.

Further, throughout the paper, we assume the system $\{a, b\}$ to be *n-dimensional,* i.e., $\dim v(L) = n$. For a given n-dimensional system $\{a, b\}$, consider the sequence of subspaces $D^r_{a,b} = v(L \cap (\mathcal{A}_1 + \cdots + \mathcal{A}_r)) \subset \mathbb{R}^n$, denote $d_r = \dim D^r_{a,b}$, $r \geq 1$, and put $r_{a,b} = \min\{r : d_r = n\}$. For any $1 \leq r \leq r_{a,b}$ there exist exactly $c_r = \dim(L \cap \mathcal{A}_r) - (d_r - d_{r-1})$ linearly independent elements $p^r_i \in L \cap \mathcal{A}_r$, $1 \leq i \leq c_r$, such that $v(p^r_i) \in D^{r-1}_{a,b}$. (We assume $D^0_{a,b} = \{0\}$, $d_0 = 0$.) We define the right ideal as

$$J_{a,b} = J = \operatorname{Lin}\{p^r_i x, \ 1 \leq i \leq c_r, \ 1 \leq r \leq r_{a,b}, \ x \in \mathcal{A} + \mathbb{R}\}.$$

If $c_1 = \cdots = c_{r_{a,b}} = 0$, then we assume $J = J_{a,b} = \{0\}$.

Thus (2.14) implies the following property of elements of the ideal $J = J_{a,b}$:

$$\text{if } z \in J \cap \mathcal{A}_m, \quad \text{then } v(z) \in v(\mathcal{A}_1 + \cdots + \mathcal{A}_{m-1}). \tag{2.15}$$

We note also that if a nonsingular transformation reduces the system $\{a, b\}$ to the system $\{a', b'\}$, then $J_{a,b} = J_{a',b'}$.

Further on, *the "tilde" symbol means the projection on* $J^\perp$; i.e., $\widetilde{x}$ and $\widetilde{L}$ denote the projections of the element x and the Lie algebra L on $J^\perp$, respectively. We see that $\dim(\sum_{r=1}^{r_{a,b}}(\widetilde{L} \cap \mathcal{A}_r)) = \sum_{r=1}^{r_{a,b}}(\dim(L \cap \mathcal{A}_r) - \dim(J \cap L \cap \mathcal{A}_r)) = d_{r_{a,b}} = n$.

On the other hand, we consider any $r_0 \geq 1$ and an arbitrary sequence of subspaces $M = \{M_r\}_{r=1}^{r_0}$, $M_r \subset L \cap \mathcal{A}_r$, such that $\sum_{r=1}^{r_0}(\dim(L \cap \mathcal{A}_r) - \dim(J_M \cap L \cap \mathcal{A}_r)) = n$, where $J_M = \mathrm{Lin}\{px : p \in \sum_{r=1}^{r_0} M_r, x \in \mathcal{A} + \mathbb{R}\}$. We denote by $\mathcal{J}$ the family of all such right ideals J_M. Obviously, the right ideal $J = J_{a,b}$ corresponding to the system $\{a, b\} \in \mathcal{U}$ belongs to $\mathcal{J}$.

The second main result of the present paper is the following theorem, which describes *all possible canonical forms* of the series of nonlinear power moments.

THEOREM 2.6 (on classification of canonical forms).

(i) *Let* $\{a, b\} \in \mathcal{U}$ *be* n*-dimensional, and let* $\widetilde{\ell}_1, \ldots, \widetilde{\ell}_n$ *be a basis Lof* $\sum_{r=1}^{r_{a,b}}(\widetilde{L} \cap \mathcal{A}_r)$ *such that* $\mathrm{ord}(\widetilde{\ell}_i) \leq \mathrm{ord}(\widetilde{\ell}_j)$ *as* $i < j$. *Then there exists a nonsingular analytic transformation* F *of a neighborhood of the origin which reduces* $S_{a,b}$ *to the canonical form*

$$F(S_{a,b}) = \begin{pmatrix} \widetilde{\ell}_1 \\ \cdots \\ \widetilde{\ell}_n \end{pmatrix} + \rho, \tag{2.16}$$

where the components of ρ *equal*

$$\rho_q = \sum_{m=\mathrm{ord}(\widetilde{\ell}_q)+1}^{\infty} \sum_{\substack{m_1+\cdots+m_k+k=m \\ k\geq 1, m_j \geq 0}} \mu_{q,m_1\ldots m_k} \xi_{m_1\ldots m_k}, \qquad \mu_{q,m_1\ldots m_k} \in \mathbb{R}, \tag{2.17}$$

and $|\rho_q(\theta, u)| \leq C\theta^{\mathrm{ord}(\widetilde{\ell}_q)+1}$ *as* $|u(t)| \leq 1$ *a.e.,* $t \in [0, \theta]$, $q = 1, \ldots, n$.

(ii) *Let* $J = J_M \in \mathcal{J}$. *Then there exists a (homogeneous) system* $\{a^*, b^*\} \in \mathcal{U}$ *such that* $J = J_{a^*,b^*}$ *and, moreover,*

$$S_{a^*,b^*} = \begin{pmatrix} \widetilde{\ell}_1 \\ \cdots \\ \widetilde{\ell}_n \end{pmatrix},$$

where $\widetilde{\ell}_1, \ldots, \widetilde{\ell}_n$ *is a basis of* $\sum_{r=1}^{r_0}(\widetilde{L} \cap \mathcal{A}_r)$.

One can find the proof of this theorem in section 5. It is based on the properties of the algebra of nonlinear power moments discussed in section 4.

COROLLARY 2.7. *For any* n*-dimensional system* $\{a, b\} \in \mathcal{U}$ *there exists a homogeneous system* $\{a^*, b^*\} \in \mathcal{U}$ *(canonical approximation) whose series coincides with the principal part of the canonical form of* $\{a, b\}$.

COROLLARY 2.8. *Two* n*-dimensional systems* $\{a, b\}, \{\widetilde{a}, \widetilde{b}\} \in \mathcal{U}$ *have the same canonical approximation iff* $J_{a,b} = J_{\widetilde{a},\widetilde{b}}$.

COROLLARY 2.9. *The set of all possible principal parts of series of nonlinear power moments representing systems from* $\mathcal{U}$ *(up to changes of variables) is in one-to-one correspondence with the family of right ideals* $\mathcal{J}$.

3. Proof of Theorem 2.4. Consider moment min-problem (2.11), and introduce the operator $H_\varepsilon(x) = (\varepsilon^{r_1}x_1, \dots, \varepsilon^{r_n}x_n)$. Obviously, if $(\theta_s^*, u_s^*(t))$ is the solution of this problem for the point $s \in \Omega$, then $(\varepsilon\theta_s^*, u_s^*(t/\varepsilon))$ is the solution of (2.11) for the point $H_\varepsilon(s)$. Denote $\omega = \{\Xi(1, u_s^*(t\theta_s^*)) : s \in \Omega\}$. Then conditions (i)–(iii) hold in the set $\cup_{\varepsilon>0} H_\varepsilon(\omega)$. Hence, without loss of generality, we assume $\Omega = \cup_{\varepsilon>0} H_\varepsilon(\omega)$. Now introduce $\omega(\delta) = \{s \in \omega : s + U_\delta \subset \Omega\}$, where $U_\delta = \{x : |x_k| \le \delta, k = 1, \dots, n\}$ and $\Omega(\delta) = \cup_{\varepsilon>0} H_\varepsilon(\omega(\delta))$. Then $\cup_{\varepsilon>0} H_\varepsilon(\omega(\delta) + U_{\delta/2}) \subset \cup_{\varepsilon\ge 0} H_\varepsilon(\overline{\omega(\delta) + U_{\delta/2}}) = \widehat{\Omega} \subset \Omega \cup \{0\}$, where $\widehat{\Omega}$ is closed.

We put $\Phi = F^{-1}$. Then the set of solutions $(\theta_{\Phi(s)}, u(t)) : u(t) \in U^{a,b}_{\Phi(s)}(\theta_{\Phi(s)})$ of the time-optimal problem for the system $\{a, b\}$ coincides with the set of solutions of the moment min-problem $s = \Xi(\theta, u) + \rho(\theta, u)$. Following [15], [24], we introduce the operator $D : \Omega \to (\mathbb{R}, L_2[0, \infty))$ which associates the pair $(\theta_x^*, u_x^*(t))$ with the point $x \in \Omega$ and consider the operator $G_s(x) = s - \rho(D(x)) : \Omega \to \mathbb{R}^n$. We show that the operator G_s has a stationary point in a certain domain when $s \in \Omega(\delta)$ is rather small.

We prove first that the operator $G_s = G_s(x)$ is continuous in Ω. Let $x_p \to x$, where $x_p, x \in \Omega$. Consider the sequence $\hat{u}_{x_p}(t) = u^*_{x_p}(t\theta^*_{x_p}) \in K$. Since $|\hat{u}_{x_p}(t)| \le 1$ and due to condition (C), there exists a subsequence $\hat{u}_{x_{p_q}}(t)$ which strongly converges in $L_2[0, 1]$ to some function $\hat{v}(t)$, $|\hat{v}(t)| \le 1$, $t \in [0, 1]$. Since $\theta^*_{x_{p_q}} \to \theta^*_x$, we get

$$\xi_{m_1\dots m_k}(\theta^*_{x_{p_q}}, u^*_{x_{p_q}}) = {\theta^*_{x_{p_q}}}^m \xi_{m_1\dots m_k}(1, \hat{u}_{x_{p_q}}) \to {\theta^*_x}^m \xi_{m_1\dots m_k}(1, \hat{v}) = \xi_{m_1\dots m_k}(\theta^*_x, v)$$

as $q \to \infty$, where $v(t) = \hat{v}(t\theta^*_x)$, $m = m_1 + \dots + m_k + k$. In particular, this means that $x_{p_q} = \Xi(\theta^*_{x_{p_q}}, u^*_{x_{p_q}}) \to \Xi(\theta^*_x, v) = x$, and therefore, $v(t) = \hat{v}(t\theta^*_x) = u^*_x(t)$. This yields that the strong limit of the sequence $\hat{u}_{x_p}(t) = u^*_{x_p}(t\theta^*_{x_p})$ exists and equals $\hat{v}(t) = u^*_x(t\theta^*_x)$. Thus we get $\rho(D(x_p)) \to \rho(D(x))$; hence the operator $G_s = G_s(x)$ is continuous at any $x \in \Omega$ for an arbitrary s.

Further on, we denote $V^\varepsilon = \{x : |x_k| \le \varepsilon^{r_k}, k = 1, \dots, n\}$. Now we show that the operator G_s maps the set $V^\varepsilon \cap \widehat{\Omega}$ into itself when ε is rather small. Denote $C_1 = \max\{\theta^*_z : z \in V^1 \cap \Omega\}$, $C_2 = \min\{\theta^*_z : z \in \Omega, z \notin V^{1/3}\} > 0$. Then for any $z \in V^\varepsilon \backslash V^{\varepsilon/3}$ we get $C_2\,\varepsilon \le \theta^*_z \le C_1\varepsilon$. Due to (2.5), one has $|\rho_k(D(x))| \le C(\theta^*_x)^{r_k+1}$, $C > 0$, as $\theta^*_x < T_0$. Now, we fix $\varepsilon \in (0, T_0/C_1)$ such that $\varepsilon \le \min_{1\le k\le n}\{\frac{\delta C_2^{r_k}}{2CC_1^{r_k+1}}, \frac{1}{2CC_1^{r_k+1}}\}$ and consider any $x \in V^\varepsilon \cap \widehat{\Omega}$. Then $|\rho_k(D(x))| \le C(C_1\varepsilon)^{r_k+1} \le \frac{1}{2}\varepsilon^{r_k}$. Therefore, for $s \in \frac{1}{2}V^\varepsilon$ we get $|s_k + \rho_k(D(x))| \le \varepsilon^{r_k}$; hence $G_s(x) \in V^\varepsilon$. Now let us give conditions under which $G_s(x) \in \widehat{\Omega}$. Namely, choose $s \in \Omega\backslash V^{\varepsilon/3}$; then $|(H^{-1}_{\theta^*_s}(\rho(D(x))))_k| \le \frac{C(\theta^*_x)^{r_k+1}}{(\theta^*_s)^{r_k}} \le \frac{C(C_1\varepsilon)^{r_k+1}}{(C_2\varepsilon)^{r_k}} \le \delta/2$. Hence we get $H^{-1}_{\theta^*_s}(s - \rho(D(x))) \in \omega(\delta) + U_{\delta/2}$, so $G_s(x) = s - \rho(D(x)) \in \widehat{\Omega}$.

Thus we claim that if ε is rather small and $s \in (\frac{1}{2}V^\varepsilon \backslash V^{\varepsilon/3}) \cap \Omega$, then the operator $G_s = G_s(x)$ is continuous and maps the closed set $V^\varepsilon \cap \widehat{\Omega}$ into itself. Hence this operator has a stationary point $s^1 \in V^\varepsilon \cap \widehat{\Omega}$, i.e., $G_s(s^1) = s^1$. That means that $s = \Xi(\theta^*_{s^1}, u^*_{s^1}) + \rho(\theta^*_{s^1}, u^*_{s^1})$. In particular, this yields that $u^*_{s^1}(t) \in U^{a,b}_{\Phi(s)}(\theta^*_{s^1})$; hence, due to [7, 8], there exists the solution of the time-optimal problem for the system $\{a, b\}$, that is, $s = \Xi(\theta_{\Phi(s)}, u) + \rho(\theta_{\Phi(s)}, u)$, $u \in U^{a,b}_{\Phi(s)}(\theta_{\Phi(s)}) \ne \emptyset$. Therefore,

$$\theta_{\Phi(s)} \le \theta^*_{s^1}. \tag{3.1}$$

Note that $s^1 \to 0$ as $s \to 0$. On the other hand, denote $s^0 = s - \rho(\theta_{\Phi(s)}, u)$, where $u \in U^{a,b}_{\Phi(s)}(\theta_{\Phi(s)})$. Then due to (3.1) for s indicated above we have $|(H^{-1}_{\theta^*_s}(\rho(\theta_{\Phi(s)}, u)))_k| \le$

$\delta/2$; hence $s^0 \in \Omega$, and, therefore, the equality $s^0 = \Xi(\theta_{\Phi(s)}, u)$ gives

$$\theta^*_{s^0} \le \theta_{\Phi(s)}. \tag{3.2}$$

In addition, by the arguments analogous to [24], one obtains

$$\frac{\theta^*_s}{\theta^*_{s^1}} \to 1, \qquad \frac{\theta^*_{s_0}}{\theta^*_{s^1}} \to 1 \qquad \text{as} \quad s \to 0,\ s \in \Omega(\delta). \tag{3.3}$$

Denote $\tilde\theta_s = \theta^*_{s^1}$, $\tilde u_s = u^*_{s^1} \in U^{a,b}_{\Phi(s)}(\tilde\theta_s)$, $s \in \Omega(\delta)$. Then (3.1)–(3.3) give (2.8).

We prove (2.9) now. Let $s_p \to 0$, $s_p \in \Omega(\delta)$. Then there exists a subsequence s_{p_q} such that the sequences $u^*_{s_{p_q}}(t\theta^*_{s_{p_q}}), \tilde u_{s_{p_q}}(t\tilde\theta_{s_{p_q}}) \in K$ strongly converge to $v^*(t)$ and $\tilde v(t)$, respectively. Since $s_{p_q} = \Xi(\theta^*_{s_{p_q}}, u^*_{s_{p_q}}) = \Xi(\tilde\theta_{s_{p_q}}, \tilde u_{s_{p_q}}) + \rho(\tilde\theta_{s_{p_q}}, \tilde u_{s_{p_q}})$, we get

$$\Xi_k(1, u^*_{s_{p_q}}(t\theta^*_{s_{p_q}})) = \left(\frac{\tilde\theta_{s_{p_q}}}{\theta^*_{s_{p_q}}}\right)^{r_k} \Xi_k(1, \tilde u_{s_{p_q}}(t\tilde\theta_{s_{p_q}})) + \theta^*_{s_{p_q}} \left(\frac{\rho_k(\tilde\theta_{s_{p_q}}, \tilde u_{s_{p_q}})}{(\theta^*_{s_{p_q}})^{r_k+1}}\right),$$

$k = 1, \dots, n$. Note that $\Xi(1, u^*_{s_{p_q}}(t\theta^*_{s_{p_q}})) \in \widehat\Omega$. Hence, by use of (2.8) and putting $q \to \infty$, we get $\Xi(1, v^*) = \Xi(1, \tilde v) \in \Omega$, which gives $v^*(t) = \tilde v(t)$. We denote $v = v^* = \tilde v$ and then $\int_0^1 |u^*_{s_{p_q}}(t\theta^*_{s_{p_q}}) - v(t)|dt \to 0$ and $\int_0^1 |\tilde u_{s_{p_q}}(t\tilde\theta_{s_{p_q}}) - v(t)|dt \to 0$. However, it is easy to see that $\int_0^1 |v(t) - v(t\theta^*_{s_{p_q}}/\tilde\theta_{s_{p_q}})|dt \to 0$ as $q \to \infty$. (We continue $v(t)$ by zero on $[1, \theta^*_{s_{p_q}}/\tilde\theta_{s_{p_q}}]$ if $\tilde\theta_{s_{p_q}} < \theta^*_{s_{p_q}}$.) Finally, $\frac{1}{\theta}\int_0^\theta |u^*_{s_{p_q}}(t) - \tilde u_{s_{p_q}}(t)|dt \to 0$, where $\theta = \min\{\theta^*_{s_{p_q}}, \tilde\theta_{s_{p_q}}\}$. Since for any sequence $s_p \to 0$, $s_p \in \Omega(\delta)$, there exists such a subsequence s_{p_q}, we get (2.9). □

Note that for the class K of the bang-bang functions condition (C) holds. Hence we obtain the following corollary.

COROLLARY 3.1. *In the case when K contains only the bang-bang functions, the statement of Theorem* 2.4 *holds if* (i) *and* (ii) *are valid. In particular, it occurs when Ξ contains linear moments only.*

Remark. If the moment min-problem (2.11) corresponds to the time-optimal control problem for a certain system $\{a^*, b^*\}$ (such a construction is given by Theorem 2.6), it is sufficient to check that the condition (C) holds for the set of controls which satisfy the maximum principle. The class of such controls can be described constructively in a number of cases.

If the system $\{a^*, b^*\}$ is autonomous and the origin belongs to the interior of its controllability set, then the function θ^*_s is continuous in a neighborhood of the origin [14].

If there exists a class K_1 which satisfies condition (C) and includes the time-optimal controls for both systems $\{a, b\}$ and $\{a^*, b^*\}$ and conditions (i), (ii) are satisfied, then one can choose $\tilde\theta_s = \theta_{\Phi(s)}$, $\tilde u_s = u_{\Phi(s)}$. As it follows from [24], if Ξ contains only linear moments, then one can choose $\tilde\theta_s = \theta_{\Phi(s)}$, $\tilde u_s = u_{\Phi(s)}$ as well.

4. Generalization of Ree's theorem. Consider now the algebra $\mathcal{A}$ of nonlinear power moments introduced in section 2. Observe that Ree's theorem (our Theorem 2.5) is equivalent to the following theorem.

THEOREM 4.1 (decomposition theorem).

(i) *The algebra $\mathcal{A}$ admits the following orthogonal decomposition:*

$$\mathcal{A} = L \oplus L^{\mathrm{sh}}.$$

(ii) *Let* $\{\ell_j\}_{j=1}^{\infty}$ *be a basis of* L. *Then the elements*

$$\{\ell_{j_1} * \cdots * \ell_{j_s} : \ s \geq 2, \ \ 1 \leq j_1 \leq \cdots \leq j_s\} \tag{4.1}$$

form a basis of L^{sh}.

Proof. Really, consider any basis element $\xi_{m_1 \ldots m_k} \in \mathcal{A}_m$. Due to Theorem 2.5, it admits the (orthogonal) decomposition $\xi_{m_1 \ldots m_k} = \ell_1 + \sum x'_{1j} * x''_{1j}$, where $\ell_1 \in L$ and x'_{1j}, x''_{1j} are basis elements (up to the constant coefficients). However, it is easy to see that $\ell_1 \in \mathrm{Lin}\{\xi_{j_1 \ldots j_k} : \{j_i\}_{i=1}^k = \{m_i\}_{i=1}^k\}$; hence $\ell_1 \in \mathcal{A}_m$, and therefore, $\mathrm{ord}(x'_{1j}) \leq m-1$, $\mathrm{ord}(x''_{1j}) \leq m-1$. Then, decomposing elements x'_{1j}, x''_{1j}, we get $\xi_{m_1 \ldots m_k} = \ell_1 + \sum \ell'_{2j} * \ell''_{2j} + \sum x'_{2j} * x''_{2j} * x'''_{2j}$, where $\ell'_{2j}, \ell''_{2j} \in L$ and $\mathrm{ord}(x'_{2j}) \leq m-2$, $\mathrm{ord}(x''_{2j}) \leq m-2$, $\mathrm{ord}(x'''_{2j}) \leq m-2$, and so on. After m such steps, we decompose $\xi_{m_1 \ldots m_k}$ in the linear combination of $\ell_1 \in L$ and elements of (4.1).

On the other hand, consider any subspace $\mathcal{A}_m$, and note that due to the theorem of Birkhoff and Witt concerning the basis of the associative algebra, the number of all such elements from (4.1) that $\sum \mathrm{ord}(\ell_{j_k}) = m$ equals $\dim(\mathcal{A}_m) - \dim(L \cap \mathcal{A}_m)$. ∎

Fix N linearly independent *Lie elements* $p_1, \ldots, p_N \in L$, each of which has an order, and denote by J the right ideal generated by them,

$$J = \mathrm{Lin}\,\{p_j x : \ 1 \leq j \leq N, \ x \in \mathcal{A} + \mathbb{R}\}, \quad p_j \in L \cap \mathcal{A}_{\mathrm{ord}(p_j)}, j = 1, \ldots, N. \tag{4.2}$$

Recall that we denote by $\widetilde{x}$ the projection of $x \in \mathcal{A}$ onto $J^{\perp}$ and by $\widetilde{L}$ the projection of L onto $J^{\perp}$. We observe the properties of $J^{\perp}$ in the following three lemmas.

LEMMA 4.2. *Let* J *be a right ideal of the form* (4.2). *Then*

(i) *it is represented as* $J = \sum_{m=1}^{\infty}(J \cap \mathcal{A}_m)$; *therefore,* $J^{\perp} = \sum_{m=1}^{\infty}(J^{\perp} \cap \mathcal{A}_m)$ *and* $\mathcal{A} = J \oplus J^{\perp}$;

(ii) *an element* $x \in \mathcal{A}$ *belongs to* $J^{\perp}$ *iff for any* $j = 1, \ldots, N$ *it admits the representation* $x = \sum_{k=1}^{n_j} p'_{kj} x'_{kj} + x''_j$, *where* $p'_{kj} \in \mathcal{A}_{\mathrm{ord}(p_j)}$, $(p'_{kj}, p_j) = 0$, $x'_{kj} \in \mathcal{A} + \mathbb{R}$, $x''_j \in Z_j = \{a \in \mathcal{A} : \ (a, yb) = 0$ *for all* $y \in \mathcal{A}_{\mathrm{ord}(p_j)}, b \in \mathcal{A} + \mathbb{R}\}$.

The proof of the lemma follows from the definitions immediately.

LEMMA 4.3. *Suppose* $a, b \in J^{\perp}$. *Then* $a * b \in J^{\perp}$.

Proof. Without loss of generality, assume $J = \{p\,x : \ x \in \mathcal{A} + \mathbb{R}\}$, where $\mathrm{ord}(p) = d$. We note that

$$\xi_{m_1 \ldots m_n} * \xi_{k_1 \ldots k_s} = \textstyle\sum_{r,q}(\xi_{m_1 \ldots m_r} * \xi_{k_1 \ldots k_q})(\xi_{m_{r+1} \ldots m_n} * \xi_{k_{q+1} \ldots k_s}) + \widetilde{z}, \tag{4.3}$$

where $\widetilde{z} \in Z = \{a \in \mathcal{A} : \ (a, yb) = 0$ for all $y \in \mathcal{A}_d, b \in \mathcal{A} + \mathbb{R}\}$ and the sum is taken over all $r \geq 0$, $q \geq 0$ such that $m_1 + \cdots + m_r + r + k_1 + \cdots + k_q + q = d$. The following cases are possible.

Case 1. $\xi_{m_1 \ldots m_n}, \xi_{k_1 \ldots k_s} \in Z$. Then obviously $r \geq 1$ and $q \geq 1$ in the sum in (4.3). In this case $(\xi_{m_1 \ldots m_r} * \xi_{k_1 \ldots k_q}, \ p) = 0$ due to Theorem 2.5; hence $\xi_{m_1 \ldots m_n} * \xi_{k_1 \ldots k_s} \in J^{\perp}$ due to Lemma 4.2.

Case 2. $\xi_{m_1 \ldots m_{r_0}} \in \mathcal{A}_d$ for a certain $1 \leq r_0 \leq n$, and $\xi_{k_1 \ldots k_s} \in Z$. Then $r \geq 1$ and q can equal 0 in (4.3); hence $\xi_{m_1 \ldots m_n} * \xi_{k_1 \ldots k_s} = \xi_{m_1 \ldots m_{r_0}} y + z'$, where $y \in \mathcal{A} + \mathbb{R}$ and $z' \in J^{\perp}$ as in Case 1.

Case 3. $\xi_{m_1 \ldots m_{r_0}} \in \mathcal{A}_d$ and $\xi_{k_1 \ldots k_{q_0}} \in \mathcal{A}_d$ for certain $1 \leq r_0 \leq n$ and $1 \leq q_0 \leq s$. Then r and q can equal 0 in (4.3); hence $\xi_{m_1 \ldots m_n} * \xi_{k_1 \ldots k_s} = \xi_{m_1 \ldots m_{r_0}} y_1 + \xi_{k_1 \ldots k_{q_0}} y_2 + z''$, where $y_1, y_2 \in \mathcal{A} + \mathbb{R}$, $z'' \in J^{\perp}$.

Suppose now $a, b \in J^{\perp}$; hence, due to Lemma 4.2, $a = \sum_{k=1}^{n_p} p'_k a'_k + a''$, $b = \sum_{k=1}^{n_p} p'_k b'_k + b''$, where $p'_k \in \mathcal{A}_d$, $(p'_k, p) = 0$, $a'_k, b'_k \in \mathcal{A} + \mathbb{R}$, $a'', b'' \in Z$. Then from

Cases 1–3 it follows that $a * b = \sum_{k=1}^{n_p} p'_k y'_k + z$, where $y'_k \in \mathcal{A} + \mathbb{R}$, $z \in J^\perp$, which yields $a * b \in J^\perp$ due to Lemma 4.2. □

LEMMA 4.4. *Denote $L_p = L \cap J$. Then*

(i) *the following direct decomposition is valid:*

$$\mathcal{A} = (L_p \oplus \widetilde{L}) \dot{+} (L_p \oplus \widetilde{L})^{\mathrm{sh}};$$

(ii) *if $B_p = \{b_m\}_{m=-\infty}^{-1}$ is a basis of L_p ($M = 0, 1$, or ∞, if $N = 0, 1$ or ≥ 2, respectively) and $B = \{b_j\}_{j=1}^{\infty}$ complements B_p to the basis of L, then the elements*

$$\begin{gathered} \{b_{m_1} * \cdots * b_{m_n} * \widetilde{b}_{j_1} * \cdots * \widetilde{b}_{j_s} \ : \ n, s \geq 0, n + s \geq 1, \\ m_1 \leq \cdots \leq m_n < 0 < j_1 \leq \cdots \leq j_s\} \end{gathered} \tag{4.4}$$

form a basis of $\mathcal{A}$.

Proof. Note that L_p equals the Lie subalgebra generated by $\{p_1, \ldots, p_N\}$ (the linear span of all Lie brackets including $p_1, \ldots, p_N$ only). Obviously, we can choose the basis B_p so that its elements have an order. Without loss of generality, we assume that elements of B also have an order as well as elements of $\widetilde{B}$.

In view of Theorem 4.1, it is sufficient to prove that any element $\ell \in L$ having an order can be represented as a linear combination of elements from (4.4) in the unique way. Since $\mathcal{A}_1 = \mathrm{Lin}\,\{\xi_0\}$, for elements from $\mathcal{A}_1$ the mentioned fact is trivial. Assume it is valid for any $\ell' \in L$ such that $\mathrm{ord}(\ell') < m$, and consider an arbitrary element $\ell \in L \cap \mathcal{A}_m$. Then due to Lemma 4.2,

$$\ell = \widetilde{\ell} + x, \tag{4.5}$$

where $\widetilde{\ell} \in \widetilde{L} \cap \mathcal{A}_m$ and $x \in J \cap \mathcal{A}_m$. Further, due to Theorem 4.1,

$$x = \ell^* + y, \tag{4.6}$$

where $\ell^* \in L \cap \mathcal{A}_m$, $y \in L^{\mathrm{sh}} \cap \mathcal{A}_m$ are defined uniquely. Thus

$$\ell - \ell^* = \widetilde{\ell} + y. \tag{4.7}$$

The condition $y \in L^{\mathrm{sh}} \cap \mathcal{A}_m$ means that y is a linear combination of elements of the form $\ell_1 * \cdots * \ell_s$, $s \geq 2$, $\ell_i \in L$, and $\mathrm{ord}(\ell_1) + \cdots + \mathrm{ord}(\ell_s) = m$. Hence $\mathrm{ord}(\ell_i) < m$, and due to the induction assumption the right part of (4.7) is represented as a linear combination of elements of (4.4).

On the other hand, formulas (4.5) and (4.6) associate to any basis element $b_i \in (B_p \cup B) \cap \mathcal{A}_m$ the unique element $b_i - b_i^* \in L \cap \mathcal{A}_m$ which equals 0 iff $b_i \in B_p$. Suppose $B_p \cap \mathcal{A}_m = \{b_{m_i}\}_{i=1}^q$ and $B \cap \mathcal{A}_m = \{b_{j_i}\}_{i=1}^r$, and consider the subset of $L \cap \mathcal{A}_m$

$$\{b_{m_i}\}_{i=1}^q \cup \{b_{j_i} - b_{j_i}^*\}_{i=1}^r. \tag{4.8}$$

Note that the number of elements in this set is equal to $\dim(L \cap \mathcal{A}_m)$. Let us show that these elements are linearly independent. Suppose the contrary; then

$$\sum_{i=1}^r \mu_i (b_{j_i} - b_{j_i}^*) \in L_p = L \cap J,$$

where $\sum_{i=1}^r \mu_i^2 > 0$. Since $b_{j_i} - b_{j_i}^* = \widetilde{b}_{j_i} + y_i$, where $\widetilde{b}_{j_i} \in J^\perp \subset L_p^\perp$, $y_i \in L^\perp \subset L_p^\perp$,

$$\sum_{i=1}^r \mu_i (b_{j_i} - b_{j_i}^*) = \sum_{i=1}^r \mu_i \widetilde{b}_{j_i} + \sum_{i=1}^r \mu_i y_i = 0.$$

This yields that $\sum_{i=1}^{r}\mu_i\widetilde{b}_{j_i}\in L^{\perp}$. In particular, the element $\sum_{i=1}^{r}\mu_i b_{j_i}\in L$ is orthogonal to its projection $\sum_{i=1}^{r}\mu_i\widetilde{b}_{j_i}$ on $J^{\perp}$. Hence $\sum_{i=1}^{r}\mu_i b_{j_i}\in J\cap L=L_p$, which gives $\mu_1=\cdots=\mu_r=0$. So we have that (4.8) is a basis of $L\cap\mathcal{A}_m$, and due to (4.7) all elements of this basis are represented as certain linear combinations of elements of (4.4), which proves the lemma. □

The following theorem is based on Lemmas 4.2–4.4 and generalizes Theorem 4.1.

THEOREM 4.5 (generalized decomposition theorem).

(i) *The following orthogonal decomposition is valid:*

$$J^{\perp}=\widetilde{L}\oplus\widetilde{L}^{\mathrm{sh}}.$$

(ii) *Let $B=\{b_j\}_{j=1}^{\infty}$ complement a basis of $L_p=L\cap J$ to a basis of L. Then the elements*

$$\{\widetilde{b}_{j_1}*\cdots*\widetilde{b}_{j_n}:\ n\geq 2,\quad 0<j_1\leq\cdots\leq j_n\}$$

form a basis of $\widetilde{L}^{\mathrm{sh}}$.

Proof. Assume $B_p=\{b_m\}_{m=-M}^{-1}$ is a basis of L_p and $B=\{b_j\}_{j=1}^{\infty}$ complements it to a basis of L. Let the elements B_p have an order as well as the elements of B. Then $\mathrm{ord}(\widetilde{b}_j)=\mathrm{ord}(b_j)$, $j>0$. Due to Lemma 4.4, $\dim(\mathcal{A}_m)$ equals the number of elements in the set

$$\left\{\begin{array}{l} b_{m_1}*\cdots*b_{m_n}*\widetilde{b}_{j_1}*\cdots*\widetilde{b}_{j_s}:n,s\geq 0,\ n+s\geq 1,\\ m_1\leq\cdots\leq m_n<0<j_1\leq\cdots\leq j_s,\sum_{i=1}^{n}\mathrm{ord}(b_{m_i})+\sum_{i=1}^{s}\mathrm{ord}(b_{j_i})=m\end{array}\right\}. \tag{4.9}$$

Further, by use of the theorem of Birkhoff and Witt on the basis of the associative algebra one can prove that $\dim(J\cap\mathcal{A}_m)$ equals the number of elements of (4.9) such that $n\geq 1$. Hence (see Lemma 4.2) $\dim(J^{\perp}\cap\mathcal{A}_m)$ equals the number of elements of (4.9) such that $n=0$,

$$\left\{\widetilde{b}_{j_1}*\cdots*\widetilde{b}_{j_s}:\ s\geq 1,\ 0<j_1\leq\cdots\leq j_s,\sum_{i=1}^{s}\mathrm{ord}(b_{j_i})=m\right\}. \tag{4.10}$$

On the other hand, elements (4.10) are linearly independent (see Lemma 4.4) and belong to $J^{\perp}\cap\mathcal{A}_m$ (see Lemma 4.3); hence, they form a basis of $J^{\perp}\cap\mathcal{A}_m$, which proves (ii) and the fact that $J^{\perp}=\widetilde{L}\dot{+}\widetilde{L}^{\mathrm{sh}}$.

It remains to prove that $\widetilde{L}$ is orthogonal to $\widetilde{L}^{\mathrm{sh}}$. Consider any $\widetilde{b}_j\in\widetilde{B}$; under our construction it is a projection of the element $b_j\in B$ on $J^{\perp}$; that is (see Lemma 4.2), $\widetilde{b}_j=b_j+x$, $x\in J$. Since $b_j\in L$, then b_j is orthogonal to $\widetilde{L}^{\mathrm{sh}}$ (see Theorem 2.5). At the same time, $\widetilde{L}^{\mathrm{sh}}\subset J^{\perp}$ (see Lemma 4.3); hence $x\in(\widetilde{L}^{\mathrm{sh}})^{\perp}$. Thus $\widetilde{b}_j\in(\widetilde{L}^{\mathrm{sh}})^{\perp}$, which completes the proof of the theorem. □

5. Proof of Theorem 2.6. (i) Without loss of generality we assume that $\{\widetilde{\ell}_i\}_{i=1}^{n}$ is an orthonormal basis of $\sum_{r=1}^{r_{a,b}}(\widetilde{L}\cap\mathcal{A}_r)$. We note that $\mathrm{ord}(\widetilde{\ell}_q)=r$ for any q such that $d_{r-1}+1\leq q\leq d_r$, $q=1,\ldots,n$, $r=1,\ldots,r_{a,b}$.

We construct the required transformation F by $r_{a,b}$ steps.

Let $S^{(0)} = S_{a,b}$ and $F^{(0)}$ be the identical map. Suppose after $r-1$ steps, where $1 \le r \le r_{a,b}$, the series $S_{a,b}$ has been transformed to the form $S^{(r-1)} = F^{(r-1)}(S_{a,b})$,

$$\begin{aligned} S_1^{(r-1)} &= \widetilde{\ell}_1 + \rho_1 \\ \cdots &\quad \cdots \\ S_{d_{r-1}}^{(r-1)} &= \widetilde{\ell}_{d_{r-1}} + \rho_{d_{r-1}}, \\ S_{d_{r-1}+1}^{(r-1)} &= \rho_{d_{r-1}+1}^{(r-1)} \\ \cdots &\quad \cdots \\ S_n^{(r-1)} &= \rho_n^{(r-1)}, \end{aligned}$$

where $\rho_1, \dots, \rho_{d_{r-1}}$ are of the form (2.17), and

$$\rho_q^{(r-1)} = \sum_{m=r}^{\infty} \sum_{\substack{m_1+\cdots+m_k+k=m \\ k\ge 1, m_j \ge 0}} (v_{m_1\dots m_k}^{(r-1)})_q \xi_{m_1\dots m_k}, \qquad q = d_{r-1}+1, \dots, n.$$

Here $(v)_q$ denotes the qth component of the vector v and $v_{m_1\dots m_k}^{(r-1)}$ are of the form (2.4) with $a = a^{(r-1)}$, $b = b^{(r-1)}$, where vector fields $a^{(r-1)}(t,z)$, $b^{(r-1)}(t,z)$ are such that the system $\dot z = a^{(r-1)}(t,z) + ub^{(r-1)}(t,z)$ is obtained from the system $\{a,b\}$ after substitution of the variables $z = F^{(r-1)}(x)$. Represent $\rho_q^{(r-1)} = \varphi_q + \hat\rho_q$, where

$$\varphi_q = \sum_{\substack{m_1+\cdots+m_k+k=r \\ k\ge 1, m_j \ge 0}} (v_{m_1\dots m_k}^{(r-1)})_q \xi_{m_1\dots m_k}, \qquad q = d_{r-1}+1, \dots, n.$$

Due to Theorem 4.5, the subspace $\mathcal{A}_r$ admits the orthogonal decomposition $\mathcal{A}_r = (J \cap \mathcal{A}_r) \oplus (\widetilde{L} \cap \mathcal{A}_r) \oplus (\widetilde{L}^{\mathrm{sh}} \cap \mathcal{A}_r)$. Let $\{x_i\}$ and $\{y_i\}$ be the orthonormal basis of $J \cap \mathcal{A}_r$ and $\widetilde{L}^{\mathrm{sh}} \cap \mathcal{A}_r$, respectively. Then

$$\varphi_q = \sum_i \alpha_{qi} x_i + \sum_{j=d_{r-1}+1}^{d_r} \beta_{qj} \widetilde{\ell}_j + \sum_i \gamma_{qi} y_i,$$

where $\alpha_{qi} = \big(v^{(r-1)}(x_i)\big)_q$, $\beta_j = \big(v^{(r-1)}(\widetilde{\ell}_j)\big)_q$, $\gamma_{qi} = \big(v^{(r-1)}(y_i)\big)_q$. It follows from (2.15) that $v^{(r-1)}(x_i) \in v^{(r-1)}\big(\sum_{k=1}^{r-1} \mathcal{A}_k\big) = \operatorname{Lin}\{e_k\}_{k=1}^{d_{r-1}}$ (e_k is the unit vector with 1 on the kth place), which yields $\alpha_{qi} = 0$, $d_{r-1}+1 \le q \le n$. Further, due to (2.15), one has $v^{(r-1)}(\widetilde{\ell}_j) - v^{(r-1)}(\ell_j) = v^{(r-1)}(\widetilde{\ell}_j - \ell_j) \in \operatorname{Lin}\{e_k\}_{k=1}^{d_{r-1}}$; hence $\beta_{qj} = (v^{(r-1)}(\ell_j))_q$, $d_{r-1}+1 \le j \le d_r$, $d_{r-1}+1 \le q \le n$. Finally, $\sum_i \gamma_{qi} y_i$ equals the linear combination of elements of the form $\widetilde{\ell}_{i_1} * \cdots * \widetilde{\ell}_{i_k}$, $k \ge 2$, where $\sum_{j=1}^{k} \operatorname{ord}(\widetilde{\ell}_{i_j}) = r$; hence $\operatorname{ord}(\widetilde{\ell}_{i_j}) < r$. Therefore, there exist polynomials $P_q = P_q(x_1, \dots, x_{d_{r-1}})$ such that $P_q(\widetilde{\ell}_1, \dots, \widetilde{\ell}_{d_{r-1}}) = \sum_i \gamma_{qi} y_i$, $q = d_{r-1}+1, \dots, n$.

Define $\widehat{F}^{(r)}(x)_q = x_q - P_q(x_1, \dots, x_{d_{r-1}})$, $q = d_{r-1}+1, \dots, n$; then

$$\widehat{F}^{(r)}(S^{(r-1)})_q = S_q^{(r-1)} - P_q\big(S_1^{(r-1)}, \dots, S_{d_{r-1}}^{(r-1)}\big) = \sum_{j=d_{r-1}+1}^{d_r} (v^{(r-1)}(\ell_j))_q \widetilde{\ell}_j + \widetilde{\rho}_q,$$

where $\widetilde{\rho}_q$ contains terms of order greater than r. Complete the definition of $\widehat{F}^{(r)}$, putting $\widehat{F}^{(r)}(x)_q = x_q$, $q = 1, \dots, d_{r-1}$. Since $\{v^{(r-1)}(\ell_j)\}_{j=d_{r-1}+1}^{d_r}$ are linearly independent, we can find a nonsingular matrix $H^{(r)}$ such that $H^{(r)} e_j = e_j$, $j = 1, \dots, d_{r-1}$,

and $H^{(r)}v^{(r-1)}(\ell_j) = e_j$, $j = d_{r-1}+1, \dots, d_r$. Choosing finally $F^{(r)} = H^{(r)}\widehat{F}^{(r)}$, we obtain that the series $S^{(r)} = F^{(r)}(S^{(r-1)})$ is of the form

$$\begin{aligned} S_1^{(r)} &= \widetilde{\ell}_1 + \rho_1 \\ \dots &\quad \dots \\ S_{d_r}^{(r)} &= \widetilde{\ell}_{d_r} + \rho_{d_r}, \\ S_{d_r+1}^{(r)} &= \rho_{d_r+1}^{(r)} \\ \dots &\quad \dots \\ S_n^{(r)} &= \rho_n^{(r)}, \end{aligned}$$

where $\rho_1, \dots, \rho_{d_r}$ are of the form (2.17), and

$$\rho_q^{(r)} = \sum_{m=r+1}^{\infty} \sum_{\substack{m_1+\dots+m_k+k=m \\ k\ge 1, m_j\ge 0}} (v_{m_1\dots m_k}^{(r)})_q \xi_{m_1\dots m_k}, \qquad q = d_r+1, \dots, n.$$

Obviously, the series $F(S_{a,b}) = F^{(r_{a,b})}\cdots F^{(1)}(S_{a,b})$ constructed after $r_{a,b}$ steps has the form (2.16).

(ii) We construct *a time-optimal canonical approximation* $\{a^*, b^*\} \in \mathcal{U}$ such that $a^*(t,x) \equiv 0$. Consider the linear span of integrals

$$\xi_{m_1\dots m_k}(t) = \xi_{m_1\dots m_k}(t,\theta,u) = \int_t^\theta \int_t^{\tau_1} \cdots \int_t^{\tau_{k-1}} \tau_1^{m_1}\tau_2^{m_2}\cdots\tau_k^{m_k} \prod_{j=1}^{k} u(\tau_j) d\tau_k \cdots d\tau_2 d\tau_1.$$

It also may be considered as the realization of $\mathcal{A}$, and the shuffle product in $\mathcal{A}$ also corresponds to the product of $\xi_{m_1\dots m_k}(t)$ as functionals of u. Further, for any element $y = \sum \mu_{m_1\dots m_k}\xi_{m_1\dots m_k} \in \mathcal{A}$, we denote by $y(t)$ the functional $\sum \mu_{m_1\dots m_k}\xi_{m_1\dots m_k}(t)$.

We construct $b^*(t,x)$ by r_0 steps. Let $1 \le r \le r_0$ and the first d_{r-1} components of the vector function $b^*(t,x)$ be chosen already so that the trajectory $x(t)$ of the system $\{a^*, b^*\}$ satisfies $x_q(t) = \widetilde{\ell}_q(t)$, $q = 1, \dots, d_{r-1}$. Then $(S_{a^*,b^*})_q = x_q(0) = \widetilde{\ell}_q$, $q = 1, \dots, d_{r-1}$. On the rth step let us construct the $d_{r-1}+1, \dots, d_r$th components of $b^*(t,x)$ to satisfy the equalities $(S_{a^*,b^*})_q = x_q(0) = \widetilde{\ell}_q$, $q = d_{r-1}+1, \dots, d_r$.

Consider the elements $\widetilde{\ell}_q$, $q = d_{r-1}+1, \dots, d_r$. Since $\widetilde{\ell}_q \in \mathcal{A}_r$, they allow the representation $\widetilde{\ell}_q = \sum_{m=0}^{r-2} y_m^q \xi_m + \alpha^q \xi_{r-1}$, where $\alpha^q \in \mathbb{R}$ and $y_m^q \in \mathcal{A}_{r-m-1}$. Let us show that $y_m^q \in J^\perp$. Really, for any $x \in J \cap \mathcal{A}_{r-m-1}$, we have $(y_m^q, x) = (y_m^q\xi_m, x\xi_m) = (\widetilde{\ell}_q, x\xi_m) = 0$ since $\widetilde{\ell}_q \in J^\perp$. Due to Theorem 4.5 and since $\{\widetilde{\ell}_j\}_{j=1}^{d_{r-1}}$ form a basis of $\widetilde{L} \cap (\mathcal{A}_1 + \dots + \mathcal{A}_{r-1})$, we obtain $y_m^q = P_m^q(\widetilde{\ell}_1, \dots, \widetilde{\ell}_{d_{r-1}})$, where P_m^q are polynomials. By our assumptions, $y_m^q(t) = P_m^q(x_1(t), \dots, x_{d_{r-1}}(t))$.

Put $(b^*(t,x))_q = -\sum_{m=0}^{r-2} t^m P_m^q(x_1, \dots, x_{d_{r-1}}) - \alpha^q t^{r-1}$, $q = d_{r-1}+1, \dots, d_r$; then

$$\begin{aligned} x_q(t) = -\int_t^\theta u(\tau)(b^*(\tau, x(\tau)))_q d\tau &= \int_t^\theta u(\tau) \sum_{m=0}^{r-2} \tau^m P_m^q(x_1(\tau), \dots, x_{d_{r-1}}(\tau)) d\tau \\ &\quad + \alpha^q \int_t^\theta \tau^{r-1} u(\tau) d\tau = \sum_{m=0}^{r-2} \int_t^\theta \tau^m u(\tau) y_m^q(\tau) d\tau + \alpha^q \xi_{r-1}(t). \end{aligned}$$

It follows from the obvious equality $\int_t^\theta \tau^m u(\tau)\xi_{j_1\dots j_k}(\tau)d\tau = \xi_{j_1\dots j_k m}(t)$ that $x_q(t) = \sum_{m=0}^{r-2}(y_m^q\xi_m)(t) + \alpha^q\xi_{r-1}(t) = \widetilde{\ell}_q(t)$. Hence $(S_{a^*,b^*})_q = x_q(0) = \widetilde{\ell}_q$, $q = d_{r-1}+1, \dots, d_r$, and after r_0 steps we construct the required system. ∎

6. Example of a series and its canonical form. Consider the system of the Euler equations for a spacecraft [21, Example 3.24.]:

$$\dot\omega_1 = \omega_2\omega_3 + u, \qquad \dot\omega_2 = -\omega_1\omega_3 + u, \qquad \dot\omega_3 = u. \tag{6.1}$$

One has

$$a(\omega) = \begin{pmatrix} \omega_2\omega_3 \\ -\omega_1\omega_3 \\ 0 \end{pmatrix}, \qquad b(\omega) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix};$$

hence $a(0) = 0$. Denote $\delta_m = (-1)^{[\frac{m}{2}]}$. It is easy to see that

$$\mathrm{ad}_{R_a} R_b E(\omega) = \begin{pmatrix} -\omega_2 - \omega_3 \\ \omega_1 + \omega_3 \\ 0 \end{pmatrix}, \qquad \mathrm{ad}^m_{R_a} R_b E(\omega) = \begin{pmatrix} \delta_{m+1}\omega_3^m \\ \delta_m\omega_3^m \\ 0 \end{pmatrix}, \quad m \ge 2.$$

Hence $\mathrm{ad}^{m_1}_{R_a} R_b \circ \mathrm{ad}^{m_2}_{R_a} R_b D\omega = 0$ as $m_1 \ge 1$ and $m_2 \ge 2$ for any matrix D. That means that all the coefficients of the series (2.2) equal zero except $v_{0\dots0i1\dots1}$. Further,

$$(R_b)^q \circ \mathrm{ad}^i_{R_a} R_b \circ (\mathrm{ad}_{R_a} R_b)^j E(\omega) = \begin{pmatrix} \frac{i!}{(i-q)!}\delta_{i+j+1}\omega_3^{i-q} \\ \frac{i!}{(i-q)!}\delta_{i+j}\omega_3^{i-q} \\ 0 \end{pmatrix}, \quad i \ge q;$$

that is, $v_{\underbrace{0\dots0}_{q} i \underbrace{1\dots1}_{j}} = 0$ as $i \ne q$. Hence all nonzero coefficients are as follows:

$$v_0, v_{\underbrace{01\dots1}_{j}} = (-1)^{j+1}\begin{pmatrix} 2\delta_{j+1} \\ 2\delta_j \\ 0 \end{pmatrix}, \quad v_{\underbrace{0\dots0}_{i} i \underbrace{1\dots1}_{j}} = (-1)^{i+j+1}\begin{pmatrix} \delta_{i+j+1} \\ \delta_{i+j} \\ 0 \end{pmatrix}, \qquad i > 1.$$

That implies that series (2.2) corresponding to system (6.1) is of the form

$$S_{a,b} = -\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\xi_0 + \sum_{N=1}^{\infty}(-1)^{N+1}\begin{pmatrix} \delta_{N+1} \\ \delta_N \\ 0 \end{pmatrix}\left(\sum_{j=1}^{N}\xi_{\underbrace{0\dots0}_{j} j \underbrace{1\dots1}_{N-j}} + \xi_{\underbrace{01\dots1}_{N}}\right).$$

Further, let us separate the first terms of the series,

$$S_{a,b} = -\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\xi_0 + 2\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}\xi_{01} + \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}(2\xi_{011} + \xi_{002}) + \rho,$$

where ρ includes terms of order greater than 5. Then

$$\begin{aligned}
&D^1_{a,b} = \mathrm{Lin}\,\{(-1,-1,-1)\},\ d_1 = 1; \qquad D^2_{a,b} = D^1_{a,b},\ d_2 = 1, \quad p^2_1 = \xi_1;\\
&D^3_{a,b} = D^1_{a,b} + \mathrm{Lin}\,\{(-1,1,0)\},\ d_3 = 2, \quad p^3_1 = \xi_2;\\
&D^4_{a,b} = D^3_{a,b},\ d_4 = 2, \quad p^4_1 = \xi_3, \quad p^4_2 = [\xi_0,\xi_2], \quad p^4_3 = [\xi_0,[\xi_0,\xi_1]];\\
&D^5_{a,b} = D^3_{a,b} + \mathrm{Lin}\,\{(1,1,0)\},\ d_5 = 3,\ r_{a,b} = 5,\\
&\quad p^5_1 = \xi_4, \quad p^5_2 = [\xi_0,\xi_3], \quad p^5_3 = [\xi_0,[\xi_0,[\xi_0,\xi_1]]], \quad p^5_4 = 2[\xi_0,[\xi_0,\xi_2]] - [\xi_1,[\xi_1,\xi_0]].
\end{aligned}$$

Hence the right ideal $J_{a,b}$ for system (6.1) is generated by the elements $p_1^2, p_1^3, \{p_k^4\}_{k=1}^3$, $\{p_k^5\}_{k=1}^4$, and the principal part of the canonical form of the series $S_{a,b}$ equals

$$\Xi = \begin{pmatrix} \widetilde{\ell}_1 \\ \widetilde{\ell}_2 \\ \widetilde{\ell}_3 \end{pmatrix} = \begin{pmatrix} \xi_0 \\ \xi_{01} \\ 2\xi_{011} + \xi_{002} \end{pmatrix},$$

where $\ell_1 = \xi_0$, $\ell_2 = [\xi_0, \xi_1]$, $\ell_3 = [\xi_0, [\xi_0, \xi_2]] + 2[\xi_1, [\xi_1, \xi_0]]$. The system $\{a^*, b^*\}$ (canonical approximation) with the series $S_{a^*,b^*} = \Xi$ constructed by the method given in the proof of Theorem 2.6 is of the form

$$\dot{x}_1 = -u, \quad \dot{x}_2 = -tux_1, \quad \dot{x}_3 = -2tux_2 - \frac{1}{2}t^2ux_1^2.$$

Note that from the results of [25] it follows that there exists the autonomous system corresponding to the series S_{a^*,b^*}, namely,

$$\dot{x}_1 = -u, \quad \dot{x}_2 = -\frac{1}{2}x_1^2, \quad \dot{x}_3 = -x_1x_2.$$

7. Examples of solutions of nonlinear time-optimal control problems.

(A) Consider the system $\{a, b\}$ of the form

$$\dot{x}_1 = u, \ \dot{x}_2 = \frac{1}{2}x_1^2. \tag{7.1}$$

The series $S_{a,b}$ is of the form $S_{a,b} = (-\xi_0, -\xi_{01})$. The controllability set for this system is of the form $D = \{x : \ x_2 \le -\frac{1}{6}|x_1|^3\}$. The maximum principle gives

$$H = u\psi_1 + \frac{1}{2}x_1^2\psi_2, \qquad \dot{\psi}_1 = -x_1\psi_2, \quad \dot{\psi}_2 = 0.$$

Hence the optimal control $u_{x^0}(t)$ can be singular and, therefore, equals $u_{x^0}(t) \equiv 0$ iff $\psi_1 \equiv 0$, $\psi_2 \neq 0$, which yields $x_1 \equiv 0$, $x_2 \equiv \text{const}$. Since the optimal trajectory cannot be stationary, the optimal control is the bang-bang and such that $u_{x^0}(t) = \operatorname{sign}(\psi_1(t))$, where $\ddot{\psi}_1 = -u_{x^0}\psi_2$. Hence, if $\{t_i\}_{i=1}^N$ are the switchings of the optimal controls, then one has $t_{i+1} - t_i = t_i - t_{i-1}$, $i = 1, \dots, N-1$, and $t_1 \le t_2 - t_1$, $\theta_{x^0} - t_N \le t_2 - t_1$. It is easy to check that the optimal control really has no more than one switching and is unique except in the case $x_1^0 = 0$, when there exist two optimal controls.

Further, the optimal time equals $\theta_{x^0} = 2\left(\frac{1}{2}|x_1^0|^3 - 3x_2^0\right)^{1/3} - |x_1^0|^3$. In other words, $\theta_{x^0} = \left|x_1^0\right|(2(3\mu+1)^{1/3} - 1)$ on any curve $x_2^0 = -(\mu + \frac{1}{6})\left|x_1^0\right|^3$, $\mu \ge 0$. Hence the optimal time is continuous in the controllability set. For example, that means that due to Theorem 2.4 the time-optimal problem for system (7.1) approximates the time-optimal problem for the locally controllable system

$$\dot{x}_1 = u, \qquad \dot{x}_2 = \frac{1}{2}x_1^2 + x_1^3$$

in the domains $\Omega_1 = \{x : -\delta_1 x_1^3 < x_2 < -\delta_2 x_1^3, x_1 > 0\}$ and $\Omega_2 = \{x : \delta_1 x_1^3 < x_2 < \delta_2 x_1^3, x_1 < 0\}$ for arbitrary $\delta_1 > \delta_2 > \frac{1}{6}$.

(B) Consider the system $\{a, b\}$ of the form

$$\dot{x}_1 = u, \ \dot{x}_2 = x_1, \ \dot{x}_3 = \frac{1}{2}x_1^2. \tag{7.2}$$

Its series is of the form $S_{a,b} = (-\xi_0, \xi_1, -\xi_{01})$. The canonical form mentioned in Theorem 2.6 equals $F(S_{a,b}) = (\xi_0, \xi_1, \xi_{01} - \xi_{10})$, where $F(z) = (-z_1, z_2, -2z_3 + z_1 z_2)$.

For the sake of convenience, we introduce the following notation.

Notation. Denote by $u(t; a^+)$, $u(t; a^-)$, and $u(t; a^0)$ the constant functions defined on the interval $(0, a)$, and taking values $+1$, -1, and 0, respectively. For $f_i(t)$ defined on the intervals $(0, a_i)$, $i = 1, 2$, let $f = f_1 \circ f_2$ be the function defined on the interval $(0, a_1 + a_2)$ and satisfying the equalities $f(t) = f_1(t)$, $t \in (0, a_1)$, and $f(t) = f_2(t - a_1)$, $t \in (a_1, a_1 + a_2)$. Finally, for $a_1, \ldots, a_m \geq 0$, $p_1, \ldots, p_m \in \{+, -, 0\}$, denote

$$u(\cdot; a_1^{p_1}, \ldots, a_m^{p_m}) = u(\cdot; a_1^{p_1}) \circ \ldots \circ u(\cdot; a_m^{p_m}).$$

We formulate the properties of system (7.2) in the following statements.

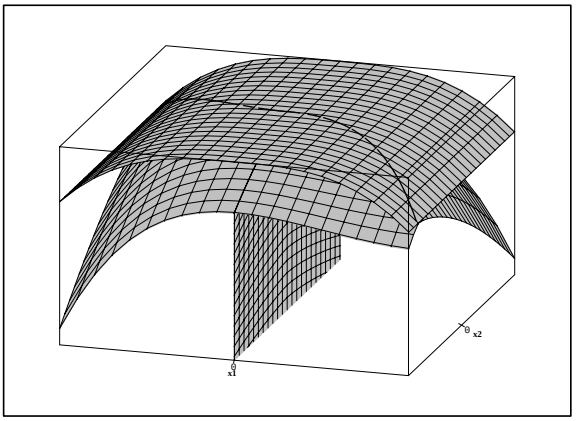


FIG. 7.1. *The upper surface is* $\partial D = \{x : x_3 = -\frac{1}{6}|x_1|^3\}$. *The lower surface is* M. *The vertical plane is* $\{x : x_1 = 0\}$. *The heavily drawn curve is* $M \cap \partial D$.

PROPOSITION 7.1. *The controllability set* D *for the system* (7.2) *equals*

$$D = \left\{x : x_3 < -\frac{1}{6}|x_1|^3\right\} \cup \left\{x = \left(x_1, -\frac{1}{2}x_1|x_1|, -\frac{1}{6}|x_1|^3\right)\right\}.$$

Proof. Really, $\frac{d}{dt}(x_3 \pm \frac{1}{6}x_1^3) = \frac{1}{2}x_1^2(1 \pm u) \geq 0$; hence $x_3^0 \pm \frac{1}{6}(x_1^0)^3 \leq 0$. Moreover, $x_3^0 = \frac{1}{6}(x_1^0)^3$ iff $u \equiv 1$ and $x_3^0 = -\frac{1}{6}(x_1^0)^3$ iff $u \equiv -1$. In these cases $x_2^0 = -\frac{1}{2}x_1^0|x_1^0|$, $x_3^0 = -\frac{1}{6}|x_1^0|^3$. On the other hand, consider the surface

$$M = \left\{x : x_3 = -\frac{1}{6}\sigma x_1^3 - \frac{1}{3}\left(\frac{1}{2}x_1^2 + \sigma x_2\right)^{3/2}\right\},$$

where $\sigma = \operatorname{sign}(x_2 + \frac{1}{2}x_1|x_1|)$, which intersects the ∂D at the points $x = (x_1, -\frac{1}{2}x_1|x_1|, -\frac{1}{6}|x_1|^3)$ (see Figure 7.1). This surface consists of the points from which the origin can be reached by the bang-bang control with no more than one switching. The surface M breaks the set $\{x : x_3 < -\frac{1}{6}|x_1|^3\}$ into two parts:

$$D_0 = \left\{x : -\frac{1}{6}\sigma x_1^3 - \frac{1}{3}\left(\frac{1}{2}x_1^2 + \sigma x_2\right)^{3/2} < x_3 < -\frac{1}{6}|x_1|^3\right\},$$

$$D_1 = \left\{x : x_3 < -\frac{1}{6}\sigma x_1^3 - \frac{1}{3}\left(\frac{1}{2}x_1^2 + \sigma x_2\right)^{3/2}\right\}.$$

Let $x' \in M$ be steered to the origin by the control $u'(t) = u(t; a^+, b^-)$, where $\sigma = -1$, $b = \left(\frac{1}{2}(x_1')^2 - x_2'\right)^{1/2}$, $a = b - x_1'$. Then by the controls $u_\alpha(t) = u(t; a^+, (b+\alpha)^-, (2\alpha)^+, \alpha^-)$ with $\alpha > 0$, we can achieve the origin from the points of the form $(x_1', x_2', x_3' - \frac{2}{3}\alpha^3)$, i.e., from all points of D_1.

For $x' \in D_0$ consider two cases. For $x_1' \le 0$ (i.e., $a \ge b$) introduce controls $u_{\alpha,\beta}(t) = u(t; (a-\alpha)^+, (2\alpha)^0, (b-\alpha)^-, \beta^+, \gamma^0, \beta^-)$ with $0 < \alpha \le b$, $\beta > 0$, $\gamma = \alpha^2/\beta - \beta > 0$. For $x_1' > 0$ (i.e., $a < b$) consider

$$u_{\alpha,\beta}(t) = \begin{cases} u(t; (a-\alpha)^+, (2\alpha)^0, (b-\alpha)^-, \beta^+, \gamma^0, \beta^-) & \text{for } 0 < \alpha \le a, \\ \quad \text{where } \beta > 0, \gamma = \alpha^2/\beta - \beta > 0, & \\ u(t; (\alpha-a)^-, (2a)^0, (b-\alpha)^-, \beta^+, \gamma^0, \beta^-) & \text{for } a < \alpha \le b, \\ \quad \text{where } \beta > 0, \gamma = (\alpha^2 - (\alpha-a)^2)/\beta - \beta > 0. & \end{cases}$$

By these controls we achieve the origin from the points $(x_1', x_2', x_3' + \phi(\alpha) + \varepsilon)$ with $\phi(\alpha) = \alpha^2(b - \frac{2}{3}\alpha)$, $0 < \alpha \le b$, if $x_1' \le 0$ and

$$\phi(\alpha) = \begin{cases} \alpha^2\left(b - \dfrac{2}{3}\alpha\right) & \text{as} \quad 0 < \alpha \le a \\ -\alpha^2 a + 2ab\alpha - ba^2 + \dfrac{1}{3}a^3 & \text{as} \quad a < \alpha \le b \end{cases} \qquad \text{if } x_1' > 0,$$

and $\varepsilon = \frac{1}{3}\beta^3 + \frac{1}{2}\beta^2\gamma$. We note that $\varepsilon \to 0$ as $\beta \to 0$. Hence the value $\phi(\alpha) + \varepsilon$ runs through the interval $(0, \frac{1}{3}b^3)$ for $x_1' \le 0$ and $(0, \frac{1}{3}b^3 - \frac{1}{3}(b-a)^3)$ for $x_1' > 0$. Since $b = \left(\frac{1}{2}(x_1')^2 - x_2'\right)^{1/2}$, $x_3' = \frac{1}{6}(x_1')^3 - \frac{1}{3}(\frac{1}{2}(x_1')^2 - x_2')^{3/2}$, and $b - a = x_1'$, we achieve the origin from any point of D_0. □

PROPOSITION 7.2. *Points of D_0 satisfy the following property: if the control u steers $x^0 \in D_0$ to the origin in time T and there exists a control u' which steers the point $(x_1^0, x_2^0, x_3^0 + \varepsilon)$ to the origin in time T where $\varepsilon > 0$, then T is not the optimal time for the point x^0.*

Proof. Denote by x' the trajectory of system (7.2) beginning at the point x^0 with the control u'. Then obviously $x_3'(T) = -\varepsilon < 0$. Introduce the family of controls

$$u_\lambda(t) = \begin{cases} u'(t) & \text{as} \quad t \le \lambda, \\ \hat{u}_\lambda(t - \lambda) & \text{as} \quad \lambda < t \le \lambda + \hat{\theta}_\lambda, \end{cases}$$

where $\hat{\theta}_\lambda$ and $\hat{u}_\lambda(t)$ are the optimal time and the time-optimal control which steers the point $(x_1'(\lambda), x_2'(\lambda)) \in \mathbb{R}^2$ to the origin by virtue of the two-dimensional linear system $\dot{x}_1 = u$, $\dot{x}_2 = x_1$.

Consider the trajectory $x(t; \lambda)$ of system (7.2) with the control u_λ, and denote $f(\lambda) = x_3(\lambda + \hat{\theta}_\lambda; \lambda)$. Then f is a continuous function and $f(0) = x_3^0 + \frac{1}{6}\sigma(x_1^0)^3 + \frac{1}{3}\left(\frac{1}{2}(x_1^0)^2 + \sigma x_2^0\right)^{3/2} > 0$, $f(T) = -\varepsilon < 0$. Then there exists λ' such that $f(\lambda') = 0$; hence $x(\lambda' + \hat{\theta}_{\lambda'}; \lambda') = 0$. Note that $\lambda' + \hat{\theta}_{\lambda'} < T$ since $\lambda' + \hat{\theta}_{\lambda'} = T$ iff $u'(t) \equiv \hat{u}_{\lambda'}(t)$ as $t > \lambda'$, which gives $f(\lambda') = f(T) < 0$. Thus we construct the control $u_{\lambda'}$, which steers the point x^0 to the origin in time less than T. □

COROLLARY 7.3. *For $x^0 \in D_0$ the optimal control is unique.*

The proof follows arguments completely analogous to [20, p. 447].

PROPOSITION 7.4. *In D_1 the following property holds: if the control u steers $x^0 \in D_1$ to the origin in time T and there exists a control u' which steers the point $(x_1^0, x_2^0, x_3^0 - \varepsilon)$ to the origin in time T where $\varepsilon > 0$, then T is not the optimal time for the point x^0.*

The proof is analogous to the proof of Proposition 7.2.

PROPOSITION 7.5. *The maximum principle gives the two possible kinds of the optimal control:*

$$\begin{aligned} &\textit{singular type:} && u(t) = u(t; a^{\pm}, b^0, c^{\pm}); \\ &\textit{bang-bang type:} && u(t) = \pm u(t; a^+, b^-, b^+, \cdots, b^+, c^-) \\ &\textit{or} && u(t) = \pm u(t; a^+, b^-, b^+, \ldots, b^+, b^-, c^+), \quad a \le b,\ c \le b. \end{aligned}$$

Proof. For the maximum principle we consider

$$H = u\psi_1 + x_1\psi_2 + \frac{1}{2}x_1^2\psi_3, \qquad \dot\psi_1 = -\psi_2 - x_1\psi_3, \quad \dot\psi_2 = 0, \quad \dot\psi_3 = 0.$$

Hence the extremal control equals $u = \operatorname{sign}(\psi_1)$, and its singular value is $u = 0$. Moreover, note that $\ddot\psi_1 = -\psi_3 u$, $\psi_3 = \text{const}$. So, the smooth function $\psi_1(t)$, $t \in (0, T)$, can be of one of the following forms:

$$\psi_1(t) = \begin{cases} \pm\dfrac{1}{2}(t-a)^2 & \text{as} \quad t \in [0, a], \\ 0 & \text{as} \quad t \in (a, b), \\ \pm\dfrac{1}{2}(t-b)^2 & \text{as} \quad t \in (b, T), \end{cases}$$

or $\psi_1(t) = \pm\frac{(-1)^i}{2}(t-(a+(i-1)b))(t-(a+ib))$ as $t \in (a+(i-1)b, a+ib]$, $i = 0, \ldots, N$, where $a - b \le 0$ and $a + Nb \ge T$. These correspond to the two possible kinds of the optimal control given in the proposition. ∎

PROPOSITION 7.6. *For $x^0 \in D_0$ the optimal control is of singular type, while for $x^0 \in D_1$ the optimal control is of the bang-bang type.*

Proof. Let the control u of *singular type* steer the point x^0 to the origin, $b > 0$. Denote $A = |x_1^0 + u(+0)a| > 0$, and consider the control $u'(t) = u(t) + v_\alpha(t)$, where $v_\alpha(t) = u(t; a^0, \alpha^+, (2\alpha)^-, \alpha^+, (b + c - 4\alpha)^0)$ as $0 < \alpha < \min\{A, \frac{1}{4}b\}$. One can see that the control $u'(t)$ steers the point $x' = (x_1^0, x_2^0, x_3^0 - \frac{2}{3}\alpha^3)$ to the origin in the same time. Hence, due to Proposition 7.4, the control $u(t)$ cannot be optimal in D_1.

Let the control $u(t)$, $t \in (0, T)$, be of bang-bang type and have no less than two switchings. Without loss of generality, assume $u(+0) = +1$. Denote $A = x_1^0 + a \ge 0$; it follows from the maximum principle that $b \ge A$. Consider the control $u'(t) = u(t) + v_\alpha(t)$, where $v_\alpha(t) = u(t; (a-\alpha)^0, \alpha^-, \alpha^+, (b-2\alpha)^0, \alpha^+, \alpha^-, (T-a-b-\alpha)^0)$ as $0 < \alpha < \min\{a, b - A\}$ for the case $b > A$ and $0 < \alpha < a$ when $b = A$. One can see that the control $u'(t)$ steers the point $x' = (x_1^0, x_2^0, x_3^0 + \varepsilon)$ to the origin in the same time, where $\varepsilon = \alpha^2(b - \frac{4}{3}\alpha) > 0$. Hence, due to Proposition 7.2, the control $u(t)$ is not optimal in D_0. ∎

PROPOSITION 7.7. *In D_1 the optimal controls (which are of bang-bang type) have two switchings.*

Proof. The proof follows [3]. Really, consider the control $u'(t) = u(t; c^-, b^+, a^-)$, where $b \ge a > 0$, $b \ge c > 0$, which steers the point x' to the origin in time $T = a+b+c$. As it can be calculated,

$$\begin{aligned} &x_1^0 = a - b + c, \quad x_2^0 = -\frac{1}{2}(a+b+c)^2 + (b+c)^2 - c^2, \\ &x_3^0 = -\frac{1}{3}\left(a^3 + (b-a)^3 + \frac{1}{2}(a-b+c)^3\right). \end{aligned}$$

Construct the control $u''(t) = u(t; \gamma^+, \beta^-, \alpha^+)$, which steers the point $x'' = (x_1', x_2', x_3'')$ to the origin in the same time. As it is calculated in [3, p.182], these conditions give $x_3' - x_3'' = (a - b + c)abc/(a + c)$. Obviously, $x_3'' < x_3'$ iff $a + c > b$, and in this case Proposition 7.4 yields that u' cannot be optimal for the point x'.

Assume now that the optimal control $\hat{u}$ which steers the point x^0 to the origin along the trajectory $x(t)$ has three switchings, $\hat{u}(t) = u(t; d^+, c^-, b^+, a^-)$. Since it is of bang-bang type, one has $b = c$, $a \le b$, $d \le b$. Consider the point on the optimal trajectory $x' = x(d)$. Since $a + c = a + b > b$, the control $u'(t) = \hat{u}(t + d)$ is not optimal for this point. Hence the control $\hat{u}$ cannot be optimal for the point x^0. □

COROLLARY 7.8. *If $x_1^0 \in D_1$ and $x_1^0 > 0$, then the optimal control initially equals $+1$; if $x_1^0 < 0$, then the optimal control initially equals -1.*

Proof. Really, it follows from the proof of Proposition 7.7 that if the optimal control equals $u(t; c^-, b^+, a^-)$, then $x_1^0 = a - b + c$ and $b \ge a + c$. Hence this control can be optimal in the case $x_1^0 \le 0$ only. □

COROLLARY 7.9. *In D_1 the optimal control is unique except at points x^0 such that $x_1^0 = 0$, for which there exist two optimal controls.*

Proof. One can easily see that there exists only one set of numbers a, b, c such that the control $u(t; c^-, b^+, a^-)$ steers the point x^0 to the origin in time $T = a + b + c$. Further, there exists the control $u(t; \gamma^+, \beta^-, \alpha^+)$ which steers this point to the origin in the same time iff $b = a + c$, which corresponds to the case $x_1^0 = 0$. In this case, $\alpha = c$, $\beta = b$, and $\gamma = a$. □

PROPOSITION 7.10. *The optimal time θ_x is a continuous function as $x \in \operatorname{int} D$.*

Proof. (i) Consider a point $x^0 \in D_0$, and assume $x_1^0 \ge 0$. Then the optimal control is described by three parameters a, b, c, where $a \ge 0$, $b > 0$, $c > 0$, and is of one of the following three forms: $u_1(t) = u(t; a^-, b^0, c^-)$, $u_2(t) = u(t; a^+, b^0, c^-)$, $u_3(t) = u(t; a^-, b^0, c^+)$. In each of these cases, one can calculate the Jacobian of the function $x^0 = x^0(a, b, c)$. It is easy to see that it is nonzero; hence this function is nonsingular. That means that for any $\varepsilon > 0$ there exists a neighborhood $V \subset D_0$ of x^0 such that for any point $x \in V$ one has $\theta_x \le a + b + c + \varepsilon = \theta_{x^0} + \varepsilon$. The lower semicontinuity of θ_x is evident; hence the function θ_x is continuous in $x^0 \in D_0$.

(ii) The analogous arguments prove the continuity of θ_{x^0} for any point $x^0 \in D_1$ such that $x_1^0 \ne 0$. Let now $x^0 \in D_1$ and $x_1^0 = 0$. Then the optimal control $u_{x^0}(t)$ is of the form $u_{x^0}(t) = \pm u(t; a^-, (a + c)^+, c^-)$, and the optimal time equals $\theta_{x^0} = 2(a + c)$. Since the function $(x_2^0, x_3^0) = (x_2^0(a, c), x_3^0(a, c))$ is nonsingular as $a > 0$, $c > 0$, then for any $\varepsilon > 0$ one can choose a neighborhood $V \subset D_1$ of x^0 such that for any $x' \in V$, $x_1' = 0$, one has $\theta_{x'} \le \theta_{x^0} + \varepsilon$. Let $V_1 \subset V$ be such a subneighborhood and $\tau = \tau(\varepsilon) > 0$ be such a number that for any $x \in V_1$ the point $x' = (0, x_2 - \frac{1}{2}\tau^2, x_3 - \frac{1}{6}\tau^3) \in V$. Consider any $x \in V_1$; without loss of generality, assume $x_1 > 0$. Then the origin is reached from the point x in time $\theta_{x'} + \tau$ by the control $u(t) = u(t; \tau^-) \circ u'(t)$, where u' steers x' to the origin. Hence $\theta_x \le \theta_{x^0} + \varepsilon + \tau$, which proves the continuity of θ_x when $x \to x^0$, $x_1 > 0$. The same can be proved for $x \to x^0$, $x_1 < 0$.

(iii) Finally, consider $x^0 \in M \backslash \partial D$. The continuity of θ_x as $x \to x^0$ for $x \in D_0$ and $x \in D_1$ can be shown by means of the families of controls used in the proof of Proposition 7.1. □

Remark. Note that the optimal time is obviously discontinuous at points of the form $(x_1, -\frac{1}{2}x_1|x_1|, -\frac{1}{3}|x_1|^3) \in \partial D \cap M$.

Thus, Propositions 7.1, 7.2, 7.4, 7.5, 7.6, 7.7, and 7.10 and Corollaries 7.3, 7.8, and 7.9 yield all conditions of Theorem 2.4. For example, that means that the time-optimal problem for system (7.2) approximates the time-optimal problem for the

locally controllable system

$$\dot{x}_1 = u, \qquad \dot{x}_2 = x_1, \qquad \dot{x}_3 = \frac{1}{2}x_1^2 + x_1^3$$

in the domains $\Omega_1 = \{x : -\delta_1 x_1^3 < x_3 < -\delta_2 x_1^3, x_1 > 0\}$ and $\Omega_2 = \{x : \delta_1 x_1^3 < x_3 < \delta_2 x_1^3, x_1 < 0\}$ for arbitrary $\delta_1 > \delta_2 > \frac{1}{6}$.

REFERENCES

[1] A. A. Agrachev and R. V. Gamkrelidze, *The shuffle product and symmetric groups*, in Geometry of Feedback and Optimal Control, Monogr. Textbooks Pure Appl. Math. 207, Dekker, New York, 1998, pp. 365–382.

[2] R. M. Bianchini and G. Stefani, *Graded approximations and controllability along a trajectory*, SIAM J. Control Optim., 28 (1990), pp. 903–924.

[3] A. Bressan, *The generic local time-optimal stabilizing controls in dimension* 3, SIAM J. Control Optim., 24 (1986), pp. 177–190.

[4] R. W. Brockett, *Volterra series and geometric control theory*, Automatica, J. IFAC 12 (1976), pp. 167–176.

[5] P. E. Crouch and F. Lamnabhi-Lagarrigue, *Algebraic and multiple integral identities*, Acta Appl. Math., 15 (1989), pp. 235–274.

[6] P. E. Crouch, *Solvable approximations to control systems*, SIAM J. Control Optim., 22 (1984), pp. 40–54.

[7] A. F. Filippov, *On some questions in the theory of optimal regulation: Existence of a solution of the problem of optimal regulation in the class of bounded measurable functions*, Vestnik Moskov. Univ. Ser. 1 Mat. Mec. Astr. Fiz. Him., 1959, No. 2, pp. 25–32 (in Russian).

[8] A. F. Filippov, *On certain questions in the theory of optimal control*, J. SIAM Control Ser. A, 1 (1962), pp. 76–84.

[9] M. Fliess, *Fonctionnelles causales non linéaires et indéterminées non commutatives*, Bull. Soc. Math. France, 109 (1981), pp. 3–40.

[10] H. Hermes, *Control systems which generate decomposable Lie algebras*, J. Differential Equations, 44 (1982), pp. 166–187.

[11] H. Hermes, *Nilpotent approximations of control systems*, in Modern Optimal Control, Lecture Notes in Pure and Appl. Math. 119, Dekker, New York, 1989, pp. 157–172.

[12] H. Hermes, *Asymptotic stabilization via homogeneous approximation*, in Geometry of Feedback and Optimal control, Monogr. Textbooks Pure Appl. Math. 207, Dekker, New York, 1998, pp. 205–218.

[13] M. Kawski, *Nonlinear control and combinatorics of words*, in Geometry of Feedback and Optimal control, Monogr. Textbooks Pure Appl. Math. 207, Dekker, New York, 1998, pp. 305–346.

[14] V. I. Korobov, *On continuous dependence of the solution of the optimal control problem with free time on initial data*, Differ. Uravn., 7 (1971), pp. 1120–1123 (in Russian).

[15] V. I. Korobov and G. M. Sklyar, *The Markov moment problem on the smallest possible interval*, Dokl. Akad. Nauk SSSR, 308 (1989), pp. 525–528 (in Russian); translation in Soviet Math. Dokl., 40 (1990), pp. 334–337.

[16] V. I. Korobov and G. M. Sklyar, *Time-optimality and the power moment problem*, Mat. Sb. (N.S.), 134 (1987), pp. 186–206 (in Russian); translation in Sb. Math., 62 (1989), pp. 185–206.

[17] V. I. Korobov and G. M. Sklyar, *Time-optimality and the trigonometric moment problem*, Izv. Akad. Nauk SSSR Ser. Mat., 53 (1989), pp. 868–885 (in Russian); translation in Math. USSR-Izv., 35 (1990), pp. 203–220.

[18] V. I. Korobov and G. M. Sklyar, *Markov power min-problem with periodic gaps*, J. Math. Sci., 80 (1996), pp. 1559–1581.

[19] A. J. Krener, *Bilinear and nonlinear realizations of input-output maps*, SIAM J. Control, 13 (1975), pp. 827–834.

[20] C. Marchal, *Chattering arcs and chattering controls*, J. Optim. Theory Appl., 11 (1973), pp. 441–468.

[21] H. Nijmeijer and A. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.

[22] R. Ree, *Lie elements and an algebra associated with shuffles*, Annals of Math. (2), 68 (1958), pp. 210–220.

[23] G. M. Sklyar and S. Yu. Ignatovich, *A classification of linear time-optimal control problems in a neighborhood of the origin*, J. Math. Anal. Appl., 203 (1996), pp. 791–811.

[24] G. M. Sklyar and S. Yu. Ignatovich, *Moment approach to nonlinear time optimality*, SIAM J. Control Optim., 38 (2000), pp. 1707–1728.

[25] G. M. Sklyar and S. Yu. Ignatovich, *Representations of control systems in the Fliess algebra and in the algebra of nonlinear power moments*, Systems Control Lett., 47 (2002), pp. 227–235.

[26] H. J. Sussmann, *Lie brackets and local controllability: A sufficient condition for scalar-input systems*, SIAM J. Control Optim., 21 (1983), pp. 686–713.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1347–1373

ORTHONORMAL BASIS FUNCTIONS IN TIME AND FREQUENCY DOMAIN: HAMBO TRANSFORM THEORY*

PETER S. C. HEUBERGER†, THOMAS J. DE HOOG‡, PAUL M. J. VAN DEN HOF†, AND BO WAHLBERG§

Abstract. The class of finite impulse response (FIR), Laguerre, and Kautz functions can be generalized to a family of rational orthonormal basis functions for the Hardy space H_2 of stable linear dynamical systems. These basis functions are useful for constructing efficient parameterizations and coding of linear systems and signals, as required in, e.g., system identification, system approximation, and adaptive filtering. In this paper, the basis functions are derived from a transfer function perspective as well as in a state space setting. It is shown how this approach leads to alternative series expansions of systems and signals in time and frequency domain. The generalized basis functions induce signal and system transforms (Hambo transforms), which have proved to be useful analysis tools in various modelling problems. These transforms are analyzed in detail in this paper, and a large number of their properties are derived. Principally, it is shown how minimal state space realizations of the system transform can be obtained from minimal state space realizations of the original system and vice versa.

Key words. orthogonal basis functions, Hambo transform, cascade inner network, expansion coefficients

AMS subject classifications. 41A20, 42C10, 42C20, 47B35

DOI. 10.1137/S0363012902405340

1. Introduction.

Orthonormal bases and the transformations that are related to them are useful tools in many branches of science. Well-known examples are the trigonometric bases which induce the various Fourier transforms or the more recently developed orthonormal wavelet bases and their associated transforms. Within the field of systems and control theory, *rational* orthonormal bases play an important role. By approximating the impulse response of a linear time-invariant (LTI) system by a finite sum of exponentials, the problem of modelling and identification is considerably simplified. This comes down to using rational basis functions in the model structure.

Over the last years a general theory has been developed for the construction and analysis of generalized orthonormal rational basis functions for the class of stable linear systems, which extends the work on Laguerre filters by Wiener in the thirties [19]. The corresponding filters are parameterized in terms of prespecified poles, which makes it possible to incorporate a priori information about time constants in the model structure. The main applications are in system identification and adaptive signal processing, where the parameterization of models in terms of finite expansion coefficients is attractive because it is linear-in-the-parameters. This allows the use of simple linear regression estimation techniques to identify the system from observed data, thus avoiding nonconvex optimization problems. Orthonormality is associated with white

*Received by the editors April 11, 2002; accepted for publication (in revised form) March 6, 2003; published electronically October 2, 2003.

http://www.siam.org/journals/sicon/42-4/40534.html

†Delft Center for Systems and Control, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands (p.s.c.heuberger@dcsc.tudelft.nl).

‡Philips Research Laboratories, Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands (thomas.de.hoog@philips.com).

§S3—Automatic Control, Royal Institute of Technology, S-100 44 Stockholm, Sweden (bo@s3.kth.se).

noise input signals. However, the special shift structure of generalized orthonormal basis functions gives a certain Toeplitz structure for general quasi-stationary input signals, which can be utilized to construct efficient algorithms and to derive statistical performance results. The use of orthogonal basis functions has also resulted in intuitive expressions for the variance of estimated transfer functions and noise models. Here the basis functions and related reproducing kernels are used to analyze and simplify complicated variance expressions. See [46, 27, 28] for the most recent contributions. For the field of adaptive filtering, see, for instance, [2, 9, 21].

The application potentials of orthogonal basis functions go beyond the areas of system identification and adaptive signal processing. Many problems in circuit theory, signal processing, telecommunication, systems and control theory, estimation, and optimization theory benefit from an efficient representation or parameterization of particular classes of signals/systems. See, for instance, [31, 5] for applications in audio processing and [24, 23, 36] for the use of orthogonal basis functions in nonlinear modelling and estimation.

By exploiting prior knowledge of the object (signal/system) to be described, a decomposition of signals/systems in terms of flexibly chosen orthogonal (independent) components leads to efficient and robust estimation and prediction algorithms. Orthogonality is the key principle in linear estimation; see [16]. Orthogonal filters, which correspond to orthogonal rational functions, are of capital importance in filter design and robust filter implementation, as discussed in, e.g., [32].

In this paper a comprehensive account is given of the unitary transforms that result when considering series expansion representations of signals and systems in terms of a special class of generalized rational orthonormal basis functions, the so-called Hambo[1] functions. This transform generalizes the Z- and the Laguerre transforms and will be shown to have very intriguing structural properties. Preliminary results on this transform have appeared earlier in the analysis of system identification algorithms [39], in system approximation [13], and in minimal partial realization [37, 8]. In these papers, the transform results were shown to be instrumental in the statistical analysis of system identification and in solving partial realization problems. The present paper is the first to give a comprehensive account of the development and the properties of the considered transform, including analysis and algorithms in state space form.

The technique of transformation, or, equivalently, the choice of an alternative domain of representation, has been used successfully for the solution of a wide range of problems in various scientific areas; cf. Laplace and Fourier transformations in the fields of system and control theory or signal processing. It is expected that the transformation which is proposed in this paper and that has the powerful property that it can be adapted to the dynamics of a specific problem will open new possibilities for the solution of a broad class of problems.

The remainder of the paper is constructed as follows. First, in section 2, the considered basis functions will be specified and reviewed. After considering series expansion expressions in section 3, the related signal and system transforms are presented in section 4. In section 5, the constituting expressions for calculating the transforms are presented. Additional properties are discussed in section 6, while in section 7 some extensions are briefly indicated.

[1]The word Hambo originated as an acronym for Hankel matrix based orthogonality. In the remainder of the paper, these Hambo functions will also be referred to as generalized basis functions.

Notation.

$A^T, \overline{A}, A^*$ Transpose, respectively, complex conjugate and complex conjugate transpose of the matrix A.

$\mathbb{T}$ Unit circle.

$L_2^{p\times m}(\mathbb{T})$ Hilbert space of complex matrix functions of dimension $p\times m$ that are square integrable on the unit circle. The superscript $p\times m$ will be suppressed if $p=m=1$.

$H_2^{p\times m}$ Hardy space of all functions which are analytic in the exterior of the unit disc such that[2]

$$\lim_{r\to 1}\frac{1}{2\pi}\int_0^{2\pi}\mathrm{Trace}(f(re^{i\omega})f(re^{i\omega})^*)d\omega<\infty.$$

$RH_2^{p\times m}$ Subspace of *rational* transfer functions of $H_2^{p\times m}$.

$H_2^\perp$ The orthogonal complement of H_2 in L_2.

$H_{2-}^{p\times m}$ The same as $H_2^{p\times m}$, with the restriction that the functions must be zero at infinity (i.e., $f_0=0$).

$RH_{2-}^{p\times m}$ Subspace of *rational* transfer functions of $H_{2-}^{p\times m}$.

$\ell_2^n(J)$ The space of square summable vector sequences, of vector dimension n, where J denotes the index set of the sequence. The superscript n will be omitted if $n=1$.

$\langle F,G\rangle$ Inner product of F and G in $L_2^{p\times m}(\mathbb{T})$:

$$\frac{1}{2\pi i}\int_0^{2\pi}\mathrm{Trace}\{F^T(e^{i\omega})\overline{G(e^{i\omega})}\}d\omega.$$

$\langle x,y\rangle$ Inner product of x and y in $\ell_2^n(J)$: $\sum_{k\in J}x^T(k)\overline{y(k)}$.

$[\![x,y]\!]$ ℓ_2 Matrix "inner product" $\sum_{k\in J}x(k)y^T(k)$, with $x\in\ell_2^{n\times p}(J)$, $y\in\ell_2^{m\times p}(J)$.

$[\![X,Y]\!]$ L_2 Matrix "inner product" $\frac{1}{2\pi i}\oint X(z)Y^*(1/z)\frac{dz}{z}$, with $X\in L_2^{n\times p}(\mathbb{T})$, $Y\in L_2^{m\times p}(\mathbb{T})$.[3]

$\mathbf{P}_X$ Orthogonal projection onto the subspace X.

$\mathbf{e}_i$ ith canonical Euclidean basis (column) vector.

q shift operator; for $x\in\ell_2$, $n\in\mathbb{Z}$: $(q^nx)(t)=x(t+n)$.

In this paper, ℓ_2 signals will be generally denoted by small characters, whereas capitals will be used for their Z-transforms, i.e., $x(t)$, respectively, $X(z)$. Expansion coefficients of a signal in a nonstandard basis are characterized with the $\breve{}$ symbol, as in $x(t)=\sum_k\breve{x}(k)f_k(t)$. By abuse of notation, systems and operators will generally be denoted with arguments; for instance, $x(t), G(z)$ will denote elements of ℓ_2, respectively, H_2.

Unless otherwise mentioned, the notion of orthonormality will be used with respect to the ℓ_2 or L_2 inner products, as defined above.

2. Basis construction. In this section, we will present the basis functions under consideration, first in transfer function form, followed by an interpretation in a state space setting.

[2]Here H_2 is identified with the subspace of L_2 with vanishing negative Fourier coefficients. More precisely, for $F\in H_2, F(z)=f(0)+f(1)z^{-1}+f(2)z^{-2}+\cdots$, and $\sum_{k=0}^\infty|f(k)|^2<\infty$.

[3]Here $Y^*(1/z)=\sum_k y(k)^*z^{-k}$.

2.1. Transfer function approach. The main idea of constructing rational orthonormal basis functions is to generate a set of orthonormal functions that have exponential decay. A straightforward approach to this problem is to orthonormalize the set of functions

$$F_{i,j}(z) = \frac{1}{(z-a_i)^j}, \quad i \in \mathbb{N}, \quad 1 \le j \le m_i, \tag{2.1}$$

where the poles a_i can be any complex number with $|a_i| < 1$, such that $a_i \neq a_k$, $i \neq k$, and where m_i is the multiplicity of pole a_i. Obviously any rational function in H_{2-} can be described as a weighted sum of these functions if the poles a_i are chosen appropriately.

PROPOSITION 2.1. *Application of the Gram–Schmidt procedure to the sequence of functions, given by* (2.1)*, yields the orthonormal functions*

$$\Phi_k(z) = \frac{\sqrt{1-|\xi_k|^2}}{z-\xi_k} \prod_{j=1}^{k-1} \frac{1-\bar{\xi}_j z}{z-\xi_j}, \quad k \in \mathbb{N}, \tag{2.2}$$

where $\xi_{N_i+l} = a_i$, $1 \le l \le m_i$, *with* $N_i = \sum_{j=1}^{i-1} m_j$.

According to [45], this sequence of orthonormal functions was originally derived in the 1920s by Takenaka [38] and Malmquist [20] and will henceforth be referred to as the Takenaka–Malmquist functions. In the 1950s, the continuous-time version of these functions was derived by Kautz [18] in the context of network synthesis. They emerged again in the work of Ninness and Gustafsson [26] in the context of system identification. See also [4]. Orthonormality of these functions can easily be established using residue calculus. A more fundamental question is whether the orthonormal set is complete in H_{2-}. The following result, already given in [38] and [20], gives necessary and sufficient conditions for completeness.

PROPOSITION 2.2. *Let* $\{\xi_k\}_{k\in\mathbb{N}}$ *be such that* $|\xi_k| < 1$ *for all* $k \in \mathbb{N}$. *The set of Takenaka–Malmquist functions* $\{\Phi_k(z)\}_{k\in\mathbb{N}}$, *as given in* (2.2)*, is complete in* H_{2-} *if and only if*

$$\sum_{k=1}^{\infty} (1-|\xi_k|) = \infty. \tag{2.3}$$

In other words, if the sequence of poles does not converge to the unit circle "too fast," then the set of Takenaka–Malmquist functions constitutes an orthonormal basis for H_{2-}. Until the early 1990s, only special cases of these functions have been used extensively, especially in the context of system identification and signal processing. Of these special cases, the pulse and Laguerre functions are the best known examples. Consider the case where for all $k, \xi_k = a \in \mathbb{R}$, with $|a| < 1$. The corresponding basis functions are the discrete Laguerre functions

$$\Phi_k(z) = \frac{\sqrt{1-a^2}}{z-a} \left[\frac{1-az}{z-a}\right]^{k-1} \tag{2.4}$$

that reduce to the pulse functions $\Phi_k(z) = z^{-k}$ for $a = 0$.

A second special case that is discussed in detail in this paper considers the situation where all poles are taken in a repetitive manner from a finite set $\{\xi_1, \xi_2, \dots, \xi_{n_b}\}$, such that $\xi_{k\cdot n_b+j} = \xi_j$, where $k \in \mathbb{N}$ and $j = 1, \dots, n_b$. When the poles appear in

complex conjugate pairs, this results in the class of so-called generalized orthonormal basis functions, or Hambo functions [13]. For ease of notation, we introduce the inner (stable all-pass) function $G_b(z) = \prod_{i=1}^{n_b}[\frac{1-\bar{\xi}_i z}{z-\xi_i}]$. Now since $\xi_{n_b+1} = \xi_1$, it follows that $\Phi_{n_b+1}(z) = \frac{\sqrt{1-|\xi_1|^2}}{z-\xi_1}G_b(z) = \Phi_1(z)G_b(z)$, and it is easy to see that an equivalent relation holds for the next functions, $\Phi_{n_b+j}(z) = \Phi_j(z)G_b(z), j \in \mathbb{N}$. From these relations it is straightforward to derive the so-called generalized shift property:

$$\Phi_{k\cdot n_b+j}(z) = \Phi_j(z)G_b^{k-1}(z), \quad k \in \mathbb{N}, \quad j = 1, \ldots, n_b.$$

For convenience of notation, these functions are often grouped into vector functions

$$V_k(z) = \begin{bmatrix}\Phi_{(k-1)\cdot n_b+1}(z) & \Phi_{(k-1)\cdot n_b+2}(z) & \cdots & \Phi_{k\cdot n_b}(z)\end{bmatrix}^T, \tag{2.5}$$

in which case the shift property comes down to $V_k(z) = V_1(z)G_b^{k-1}(z)$. This shift property will be of paramount importance in the remainder of this paper.

In the context of system approximation and identification, it is often desired that the system responses are real-valued, and for that reason it will be advantageous to restrict the basis functions to being real-valued as well. Ninness and Gustafsson [26] showed that if the poles appear in complex conjugate pole pairs, all basis functions can be made real-valued by a simple unitary transformation of the set of basis functions.

2.2. State space interpretation. An alternative way to interpret or derive these basis functions employs state space models. Consider a (single input) stable state space model

$$x(t+1) = Ax(t) + Bu(t). \tag{2.6}$$

The function $V(z) = [zI - A]^{-1}B$ is the transfer function from the input $u(t)$ to the states $x(t)$. Now assume that the input signal $u(t)$ is a zero mean white noise process with variance 1, i.e., $\mathbb{E}\{u(t)u(t+k)\} = \delta_k$. The state covariance matrix $P = \mathbb{E}\{x(t)x^T(t)\}$ satisfies the Lyapunov equation $P = APA^T + BB^T$. P also equals the so-called controllability Gramian of the state space model. The reason why we are interested in the state covariance matrix is that

$$P = \frac{1}{2\pi i}\oint_{\mathbb{T}} V(z)V^T(1/z)\frac{dz}{z} = [\![V, V]\!]. \tag{2.7}$$

The basic idea now is to find a new state space realization for which the state covariance equals the identity matrix, $P = I$. The corresponding input to state transfer functions will then be orthonormal and will span the same space as the original functions, as only linear transformations are considered. A state space realization for which $P = I$ is called input balanced [22].

In order to extend this resulting finite set of orthonormal functions, we consider the class of square inner functions, i.e., stable transfer functions $G_b(z)$ that satisfy

$$G_b(z)G_b^T\left(\frac{1}{z}\right) = I.$$

It was shown in [33] that square inner functions can be realized by so-called orthogonal state space realizations; i.e., they satisfy $G_b(z) = D + C(zI - A)^{-1}B$, where

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^T \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix}^T = I. \tag{2.8}$$

From this orthogonality property, it directly follows that the controllability Gramian P and the observability Gramian Q, which are defined as $P = APA^T + BB^T$ and $Q = A^TQA + C^TC$, satisfy $P = Q = I$, and so realizations with this property are balanced in the sense of [22]. Thus it follows that the input-to-state functions (i.e., the elements of $V(z) = [zI - A]^{-1} B$) are mutually orthonormal with respect to the H_{2-} inner product (assuming $G_b(z)$ is scalar).

Example 2.3. We consider first and second order inner functions.

1. Let $G_b(z) = \frac{1-az}{z-a}$, with $|a| < 1$. Then $\{a, \sqrt{1-a^2}, \sqrt{1-a^2}, -a\}$ is a balanced realization for G_b, and the input to state transfer is $\frac{\sqrt{1-a^2}}{z-a}$, the first Laguerre function with pole in a.
2. Let $G_b(z) = \frac{-cz^2+b(c-1)z+1}{z^2+b(c-1)z-c}$ with some real-valued b, c satisfying $|c|, |b| < 1$. A balanced realization (see, e.g., [39]) results in $V(z) = \frac{\sqrt{1-c^2}}{z^2+b(c-1)z-c}[(z-b) \cdot \sqrt{(1-b^2)}]^T$, which represents the first two functions of the so-called 2-parameter Kautz construction.

On the other hand, when given an arbitrary pair (A, B) with controllability Gramian $P = I$, it is easy to show that there exist matrices (C, D) such that the transfer function $G(z) = D + C(zI - A)^{-1}B$ is an inner function [12]. Note that this realization is automatically balanced.

Hence, when the state space approach is used to create orthonormal functions, these functions can be considered as the input-to-state functions of a balanced realization of an inner function.

A second result from [33] as indicated in [3] is that for two inner functions $G_i(z) \in H_2$ $(i = 1, 2)$, with corresponding balanced realizations (A_i, B_i, C_i, D_i), the product $G_2(z)G_1(z)$ has a balanced realization (A, B, C, D) with

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{cc|c} A_1 & 0 & B_1 \\ B_2C_1 & A_2 & B_2D_1 \\ \hline D_2C_1 & C_2 & D_2D_1 \end{array}\right]. \tag{2.9}$$

For any input signal $u(t)$, the state sequence $x(t)$ related to this realization can be decomposed by $x(t) = [x_1(t) \quad x_2(t)]^T$, where $x_1(t)$ is the state trajectory related to the realization of $G_1(z)$ separately:

$$x_1(t) = [qI - A_1]^{-1}B_1u(t) \qquad \text{and} \qquad x_2(t) = [qI - A_2]^{-1}B_2G_1(q)u(t).$$

Here q denotes the shift operator, as defined in our notation.

In other words, there exists a recursive structure, where concatenating inner functions provide an increasing number of state functions that are orthogonal to each other with respect to the standard ℓ_2 inner product, i.e., $\sum_t x_k^T(t)\overline{x_j(t)} = \delta_{kj}$ or, equivalently, $\frac{1}{2\pi i}\int_0^{2\pi} X_k^T(e^{i\omega})\overline{X_j(e^{i\omega})}d\omega = \delta_{kj}$. This leads to the following construction.

PROPOSITION 2.4. *Given a sequence of inner functions $G_i(z), i = 1, 2, \dots,$ each with balanced realization (A_i, B_i, C_i, D_i), the collection of functions $\{X_i(z), i = 1, 2, \dots\}$ with*

$$X_1(z) = [zI - A_1]^{-1}B_1, \qquad X_i(z) = [zI - A_i]^{-1}B_iG_1(z)G_2(z)\cdots G_{i-1}(z),$$

is mutually orthonormal.

With this property and the balanced realizations of Example 2.3, it is straightforward to rederive the Takenaka–Malmquist functions (2.2) as well as the Laguerre

functions and the Hambo functions (see (2.5)). Both approaches lead to the same class of functions. Hence the completeness condition (2.3) is valid for both approaches. The special case of Proposition 2.4, where all $G_i(z)$ are equal to the same second order inner function with a complex conjugate pole pair (see Example 2.3 2) is known in the literature as the 2-parameter Kautz construction [42, 14, 26].

3. Related bases and series expansions. Since the Takenaka–Malmquist functions constitute a basis for H_{2-}, a basis for the related space $\ell_2(\mathbb{N})$ follows by considering the inverse Z-transform, which is isomorphic. With $\{\phi_k(t)\}$ the impulse response (Fourier coefficients) of $\Phi_k(z)$, according to $\Phi_k(z) := \sum_{t=1}^{\infty} \phi_k(t) z^{-t}$, the functions $\{\phi_k(t)\}$ will constitute an orthonormal basis for $\ell_2(\mathbb{N})$. Note that these basis functions exhibit the property that they can incorporate system dynamics in a very general way. One can construct inner functions from any given set of poles, and thus the resulting basis can incorporate dynamics of any complexity, combining, e.g., both fast and slow dynamics in damped and resonant modes. Considering the Takenaka–Malmquist basis functions, for any system $H(z) \in H_{2-}$ or signal $y(t) \in \ell_2(\mathbb{N})$, there exist unique series expansions:

$$H(z) = \sum_{k=1}^{\infty} \langle H, \Phi_k \rangle \Phi_k(z), \tag{3.1a}$$

$$y(t) = \sum_{k=1}^{\infty} \langle y, \phi_k \rangle \phi_k(t). \tag{3.1b}$$

In the remainder of this paper, attention will be focused on the Hambo functions, as introduced in section 2.1, i.e., the subclass of Takenaka–Malmquist functions where the basis poles are taken in a repetitive manner from a finite set $\{\xi_1, \ldots, \xi_{n_b}\}$. When these poles $\{\xi_i\}_{i=1}^{n_b}$ are stable, i.e., $|\xi_i| < 1$, it follows from Proposition 2.2 that the set of Hambo functions constitutes a basis for H_{2-}. In what follows, we will also assume that the basis poles appear in complex conjugate pairs only. Furthermore, we will primarily consider the real-rational form of these functions that results from the application of Proposition 2.4, using a real-valued state space realization of the inner function

$$G_b(z) = \prod_{i=1}^{n_b} \frac{1 - \bar{\xi}_i z}{z - \xi_i}. \tag{3.2}$$

Definition 3.1. *Let $G_b(z)$ be a real-rational inner function, with real-valued minimal balanced realization (A_b, B_b, C_b, D_b). Let for $k \in \mathbb{N}$ the vector functions $V_k(z)$ be defined as $V_k(z) = [zI - A_b]^{-1} B_b G_b^{k-1}(z)$. Then the collection of all scalar elements of the vectors $V_k(z)$, $\Phi_{k,i}(z) = e_i^T V_k(z)$, $k \in \mathbb{N}$, $1 \le i \le n_b$, is referred to as a Hambo basis of H_{2-}. The corresponding vectors with basis functions for $\ell_2(\mathbb{N})$ will be denoted by $\{v_k(t)\}$.*

It is straightforward to recognize the shift structure in the functions $v_k(t)$:

$$v_{k+1}(t) = G_b(q) \cdot v_k(t), \quad k = 1, 2, \ldots, \tag{3.3a}$$

$$v_1(t) = A_b^{t-1} B_b. \tag{3.3b}$$

For the class of Hambo functions, based on an inner function $G_b(z)$, the series expansions (3.1) can be rewritten such that the vector structure is maintained:

$$H(z) = \sum_{k=1}^{\infty} \breve{h}^T(k) V_k(z), \quad \breve{h}(k) = [\![V_k, H]\!], \tag{3.4a}$$

$$y(t) = \sum_{k=1}^{\infty} \breve{y}^T(k) v_k(t), \quad \breve{y}(k) = [\![v_k, y]\!]. \tag{3.4b}$$

The vector coefficient sequence $\breve{y} = \{\breve{y}(k)\}_{k\in\mathbb{N}}$ in (3.4) is called the Hambo signal transform of y. This transform will play a fundamental role in this paper. A formal definition will be given in section 4. The next proposition shows that the Parseval identity holds for this transform.

PROPOSITION 3.2 (Parseval's identity). *For any pair $x(t), y(t) \in \ell_2(\mathbb{N})$ and corresponding expansion coefficient sequences $\breve{x}, \breve{y}$, taken with respect to the basis vectors $\{v_k(t)\}_{k\in\mathbb{N}}$ as in* (3.4)*, it holds that $\langle x, y\rangle = \langle \breve{x}, \breve{y}\rangle$.*

Proof. $\langle x, y\rangle = [\![\sum_k \breve{x}^T(k) v_k, \sum_{k'} \breve{y}^T(k') v_{k'}]\!] = \sum_k \sum_{k'} \breve{x}^T(k) [\![v_k, v_{k'}]\!] \breve{y}(k') = \sum_k \breve{x}^T(k)\breve{y}(k)$. ∎

A dual orthonormal basis of $\ell_2^{n_b}(\mathbb{N})$. One consequence of Proposition 3.2 is that an orthonormal basis of $\ell_2^{n_b}(\mathbb{N})$ can be obtained by taking the signal transform of the standard orthonormal basis functions of $\ell_2(\mathbb{N})$: $\delta(t-k)$, $k > 0$. The resulting basis functions w_l are given by

$$w_l(k) = [\![v_k(t), \delta(t-l)]\!] = \sum_{t=1}^{\infty} v_k(t)\delta(t-l) = v_k(l). \tag{3.5}$$

Therefore, we can state the following.

PROPOSITION 3.3 (dual orthonormal basis). *Consider the basis function vectors $v_k(t)$ with $k \in \mathbb{N}$, as defined in Definition* 3.1*. The vector functions $w_t(k) \in \ell_2^{n_b}$, $t \in \mathbb{N}$, defined by $w_t(k) = v_k(t)$, constitute an orthonormal basis of the space $\ell_2^{n_b}(\mathbb{N})$.*

It turns out that—as is the case with $v_k(t)$ (see (3.3))— these functions $w_k(t)$ can be calculated using a shift structure.

PROPOSITION 3.4. *Let $G_b(z)$ be a scalar inner function with McMillan degree $n_b > 0$, having a minimal balanced realization (A_b, B_b, C_b, D_b). Consider $v_k(t), w_k(t)$ as before, and let $N(z) = A_b + B_b[zI - D_b]^{-1}C_b$. Then*

$$w_{k+1}(t) = N(q) \cdot w_k(t), \quad k = 1, 2, \ldots, \tag{3.6a}$$

$$w_1(t) = B_b D_b^{t-1}, \tag{3.6b}$$

where the shift operator q operates on the time sequence w_k, i.e., $(qw_k)(t) = w_k(t+1)$.

Proof. The proof uses the balanced state space realization $(A_{k+1}, B_{k+1}, C_{k+1}, D_{k+1})$ of $G_b^{k+1}(z)$ (see (2.9)), where

$$A_{k+1} = \begin{bmatrix} A_b & 0 & \cdots & \cdot & 0 \\ B_bC_b & A_b & 0 & \cdot & 0 \\ B_bD_bC_b & B_bC_b & \cdot & \cdot & 0 \\ \vdots & \vdots & \cdot & \ddots & 0 \\ B_bD_b^{k-1}C_b & B_bD_b^{k-2}C_b & \cdots & B_bC_b & A_b \end{bmatrix},$$

$$B_{k+1} = \begin{bmatrix} B_b \\ B_bD_b \\ B_bD_b^2 \\ \vdots \\ B_bD_b^k \end{bmatrix}. \tag{3.7}$$

It is straightforward that $\left[v_1^T(t) \ \cdots \ v_{k+1}^T(t)\right]^T = A_{k+1}^{t-1} B_{k+1}$, and hence $w_1(t) = B_b D_b^{t-1}$. For $t \geq 1$

$$\begin{aligned} w_{k+1}(t+1) = v_{t+1}(k+1) &= A_b v_{t+1}(k) + B_b C_b v_t(k) + \cdots + B_b D_b^{k-1} C_b v_1(k) \\ &= A_b w_k(t+1) + \sum_{i=1}^{t} B_b D_b^{i-1} C_b w_k(t+1-i), \end{aligned}$$

which proves the result. □

We will denote the Z-transform of the functions $w_k(t)$ by $W_k(z) := \sum_{t=1}^{\infty} w_k(t) z^{-t}$, while as a direct result of Proposition 3.4 it holds that $W_k(z) = N^{k-1}(z) \cdot W_1(z)$, with $W_1(z) := (zI - D_b)^{-1} B_b$. Note the duality between the functions $G_b(z)$ and $N(z)$, which are simply related by ordering the state space realizations in reverse.

As a consequence, for any strictly proper system $\breve{H}(z) \in H_{2-}^{n_b}$ or signal $\breve{y}(t) \in \ell_2^{n_b}[1, \infty)$, there exist unique series expansions:

$$\breve{H}(z) = \sum_{k=1}^{\infty} h(k) W_k(z), \quad h(k) = \langle \breve{H}, W_k \rangle, \tag{3.8}$$

$$\breve{y}(t) = \sum_{k=1}^{\infty} y(k) w_k(t), \quad y(k) = \langle \breve{y}, w_k \rangle. \tag{3.9}$$

In fact, these are exactly the inverses of the expansions given by (3.4).

Extension to L_2. The bases for H_{2-} that we introduced can be extended to $L_2(\mathbb{T})$, i.e., to include $(H_{2-})^{\perp}$ (see, e.g., [1]). First observe that given a basis $\{F_k(z)\}$ for H_{2-}, $\{z^{-1} F_k(\frac{1}{z})\}$ is a basis for $(H_{2-})^{\perp}$. In fact, given two bases for H_{2-}, say, $\{F_k(z)\}$ and $\{G_k(z)\}$, the set of functions $\{F_k(z), z^{-1} G_k(\frac{1}{z}), k = 1, 2, \dots\}$ is a basis for $H_{2-} \cup (H_{2-})^{\perp} = L_2(\mathbb{T})$. Using an inner function $G_b(z)$ with balanced realization (A_b, B_b, C_b, D_B), the Hambo functions have been defined as $\{V_1(z) G_b(z)^{k-1}, k \in \mathbb{N}\}$, where $V_1(z) = [zI - A_b]^{-1} B$. Another Hambo basis is created by $\{U_1(z) G_b(z)^{k-1}, k \in \mathbb{N}\}$, where $U_1(z) = [zI - A_b^T]^{-1} C_b^T$. In line with the forgoing, it follows that $\{z^{-1} U_1(\frac{1}{z}) G_b^{k-1}(\frac{1}{z}), k \in \mathbb{N}\}$ is a basis for $H_{2-}^{\perp}$. Now an interesting observation is given by the following lemma.

LEMMA 3.5. *Let $G_b(z), V_1(z), U_1(z)$ be defined as above. Then $U_1(z)$ and $V_1(z)$ are related by $z^{-1} U_1(\frac{1}{z}) = V_1(z) G_b(\frac{1}{z})$.*

Proof. Using (2.9), it is easy to show that $C_b^T G_b(z) = \left(I - z A_b^T\right) [zI - A_b]^{-1} B_b$. Substituting this relation in the expression $U_1(\frac{1}{z}) G_b(z)$ yields $U_1(\frac{1}{z}) G_b(z) = z[I - zA_b^T]^{-1} \left(I - zA_b^T\right) [zI - A_b]^{-1} B_b = z[zI - A_b]^{-1} B_b = z V_1(z)$. □

COROLLARY 3.6. *Let $G_b(z)$ and $V_1(z)$ be defined as above. The set $\{V_1(z) G_b^k(z), k \in J\}$ defines a basis, respectively, for H_{2-} if $J = \mathbb{N}$, for $H_{2-}^{\perp}$ if $J = \mathbb{Z} \setminus \mathbb{N}$, and for $L_2(\mathbb{T})$ if $J = \mathbb{Z}$.*

Analogously the dual Hambo basis of $H_{2-}^{n_b}$ can be complemented with a set of basis functions of $H_{2-}^{n_b \perp}$ such that a basis of $L_2^{n_b}(\mathbb{T})$ is obtained. A dual basis of $H_{2-}^{n_b \perp}$ is given by the functions $W_{-t}(z) = \left(N^T(\frac{1}{z})\right)^t W_0(z)$, $t > 0$, with $W_0(z)$ given by $C_b^T z^{-1} \ (z^{-1} \ I - D_b^T)^{-1}$. The vector $W_0(z)$ can be related to $W_1(z)$ (the first basis element of the dual basis of $H_{2-}^{n_b}$) as follows.

LEMMA 3.7. *With $N(z)$ a square inner function with orthogonal realization (D_b, C_b, B_b, A_b) and $W_1(z) = B_b(zI - D_b)^{-1}$, it holds that $W_0(z) = C_b^T z^{-1} \ (z^{-1} \ I - D_b^T)^{-1} = N^T(\frac{1}{z}) W_1(z)$.*

Proof. The proof is similar to that of Lemma 3.5. It is straightforward to show that $N(z)W_0(z) = W_1(z)$, using the fact that $N(z)$ is inner. ∎

As a consequence, the inner function $N(z)$ generates a basis of $L_2^{n_b}(\mathbb{T})$ in the same way that G_b generates a basis of $L_2(\mathbb{T})$. We use that $N^T(\frac{1}{z})$ is the inverse of $N(z)$.

PROPOSITION 3.8. *The set of vector functions* $\{W_k(z), k \in J\}$*, with* $W_k \in L_2^{n_b}(\mathbb{T})$*, defined as* $W_k(z) = N(z)^{k-1}B_b(zI - D_b)^{-1}$*, constitutes an orthonormal basis of* $H_{2-}^{n_b}$ *if* $J = \mathbb{N}$*, of* $H_{2-}^{n_b\perp}$ *if* $J = \mathbb{Z} \setminus \mathbb{N}$*, and of* $L_2^{n_b}(\mathbb{T})$ *if* $J = \mathbb{Z}$*.*

4. Signal and operator transforms. In this section, the fundamentals of the transform theory that underlies expansions in the generalized basis are given. It is an extension of the work that was started in [13, 39] and can be viewed as a generalization of the Laguerre transform theory for signals and systems that was developed in [30] and [29].

4.1. Signals. In the previous section, it was shown how ℓ_2 signals can be expanded in terms of general rational orthonormal basis functions that are generated by an inner function $G_b(z)$ in balanced state space form.

It will turn out to be expedient to have a definition of the Hambo signal transform that also applies to multivariable signals. Also, we will need a definition that not only applies to the Hambo basis of $\ell_2(\mathbb{N})$ but also to the Hambo bases of $\ell_2(\mathbb{Z} \setminus \mathbb{N})$ and $\ell_2(\mathbb{Z})$, as discussed in section 3. Therefore, the definitions in this section will be given for $\ell_2(J)$ signals, where J is either $\mathbb{N}, \mathbb{Z}$ or $\mathbb{Z} \setminus \mathbb{N}$.

Consider a vector signal $x(t) \in \ell_2^n(J)$ such that $x(t) = \begin{bmatrix} x_1(t) & x_2(t) & \cdots & x_n(t) \end{bmatrix}^T$. Each scalar signal $x_i(t)$ can be expanded in the corresponding Hambo basis, yielding the expansion sequences $\breve{x}_i(k)$ which are elements of $\ell_2^{n_b}(J)$. Hence it holds that

$$x(t) = \sum_{k=1}^{\infty} \begin{bmatrix} \breve{x}_1(k) & \breve{x}_2(k) & \cdots & \breve{x}_n(k) \end{bmatrix}^T v_k(t) = \sum_{k\in J} \breve{x}^T(k) v_k(t). \tag{4.1}$$

DEFINITION 4.1 (multivariable Hambo signal transform). *Given a signal* $x(t) \in \ell_2^n(J)$*, its Hambo signal transform is defined as the matrix sequence* $\{\breve{x}(k)\}_{k\in J}$*, with* $\breve{x}(k) \in \mathbb{R}^{n_b \times n}$ *given by*

$$\breve{x}(k) = [\![v_k, x]\!]. \tag{4.2}$$

Furthermore, we define the λ*-domain representation of the Hambo signal transform as*

$$\breve{X}(\lambda) = \sum_{k\in J} \breve{x}(k)\lambda^{-k}.$$

Note that $\breve{X}(\lambda)$ is simply the Z-transform of $\breve{x}(k)$ with Z replaced by λ to avoid confusion. As $\breve{X}(\lambda)$ is just a representation of the Hambo signal transform $\breve{x}(k)$ in an alternative domain, it is also commonly called the Hambo signal transform [13].

For purposes of calculation, we will also need a definition for the Hambo transform of a signal $y(t) \in \ell_2^{1\times n_b}$. This is defined through Definition 4.1 by using $x(t) = y^T(t)$ and defining

$$\breve{Y}(\lambda) := \breve{X}^T(\lambda).$$

With the multivariable signal transform as defined above, the following isomorphic relation holds.

PROPOSITION 4.2 (multivariable Hambo signal transform isomorphism). *With* $X(z) \in L_2^{n_x}(\mathbb{T})$ *and* $Y(z) \in L_2^{n_y}(\mathbb{T})$, *it holds that* $[\![X, Y]\!] = [\![\breve{X}^T, \breve{Y}^T]\!]$.

Proof. The (i, j) element of $[\![X, Y]\!]$ is equal to $\langle X_i, Y_j \rangle$. By the isomorphism of the Hambo signal transform for scalar signals, it holds that this is equal to $\langle \breve{X}_i, \breve{Y}_j \rangle$. Then, with $\breve{X}(\lambda)$ and $\breve{Y}(\lambda)$ as defined before, it follows that $[\![X, Y]\!] = \frac{1}{2\pi} \int_0^{2\pi} \breve{X}^T(e^{i\omega}) \overline{\breve{Y}(e^{i\omega})} d\omega$ $= [\![\breve{X}(\lambda)^T, \breve{Y}^T(\lambda)]\!]$. ∎

4.2. Systems. A system $G(z) \in L_2^n(\mathbb{T})$ is uniquely described by its impulse response $\{g(k)\} \in \ell_2^n$. We will use this property to define the *Hambo signal transform of a system* as the Hambo signal transform of the impulse response of the system.

DEFINITION 4.3. *Consider a system* $G(z) \in L_2^n(\mathbb{T})$ *and a Hambo basis* $\{V_k(z)\}_{k \in \mathbb{Z}}$. *The Hambo signal transform of* $G(z)$, *denoted as* $\breve{G}(\lambda)$, *is defined as*

$$\breve{G}(\lambda) = \sum_{k=-\infty}^{\infty} \breve{g}(k) \lambda^{-k}, \quad \textit{where} \quad \breve{g}(k) = [\![V_k, G]\!].$$

Example 4.4. Consider the Hambo signal transform of the basis function vector $G(z) = V_j(z)$. Obviously, in this simple case, the expansion vector coefficients are given by $\breve{g}(k) = \delta(k - j)I$. Hence it holds that the Hambo signal transform of $V_j(z)$ is equal to $\lambda^{-j} I$.

Another transform of the system $G(z)$ that is closely related to the signal transform but essentially different is the so-called *Hambo operator transform*, which describes the relationship between the signal transforms of the input and output signals of a *scalar* stable and causal system.

DEFINITION 4.5 (Hambo operator transform). *Consider a system* $G(z) \in H_2$ *and a Hambo basis* $\{V_k(z)\}_{k \in \mathbb{N}}$, *associated with the inner function* $G_b(z)$. *We define the* Hambo operator transform *of* $G(z)$, *denoted by* $\widetilde{G}(\lambda)$, *as*

$$\widetilde{G}(\lambda) = \sum_{\tau=0}^{\infty} M_\tau \lambda^{-\tau}, \tag{4.3}$$

$$\textit{where} \quad M_\tau = [\![V_1(z) G_b^\tau(z), V_1(z) G(z)]\!]. \tag{4.4}$$

PROPOSITION 4.6. *Consider signals* $u(t), y(t) \in \ell_2(\mathbb{N})$ *and a system* $G(z) \in H_2$ *such that* $y(t) = G(q)u(t)$. *With* $\widetilde{G}(\lambda)$ *the Hambo operator transform of* $G(z)$, *it holds that* $\breve{Y}(\lambda) = \widetilde{G}(\lambda) \breve{U}(\lambda)$.

Proof. Let $\breve{u}(k), \breve{y}(k)$ be the expansion coefficients of $u(t)$ and $y(t)$. $\breve{y}(k)$ can be expressed as $\breve{y}(k) = [\![V_k, G \sum_{j=1}^{\infty} \breve{u}^T(j) V_j]\!] = \sum_{j=1}^{\infty} [\![V_k, V_j G]\!] \breve{u}(j) = \sum_{j=1}^{\infty} [\![V_1 G_b^{k-1}, V_1 G_b^{j-1} G]\!] \breve{u}(j)$. Consider the inner product term for the case where $j \leq k$. Use is made of the fact that the adjoint of $G_b(z)$ by its inner property is equal to $G_b^{-1}(z)$. Hence $[\![V_1 G_b^{k-1}, V_1 G_b^{j-1} G]\!] = [\![V_1 G_b^{k-j}, V_1 G]\!]$. Now consider the inner product term for the case where $j > k$. Then, with the same argument, one finds that it holds that $[\![V_1 G_b^{k-1}, V_1 G_b^{j-1} G]\!] = [\![V_1, V_1 G_b^{j-k} G]\!]$. This latter expression is equal to zero, which follows from the fact that the elements of the transfer function $V_1(z)$ constitute an orthonormal set which exactly spans the orthogonal complement in H_2 of the shift-invariant subspace $G_b(z) H_2$. The right-hand side argument of the inner product is an element of that subspace. Applying the signal transform of Definition 4.1 to $\breve{y}(k)$

(with $J = \mathbb{N}$) reveals that it holds that

$$\breve{Y}(\lambda) = \left(\sum_{\tau=0}^{\infty} M_\tau \lambda^{-\tau} \right) \breve{U}(\lambda). \qquad \square \tag{4.5}$$

The parameters M_τ are matrices of dimension $n_b \times n_b$. They can be viewed as the Markov parameters of the multivariable transfer function $\widetilde{G}(\lambda)$. The expansion coefficients $\{\breve{y}(k)\}$ and the Markov parameters $\{M_\tau\}$, as given by Definitions 4.3 and 4.5, are closely connected through a linear relation; see [37, 8, 7] for details.

The Hambo operator transform of the system $G_b(z)$ has a particularly simple form. It holds for all $U \in H_{2-}$ that

$$G_b(z)U(z) = \sum_{k=1}^{\infty} \breve{u}^T(k) V_k(z) G_b(z) = \sum_{k=1}^{\infty} \breve{u}^T(k) V_{k+1}(z).$$

Hence, with $Y(z) = G_b(z)U(z) = \sum_{k=1}^{\infty} \breve{y}^T(k) V_k(z)$, it follows that $\breve{y}(k) = \breve{u}(k-1)$ for $k > 1$ and $\breve{y}(1) = 0$. Therefore, it holds that $M_1 = I$ and $M_\tau = 0$ for all $\tau \neq 1$, and consequently

$$\widetilde{G_b}(\lambda) = \lambda^{-1} I. \tag{4.6}$$

We can hence conclude that a multiplication with $G_b(z)$ in the Z-domain corresponds to applying a canonical shift in the λ-domain.

Although the Hambo operator transform is defined only for SISO systems, there is a simple multivariable case in which it can also be used. We will need it in the next section.

PROPOSITION 4.7. *Consider a signal $u(t) \in \ell_2^m(J)$ and an SISO system $G(z) \in H_2$. Let $y(t) \in \ell_2^m(J)$ be given by $y(t) = G(q) \cdot I\ u(t)$. Then it holds that $\breve{Y}(\lambda) = \widetilde{G}(\lambda)\breve{U}(\lambda)$.*

Proof. Denoting the elements of $U(z)$ and $Y(z)$ as $U_i(z)$ and $Y_i(z)$ according to $U(z) = \begin{bmatrix} U_1(z) & U_2(z) & \cdots & U_m(z) \end{bmatrix}^T$ and $Y(z) = \begin{bmatrix} Y_1(z) & Y_2(z) & \cdots & Y_m(z) \end{bmatrix}^T$, we have that $Y_i(z) = G(z)U_i(z)$ for $1 \leq i \leq m$. Then the Hambo signal transform of $Y_i(z)$ satisfies, by definition of the Hambo operator transform, $\breve{Y}_i(\lambda) = \widetilde{G}(\lambda)\breve{U}_i(\lambda)$. The result then follows from the fact that

$$\breve{Y}(\lambda) = \begin{bmatrix} \breve{Y}_1(\lambda) & \breve{Y}_2(\lambda) & \cdots & \breve{Y}_m(\lambda) \end{bmatrix} = \widetilde{G}(\lambda)\breve{U}(\lambda). \qquad \square$$

5. Operator transform expressions. As shown, the Hambo operator transform of a system $G(z) \in H_2$ is a causal LTI system. Furthermore, the transform of a rational transfer function is again rational. We will now derive expressions by which the operator transform can actually be computed. First it is shown that an expression for $\widetilde{G}(\lambda)$ is obtained by making a variable substitution in the Laurent expansion of $G(z)$. Next it is shown how a state space realization of $\widetilde{G}(\lambda)$ can be derived on the basis of a state space realization of $G(z)$.

5.1. Variable substitution property. The Hambo operator transform, as defined in Definition 4.5, can be obtained from the original transfer function $G(z) \in H_2$ by applying a variable substitution in its Laurent expansion, which is given by

$$G(z) = \sum_{\tau=0}^{\infty} g(\tau) z^{-\tau}. \tag{5.1}$$

This variable substitution consists of a replacement of the shift operation z^{-1} by the causal linear time-invariant operator $N(\lambda)$.

PROPOSITION 5.1 (variable substitution property [39]). *Let $N(\lambda)$ be as in Proposition* 3.4. *Then the Hambo operator transform $\widetilde{G}(\lambda)$ of a given system $G(z) \in H_2$ is equal to*

$$\widetilde{G}(\lambda) = \sum_{\tau=0}^{\infty} g(\tau) N^{\tau}(\lambda). \tag{5.2}$$

With slight abuse of notation, (5.2) is sometimes stated as $\widetilde{G}(\lambda) = G(z)|_{z^{-1}=N(\lambda)}$. An immediate consequence of Proposition 5.1 is that the operator transform of the canonical shift z^{-1} is equal to $N(\lambda)$. This means that a shift in the time domain corresponds to the application of the operator $N(\lambda)$ in the signal transform domain. Another immediate consequence of this proposition is that $N(\lambda)$ and $\widetilde{G}(\lambda)$ are commuting operators. A third consequence of Proposition 5.1 is the following relation between the Hambo signal transform and the Hambo operator transform.

COROLLARY 5.2. *The Hambo signal transform $\breve{G}(\lambda)$ and Hambo operator transform $\widetilde{G}(\lambda)$ of a given system $G(z) \in H_{2-}$ are related through $\breve{G}(\lambda) = \widetilde{G}(\lambda) W_0(\lambda)$, with $W_0(\lambda) \in H_2^{n_b \perp}$ equal to $C_b^T \frac{1}{\lambda} (\frac{1}{\lambda} I - D_b^T)^{-1}$, in accordance with Proposition* 3.8.

Proof. As the functions $\{W_t(\lambda)\}_{t\in\mathbb{N}}$ constitute the dual Hambo basis, $\breve{G}(\lambda)$ satisfies $\breve{G}(\lambda) = \sum_{t=1}^{\infty} g(t) W_t(\lambda)$, with $g(t)$ the impulse response coefficients of $G(z)$. By Proposition 3.4 and the fact that $N(\lambda)$ is inner, we can write $\breve{G}(\lambda) = \sum_{t=1}^{\infty} g(t) N(\lambda)^t \cdot N^T(\frac{1}{\lambda}) W_1(\lambda)$. By Lemma 3.7 and Proposition 5.1, it then follows that $\breve{G}(\lambda) = \sum_{t=1}^{\infty} g(t) N(\lambda)^t W_0(\lambda) = \widetilde{G}(\lambda) W_0(\lambda)$. ∎

It was shown in [13] that, inversely, $G(z)$ can also be obtained from $\widetilde{G}(\lambda)$ by means of a variable substitution:

$$G(z) = z V_1^T(z)\, \widetilde{G}(\lambda) W_1(\lambda) \lambda \Big|_{\lambda^{-1}=G_b(z)}. \tag{5.3}$$

Using the multivariable signal transform Definition 4.1 one can establish an isomorphic relation that involves the Hambo operator transform.

PROPOSITION 5.3 (Hambo operator transform isomorphism). *Consider the Hambo basis of $L_2(\mathbb{T})$, generated by an inner function $G_b(z)$. Hence we have that $V_k(z) = V_1(z) G_b(z)^{k-1}$ and $W_k(\lambda) = N(\lambda)^{k-1} W_1(\lambda)$. Then for all $G_1(z), G_2(z) \in H_2$, $k \in \mathbb{Z}$,*

$$[\![V_k G_1, V_k G_2]\!] = [\![\widetilde{G}_1^T, \widetilde{G}_2^T]\!], \tag{5.4}$$

$$\textit{and} \qquad \langle G_1, G_2 \rangle = \langle \widetilde{G}_1 W_k, \widetilde{G}_2 W_k \rangle. \tag{5.5}$$

Proof. We will prove both assertions for the case $k = 1$. The other cases follow immediately from the inner property of $G_b(z)$, and $N(\lambda)$. By Proposition 4.2, it holds that $[\![V_1 G_1, V_1 G_2]\!] = [\![(V_1 \breve{G}_1)^T, (V_1 \breve{G}_2)^T]\!]$. The elements of the vector $V_1(z) G_k(z)$, $k = 1, 2$, are equal to $G_k(z) \Phi_{1,i}(z)$, $1 \le i \le n_b$, where $\Phi_{1,i}(z)$ are the first n_b scalar basis functions. The Hambo signal transform of $G_k(z)\Phi_{1,i}(z)$ is, by definition of the operator transform, equal to $\widetilde{G}_k(\lambda) \breve{\Phi}_{1,i}(\lambda) = \widetilde{G}_k(\lambda) e_i^T \lambda^{-1}$. By Definition 4.1, it then follows that $(V_1 \breve{G}_k) = \widetilde{G}_k(\lambda) \lambda^{-1}$. Hence $[\![V_1 G_1, V_1 G_2]\!] = [\![\widetilde{G}_1^T(\lambda) \lambda^{-1}, \widetilde{G}_2^T(\lambda) \lambda^{-1}]\!] = [\![\widetilde{G}_1^T, \widetilde{G}_2^T]\!]$. The second assertion is proved as follows. It holds that $\langle G_1, G_2 \rangle = \langle G_1 z^{-1}, G_2 z^{-1} \rangle = \langle (G_1 \breve{z}^{-1}), (G_2 \breve{z}^{-1}) \rangle$. The last equality follows from the isomorphism of the signal transform. Using the fact that $W_1(\lambda)$ is the Hambo signal transform of z^{-1} and by definition of the Hambo operator transform, the result follows. ∎

5.2. Hankel operator representations. The Hankel operator associated with an LTI system $G(z)$ can be represented in a number of ways, depending on the (orthonormal) coordinate systems that are used for the input and output signal spaces. The Hankel operator of a scalar system maps from $\ell_2(-\infty, 0]$ to $\ell_2[1, \infty)$. Usually, the canonical bases of these spaces are employed to represent the input and output signals. In that case, the Hankel operator can be represented as a Hankel matrix $\mathbf{H}$ that contains the Markov parameters $g(k), k > 0$, of $G(z)$, as $\mathbf{H}_{i,j} = g(i+j-1)$. Now define $\mathbf{y} = \begin{bmatrix} y(1) & y(2) & \cdots \end{bmatrix}^T$, $\mathbf{u} = \begin{bmatrix} u(0) & u(-1) & \cdots \end{bmatrix}^T$. Then it holds that

$$\mathbf{y} = \mathbf{H}\mathbf{u}. \tag{5.6}$$

Alternative representations of the Hankel operator would be obtained if one were to use other orthonormal bases for the representation of the input and output signals. A particularly interesting case occurs when we use a Hambo basis for the output space $\ell_2[1, \infty)$ and the complementary Hambo basis for $\ell_2(-\infty, 0]$ for the input space. Consider the expansion of the output signal $y(t) \in \ell_2[1, \infty)$ and the input signal $u(t) \in \ell_2(-\infty, 0]$ in terms of a Hambo basis. We then obtain the coefficients $\breve{y}(k) = [\![y, v_k]\!]^T$ with $k \in \mathbb{N}$ and $\breve{u}(k) = [\![u, v_k]\!]^T$ with $k \in \mathbb{Z} \setminus \mathbb{N}$. We collect these coefficients in column vectors $\breve{\mathbf{y}}, \breve{\mathbf{u}}$ defined as

$$\breve{\mathbf{y}}^T = \begin{bmatrix} \breve{y}^T(1) & \breve{y}^T(2) & \breve{y}^T(3) & \cdots \end{bmatrix}, \tag{5.7}$$

$$\breve{\mathbf{u}}^T = \begin{bmatrix} \breve{u}^T(0) & \breve{u}^T(-1) & \breve{u}^T(-2) & \cdots \end{bmatrix}. \tag{5.8}$$

Defining the block row vectors $\mathbf{v}_k$ with $k \in \mathbb{Z}$ as

$$\mathbf{v}_k = \begin{cases} \begin{bmatrix} v_k(1) & v_k(2) & v_k(3) & \cdots \end{bmatrix}, & k \geq 1, \\ \begin{bmatrix} v_k(0) & v_k(-1) & v_k(-2) & \cdots \end{bmatrix}, & k < 1, \end{cases}$$

and defining $V_f = \begin{bmatrix} \mathbf{v}_1^T & \mathbf{v}_2^T & \cdots \end{bmatrix}^T$ and $V_p = \begin{bmatrix} \mathbf{v}_0^T & \mathbf{v}_{-1}^T & \cdots \end{bmatrix}^T$, we can write

$$\breve{\mathbf{y}} = \mathbf{V}_f \mathbf{y} \quad \text{and} \quad \breve{\mathbf{u}} = \mathbf{V}_p \mathbf{u}. \tag{5.9}$$

It is clear that the infinite dimensional matrices $\mathbf{V}_f$ and $\mathbf{V}_p$ are unitary (orthogonal) matrices as their rows are orthogonal vectors. It hence follows that we can also write $\mathbf{y} = \mathbf{V}_f^T \breve{\mathbf{y}}$ and $\mathbf{u} = \mathbf{V}_p^T \breve{\mathbf{u}}$. Substituting this in (5.6) gives the relation $\mathbf{V}_f^T \breve{\mathbf{y}} = \mathbf{H}\mathbf{V}_p^T \breve{\mathbf{u}}$. Again using the fact that $\mathbf{V}_f$ is orthogonal, this can be rephrased as $\breve{\mathbf{y}} = \widetilde{\mathbf{H}} \breve{\mathbf{u}}$, with $\widetilde{\mathbf{H}} = \mathbf{V}_f \mathbf{H}\, \mathbf{V}_p^T$. The matrix operator $\widetilde{\mathbf{H}}$ is an alternative representation of the Hankel operator of $G(z)$. If we partition the matrix $\widetilde{\mathbf{H}}$ in blocks of dimension $n_b \times n_b$ corresponding to the partitioning of $\breve{\mathbf{u}}$ and $\breve{\mathbf{y}}$, then we find that the (i, j) block element equals $\widetilde{\mathbf{H}}_{(i,j)} = \mathbf{v}_i \mathbf{H}\, \mathbf{v}_{-j+1}^T$, with $\mathbf{v}_k$ the vector representations of the basis functions v_k as defined above. It is then clear that $\widetilde{\mathbf{H}}_{(i,j)}$ is equal to the matrix inner product between $V_i(z)$ and the Z-transform expression for the vector $\mathbf{H}\mathbf{v}_{-j+1}^T$. This leads to the following proposition.

PROPOSITION 5.4. *Let $\widetilde{\mathbf{H}}$ be the matrix representation of the Hankel operator of a system $G(z) \in H_2$, in terms of a Hambo basis associated with an inner function $G_b(z)$, such that $\breve{\mathbf{y}} = \widetilde{\mathbf{H}} \breve{\mathbf{u}}$, where $\breve{\mathbf{y}}$ and $\breve{\mathbf{u}}$ are as defined by* (5.7) *and* (5.8). *Let $\widetilde{\mathbf{H}}$ be partitioned in blocks of dimension $n_b \times n_b$, and let $\widetilde{\mathbf{H}}_{(i,j)}$ denote the (i, j)th block*

element. Then it holds that $\widetilde{\mathbf{H}}_{(i,j)} = M_{i+j-1}$*, where* M_k *represents the* k*th Markov parameter of* $\widetilde{G}(\lambda)$*, as defined in* (4.4).

Proof. By definition of the Hankel map, the term $\mathbf{H}\mathbf{v}_{-j+1}^T$ is the output of the system $G(z)$ in response to the input $v_{-j+1} \in \ell_2^{n_b}(-\infty, 0]$, restricted to the space of future signals $\ell_2^{n_b}[1, \infty)$. In Z-transform notation, this output can be expressed as $\mathbf{P}_{H_{2-}^{n_b}} G(z) V_{-j+1}^T(z) = \mathbf{P}_{H_{2-}^{n_b}} G(z) G_b^{-j} V_1^T(z)$. The last equality follows from the fact that $V_k(z) = G_b^{k-1}(z) V_1(z)$ for all $k \in \mathbb{Z}$. It then follows that $\widetilde{\mathbf{H}}_{(i,j)} = [\![V_i, \mathbf{P}_{H_{2-}^{n_b}} G G_b{}^{-j} V_1]\!] = [\![G_b^{i-1} V_1, G G_b^{-j} V_1]\!]$. Because $G_b(z)$ is inner, this expression simplifies to $\widetilde{\mathbf{H}}_{(i,j)} = [\![G_b^{i+j-1} V_1, G V_1]\!]$, which is equal to M_{i+j-1}, as was established earlier; see (4.4). □

Proposition 5.4 shows that $\widetilde{\mathbf{H}}$ has a block Hankel form, which coincides with the standard block Hankel matrix representation of the Hambo operator transform $\widetilde{G}(\lambda)$. One consequence of this observation is that Hankel singular values and the McMillan degree are invariant under Hambo operator transformation.

5.3. State space expressions for the Hambo operator transform and its inverse. In this section, we will derive the expressions by which a minimal realization of the Hambo operator transform can be obtained from a minimal state space realization of the original system and vice versa. The derivation is based on the isomorphic relation that exists between such state space realizations. We will first establish this relation. Consider the (block) Hankel matrix representation $\mathbf{H}$ of the Hankel operator of an LTI system $G(z)$. It is a well-known result from realization theory that *any* full rank decomposition $\mathbf{H} = \mathbf{\Gamma}\mathbf{\Delta}$ corresponds to a minimal realization of $G(z)$ [15, 17]. That is, there exists a minimal realization (A, B, C, D) of $G(z)$ such that $\mathbf{\Gamma} = \begin{bmatrix} C^T & (CA)^T & (CA^2)^T & \cdots \end{bmatrix}^T$ and $\mathbf{\Delta} = \begin{bmatrix} B & AB & A^2B & \cdots \end{bmatrix}$. We define the transfer functions $\Gamma(z) \in H_{2-}^n$ and $\Delta(z) \in H_{2-}^{n\,\perp}$ as

$$\Gamma(z) = \sum_{k=1}^{\infty} CA^{k-1} z^{-k} = C(zI - A)^{-1}, \qquad \Delta(z) = \sum_{k=0}^{\infty} A^k B z^k = z^{-1}\ (z^{-1}\ I - A)^{-1} B.$$

The following lemma establishes an important relation between these functions and their counterparts in the transform domain.

LEMMA 5.5. *Consider a system* $G(z) \in RH_2$ *with minimal realization* (A, B, C, D)*. Let* $\Gamma(z)$ *and* $\Delta(z)$ *be defined as* $\Gamma(z) = C(zI - A)^{-1}$*,* $\Delta(z) = z^{-1}\ (z^{-1}\ I - A)^{-1} B$*. Then the Hambo operator transform* $\widetilde{G}(\lambda)$ *of* $G(z)$ *has a minimal state space realization* $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ *such that it holds that*

$$\tilde{C}(\lambda I - \tilde{A})^{-1} = \breve{\Gamma}^T(\lambda) \ \text{ and } \ \lambda^{-1}\ (\lambda^{-1}\ I - \tilde{A})^{-1} \tilde{B} = \breve{\Delta}^T(\lambda), \tag{5.10}$$

where $\breve{\Gamma}(\lambda)$ *and* $\breve{\Delta}(\lambda)$ *are the (multivariable) Hambo signal transforms of* $\Gamma(z)$*, respectively* $\Delta(z)$*, as defined in Definition* 4.3.

Conversely, any Hambo operator transform $\widetilde{G}(\lambda)$ *with minimal state space realization* $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ *has a preimage* $G(z)$ *with minimal realization* (A, B, C, D) *such that* (5.10) *holds.*

Proof. From the analysis in the previous section, it follows that, given a full rank factorization $\mathbf{H} = \mathbf{\Gamma}\mathbf{\Delta}$, a full rank factorization of $\widetilde{\mathbf{H}}$ can be obtained according to $\widetilde{\mathbf{H}} = (\mathbf{V}_f \mathbf{\Gamma})(\mathbf{\Delta} \mathbf{V}_p^T)$. Denote the minimal state space realization of $\widetilde{G}(\lambda)$ that corresponds

to this realization by $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$. We then denote $(\mathbf{V}_f\mathbf{\Gamma})_{(k)} = \tilde{C}\tilde{A}^{k-1}$, $k \geq 1$, and $(\mathbf{\Delta}\mathbf{V}_p^T)_{(-k)} = \tilde{A}^k\tilde{B}$, $k \geq 0$. It holds that $(\mathbf{V}_f\mathbf{\Gamma})_{(k)} = \mathbf{v}_k\mathbf{\Gamma}$, $k \geq 1$, and $(\mathbf{\Delta}\mathbf{V}_p^T)_{(k)} = \mathbf{\Delta}\mathbf{v}_k^T$, $k < 1$. With $\Gamma(z)$ and $\Delta(z)$ as defined above, we then see that

$$(\mathbf{V}_f\mathbf{\Gamma})_{(k)} = [\![\Gamma^T(z), V_k(z)]\!], \quad k \geq 1, \quad \text{and} \quad (\mathbf{\Delta}\mathbf{V}_p^T)_{(k)} = [\![\Delta(z), V_k(z)]\!], \quad k < 1,$$

where the last equation holds under the assumption that the realization of $\Delta(z)$ is real. This shows, using (4.2), that $\{(\mathbf{V}_f\mathbf{\Gamma})_{(k)}\}$ and $\{(\mathbf{\Delta}\mathbf{V}_p^T)_{(k)}\}$ constitute the multivariable Hambo signal transforms of $\Gamma^T(z)$ and $\Delta(z)$, respectively. Since any minimal realization of $G(z)$ corresponds to a full rank factorization of $\mathbf{H}$, the first part of the lemma is proven. The last statement of the lemma follows from the fact that the Hambo signal transform is a bijective map. ∎

Lemma 5.5 is a very powerful result as it permits us to derive very compact expressions for computing the Hambo operator transform and its inverse, using the isomorphism relation for the multivariable Hambo signal transform given in Proposition 4.2.

Suppose that the realizations (A, B, C, D) and $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ are linked to each other via the Hambo signal transform as described in Lemma 5.5. Let us denote the controllability Gramians associated with these realizations as X_c and $\tilde{X}_c$ and the observability Gramians as X_o and $\tilde{X}_o$, respectively. Then, by the Hambo signal transform isomorphism, it holds for the functions $\Gamma(z)$ and $\Delta(z)$ that

$$X_o = [\![\Gamma^T(z), \Gamma^T(z)]\!] = [\![\breve{\Gamma}^T(\lambda), \breve{\Gamma}^T(\lambda)]\!] = \tilde{X}_o, \tag{5.11}$$

$$X_c = [\![\Delta(z), \Delta(z)]\!] = [\![\breve{\Delta}(\lambda), \breve{\Delta}(\lambda)]\!] = \tilde{X}_c. \tag{5.12}$$

Using the Hambo signal transform isomorphism, we can now establish a matrix inner product expression for the realization $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$, as follows.

PROPOSITION 5.6. *With $\Gamma(z)$ and $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ as defined in Lemma 5.5 and X_o the controllability Gramian of this realization, it holds that*

$$\begin{bmatrix} X_o & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{bmatrix} = \left[\!\!\left[\begin{bmatrix} \Gamma^T(z)G_b(z) \\ V_1(z) \end{bmatrix}, \begin{bmatrix} \Gamma^T(z) \\ V_1(z)G(z) \end{bmatrix} \right]\!\!\right]. \tag{5.13}$$

Proof. The system $\widetilde{G}^T(\lambda)$ is described by the equation

$$\begin{bmatrix} X(\lambda)\lambda \\ Y(\lambda) \end{bmatrix} = \begin{bmatrix} \tilde{A}^T & \tilde{C}^T \\ \tilde{B}^T & \tilde{D}^T \end{bmatrix} \begin{bmatrix} X(\lambda) \\ U(\lambda) \end{bmatrix}.$$

It holds that

$$\left[\!\!\left[\begin{bmatrix} X(\lambda) \\ U(\lambda) \end{bmatrix}, \begin{bmatrix} X(\lambda)\lambda \\ Y(\lambda) \end{bmatrix} \right]\!\!\right] = \left[\!\!\left[\begin{bmatrix} X(\lambda) \\ U(\lambda) \end{bmatrix}, \begin{bmatrix} X(\lambda) \\ U(\lambda) \end{bmatrix} \right]\!\!\right] \begin{bmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{bmatrix}.$$

Let the input $u(t)$ be equal to $e_i\delta(t)$, with e_i the ith Euclidean basis vector of $\mathbb{R}^{n_b}$. Then $X(\lambda) = [\lambda I - \tilde{A}^T]^{-1}\tilde{C}^T e_i = \breve{\Gamma}(\lambda)e_i$, and by Lemma 5.5 this last equation can be written as

$$\left[\!\!\left[\begin{bmatrix} \breve{\Gamma}(\lambda)e_i \\ e_i \end{bmatrix}, \begin{bmatrix} \breve{\Gamma}(\lambda)e_i\lambda \\ \widetilde{G}^T(\lambda)e_i \end{bmatrix} \right]\!\!\right] = \left[\!\!\left[\begin{bmatrix} \breve{\Gamma}(\lambda)e_i \\ e_i \end{bmatrix}, \begin{bmatrix} \breve{\Gamma}(\lambda)e_i \\ e_i \end{bmatrix} \right]\!\!\right] \begin{bmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{bmatrix}.$$

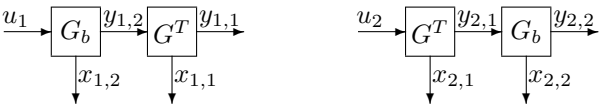


FIG. 5.1. *Systems for proof of Corollary* 5.7.

Because this holds for all i such that $1 \le i \le n_b$, we can also write (after summation of the latter equation over all $i = 1, \dots, n_b$)

$$\left[\!\left[\begin{bmatrix}\breve{\Gamma}(\lambda)\\ I\end{bmatrix}, \begin{bmatrix}\breve{\Gamma}(\lambda)\lambda\\ \widetilde{G}^T(\lambda)\end{bmatrix}\right]\!\right] = \left[\!\left[\begin{bmatrix}\breve{\Gamma}(\lambda)\\ I\end{bmatrix}, \begin{bmatrix}\breve{\Gamma}(\lambda)\\ I\end{bmatrix}\right]\!\right] \begin{bmatrix}\tilde{A} & \tilde{B}\\ \tilde{C} & \tilde{D}\end{bmatrix} = \begin{bmatrix}\tilde{X}_o & 0\\ 0 & I\end{bmatrix} \begin{bmatrix}\tilde{A} & \tilde{B}\\ \tilde{C} & \tilde{D}\end{bmatrix}.$$

The term on the left-hand side of this equation equals

$$\left[\!\left[\begin{bmatrix}\breve{\Gamma}(\lambda)\lambda^{-1}\\ I\lambda^{-1}\end{bmatrix}, \begin{bmatrix}\breve{\Gamma}(\lambda)\\ \widetilde{G}^T(\lambda)\lambda^{-1}\end{bmatrix}\right]\!\right].$$

We observe that $\lambda^{-1}I$ is equal to the Hambo operator transform of $G_b(z)$ (see (4.6)). Further, $\lambda^{-1}I$ is the Hambo signal transform of $V_1(z)$, as was demonstrated in Example 4.4. From Proposition 4.7 it then follows that $\lambda^{-1}I\,\breve{\Gamma}^T(\lambda)$ is the Hambo signal transform of $G_b(z)\cdot I\Gamma^T(z)$. Similarly, $\lambda^{-1}I\widetilde{G}(\lambda)$ is the signal transform of $V_1(z)G(z)$. Using the Hambo signal transform isomorphism (Proposition 4.2), it therefore holds that

$$\left[\!\left[\begin{bmatrix}\breve{\Gamma}(\lambda)\lambda^{-1}\\ I\lambda^{-1}\end{bmatrix}, \begin{bmatrix}\breve{\Gamma}(\lambda)\\ \widetilde{G}^T(\lambda)\lambda^{-1}\end{bmatrix}\right]\!\right] = \left[\!\left[\begin{bmatrix}\Gamma^T(z)G_b(z)\\ V_1(z)\end{bmatrix}, \begin{bmatrix}\Gamma^T(z)\\ V_1(z)G(z)\end{bmatrix}\right]\!\right]. \qquad \square$$

Obviously, a dual formulation of this proposition that uses expressions involving $\Delta(z)$ and X_c is possible.

Proposition 5.6 can also be formulated in the form of a Sylvester equation.

COROLLARY 5.7. *Consider a system $G(z) \in RH_2$, with minimal realization (A, B, C, D) and observability Gramian X_o. Then $\widetilde{G}(\lambda)$ has a minimal realization $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ that satisfies the following Sylvester equation:*

$$\begin{bmatrix}A^T & C^TC_b\\ 0 & A_b\end{bmatrix} \begin{bmatrix}X_o\tilde{A} & X_o\tilde{B}\\ \tilde{C} & \tilde{D}\end{bmatrix} \begin{bmatrix}A & BB_b^T\\ 0 & A_b^T\end{bmatrix} + \begin{bmatrix}C^TD_b\\ B_b\end{bmatrix} \begin{bmatrix}C & DB_b^T\end{bmatrix} = \begin{bmatrix}X_o\tilde{A} & X_o\tilde{B}\\ \tilde{C} & \tilde{D}\end{bmatrix}. \tag{5.14}$$

Proof. The Sylvester equation is obtained by formulating (5.13) in the time domain using straightforward state space realizations of the transfer functions that appear in the inner product. Consider the systems shown in Figure 5.1. State equations of these systems are

$$\begin{bmatrix}x_{1,1}(t+1)\\ x_{1,2}(t+1)\end{bmatrix} = \begin{bmatrix}A^T & C^TC_b\\ 0 & A_b\end{bmatrix} \begin{bmatrix}x_{1,1}(t)\\ x_{1,2}(t)\end{bmatrix} + \begin{bmatrix}C^TD_b\\ B_b\end{bmatrix} u(t)$$

and

$$\begin{bmatrix}x_{2,1}(t+1)\\ x_{2,2}(t+1)\end{bmatrix} = \begin{bmatrix}A^T & 0\\ B_bB^T & A_b\end{bmatrix} \begin{bmatrix}x_{2,1}(t)\\ x_{2,2}(t)\end{bmatrix} + \begin{bmatrix}C^T\\ B_bD^T\end{bmatrix} u(t),$$

respectively. The solution of (5.13) is then equal to

$$\left[\!\!\left[\begin{bmatrix} x_{1,1} \\ x_{1,2} \end{bmatrix}, \begin{bmatrix} x_{2,1} \\ x_{2,2} \end{bmatrix} \right]\!\!\right],$$

which results in (5.14). ∎

Existence of a solution to this Sylvester equation (5.14) is guaranteed if the systems in the inner product expression in (5.13) are stable. This is true by assumption for $\Gamma(z)$ and $G(z)$ and by definition for $G_b(z)$ and $V_1(z)$.

Note that (5.14) can be simplified further in the case where $X_o = I$, i.e., when the realization (A, B, C, D) is output balanced.

Using the Hambo signal transform isomorphism, it is equally simple to derive a matrix inner product expression for the realization (A, B, C, D) that involves $\breve{\Gamma}(\lambda)$.

PROPOSITION 5.8. *With $\Gamma(z)$ and (A, B, C, D) as defined in Lemma 5.5, it holds that*

$$\begin{bmatrix} \tilde{X}_o & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \left[\!\!\left[\begin{bmatrix} \breve{\Gamma}^T(\lambda) N^T(\lambda) \\ W_1^T(\lambda) \end{bmatrix}, \begin{bmatrix} \breve{\Gamma}^T(\lambda) \\ W_1^T(\lambda) \widetilde{G}^T(\lambda) \end{bmatrix} \right]\!\!\right]. \tag{5.15}$$

Proof. The system $G^T(z)$ is described by the state equation

$$\begin{bmatrix} X(z)z \\ Y(z) \end{bmatrix} = \begin{bmatrix} A^T & C^T \\ B^T & D^T \end{bmatrix} \begin{bmatrix} X(z) \\ U(z) \end{bmatrix}.$$

It holds that

$$\left[\!\!\left[\begin{bmatrix} X(z) \\ U(z) \end{bmatrix}, \begin{bmatrix} X(z)z \\ Y(z) \end{bmatrix} \right]\!\!\right] = \left[\!\!\left[\begin{bmatrix} X(z) \\ U(z) \end{bmatrix}, \begin{bmatrix} X(z) \\ U(z) \end{bmatrix} \right]\!\!\right] \begin{bmatrix} A & B \\ C & D \end{bmatrix}.$$

Let the input $u(t)$ be equal to $\delta(t)$. Then this last equation can be written as

$$\left[\!\!\left[\begin{bmatrix} \Gamma^T(z) \\ 1 \end{bmatrix}, \begin{bmatrix} \Gamma^T(z)z \\ G^T(z) \end{bmatrix} \right]\!\!\right] = \left[\!\!\left[\begin{bmatrix} \Gamma(z) \\ 1 \end{bmatrix}, \begin{bmatrix} \Gamma(z) \\ 1 \end{bmatrix} \right]\!\!\right] \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} X_o & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix}.$$

The term on the left-hand side of this equation equals

$$\left[\!\!\left[\begin{bmatrix} \Gamma^T(z) z^{-1} \\ z^{-1} \end{bmatrix}, \begin{bmatrix} \Gamma^T(z) \\ G^T(z) z^{-1} \end{bmatrix} \right]\!\!\right].$$

We observe that z^{-1} is equal to the inverse Hambo operator transform of $N(\lambda)$ (as follows from Proposition 5.1). At the same time, z^{-1} is the inverse Hambo signal transform of $W_1(\lambda)$. Then it follows from Proposition 4.7 that $z^{-1}I\ \Gamma(z)$ is the Hambo inverse signal transform of $N(\lambda)\breve{\Gamma}(\lambda)$. Similarly, by definition of the Hambo operator transform, $G(z)z^{-1}$ is then the inverse signal transform of $\widetilde{G}(\lambda)W_1(\lambda)$. Using the Hambo signal transform isomorphism (Proposition 4.2), it therefore holds that

$$\left[\!\!\left[\begin{bmatrix} \Gamma^T(z) z^{-1} \\ z^{-1} \end{bmatrix}, \begin{bmatrix} \Gamma^T(z) \\ G^T(z) z^{-1} \end{bmatrix} \right]\!\!\right] = \left[\!\!\left[\begin{bmatrix} \breve{\Gamma}^T(\lambda) N^T(\lambda) \\ W_1^T(\lambda) \end{bmatrix}, \begin{bmatrix} \breve{\Gamma}^T(\lambda) \\ W_1^T(\lambda) \widetilde{G}^T(\lambda) \end{bmatrix} \right]\!\!\right]. \qquad \blacksquare$$

Again a dual formulation of this proposition is possible that uses expressions involving $\Delta(z)$ and X_c. Expression (5.15) can also be put in Sylvester equation form.

COROLLARY 5.9. *Consider a Hambo transform $\widetilde{G}(\lambda)$ of a system $G(z) \in RH_2$, with minimal state space realization $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ and observability Gramian $\tilde{X}_o$. Then*

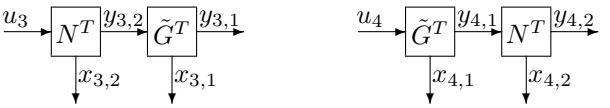


FIG. 5.2. *Systems for proof of Corollary* 5.9.

$G(z)$ has a minimal state space realization (A, B, C, D) that satisfies the following Sylvester equation:

$$\begin{bmatrix} \tilde{A}^T & \tilde{C}^T C_b^T \\ 0 & D_b^T \end{bmatrix} \begin{bmatrix} \tilde{X}_o A & \tilde{X}_o B \\ C & D \end{bmatrix} \begin{bmatrix} \tilde{A} & \tilde{B} B_b \\ 0 & D_b \end{bmatrix} + \begin{bmatrix} \tilde{C}^T A_b^T \\ B_b^T \end{bmatrix} \begin{bmatrix} \tilde{C} & \tilde{D} B_b \end{bmatrix} = \begin{bmatrix} \tilde{X}_o A & \tilde{X}_o B \\ C & D \end{bmatrix}. \tag{5.16}$$

Proof. The proof is similar to that of Corollary 5.7. Consider the systems shown in Figure 5.2. State equations of these systems are

$$\begin{bmatrix} x_{3,1}(t+1) \\ x_{3,2}(t+1) \end{bmatrix} = \begin{bmatrix} \tilde{A}^T & \tilde{C}^T C_b^T \\ 0 & D_b^T \end{bmatrix} \begin{bmatrix} x_{3,1}(t) \\ x_{3,2}(t) \end{bmatrix} + \begin{bmatrix} \tilde{C}^T A_b^T \\ B_b^T \end{bmatrix} u(t)$$

and

$$\begin{bmatrix} x_{4,1}(t+1) \\ x_{4,2}(t+1) \end{bmatrix} = \begin{bmatrix} \tilde{A}^T & 0 \\ B_b^T \tilde{B}^T & D_b^T \end{bmatrix} \begin{bmatrix} x_{4,1}(t) \\ x_{4,2}(t) \end{bmatrix} + \begin{bmatrix} \tilde{C}^T \\ B_b^T \tilde{D}^T \end{bmatrix} u(t),$$

respectively. The solution of (5.15) is then equal to

$$\left[\!\!\left[\begin{bmatrix} x_{3,1} \\ x_{3,2} \end{bmatrix}, \begin{bmatrix} x_{4,1} \\ x_{4,2} \end{bmatrix} \right]\!\!\right],$$

which results in (5.16). □

Existence of a solution to this Sylvester equation (5.16) is guaranteed if the systems in the inner product expression in (5.15) are stable. That this is true for $\breve{\Gamma}(\lambda)$ follows from the assumption that $G(z)$ is stable and the fact that $\Gamma^T(z)$ is stable. Consequently $\widetilde{G}(\lambda)$ is also stable. $W_1(\lambda)$ and $N^T(\lambda)$ are stable by definition.

Equation (5.16) can again be simplified further when $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ is output balanced.

Note that formulas (5.14) and (5.16) look very similar. Also note that the formulas are reciprocal: using a realization (A, B, C, D) in (5.14) results in a realization $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$, which, when used in (5.16), yields the original (A, B, C, D) again. This follows from the fact that the functions $\Gamma^T(z)$ and $\breve{\Gamma}(\lambda)$ correspond uniquely through the Hambo signal transform.

As stated, similar results as those given by Corollaries 5.7 and 5.9 can be given using a controllability approach. We state the results here without proof. Details can be found in [7].

COROLLARY 5.10 (Hambo system transform—controllability form [7]). *Consider a system $G(z) \in RH_2$ with minimal state space realization (A, B, C, D) and controllability Gramian X_c. Then its Hambo transform $\widetilde{G}(\lambda)$ has a minimal state*

space realization $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ *with controllability Gramian* $\tilde{X}_c = X_c$ *that satisfies the following Sylvester equation:*

$$\begin{bmatrix} A & 0 \\ C_b^T & A_b^T \end{bmatrix} \begin{bmatrix} \tilde{A}X_c & \tilde{B} \\ \tilde{C}X_c & \tilde{D} \end{bmatrix} \begin{bmatrix} A^T & 0 \\ B_b B^T & A_b \end{bmatrix} + \begin{bmatrix} B \\ C_b^T D \end{bmatrix} \begin{bmatrix} D_b B^T & C_b \end{bmatrix} = \begin{bmatrix} \tilde{A}X_c & \tilde{B} \\ \tilde{C}X_c & \tilde{D} \end{bmatrix}. \tag{5.17}$$

COROLLARY 5.11 (inverse Hambo system transform—controllability form [7]). *Consider a Hambo transform* $\widetilde{G}$ *of a system* $G(z) \in RH_2$ *with minimal state space realization* $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ *and controllability Gramian* $\tilde{X}_c$*. Then* $G(z)$ *has a minimal state space realization* (A, B, C, D) *with controllability Gramian* $X_c = \tilde{X}_c$ *that satisfies the following Sylvester equation:*

$$\begin{aligned} &\begin{bmatrix} \tilde{A} & 0 \\ C_b\tilde{C} & D_b \end{bmatrix} \begin{bmatrix} A\tilde{X}_c & B \\ C\tilde{X}_c & D \end{bmatrix} \begin{bmatrix} \tilde{A}^T & 0 \\ B_b^T\tilde{B}^T & D_b^T \end{bmatrix} + \begin{bmatrix} \tilde{B} \\ C_b\tilde{D} \end{bmatrix} \begin{bmatrix} A_b^T\tilde{B}^T & C_b^T \end{bmatrix} \\ &\qquad = \begin{bmatrix} A\tilde{X}_c & B \\ C\tilde{X}_c & D \end{bmatrix}. \end{aligned} \tag{5.18}$$

There are various formulas that can be derived in this context. For instance, it is straightforward to derive a generic formula for $\tilde{D}$ that is a direct result of substituting $\lambda = \infty$ in (5.2):

$$\tilde{D} = \sum_{k=0}^{\infty} g(k) A_b^k. \tag{5.19}$$

An equivalent relation (see [7]) can be derived for $\tilde{A}$ as defined by (5.14), (5.17) when we define $g_b(k)$ as the impulse response sequence of $G_b(z)$:

$$\tilde{A} = \sum_{k=0}^{\infty} g_b(k) A^k. \tag{5.20}$$

This expression can be verified as follows. Define $F = \sum_{k=0}^{\infty} g_b(k) A^k$, and consider the expression $A^T X_o F A$, where X_o is the observability Gramian of the realization (A, B, C, D). It follows that

$$\begin{aligned} A^T X_o F A &= \sum_{k=0}^{\infty} g_b(k)(A^T X_o A) A^k = \sum_{k=0}^{\infty} g_b(k)(X_o - C^T C) A^k \\ &= X_o F - C^T C \cdot D_b - \sum_{k=1}^{\infty} C^T C_b A_b^{k-1} B_b C A^k \\ &= X_o F - C^T D_b C - C^T C Y A, \quad \text{where} \quad A_b Y A + B_b C = Y. \end{aligned}$$

Evaluation of the terms in (5.14) yields that it must hold that $Y = \tilde{C}$ and $F = \tilde{A}$, as defined by (5.14). Analogously, evaluation of AFX_cA^T shows that $F = \tilde{A}$, as defined by (5.17).

6. Properties of Hambo transforms. We proceed with demonstrating a number of interesting properties of Hambo transforms that ensue from the theory developed in the preceding sections. These properties are of interest because they are instrumental to the application of the basis function theory in the context of system modelling [39, 8, 7].

6.1. Calculation rules. The Hambo operator transform obeys the following rules:

$$\text{(6.1)}\quad \text{if}\quad H(z) = (\alpha G_1(z) + \beta G_2(z)),\quad \text{then}\quad \widetilde{H}(\lambda) = \alpha\widetilde{G_1}(\lambda) + \beta\widetilde{G_2}(\lambda),$$

$$\text{(6.2)}\qquad \widetilde{G_1G_2}(\lambda) = \widetilde{G_1}(\lambda)\widetilde{G_2}(\lambda) = \widetilde{G_2}(\lambda)\widetilde{G_1}(\lambda),$$

$$\text{(6.3)}\qquad \widetilde{G^{-1}}(\lambda) = \big(\widetilde{G}(\lambda)\big)^{-1},$$

where $G_1(z), G_2(z), G(z), G^{-1}(z) \in H_2$, and $\alpha, \beta \in \mathbb{R}$.

Proof. (6.1): The proof follows trivially from the definition of the Hambo operator transform and the linearity of the Hambo signal transform.

(6.2): Let $Y(z) = G_1(z)G_2(z)U(z)$. Define $X(z) = G_2(z)U(z)$. By definition of the operator transform, it holds that $\breve{Y}(\lambda) = \widetilde{G_1}(\lambda)\breve{X}(\lambda) = \widetilde{G_1}(\lambda)\widetilde{G_2}(\lambda)\breve{U}(\lambda)$. Since this holds for all $U(z), Y(z)$, (6.2) follows. The second equality follows from the fact that the scalar systems $G_1(z)$ and $G_2(z)$ commute.

(6.3): Assuming that $G^{-1}(z) \in H_2$, we have by definition of the Hambo transform that $\breve{U}(\lambda) = \widetilde{G^{-1}}(\lambda)\breve{Y}(\lambda)$. We also know that $\breve{Y}(\lambda) = \widetilde{G}(\lambda)\breve{U}(\lambda)$. Hence $(\widetilde{G}(\lambda))^{-1} = \widetilde{G^{-1}}(\lambda)$. □

On the basis of these properties, it holds, for instance, that if $H(z) = (G(z)(1 + G(z))^{-1})$, then $\widetilde{H}(\lambda) = \widetilde{G}(\lambda)(I + \widetilde{G}(\lambda))^{-1} = (I + \widetilde{G}(\lambda))^{-1}\widetilde{G}(\lambda)$, assuming that $(1 + G(z))^{-1} \in H_2$.

These properties thus imply that parallel and series interconnections of systems remain unchanged under Hambo operator transformation. Feedback interconnections also remain unchanged under the condition that the inverse taken is also in H_2. It follows immediately that the same goes for linear fractional transformations (LFT), where we assume a pointwise definition of the operator transform for multivariable systems, i.e.,

$$\begin{bmatrix} \widetilde{G_{11}(z)\ G_{12}(z)} \\ G_{21}(z)\ \ G_{22}(z) \end{bmatrix} = \begin{bmatrix} \widetilde{G}_{11}(z) & \widetilde{G}_{12}(z) \\ \widetilde{G}_{21}(z) & \widetilde{G}_{22}(z) \end{bmatrix}.$$

6.2. Pole locations. It was established in section 5.2 that the McMillan degree of a Hambo operator transform is equal to the McMillan degree of the original system. Hence the number of poles of $\widetilde{G}(\lambda)$ is equal to that of $G(z)$. The locations of the poles of $\widetilde{G}(\lambda)$ are determined as follows.

PROPOSITION 6.1. *Consider a system $G(z) \in RH_2$ and a Hambo basis generated by an inner function $G_b(z)$. If $G(z)$ has a pole at $z = a_i$, its Hambo operator transform $\widetilde{G}(\lambda)$ will have a pole at $\mu_i = G_b(\frac{1}{a_i}) = G_b^{-1}(a_i)$.*

Proof. This assertion can be proved on the basis of (5.20). That is, if $G(z)$ has a state space realization (A, B, C, D), $\widetilde{G}(\lambda)$ will have a state space realization $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ with

$$\text{(6.4)}\qquad \tilde{A} = \sum_{k=0}^{\infty} g_b(k)A^k,$$

where $g_b(k)$ represents the impulse response sequence of $G_b(z)$. Consider any eigenvalue a_i of A and a corresponding eigenvector $x_i \in \mathbb{C}^n$, $Ax_i = x_i a_i$. If we multiply (6.4) from the right with x_i, we find $\tilde{A}x_i = x_i \sum_{k=0}^{\infty} g_b(k)a_i^k = x_i G_b(\frac{1}{a_i}) = x_i G_b^{-1}(a_i)$. Therefore, $\tilde{A}$ has eigenvalue $G_b(\frac{1}{a_i})$ with corresponding eigenvector x_i. □

COROLLARY 6.2. *The Hambo operator transform $\widetilde{G}(\lambda)$ of a system $G(z) \in H_2$ is stable.*

Proof. By the maximum modulus theorem [34] it holds that for an inner function $G_b(z)$, $|G_b(z)| < 1$ outside the unit disk. Hence $|G_b(\frac{1}{a})| < 1$ for all $a < 1$. Consequently, $\widetilde{G}(\lambda)$ is stable if $G(z)$ is stable. □

Corollary 5.2 showed that for $G(z) \in H_{2-}$, $\breve{G}(\lambda) = \widetilde{G}(\lambda)W_0(\lambda)$. Since this must be an element of $H_2^{n_b}$, it therefore must hold that the unstable pole of W_0 which lies at $\frac{1}{D_b}$ is cancelled by a zero at $\frac{1}{D_b}$ of $\det \widetilde{G}(\lambda)$, and hence we can immediately conclude that the poles of $\breve{G}(\lambda)$ constitute a subset of the poles of $\widetilde{G}(\lambda)$.

COROLLARY 6.3. *Let $G(z) \in H_{2-}$ have McMillan degree n with poles at a_i with $1 \leq i \leq n$. Then the Hambo signal transform $\breve{G}(\lambda)$ is stable, and its poles form a subset of $\{\mu_i\}_{1 \leq i \leq n}$ with $\mu_i = G_b(\frac{1}{a_i})$. Hence the McMillan degree of $\breve{G}(\lambda)$ is smaller than or equal to n.*

On the basis of Corollary 6.3, one can make the following statement about the convergence rate of an expansion in terms of Hambo basis functions [13].

PROPOSITION 6.4. *Let a Hambo basis function expansion of $G(z) \in H_{2-}$ be given by $G(z) = \sum_{k \in \mathbb{N}} \breve{g}^T(k) V_k(z) = \sum_{k \in \mathbb{N}} \sum_{i=1}^{n_b} \breve{g}(k)_i \Phi_{k,i}(z)$. Further, let $G(z)$ have McMillan degree n and poles $a_i, 1 \leq i \leq n$. Then with μ defined as $\mu = \max_{1 \leq i \leq n} |G_b(\frac{1}{a_i})|$, it holds that there exists a positive constant $c \in \mathbb{R}$ such that $\max_{1 \leq i \leq n_b} |\breve{g}(k)_i| \leq c\mu^{k-1}$.*

This is simply a result of the well-known fact that the convergence of an impulse response sequence is dominated by the pole with the largest modulus. If the poles of $G(z)$ are a subset of the poles $\xi_j, 1 \leq j \leq n_b$, of $G_b(z)$, then it holds that $G_b(\frac{1}{a_i}) = 0$ for $1 \leq i \leq n$. Hence it follows that in this case $\breve{g}(k) = 0$ for all $k > 1$, and the basis function expansion converges to zero in one step. This illustrates the mechanism that the convergence becomes very fast when the poles in the basis generating inner function lie close to the poles of $G(z)$.

6.3. Eigenstructure of Hambo operator transforms. In this section, we analyze some of the structural properties of Hambo operator transforms. A direct relation between the eigenvalues of a Hambo operator transform $\widetilde{G}(\lambda)$ and its preimage $G(z)$ is established. It is further shown how $\widetilde{G}(\lambda)$, evaluated on the unit circle, can be diagonalized by means of a similarity transformation with an orthogonal matrix, thus revealing information about the singular values of the Hambo operator transform. We first observe the following result which was previously shown to hold in [43, 44].

LEMMA 6.5. *Given a Hambo basis generating inner function $G_b(z)$ and its corresponding dual basis generating inner function $N(\lambda)$, for $z \neq 0$*

$$zV_1^T(z)\, N(\lambda)|_{\lambda^{-1}=G_b(z)} = V_1^T(z). \tag{6.5}$$

Proof. The proof follows by direct evaluation of $N(G_b(z))V_1(z)$ using $G_b(z) = C_b(zI - A_b)^{-1}B_b + D_b$, making the assumption that $z \notin \sigma(A_b)$:

$$\begin{aligned} N(G_b(z))V_1(z) &= \left(A_b + B_b(G_b(z) - D_b)^{-1}C_b\right)(zI - A_b)^{-1}B_b \\ &= A_b(zI - A_b)^{-1}B_b + B_b\left(C_b(zI - A_b)^{-1}B_b\right)^{-1} C_b(zI - A_b)^{-1}B_b \\ &= A_b(zI - A_b)^{-1}B_b + B_b = V_1(z)z. \end{aligned}$$

By the inner property of $N(\lambda)$ and $G_b(z)$, this latter equation can be rephrased as (6.5). Since A_b has only a finite number of eigenvalues, continuity shows that the result is valid for all $z \in \mathbb{C}$. ∎

We see that for $z \neq 0$, $V_1^T(z)$ is a left eigenvector of $N(\lambda)|_{\lambda^{-1}=G_b(z)}$, with z^{-1} the corresponding eigenvalue. This has the following consequence.

PROPOSITION 6.6. *Consider a Hambo basis generated by the inner function $G_b(z)$ and a transfer function $G(z) \in H_2$. Then the Hambo operator transform $\widetilde{G}(\lambda)$ satisfies*

$$V_1^T(z)\ \widetilde{G}(\lambda)\Big|_{\lambda^{-1}=G_b(z)} = G(z)V_1^T(z) \tag{6.6}$$

for all $z \neq 0$.

Proof. It follows by direct substitution of Lemma 6.5 in Proposition 5.1 that

$$V_1^T(z)\ \widetilde{G}(\lambda)\Big|_{\lambda^{-1}=G_b(z)} = \sum_{\tau=0}^{\infty} g(\tau)V_1^T(z)\ N(\lambda)^\tau|_{\lambda^{-1}=G_b(z)} = \sum_{\tau=0}^{\infty} g(\tau)z^{-\tau}V_1^T(z). \qquad ∎$$

Consider a certain fixed value of λ denoted as λ_0. Because $G_b(z)$ is an inner function of McMillan degree n_b, the equation ${\lambda_0}^{-1} = G_b(z)$ will have n_b solutions which we will denote as z_i. Defining the matrix $X(\{z_i\})$ as $X(\{z_i\}) = \begin{bmatrix} V_1(z_1)\ V_1(z_2)\ \cdots\ V_1(z_{n_b}) \end{bmatrix}$, one can write, using Proposition 6.6, $X^T(\{z_i\})\widetilde{G}(\lambda_0) = \operatorname{diag} G(z_i)X^T(\{z_i\})$. If the solutions z_i to ${\lambda_0}^{-1} = G_b(z)$ are distinct, it holds that $V_1^T(z_i)V_1(\frac{1}{z_j}) = 0$ for $z_i \neq z_j$. This follows directly from the following result, which is known as the *Christoffel–Darboux* formula [6, 4] for the Hambo basis. It gives an expression for the reproducing kernel of the subspace spanned by the functions $\Phi_{1,i}(z), 1 \leq i \leq n_b$, which is equal to $K(z, z') = V_1^T(z')V_1(\frac{1}{z})$.

LEMMA 6.7 (Christoffel–Darboux formula). *Consider a Hambo basis generating inner function $G_b(z)$. It holds for all $z_1, z_2 \in \mathbb{C}$, $z_1 \neq z_2$, that*

$$V_1^T(z_1)V_1\left(\frac{1}{z_2}\right) = \frac{G_b(z_1)G_b(\frac{1}{z_2}) - 1}{1 - \frac{z_1}{z_2}}. \tag{6.7}$$

Proof. The proof follows from the properties of the orthogonal realization (A_b, B_b, C_b, D_b). Using that $zV_1(z) = A_bV_1(z) + B_b$, we have that

$$\begin{aligned} z_1V_1^T(z_1)V_1\left(\frac{1}{z_2}\right)\ \frac{1}{z_2} &= (V_1^T(z_1)A_b^T + B_b^T)\left(A_bV_1\left(\frac{1}{z_2}\right) + B_b\right) \\ &= V_1^T(z_1)A_b^TA_bV_1\left(\frac{1}{z_2}\right) + V_1^T(z_1)A_b^TB_b + B_b^TA_bV_1\left(\frac{1}{z_2}\right) + B_b^TB_b. \end{aligned}$$

Substituting $A_b^TA_b = I - C_b^TC_b$, $A_b^TB_b = -C_b^TD_b$, and $B_b^TB_b = 1 - D_b^TD_b$ results in

$$z_1V_1^T(z_1)V_1\left(\frac{1}{z_2}\right)\ \frac{1}{z_2} = V_1^T(z_1)V_1\left(\frac{1}{z_2}\right) - G_b(z_1)G_b\left(\frac{1}{z_2}\right) + 1,$$

which can be rephrased as (6.7). ∎

We now have that, if the solutions z_i to $\lambda_0^{-1} = G_b(z)$ are distinct, it holds that

$$X^T(\{z_i\})\widetilde{G}(\lambda_0)X\left(\left\{\frac{1}{z_i}\right\}\right) = \operatorname{diag} G(z_i) \operatorname{diag} V_1^T(z_i)V_1\left(\frac{1}{z_i}\right).$$

The case where $|\lambda_0| = 1$ is a simple but important situation for which it holds that the solutions z_i to $\lambda_0^{-1} = G_b(z)$ are all distinct. This follows directly from the fact that any scalar inner function with McMillan degree n_b can be written as a Blaschke product $G_b(z) = \pm\prod_{k=1}^{n_b} \frac{1-\xi_k^* z}{z-\xi}$, and thus the map $e^{i\omega} \to G_b(e^{i\omega})$ will go through the unit circle n_b times as ω goes from 0 to 2π, and hence there are n_b different solutions $0 \le \omega_1 < \omega_2 < \cdots < \omega_{n_b} < 2\pi$ with $G_b(e^{i\omega_k}) = 1$.

A further consequence of the observation that $V_1^T(z)V_1(\frac{1}{z}) > 0$ if $|z| = 1$ is that for $|\lambda_0| = 1$, the matrix $X(\{\frac{1}{z_i}\})(\operatorname{diag}\sqrt{V_1^T(z_i)V_1(\frac{1}{z_i})})^{-1}$ is an orthogonal matrix. This brings us the following diagonal decomposition of $\widetilde{G}(\lambda_0)$.

PROPOSITION 6.8. *Let z_i, with $1 \le i \le n_b$, be the solutions to $\lambda_0^{-1} = G_b(z_i)$ with $|\lambda_0| = 1$. Then, defining $R = \operatorname{diag}\sqrt{V_1^T(z_i)V_1(\frac{1}{z_i})}$, it holds that*

$$R^{-1}X^T(\{z_i\})\widetilde{G}(\lambda_0)X\left(\left\{\frac{1}{z_i}\right\}\right)R^{-1} = \operatorname{diag} G(z_i).$$

$X(\{\frac{1}{z_i}\})R^{-1}$ is an orthogonal matrix. Hence the singular values of $\widetilde{G}(\lambda_0)$ are equal to $|G(z_i)|$.

This proposition also shows that $\widetilde{G}(\lambda)$ is Hermitian when $|\lambda| = 1$.

6.4. Norm invariance under Hambo operator transformation. It was shown before that the Hambo transforms of scalar stable finite dimensional LTI systems are again stable finite dimensional LTI systems, albeit that they have input/output dimension $n_b \times n_b$. For the particular case of the Hambo operator transform, it was further shown that the McMillan degree, Hankel singular values, and ℓ_2-gain are also invariant under Hambo operator transformation. This leads to the following observations.

COROLLARY 6.9. *The Hankel and H_∞-norms of a system $G(z) \in H_\infty$ are invariant under Hambo operator transformation.*

The assertion for the Hankel norm follows from invariance of the Hankel singular values. Invariance of the H_∞-norm follows from the fact that the H_∞-norm is equal to the ℓ_2-gain. Alternatively, it follows from Proposition 6.8, which shows that $\sup_{\omega\in[0,2\pi)} \bar{\sigma}(\widetilde{G}(e^{i\omega})) = \sup_{\omega\in[0,2\pi)} |G(e^{i\omega})|$. Given the definition of the Hambo operator transform, it is not surprising that these norms are invariant as they are both norms that are induced by the ℓ_2-norm for signals, which is invariant under Hambo signal transformation as follows, e.g., from Proposition 4.2. It is important to take notice of the fact that the H_2-norm is *not* invariant under Hambo *operator* transformation. On the basis of Proposition 5.3, we can, however, conclude the following.

COROLLARY 6.10. *The H_2-norm of the Hambo operator transform of $G(z) \in H_2$ satisfies*

$$\|\widetilde{G}\|_2 = \|V_k G\|_2 \quad \forall k \in \mathbb{Z}.$$

Proof. The proof follows by taking the trace of both sides of (5.4) with $G_1(z) = G_2(z) = G(z)$. ∎

7. Extensions and derivatives. In this section, we briefly discuss some closely related subjects in the context of the Hambo transform theory.

Time-varying transforms. In [7] a more generalized transform theory is developed, where the transforms are directly based on the Takenaka–Malmquist functions, as discussed in section 2. The main difference with the Hambo transforms is that the

transforms for the generalized case turn out to be scalar time-varying operators instead of multivariable time-invariant systems.

Multivariable systems. In this paper, the Hambo operator transform has been restricted to the class of scalar systems. An important issue here is that for scalar systems the transformed system turns out to be an element of $H_2^{n_b \times n_b}$. While it is straightforward (see, e.g., [25, 7]) to define Hambo transforms for multivariable $p \times m$ systems, the transform will blow up to dimensions $pn_b \times mn_b$. An alternative method which does not increase the input/output dimension, using a time-varying transformation, is discussed in [7].

Unstable systems. This paper primarily considers stable systems. It is not difficult to extend the transformation formulas of section 5.3 to unstable systems as well. In fact, the same formulas are valid with the exception of systems that contain poles that are reciprocals of basis poles. The problem in the latter case is that the resulting transform may be a noncausal system. This is explained by the following example for the Laguerre basis functions.

Let a be the (stable) pole of the Laguerre basis functions (2.4), and let $G(z) \in H_2^\perp$ be given by $G(z) = \frac{1}{z^{-1}-a}$. The Hambo transform of $G(z)$ can be calculated with (5.2), using that $N(\lambda) = \frac{1+a\lambda}{\lambda+a}$. This results in a noncausal $\widetilde{G}(\lambda) = \frac{\lambda-a}{1+a^2}$. So, while the Hambo transform is still well defined, the state space formulas cannot be used as is.

Realization. In [37, 8, 7], the problems of exact and partial realization in terms of Hambo functions have been solved. This concerns the situation where a sequence of expansion coefficients $\{\breve{g}(k), k = 1, \ldots, N\}$ is given and a system $G(z)$ of minimal degree is sought such that the first N expansion coefficients of $G(z)$ coincide with the given set. Such a situation typically arises in an identification setting, as described in [39]. In fact, the state space relations described in section 5.3 are a direct spin-off of this research.

Frequency warping. The variable substitution of (5.3) is sometimes referred to as a *frequency transformation*, as it maps $\mathbb{T}$ to $\mathbb{T}$. With $z = e^{i\omega}$ and $\lambda = e^{i\vartheta}$, it holds that this transformation, defined as $\vartheta = \beta(\omega)$, constitutes a continuously differentiable nondecreasing (hence bijective) mapping from $\omega \in [0, 2\pi)$ to $\vartheta \in [0, 2n_b\pi)$. The properties of this β mapping, and in particular its inverse β^{-1}, are analyzed in [35], where it is used in a frequency domain approach to Hambo basis function modelling. A discrete set of equidistantly distributed frequency points in the ϑ domain is mapped by β^{-1} to a nonequidistantly distributed set of frequency points in the ω domain. This frequency distortion, or "warping" property, is exploited in [41] for the case $n_b = 1$ to enable the application of the fast Fourier transform (FFT) algorithm to nonuniformly spaced samples of a discrete time Fourier transform (DTFT).

(Future) applications. The theory on Hambo transforms proved to be a powerful tool in the derivation of variance expressions for identification in terms of orthogonal basis functions [39]. Furthermore, as stated before, this theory has been instrumental in the derivation of approximate realization algorithms that are based on expansions in orthonormal basis functions. In [7] it is shown that these algorithms can also be used to solve certain classes of interpolation problems. Other promising future directions for use of the transform theory are, for instance, the application of system identification in the transform domain, extending the results of [40, 11, 10], and control design in the transform domain, utilizing the property that any linear system can be transformed into a system with all poles located at the origin.

8. Conclusions. In this paper, we have analyzed a signals and systems transform that is induced by the Hambo functions. These functions, which are a special

case of the Takenaka–Malmquist functions, are induced by the balanced states of scalar inner (stable all-pass) functions and encompass the classical pulse, Laguerre, and Kautz functions. The induced signals and systems transforms generalize the Z-transform and the Laguerre transform to a multidimensional representation. The transforms have been analyzed in detail, providing insight into their structural properties. Explicit and efficient algorithms have been provided that enable the calculation of minimal state space realizations of the operator transform and its inverse.

REFERENCES

[1] H. Akçay and P. Heuberger, *A frequency-domain iterative identification algorithm using general orthonormal basis functions*, Automatica J. IFAC, 37 (2001), pp. 663–674.

[2] H. Belt, *Orthonormal Bases for Adaptive Filtering*, Ph.D. thesis, Eindhoven University of Technology, The Netherlands, 1997.

[3] P. Bodin and B. Wahlberg, *Thresholding in high order transfer function estimation*, in Proceedings of the 33rd IEEE Conference on Decision and Control, Lake Buena Vista, FL, 1994, pp. 3400–3405.

[4] A. Bultheel, P. González-Vera, E. Hendriksen, and O. Njåstad, *Orthogonal Rational Functions*, Cambridge Monogr. Appl. Comput. Math. 5, Cambridge University Press, Cambridge, UK, 1999.

[5] M. Campi, R. Leonardi, and L. Rossi, *Generalized super-exponential method for blind equalization using Kautz filters*, in Proceedings of the IEEE Signal Processing Workshop on Higher Order Statistics, Caesarea, Italy, 1999, pp. 107–111.

[6] P. J. Davis, *Interpolation and Approximation*, Dover, New York, 1975.

[7] T. de Hoog, *Rational Orthonormal Bases and Related Transforms in Linear System Modeling*, Ph.D. thesis, Delft University of Technology, Delft, The Netherlands, 2001.

[8] T. de Hoog, Z. Szabó, P. Heuberger, P. Van den Hof, and J. Bokor, *Minimal partial realization from generalized orthonormal basis function expansions*, Automatica J. IFAC, 38 (2002), pp. 655–669.

[9] Z. Fejzo and H. Lev-Ari, *Adaptive Laguerre-lattice filters*, IEEE Trans. Signal Process., 45 (1997), pp. 3006–3016.

[10] B. Fischer, *System Identification in Alternative Shift Operators with Applications and Some Other Topics*, Doctoral thesis, Luleå University of Technology, Luleå, Sweden, 1999.

[11] B. Fischer and A. Medvedev, *Laguerre shift identification of a pressurized process*, in Proceedings of the American Control Conference, Philadelphia, PA, 1998, pp. 1933–1937.

[12] P. Heuberger, *On Approximate System Identification with System Based Orthonormal Functions*, Ph.D. thesis, Delft University of Technology, Delft, The Netherlands, 1991.

[13] P. Heuberger and P. Van den Hof, *The Hambo transform: A signal and system transform induced by generalized orthonormal basis functions*, in Proceedings of the 13th IFAC World Congress, San Francisco, CA, 1996, pp. 357–362.

[14] P. Heuberger, P. Van den Hof, and O. Bosgra, *A generalized orthonormal basis for linear dynamical systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 451–465.

[15] B. Ho and R. Kalman, *Effective construction of linear state-variable models from input/output functions*, Regelungstechnik, 14 (1966), pp. 545–592.

[16] T. Kailath, A. Sayed, and B. Hassibi, *Linear Estimation*, Prentice Hall Information and System Sciences Series, Prentice Hall, Upper Saddle River, NJ, 2000.

[17] R. Kalman, P. Falb, and M. Arbib, *Topics in Mathematical System Theory*, International Series in Pure and Applied Mathematics, McGraw-Hill, New York, 1969.

[18] W. Kautz, *Transient synthesis in the time domain*, IRE Transactions on Circuit Theory, 1 (1954), pp. 29–39.

[19] Y. Lee, *Statistical Theory of Communication*, Wiley, New York, 1960.

[20] F. Malmquist, *Sur la détermination d'une classe de fonctions analytiques par leurs valeurs dans un ensemble donné de points*, in Comptes Rendus du Sixième Congrès des Mathématiciens Scandinaves (Copenhagen, 1925), Jul. Gjellerups Forlag, Copenhagen, Denmark, 1926, pp. 253–259.

[21] R. Merched and A. Sayed, *Fast rls Laguerre adaptive filtering*, in Proceedings of the Allerton Conference on Communication, Control and Computing, Allerton, IL, 1999, pp. 338–347.

[22] B. Moore, *Principal component analysis in linear systems: Controllability, observability, and model reduction*, IEEE Trans. Automat. Control, 26 (1981), pp. 17–32.

[23] R. Murray-Smith and T. A. Johansen (Eds.), *Multiple Model Approaches to Modelling and Control*, Taylor and Francis, London, 1997.

[24] O. Nelles and M. Tomizuka, *On the dynamics of local linear model networks with orthonormal basis functions*, in Proceedings of the IFAC Symposium on System Identification SYSID 2000, Santa Barbara, CA, 2000.

[25] B. Ninness, S. Gómez, and J.-C. Weller, *MIMO system identification using orthonormal basis functions*, in Proceedings of the 34th IEEE Conference on Decision and Control, New Orleans, LA, 1995, pp. 703–708.

[26] B. Ninness and F. Gustafsson, *A unifying construction of orthonormal bases for system identification*, IEEE Trans. Automat. Control, 42 (1997), pp. 515–521. An extended version is available as Technical report EE9433, Department of Electrical and Computer Engineering, University of Newcastle, Australia, 1994.

[27] B. Ninness and H. Hjalmarsson, *Accurate quantification of variance error*, in Proceedings of the 15th IFAC World Congress, Barcelona, Spain, 2002.

[28] B. Ninness and H. Hjalmarsson, *Exact quantification of variance error*, in Proceedings of the 15th IFAC World Congress, Barcelona, Spain, 2002.

[29] Ü. Nurges, *Laguerre models in problems of approximation and identification of discrete systems*, Autom. Remote Control, 48 (1987), pp. 346–352.

[30] Y. Nurges and Y. Yaaksoo, *Laguerre state equations for a multivariable discrete system*, Autom. Remote Control, 42 (1981), pp. 1601–1603.

[31] T. Paatero, M. Karjalainen, and A. Härmä, *Modeling and equalization of audio systems using Kautz filters*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, 2001, pp. 434–437.

[32] M. Padmanabhan, K. Martin, and G. Peceli, *Feedback-Based Orthogonal Digital Filters: Theory, Applications and Implementation*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[33] R. Roberts and C. Mullis, *Digital Signal Processing*, Addison-Wesley Series in Electrical Engineering, Addison-Wesley, Reading, MA, 1987.

[34] W. Rudin, *Real and Complex Analysis*, 3rd ed., McGraw-Hill, New York, 1987.

[35] F. Schipp, L. Gianone, J. Bokor, and Z. Szabó, *Identification in generalized orthogonal basis—a frequency domain approach*, in Proceedings of the 13th IFAC World Congress, San Francisco, CA, 1996, pp. 387–392.

[36] C. Seretis and E. Zafiriou, *Nonlinear dynamical system identification using reduced Volterra models with generalized orthonormal basis functions*, in Proceedings of the American Control Conference, Albuquerque, NM, 1997, pp. 3042–3046.

[37] Z. Szabó, P. Heuberger, J. Bokor, and P. Van den Hof, *Extended Ho-Kalman algorithm for systems represented in generalized orthonormal bases*, Automatica J. IFAC, 36 (2000), pp. 1809–1818.

[38] S. Takenaka, *On the orthogonal functions and a new formula of interpolation*, Japan. J. Math., 2 (1925), pp. 129–145.

[39] P. Van den Hof, P. Heuberger, and J. Bokor, *System identification with generalized orthonormal basis functions*, Automatica J. IFAC, 31 (1995), pp. 1821–1834.

[40] M. Verhaegen, D. Westwick, and R. Kearney, *The use of a bilinear transformation of the shift operator in subspace model identification*, IEEE Trans. Automat. Control, 40 (1995), pp. 1422–1432.

[41] T. Von Schroeter, *Frequency warping with arbitrary allpass maps*, IEEE Signal Processing Letters, 6 (1999), pp. 116–118.

[42] B. Wahlberg, *System identification using Kautz models*, IEEE Trans. Automat. Control, 39 (1994), pp. 1276–1282.

[43] B. Wahlberg, *Orthonormal basis functions models: A transformation analysis*, in Proceedings of the 14th IFAC World Congress, Vol. H, Beijing, 1999, pp. 355–360.

[44] B. Wahlberg, *Orthonormal basis functions models: A transformation analysis*, SIAM Rev., 45 (2003), to appear.

[45] J. Walsh, *Interpolation and Approximation by Rational Functions in the Complex Domain*, 2nd ed., AMS, Providence, RI, 1956.

[46] L. Xie and L. Ljung, *Asymptotic variance expressions for estimated frequency functions*, IEEE Trans. Automat. Control, 46 (2001), pp. 1887–1899.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1374–1394

WHEN IS A STORAGE FUNCTION A STATE FUNCTION IN DISCRETE TIME?*

OSAMU KANEKO† AND TAKAO FUJII†

Abstract. The purpose of this paper is to investigate when a storage function is a state function *in discrete time*. As shown by Trentelman and Willems [*Systems Control Lett.*, 32 (1997), pp. 249–259], [*SIAM J. Control Optim.*, 36 (1998), pp. 1703–1749], every storage function is a state function in continuous time. At first glance, the same claim seems to hold also in discrete time. Contrary to this expectation, this is not true in general. In fact, the discrete time counterpart involves not only some different but also some more difficult issues compared with the continuous time case. This paper addresses these issues exactly and shows that every nonnegative storage function is a state function of a supply rate with a linear time-invariant dynamical system in discrete time.

Key words. dissipativeness, storage functions, state functions, discrete time, quadratic difference forms, two-variable polynomial matrices, behavioral approach, polynomial matrices

AMS subject classifications. 93A10, 93A30, 93C05, 93C55, 93C35, 15A63, 15A23

DOI. 10.1137/S0363012900371502

1. Introduction. Dissipativeness is one of the most important properties in dynamical systems (cf. [1], [2], [10], [14], [15], [16], [17], [23], [24]). The reason for this is that various important system characteristics, such as bounded realness, positive realness, and so on, can be formalized as dissipativeness. Intuitively, dissipativeness means that a dynamical system always dissipates energy at a *dissipation rate* for a given supplied power called a *supply rate*. This is equivalent to saying that an increase in stored energy, called a *storage function*, cannot exceed the supplied power. A storage function measures the amount of energy that is stored inside the system at any instant of time. Thus it is reasonable to expect that a storage function can be described by using an internal variable that stores the information of past trajectories, that is, state variables. In the standard system theory, it is presumed that *a storage function is described by a quadratic function of state variables.* By using such a static quadratic form, control system synthesis and analysis based on dissipativeness can be reduced to solve certain linear matrix inequalities, or equalities, which can be handled easily by numerical computations. From a theoretical point of view, the above presumption should be proved. In this connection, Trentelman and Willems proved the fundamental fact based on quadratic differential forms in [10] and [23] that *every storage function is a state function.* This result is very interesting in that the proof is self-contained via quadratic differential forms in a behavioral framework. Furthermore, this result has been used to develop various applications of dissipativeness via quadratic differential forms, for instance the generalized Pick matrix condition for halfline nonnegativity (cf. [23]), H^∞ control in a behavioral context (cf. [11]), J-spectral factorization (cf. [13]), deterministic Kalman filtering (cf. [3]), H^∞ filtering (cf. [12]), the KYP lemma, and so on. Similarly to the standard system theory, these applications are very effective for synthesizing desired systems or

*Received by the editors April 26, 2000; accepted for publication (in revised form) February 16, 2003; published electronically October 2, 2003. This research was supported by Grant-in-Aid for Scientific Research 13750417 from the Japan Society for the Promotion of Science.
http://www.siam.org/journals/sicon/42-4/37150.html
†Graduate School of Engineering Science, Osaka University, Machikaneyama 1, Toyonaka, 560-8531, Osaka, Japan (kaneko@sys.es.osaka-u.ac.jp, fujii@sys.es.osaka-u.ac.jp).

filters in a behavioral framework. In view of the fact that the behavioral approach is more general than the standard systems approach, these applications are also useful in more generalized settings.

The above types of results for dissipativeness are usually obtained in continuous time. The discrete time version of dissipativeness also plays a crucial role, since it is used to derive various synthesis and analysis tools for discrete time systems. In particular, the filtering problems are included in specific topics in discrete time. The reason for this is that filters are used to obtain desired signals from noisy discrete time data. If we can develop discrete time dissipativeness, it may enable us, as in the continuous time case (cf. [3], [12]), to obtain useful filtering algorithms which are applicable to the actual discrete time data in a behavioral framework. Moreover, interpolation problems like Nevanlinna–Pick are also deeply related to discrete time dissipativeness. In fact, Rapisarda and Willems considered subspace Nevanlinna–Pick interpolation problems and derived the most powerful unfalsified model based on dissipativeness and quadratic differential forms in continuous time (cf. [9]). The notion of the most powerful unfalsified model is also a topic relevant to discrete time data. Thus, discrete time dissipativeness will be useful for solving discrete time interpolation problems and obtaining effective algorithms applicable to actual data.

From this motivation, the authors suggested quadratic difference forms in [5] in order to develop discrete time dissipativeness. Although we derived basic properties there, many important properties that should be clarified still remain to be studied. In general, some of the important results in the continuous time case can be easily connected to those in the discrete time case (cf. [1], [6]). It is, however, probable that there do exist some differences in other important results related to dissipativeness between the continuous and the discrete time cases. From a theoretical point of view, they must be clarified. The relationship between states and storage functions is one such topic. In the continuous time case, as stated above, Trentelman and Willems have shown in [10] and [23] that every storage function is a state function. At first glance, it seems easy to prove that every storage function is a state function in discrete time in a similar way as in the continuous time case. If this is true, many applications of discrete time dissipativeness can be developed. The fact is, however, that every storage function is not always a state function in discrete time, as shown by a counterexample in section 4. There may be not only different issues from the continuous time case but also even more difficult issues. This paper addresses such issues by clarifying when a storage function is a state function in discrete time. With regard to this problem, we show that every *nonnegative* storage function is a static quadratic function of any state variable of a given dynamical system augmented with a given supply rate in discrete time. Similarly to the continuous time case, various applications of discrete time dissipativeness involve conditions on storage functions. Therefore, the results in this paper enable us to obtain more treatable conditions. Moreover, our results are obtained using only quadratic difference forms and some materials within behavioral system theory. Their derivations are thus self-contained in a behavioral framework, and hence interesting and meaningful from a theoretical point of view.

This paper is organized as follows. In section 2, some pertinent facts related to discrete time dynamical systems and quadratic difference forms are prepared by using the results obtained in [5], [7], [8], [10], [21], [22], and [23]. In section 3, the basic properties of discrete time dissipativeness are explained and the existence of extremal storage functions is shown for a certain class of supply rates, which is used as a lemma

to prove our main result. In section 4, we explain the problem we are trying to attack here. We then provide an illustrative example to demonstrate that *not* every storage function is a storage function for a given supply rate in discrete time. In section 5, we provide a sufficient condition for a storage function to be a state function by showing that *every nonnegative storage function is a state function in discrete time.* We also provide another sufficient condition on dissipation rates. Since the nonnegativity of storage functions is a key concept in this paper, we derive a necessary and sufficient condition for the existence of nonnegative storage functions as well as that for a storage function to be nonnegative in section 6. In section 7, we apply the results obtained in the previous sections to a more general case. Concluding remarks and open issues are stated in section 8. Notations and a brief review of behavioral system theory are given in Appendices A and B, respectively. The detailed proofs of theorems are shown in Appendix C.

2. Preliminaries.

2.1. Discrete time dynamical systems and state functions. In this paper, we assume that a dynamical system $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathfrak{B})$ is linear, time-invariant, complete, and controllable. We note that Σ is controllable if and only if $\mathfrak{B}$ can be represented by

$$w = M(\sigma)\ell, \tag{1}$$

where $M(\xi) \in \mathbb{R}^{q\times d}[\xi]$. Here, ℓ is observable from w if and only if $M(\lambda)$ is full column rank for all $\lambda \in \mathbb{C}$. In the following, we give a necessary and sufficient condition for a given $f = F(\sigma)\ell$ to be a state function of a system represented by (1).

PROPOSITION 2.1. *Let $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathfrak{B})$ denote a controllable dynamical system and* (1) *be an image representation of $\mathfrak{B}$. Assume that $M(\lambda)$ is full column rank for all $\lambda \in \mathbb{C}$, which implies that there exists a partition of $M(\xi)$ into $M(\xi) = col\big[\, U(\xi), \quad Y(\xi)\,\big]$ such that $U(\xi)$ is nonsingular and $Y(\xi)U(\xi)^{-1}$ is proper, possibly after permuting the components of w appropriately and, accordingly, the rows of $M(\xi)$. Let $\mathfrak{B}_s$ denote the full behavior of a minimal state space system whose manifest behavior is $\mathfrak{B}$. Let $F(\xi) \in \mathbb{R}^{\bullet\times d}[\xi]$ induce $f = F(\sigma)\ell$, and consider the following new image representation*:

$$\operatorname{col}\begin{bmatrix} w, & f \end{bmatrix} = \operatorname{col}\begin{bmatrix} M(\sigma), & F(\sigma) \end{bmatrix}\ell.$$

Let $\mathfrak{B}_{ext}$ be the manifest behavior of the above image representation. Then, there exists a real constant matrix H such that $f = Hx$ for all f and x for which there exists $w \in \mathfrak{B}$ such that $(w, f) \in \mathfrak{B}_{ext}$ and $(w, x) \in \mathfrak{B}_s$ if and only if $F(\xi)U(\xi)^{-1}$ is strictly proper.

Proof. The statement is straightforward from the continuous time case in section 8 in [8] and Theorem 5.2 in [10]. □

2.2. Quadratic difference forms and two-variable polynomial matrices. Quadratic difference forms are appropriate mathematical tools related to discrete time dissipativeness. They are used throughout this paper, so here we briefly introduce some necessary definitions and properties. See [23] and [5] for more details of the continuous time case and the discrete time case, respectively.

An element of $\mathbb{R}^{p\times q}[\zeta, \eta]$ is given by

$$\Phi(\zeta, \eta) = \sum_{k,l=0} \Phi_{kl}\zeta^k\eta^l. \tag{2}$$

The sum in (2) ranges over nonnegative integers and is assumed to be finite, and $\Phi_{kl} \in \mathbb{R}^{q\times q}$. For $\Phi(\zeta,\eta) \in \mathbb{R}^{p\times q}[\zeta,\eta]$, let $\mathbb{R}_s^{q\times q}[\zeta,\eta]$ denote the set of two-variable polynomial matrices satisfying $\Phi(\zeta,\eta) = \Phi(\eta,\zeta)^T$. For all $w \in (\mathbb{R}^q)^Z$, $\Phi(\zeta,\eta) \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$ induces a quadratic difference form $Q_\Phi : (\mathbb{R}^q)^Z \mapsto \mathbf{R}^Z$ as defined by

$$Q_\Phi(w)(t) := \sum_{k,l=0} w(t+k)^T \Phi_{kl} w(t+l).$$

Given $\Phi(\zeta,\eta) \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$, by replacing the indeterminates ζ and η with ξ^{-1} and ξ, respectively, we obtain a one-variable dipolynomial matrix $\Phi(\xi^{-1},\xi) \in \mathbb{R}^{q\times q}[\xi^{-1},\xi]$.

For a given arbitrary $\Phi(\zeta,\eta) \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$, we define the following infinite matrix:

$$\tilde{\Phi} := \begin{bmatrix} \Phi_{0,0} & \Phi_{0,1} & \cdots & \Phi_{0,l} & \cdots \\ \Phi_{0,1}^T & \Phi_{1,1} & \cdots & \Phi_{0,l} & \cdots \\ \vdots & & & \vdots & \\ \Phi_{k,0} & & \cdots & \Phi_{k,l} & \cdots \\ \vdots & & & \vdots & \end{bmatrix}.$$

Here all but a finite number of the elements of $\tilde{\Phi}$ are zero. Since we concentrate on the finite nonzero block of $\tilde{\Phi}$ in this paper, we regard $\tilde{\Phi}$ as an element of $\mathbb{R}^{q(N(\Phi)+1)\times q(N(\Phi)+1)}$, where $N(\Phi) := \min\{n' \in \mathbb{Z} \text{ such that } \Phi_{kl} = 0, \forall k \text{ and } l > n'\}$. We call $\tilde{\Phi}$ the coefficient matrix of $\Phi(\zeta,\eta)$.

In a similar way to the constant symmetric matrix case, the nonnegativity and the positivity of a quadratic difference form induced by $\Phi(\zeta,\eta) \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$ are defined by

$$\Phi(\zeta,\eta) \geq 0 :\Leftrightarrow Q_\Phi(w)(t) \geq 0 \ \ \forall w \in (\mathbb{R}^q)^Z \text{ and } \forall t \in Z$$

and

$$\Phi(\zeta,\eta) > 0 :\Leftrightarrow \Phi(\zeta,\eta) \geq 0 \text{ and } Q_\Phi(w) = 0 \Rightarrow w = 0,$$

respectively. It follows from Proposition 2.1 in [5] that $\Phi(\zeta,\eta) \geq 0$ is equivalent to $\tilde{\Phi} \geq 0$.

For $\Phi(\zeta,\eta) = \sum_{k,l=0}^{N(\Phi)} \Phi_{kl}\zeta^k\eta^l \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$, factorize $\tilde{\Phi} = \tilde{M}^T\Sigma_\Phi\tilde{M}$ such that $\tilde{M} \in \mathbb{R}^{\bullet\times q(N(\Phi)+1)}$ is full row rank and $\det(\Sigma_\Phi) \neq 0$, i.e., $\operatorname{rank}(\Sigma_\Phi) = \operatorname{rank}(\tilde{\Phi})$. With such a factorization of $\tilde{\Phi}$, we obtain a canonical factorization (cf. [10], [23]) of $\Phi(\zeta,\eta)$ as

$$\Phi(\zeta,\eta) = M(\zeta)^T\Sigma_\Phi M(\eta), \tag{3}$$

where $M(\xi) := \tilde{M}\operatorname{col}\begin{bmatrix} I, & \xi I, & \ldots, & \xi^{n-1}I, & \xi^n I \end{bmatrix} \in \mathbb{R}^{\bullet\times q}[\xi]$. We call $M(\xi)$ a canonical factor of $\Phi(\zeta,\eta)$. Note that $\Phi(\zeta,\eta)$ has many canonical factors, and all of them can be obtained from one canonical factor by replacing $M(\xi)$ by $PM(\xi)$, where P is a nonsingular matrix such that $\Sigma_\Phi = P^T\Sigma_\Phi P$. Moreover, for an arbitrary canonical factor $M(\xi)$ and an arbitrary factor, say $M'(\xi)$, which is not necessarily a canonical factor, it is easy to show that there exists a constant matrix L such that $M(\xi) = LM'(\xi)$.

For our last point, we explain the observability of a quadratic difference form. Let $\Phi(\zeta,\eta) = M(\zeta)^T\Sigma_\Phi M(\eta) \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$ denote a canonical factorization. It is easy to show that if one canonical factor is full column rank for all $\lambda \in \mathbb{C}$, then the other canonical factors are also full column rank for all $\lambda \in \mathbb{C}$. In view of this fact, we call $\Phi(\zeta,\eta)$ *observable* if a canonical factor is observable.

3. Discrete time dissipativeness. First, we can regard Q_Φ induced by $\Phi(\zeta,\eta) \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$ as the power entering into the physical system $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathfrak{B})$. The reason for this is that the power can be described by characterizing a quadratic expression involving system variables and shifted variables. By using quadratic difference forms, we can formalize dissipativeness of dynamical systems in discrete time as follows (cf. Definition 3.1 in [5], Definition 4.2 in [10], and Definition 5.1 in [23]).

DEFINITION 3.1. *Let $\Phi(\zeta,\eta) \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$ induce a supply rate Q_Φ.*

1. *$\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathfrak{B})$ is said to be dissipative for Q_Φ if $\sum_{t=-\infty}^{t=\infty} Q_\Phi(w)(t) \geq 0$ for all $w \in \mathfrak{B} \bigcap l_2^q$.*
2. *Q_Ψ induced by $\Psi(\zeta,\eta) \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$ is said to be a storage function for Q_Φ and $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathfrak{B})$ if $Q_\Psi(w)(t+1) - Q_\Psi(w)(t) \leq Q_\Phi(w)(t)$ (called a dissipation inequality) holds for all $t \in \mathbb{Z}$ and for all $w \in \mathfrak{B}$.*
3. *Q_Δ induced by $\Delta(\zeta,\eta) \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$ is said to be a dissipation rate for Q_Φ and $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathfrak{B})$ if $\sum_{t=-\infty}^{t=\infty} Q_\Phi(w)(t) = \sum_{t=-\infty}^{t=\infty} Q_\Delta(w)(t)$ and $Q_\Delta(w)(t) \geq 0$ for all $t \in \mathbb{Z}$ and for all $w \in \mathfrak{B} \bigcap l_2^q$.*

In the case of $\mathfrak{B} = (\mathbb{R}^q)^Z$, it follows from Lemma 3.1 in [5] that $\sum_{t=-\infty}^{t=\infty} Q_\Phi(w)(t) = \sum_{t=-\infty}^{t=\infty} Q_\Delta(w)(t)$ for all $w \in l_2^q$ is equivalent to the condition

$$\Phi(\lambda^{-1}, \lambda) = \Delta(\lambda^{-1}, \lambda) \tag{4}$$

for all nonzero $\lambda \in \mathbb{C}$.

Another important notion is "lossless," which means that the power supplied to the system can be stored as an increase of the internal energy of the system without dissipation.

DEFINITION 3.2. *$\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathfrak{B})$ is said to be lossless for a supply rate Q_Φ induced by $\Phi(\zeta,\eta) \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$ if $\sum_{t=-\infty}^{t=\infty} Q_\Phi(w)(t) = 0$ for all $w \in \mathfrak{B} \bigcap l_2^q$.*

The next theorem gives an interrelation between a supply rate, a storage function, and a dissipation rate (cf. Proposition 3.3 in [5], Theorem 4.3 in [10], and Proposition 5.2 in [23]).

THEOREM 3.3. *Let $\Phi(\zeta,\eta) \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$ induce a supply rate Q_Φ. Assume that a dynamical system $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathfrak{B})$ has an image representation $w = W(\sigma)\ell$, where $W(\lambda) \in \mathbb{R}^{q\times\bullet}[\lambda]$is full column rank for all $\lambda \in \mathbb{C}$. Then, the following four conditions are equivalent.*

1. *For all $w \in l_2^q \bigcap \mathfrak{B}$, $\sum_{t=-\infty}^{\infty} Q_\Phi(w)(t) \geq 0$.*
2. *$W(e^{-j\omega})^T \Phi(e^{-j\omega}, e^{j\omega}) W(e^{j\omega}) \geq 0$ for all $\omega \in [0, 2\pi)$.*
3. *Q_Φ and $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathfrak{B})$ admit a storage function.*
4. *Q_Φ and $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathfrak{B})$ admit a dissipation rate.*

Moreover, for the supply rate Q_Φ, the following one-one relation holds between storage functions Q_Ψ induced by $\Psi(\zeta,\eta) \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$ and dissipation rates Q_Δ induced by $\Delta(\zeta,\eta) \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$:

$$Q_\Psi(w)(t+1) - Q_\Psi(w)(t) = Q_\Phi(w)(t) - Q_\Delta(w)(t) \tag{5}$$

for all time $t \in \mathbb{Z}$ and $w \in \mathfrak{B}$ or, equivalently,

$$(\zeta\eta - 1) W(\zeta)^T \Psi(\zeta,\eta) W(\eta) = W(\zeta)^T \Phi(\zeta,\eta) W(\eta) - W(\zeta)^T \Delta(\zeta,\eta) W(\eta). \tag{6}$$

In the case of $\Sigma = (\mathbb{Z}, \mathbb{R}^q, (\mathbb{R}^q)^Z)$, (5) holds for all $w \in (\mathbb{R}^q)^Z$ and (6) is described by

$$(\zeta\eta - 1)\Psi(\zeta,\eta) = \Phi(\zeta,\eta) - \Delta(\zeta,\eta). \tag{7}$$

Here, we introduce some notation. For a supply rate Q_Φ induced by $\Phi(\zeta,\eta)$, we define

$$\mathcal{S}(\Phi) := \{\Psi(\zeta,\eta) \in \mathbb{R}^{q\times q}_s[\zeta,\eta] \mid Q_\Psi \text{ induced by } \Psi(\zeta,\eta) \text{ is a storage function for } Q_\Phi \text{ and } \Sigma = (\mathbb{Z},\mathbb{R}^q,(\mathbb{R}^q)^Z)\}$$

and

$$\mathcal{D}(\Phi) := \{\Delta(\zeta,\eta) \in \mathbb{R}^{q\times q}_s[\zeta,\eta] \mid Q_\Delta \text{ induced by } \Delta(\zeta,\eta) \text{ is a dissipation rate for } Q_\Phi \text{ and } \Sigma = (\mathbb{Z},\mathbb{R}^q,(\mathbb{R}^q)^Z)\}.$$

Next, for a fixed $\Delta(\zeta,\eta) \in \mathcal{D}(\Phi)$, let $\Psi(\Phi,\Delta)$ denote a two-variable polynomial matrix inducing the storage function corresponding to the dissipation rate induced by $\Delta(\zeta,\eta)$ for Q_Φ and Σ. Similarly, for a fixed $\Psi(\zeta,\eta) \in \mathcal{S}(\Phi)$, let $\Delta(\Phi,\Psi)$ denote a two-variable polynomial matrix inducing the dissipation rate corresponding to the storage function induced by $\Psi(\zeta,\eta)$ for Q_Φ and Σ.

Finally, we present the following theorem on the existence of extremal storage functions. This is not only a preparation of the proof of our main results (detailed discussions are given in the proof of Theorem 5.1) but also the self-standing important result.

THEOREM 3.4. *Let $\Phi(\zeta,\eta) \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ induce a supply rate Q_Φ for $\Sigma = (\mathbb{R},\mathbb{R}^q,(\mathbb{R}^q)^Z)$. Assume that $\Phi(e^{-j\omega},e^{j\omega}) > 0$ for all $\omega \in [0,2\pi)$. Then, there exist $\Psi^+(\zeta,\eta)$ and $\Psi^-(\zeta,\eta) \in \mathcal{S}(\Phi)$ satisfying $\Psi^-(\zeta,\eta) \le \Psi(\zeta,\eta) \le \Psi^+(\zeta,\eta)$ for any other $\Psi(\zeta,\eta) \in \mathcal{S}(\Phi)$. In addition, $\Psi^+(\zeta,\eta)$ and $\Psi^-(\zeta,\eta)$ are expressed by $\Psi^+(\zeta,\eta) = (\Phi(\zeta,\eta) - A(\zeta)^T A(\eta))/(\zeta\eta - 1)$ and $\Psi^-(\zeta,\eta) = (\Phi(\zeta,\eta) - H(\zeta)^T H(\eta))/(\zeta\eta - 1)$. Here $A(\xi)$ and $H(\xi)$ are spectral factors of $\Phi(\xi^{-1},\xi)$ such that $\Phi(\xi^{-1},\xi) = A(\xi^{-1})^T A(\xi) = H(\xi^{-1})^T H(\xi)$ and, moreover, $\det(A(\xi))$ is anti-Hurwitz and $\det(H(\xi))$ is Hurwitz.*

Proof. See Appendix C. □

4. Problem formulation. In the following sections, we treat a quadratic supply rate Q_Φ induced by $\Phi(\zeta,\eta) \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ and a (trivial) dynamical system $\Sigma = (\mathbb{Z},\mathbb{R}^q,(\mathbb{R}^q)^Z)$ for the purpose of simplifying discussions. In section 7, we show that simple modifications of discussions for the case of $\mathfrak{B} = (\mathbb{R}^q)^Z$ yield similar results in the case where $\mathfrak{B}$ can be described by an image representation $w = W(\sigma)\ell$.

First, we define a dynamical system induced by $\Phi(\zeta,\eta)$ and $\Sigma = (\mathbb{Z},\mathbb{R}^q,(\mathbb{R}^q)^Z)$ as

$$\Sigma_{supply} := (\mathbb{Z},\mathbb{R}^r,\mathrm{Im}(M(\sigma))),$$

where $M(\xi) \in \mathbb{R}^{r\times q}[\xi]$ is obtained from a canonical factorization $\Phi(\zeta,\eta) = M(\zeta)^T \Sigma_\Phi \times M(\eta)$, and $r := \mathrm{rank}(\Sigma_\Phi)$. In other words, the manifest behavior of Σ_{supply} can be described by an image representation $v = M(\sigma)w$ (we should note from the definition of Σ_{supply} that manifest variables "w" of Σ are regarded as latent variables of Σ_{supply}). Let $\mathfrak{B}_{\Phi,s}$ denote the full behavior of a minimal state space system whose manifest behavior is $\mathrm{Im}(M(\sigma))$. Under this setting, for $\Psi(\zeta,\eta) \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$, we investigate whether there exists a constant matrix $K = K^T \in \mathbb{R}^{\bullet\times\bullet}$ such that $(M(\sigma)w,x) \in \mathfrak{B}_{\Phi,s}$ implies $Q_\Psi(w) = x^T K x$. We denote the set of two-variable polynomial matrices inducing quadratic state functions of Σ_{supply} by

$$\mathcal{F}_s(\Phi) := \{\Psi(\zeta,\eta) \in \mathbb{R}^{q\times q}_s[\zeta,\eta] \mid \exists K = K^T \in \mathbb{R}^{\bullet\times\bullet} \text{ such that } (M(\sigma)w,x) \in \mathfrak{B}_{\Phi,s} \text{ implies } Q_\Psi(w) = x^T K x\}. \tag{8}$$

For any state x there exists a constant matrix H such that Hx is a minimal state (cf. Proposition 7.10 in [22]); thus $\Psi(\zeta,\eta) \in \mathcal{F}_s(\Phi)$ implies that Q_Ψ can also be described by a quadratic function of any state variables of Σ_{supply}.

Under the above preparations, the problem we attack in this study is formulated as follows.

PROBLEM 4.1. *Let* $\Phi(\zeta,\eta) \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$ *induce a supply rate, and consider* $\Psi(\zeta,\eta) \in \mathcal{S}(\Phi)$. *When is* $\Psi(\zeta,\eta) \in \mathcal{F}_s(\Phi)$ *in discrete time?*

Next, we make a comparison between the continuous and the discrete time cases. Let $\Psi(\zeta,\eta) = F(\zeta)^T\Sigma_\Psi F(\eta)$ and $\Delta(\zeta,\eta) = D(\zeta)^T D(\eta)$ denote canonical factorizations of two-variable polynomial matrices inducing a storage function and a dissipation rate, respectively. In continuous time, the dissipation equation can be described by

$$(\zeta+\eta)\Psi(\zeta,\eta) = \Phi(\zeta,\eta) - \Delta(\zeta,\eta) \tag{9}$$

(cf. Theorem 4.3 in [10]). To prove that a storage function is a state function, it suffices to show that $F(\xi)U(\xi)^{-1}$ is strictly proper in the case where $\Phi(\zeta,\eta)$ is observable (cf. Theorem 6.1 in [10]). First, by setting $\zeta = -\xi$ and $\eta = \xi$, (9) can be described by $M(-\xi)\Sigma_\Phi M(\xi) = D(-\xi)^T D(\xi)$. Next, premultiplying and postmultiplying it by $U(-\xi)^{-T}$ and $U(\xi)^{-1}$, respectively, we can see that $U(-\xi)^{-T}D(\xi)^T D(\xi)U(\xi)^{-1}$ is proper, which implies that $D(\xi)U(\xi)^{-1}$ is also proper. By using this properness and (9), we can also see that $U(\zeta)^{-T}F(\zeta)^T\Sigma_\Psi F(\eta)U(\eta)^{-1}$ is strictly proper with respect to ζ and η. It then follows from Theorem 6.1 in [10] that every storage function is a state function. Of course, even if $M(\xi)$ is not observable, the same property holds (cf. Theorem 6.1 in [10]). In discrete time, by using Theorem 3.3 and (4), the dissipation equation (7) can be described by $M(\xi^{-1})^T\Sigma_\Phi M(\xi) = D(\xi^{-1})^T D(\xi)$, which yields $U(\xi^{-1})^{-T}M(\xi^{-1})^T\Sigma_\Phi M(\xi)U(\xi)^{-1} = U(\xi^{-1})^{-T}D(\xi^{-1})^T D(\xi)U(\xi)^{-1}$. Here we should note that, unlike the continuous time case, the properness of $U(\xi^{-1})^{-T}D(\xi^{-1})^T \times D(\xi)U(\xi)^{-1}$ does not necessarily imply that of $D(\xi)U(\xi)^{-1}$. Hence, $U(\zeta)^{-T}F(\zeta)^T\Sigma_\Psi \times F(\eta)U(\eta)^{-1}$ is not necessarily proper with respect to ζ and η. In fact, we can give a simple counterexample as follows.

Example 4.1. Consider a supply rate induced by

$$\Phi(\zeta,\eta) = 1+\zeta+\eta+2\zeta\eta \tag{10}$$

and a dynamical system $\Sigma = (\mathbb{Z},\mathbb{R},\mathbb{R}^Z)$. One canonical factorization of $\Phi(\zeta,\eta)$ is given by $\Phi(\zeta,\eta) = M(\zeta)^T\Sigma_\Phi M(\eta)$, where $M(\xi) :=\mathrm{col}\begin{bmatrix} 1+\xi, & \xi \end{bmatrix}$ and $\Sigma_\Phi := I_2$. Defining $U(\xi) := 1+\xi$ enables us to observe that $M(\xi)U(\xi)^{-1}$is proper. Moreover, $(M(\xi^{-1})U(\xi^{-1}))^T$ is also proper, so $U(\xi^{-1})^{-T}\Phi(\xi^{-1},\xi)U(\xi)^{-1}$ is proper. Note that $u := U(\sigma)w = (1+\sigma)w$ is an input variable of $\Sigma_{supply} = (\mathbb{Z},\mathbb{R},\mathrm{Im}(M(\sigma)))$, with w being a latent variable of Σ_{supply}. Moreover, by taking $f = w$ in Proposition 2.1, we can see that w is a state variable of Σ_{supply}.

Here, as one $\Delta(\zeta,\eta)$ in $\mathcal{D}(\Phi)$, we can take $\Delta(\zeta,\eta) = 1+\zeta+\eta+\zeta\eta+\zeta^3\eta^3$. One canonical factorization of it is given by $\Delta(\zeta,\eta)=D(\zeta)^T D(\eta)$, where $D(\xi) :=\mathrm{col}[1+\xi, \xi^3]$.

It is clear that $D(\xi)U(\xi)^{-1}$ is not proper, while $U(\xi^{-1})^{-T}\Delta(\xi^{-1},\xi)U(\xi)^{-1}$ is proper. By using (7), $\Psi(\zeta,\eta) := \Psi(\Phi,\Delta)$ can be described by $\Psi(\zeta,\eta) = -\zeta\eta - \zeta^2\eta^2$. One canonical factorization of it is given by $\Psi(\zeta,\eta) = F(\zeta)^T\Sigma_\Psi F(\eta)$, where $F(\xi) :=\mathrm{col}[\xi,\xi^2]$, and $\Sigma_\Psi := -I_2$. Clearly, $F(\xi)U(\xi)^{-1} = \mathrm{col}[\frac{-1}{\xi+1}+1, \frac{1}{\xi+1}-(1-\xi)]$ is not strictly proper. Thus, it follows from Proposition 2.1 that $F(\sigma)w$ cannot be described by a static function of a state x. In addition, $F(\sigma)w$ can be written by

$F(\sigma)w = \mathrm{col}\begin{bmatrix} -x+u,\ x-(1-\sigma)u \end{bmatrix}$, implying $Q_\Psi(w) = -(-x+u)^2 - (x-(1-\sigma)u)^2$. Hence, we can see that $\Psi(\zeta,\eta) \notin \mathcal{F}_s(\Phi)$, which means that this storage function Q_Ψ cannot be described by a static quadratic function of a state x of Σ_{supply}.

5. Sufficient conditions for a storage function to be a state function. In this section, we show the sufficient conditions for Problem 4.1 as the main results of this paper. First, we give a sufficient condition for Problem 4.1 related to nonnegative storage functions.

THEOREM 5.1. *Let $\Phi(\zeta,\eta) \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ induce a supply rate Q_Φ. Assume that $\Phi(\zeta,\eta)$ is observable. Consider $\Psi(\zeta,\eta) \in \mathcal{S}(\Phi)$. If $\Psi(\zeta,\eta) \geq 0$, then $\Psi(\zeta,\eta) \in \mathcal{F}_s(\Phi)$.*

Proof. See Appendix C. □

Remark 5.1. We should consider the relationship between Theorem 5.1 and related studies on discrete time dissipativeness. In [2] (see also [1]), the notion of dissipative scattering (or passive scattering, respectively) was introduced as a particular dissipativeness and defined by $\|u(t)\|^2 - \|y(t)\|^2 \geq \|x(t+1)\|^2 - \|x(t)\|^2, (t \geq 0)$, where $u(t)$, $y(t)$, and $x(t)$ are inputs, outputs, and *states*, respectively, of the dynamical system described by $x(t+1) = Ax(t) + Bu(t)$, $y(t) = Cx(t) + Du(t)$. This inequality is a specific example of the dissipation inequality stated in Definition 3.1. In this special class of dissipativeness, a nonnegative state function was used as a starting point to characterize the storage function. Accordingly, Theorem 5.1 guarantees the validity of this notion of dissipative scattering based on the standard state space representation.

We give an example validating Theorem 5.1 as follows.

Example 5.1. Again, consider the supply rate Q_Φ induced by (10). In view of Theorem 5.1, we can take $\Psi(\zeta,\eta) = \frac{1}{2} = F(\zeta)^T F(\eta) \geq 0$, where $F(\xi) := \frac{1}{\sqrt{2}}$. Since $F(\xi)U(\xi)^{-1} = 1/(\sqrt{2}(1+\xi))$ is strictly proper, it follows from Proposition 2.1 that $\Psi(\zeta,\eta) \in \mathcal{F}_s(\Phi)$.

Next, we give another sufficient condition for Problem 4.1 by imposing a certain assumption on dissipation rates.

THEOREM 5.2. *Let $\Phi(\zeta,\eta) \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ induce a supply rate Q_Φ. Assume that $\Phi(\zeta,\eta)$ is observable. For this supply rate, consider $\Delta(\zeta,\eta) \in \mathcal{D}(\Phi)$. Assume that*

$$\mathrm{rank}(\tilde{\Delta}) = \mathrm{rank}(\Delta(0,0)), \tag{11}$$

where $\tilde{\Delta}$ is the coefficient matrix of $\Delta(\zeta,\eta)$. Then $\Psi(\Phi,\Delta) \in \mathcal{F}_s(\Phi)$.

Proof. See Appendix C. □

The following example also illustrates that a storage function whose corresponding dissipation rate satisfies the rank condition (11) is a state function.

Example 5.2. Again, consider the supply rate Q_Φ induced by (10). In view of Theorem 5.2, we can take $\Delta(\zeta,\eta) = \frac{3+\sqrt{5}}{2} + \zeta + \eta + \frac{3-\sqrt{5}}{2}\zeta\eta$ as one element of $\mathcal{D}(\Phi)$. One of the canonical factorizations of $\Delta(\zeta,\eta)$ is given by $\Delta(\zeta,\eta) = D(\zeta)^T D(\eta)$, where $D(\xi) := \frac{\sqrt{5}+1}{2} + \frac{\sqrt{5}-1}{2}\xi$. It is easy to verify that (11) holds and $D(\xi)U(\xi)^{-1}$ is proper. By using (7), $\Psi(\zeta,\eta) := \Psi(\Phi,\Delta)$ can be described by $\Psi(\zeta,\eta) = \frac{1+\sqrt{5}}{2} = F(\zeta)^T \Sigma_\Psi F(\eta)$, where $F(\xi) := \sqrt{\frac{1+\sqrt{5}}{2}}$ and $\Sigma_\Psi = 1$. Since $F(\xi)U(\xi)^{-1} = \sqrt{\frac{1+\sqrt{5}}{2}}(1+\xi)^{-1}$ is strictly proper, it follows from Proposition 2.1 that $\Psi(\zeta,\eta) \in \mathcal{F}_s(\Phi)$.

It follows from Proposition 3.2 in [5] that a dissipation rate is equal to zero in the discrete time lossless case. In this case, (11) holds automatically. Thus, we obtain the following theorem.

THEOREM 5.3. *Let $\Phi(\zeta,\eta) \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ induce a supply rate Q_Φ. Assume that $\Sigma = (\mathbb{Z}, \mathbb{R}^q, (\mathbb{R}^q)^Z)$ is lossless for Q_Φ and $\Phi(\zeta,\eta)$ is observable. Then $\mathcal{S}(\Phi) \subseteq \mathcal{F}_s(\Phi)$.*

In other words, "every storage function is a state function in the discrete time lossless case." In the lossless case, the supplied power to the system can be stored as an increase of the internal energy of the system without dissipation.

Remark 5.2. For simplicity and conciseness, we have supposed that $\Phi(\zeta,\eta)$ is observable in the above three theorems. However, it is neither true in the observable case nor in the nonobservable case that a storage function is a function of the state. Thus, we need to mention the results in the nonobservable case. First, consider the case in which the rank of a canonical factor $M(\lambda)$ is invariant for all $\lambda \in \mathbb{C}$. Similarly to the proof of Theorem 6.1 in [10], we can show that the same results as in the above three theorems also hold in this case. Second, consider the case where we have no assumption about $\Phi(\zeta,\eta)$. Similarly to (8), define

$$\mathcal{F}_{sa}(\Phi) := \{\Psi(\zeta,\eta) \in \mathbb{R}^{q\times q}_s[\zeta,\eta] \mid \exists K = K^T \in \mathbb{R}^{\bullet\times\bullet} \text{ such that } ((M(\sigma)w, w), x) \in \mathfrak{B}_{\Phi,a,s} \text{ implies } Q_\Psi(w) = x^T K x\},$$

where $\mathfrak{B}_{\Phi,a,s}$ is the full behavior of a minimal state space system whose manifest behavior is $\mathrm{Ker}(\begin{bmatrix} I & -M(\sigma) \end{bmatrix})$. Then the same statement holds in each of the above three theorems without the observability assumption on $\Phi(\zeta,\eta)$ if we replace $\mathcal{F}_s(\Phi)$ with $\mathcal{F}_{sa}(\Phi)$. The proofs are omitted because of their similarity to those in the observable case as well as space limitations.

6. Existence of nonnegative storage functions. In Theorem 5.1, the existence of nonnegative storage functions is assumed, so we need to consider when there exists a nonnegative storage function for a given supply rate. The following theorem provides a solution to this problem.

THEOREM 6.1. *Let $\Phi(\zeta,\eta) \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ induce a supply rate Q_Φ. Then, there exists a nonnegative storage function for Q_Φ if and only if $\sum_{t=-\infty}^{T} Q_\Phi(w)(t) \geq 0$ for all $T \in \mathbb{Z}$ and $w \in l_2^q$.*

Proof. See Appendix C. □

Next, we consider when a storage function is nonnegative.

THEOREM 6.2. *Let $\Phi(\zeta,\eta) \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ induce a supply rate Q_Φ. Let $\Psi(\zeta,\eta) \in \mathcal{S}(\Phi)$. Then, Q_Ψ induced by $\Psi(\zeta,\eta)$ is nonnegative if and only if $\Delta(\zeta,\eta) := \Delta(\Phi,\Psi)$ satisfies $\sum_{t=-\infty}^{T} Q_\Phi(w)(t) \geq \sum_{t=-\infty}^{T} Q_\Delta(w)(t)$ for all $T \in \mathbb{Z}$ and $w \in l_2^q$.*

Proof. This equivalence relation follows immediately by summing the dissipation equation of (5) along $w \in l_2$ from $-\infty$ to T. □

The above result means that the difference between the supplied energy to the system and the dissipated energy to its environment can be wholly stored as the internal net energy of the system. At the same time, a state indicates the internal status of the system. Thus, the above internal net energy should be described by a state function. Although Theorem 5.1 is a sufficient condition, we can observe that the statement of this theorem fits the above physical situation.

7. Discussions for a general case. In this section, we treat a general case; i.e., the behavior $\mathfrak{B}$ of a given dynamical system $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathfrak{B})$ can be described by an image representation $w = W(\sigma)\ell$ with $W(\xi) \in \mathbb{R}^{q\times d}[\xi]$ and a latent variable $\ell \in (\mathbb{R}^d)^Z$. Let $\Phi(\zeta,\eta) = M(\zeta)^T \Sigma_\Phi M(\eta)$ denote a canonical factorization. In this section, let $\mathfrak{B}_{\Phi,s}$ denote the full behavior of a minimal state space system whose manifest behavior can be represented by $\{v = M(\sigma)w, w \in \mathfrak{B}\}$ (i.e., $\mathrm{Im}(M(\sigma)W(\sigma))$). Our aim in this section is to investigate whether a storage function for Q_Φ and Σ can be described by a static quadratic function of any state of $\mathfrak{B}_{\Phi,s}$. As in the proof of Theorem 6.2 in [10], we can apply the results and discussions for the case of $\mathfrak{B} = (\mathbb{R}^q)^Z$

in the previous sections to this case. The following lemma guarantees the validity of these applications.

LEMMA 7.1. *Define* $\hat{\Phi}(\zeta,\eta):=W(\zeta)^T\Phi(\zeta,\eta)W(\eta) \in \mathbb{R}^{d\times d}_s[\zeta,\eta]$. *Let* $\hat{M}(\zeta)^T\Sigma_{\hat{\Phi}}\hat{M}(\eta)$ *denote a canonical factorization of* $\hat{\Phi}(\zeta,\eta)$. *Let* $\mathfrak{B}_{\hat{\Phi},s}$ *denote the full behavior of a minimal state space system whose manifest behavior is* $\mathrm{Im}(\hat{M}(\sigma))$. *Consider* $\Psi(\zeta,\eta) \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ *and define* $\hat{\Psi}(\zeta,\eta) := W(\zeta)^T\Psi(\zeta,\eta)W(\eta)$. *If there exists a constant matrix* $K = K^T \in \mathbb{R}^{\bullet\times\bullet}$ *such that* $(\hat{M}(\sigma)\ell, x_m) \in \mathfrak{B}_{\hat{\Phi},s}$ *implies* $Q_{\hat{\Psi}}(\ell) = x_m^T K x_m$, *then there also exists a constant matrix* $K' = K'^T \in \mathbb{R}^{\bullet\times\bullet}$ *such that* $(M(\sigma)w, x) \in \mathfrak{B}_{\Phi,s}$ *implies* $Q_\Psi(w) = x^T K' x$.

Proof. This lemma can be shown by using the proof of Theorem 6.2 in [10], and hence the proof is omitted here. □

Clearly, if $\Psi(\zeta,\eta)$ induces a storage function for Q_Φ and $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathrm{Im}(M(\sigma)))$, then $\hat{\Psi}(\zeta,\eta)$ induces a storage function for $Q_{\hat{\Phi}}$ and a (trivial) dynamical system $\hat{\Sigma} := (\mathbb{Z}, \mathbb{R}^d, (\mathbb{R}^d)^{\mathbb{Z}})$. Note that $\hat{\Phi}(\zeta,\eta)$ and $\hat{\Sigma} = (\mathbb{Z}, \mathbb{R}^d, (\mathbb{R}^d)^{\mathbb{Z}})$ correspond to $\Phi(\zeta,\eta)$ and $\Sigma = (\mathbb{Z}, \mathbb{R}^q, (\mathbb{R}^q)^{\mathbb{Z}})$, respectively, in sections 4, 5, and 6. By Lemma 7.1 and these facts, it is sufficient to investigate whether a storage function $Q_{\hat{\Psi}}$ is described by a static quadratic function of any minimal state of $\mathfrak{B}_{\hat{\Phi},s}$. It is clear that $\Psi(\zeta,\eta) \geq 0$ implies $\hat{\Psi}(\zeta,\eta) \geq 0$. Together with Theorem 5.1, these observations show that if $\Psi(\zeta,\eta) \geq 0$, then Q_Ψ is a state function of this original supply rate Q_Φ with this dynamical system Σ. This result can be formalized as the following theorem, which corresponds to Theorem 5.1.

THEOREM 7.2. *Let* $\Phi(\zeta,\eta) \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ *induce a supply rate. Consider a dynamical system* $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathrm{Im}(W(\sigma)))$. *Let* $\Psi(\zeta,\eta) \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ *induce a storage function for* Q_Φ *and* Σ. *Assume that* $W(\zeta)^T\Phi(\zeta,\eta)W(\eta)$ *is observable. If* $\Psi(\zeta,\eta) \geq 0$, *then there exists a constant matrix* $K = K^T \in \mathbb{R}^{\bullet\times\bullet}$ *such that* $(M(\sigma)w, x) \in \mathfrak{B}_{\Phi,s}$ *implies* $Q_\Psi(w) = x^T K x$.

Remark 7.1. In the case where $W(\zeta)^T\Phi(\zeta,\eta)W(\eta)$ is not observable, it is sufficient for us to slightly change the statement of theorem, similarly to that of Theorem 5.1 as stated in Remark 5.2.

In connection with the existence of nonnegative storage functions, we have the following theorem that corresponds to Theorem 6.1. The proof is an immediate consequence of that of Theorem 6.1.

THEOREM 7.3. *Let* $\Phi(\zeta,\eta) \in \mathbb{R}^{q\times q}_s[\zeta,\eta]$ *induce a supply rate* Q_Φ. *Consider a dynamical system* $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathrm{Im}(W(\sigma)))$. *Then, there exists a nonnegative storage function for* Q_Φ *and* Σ *if and only if* $\sum_{t=-\infty}^{T} Q_{W^T\Phi W}(\ell)(t) \geq 0$ *for all* $T \in \mathbb{Z}$ *and* $\ell \in l_2^d$.

Finally, the following theorem clarifies when a storage function is nonnegative, which corresponds to Theorem 6.2. The proof is also an immediate consequence of that of Theorem 6.2.

THEOREM 7.4. *Let* $\Phi(\zeta,\eta)$ *induce a supply rate* Q_Φ. *Consider a dynamical system* $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathrm{Im}(W(\sigma)))$. *Let* $\Psi(\zeta,\eta)$ *induce a storage function* Q_Ψ *for* Q_Φ *and* Σ. *Let* $\Delta(\zeta,\eta)$ *induce the dissipation rate corresponding to* Q_Ψ. *Then,* Q_Ψ *is nonnegative if and only if* $\Delta(\zeta,\eta)$ *satisfies* $\sum_{t=-\infty}^{T} Q_{W^T\Phi W}(\ell)(t) \geq \sum_{t=-\infty}^{T} Q_{W^T\Delta W}(\ell)(t)$ *for all* $T \in \mathbb{Z}$ *and* $\ell \in l_2^d$.

8. Conclusions. In this paper, for a given supply rate Q_Φ, we have investigated when a discrete time storage function is a state function of a given dynamical system augmented with a given supply rate. As the main solution to this problem, we have

shown that if a storage function is nonnegative, then it is also described by a static quadratic function of any state of the augmented system. In addition to this sufficient condition, we have also shown another sufficient condition for a storage function to be a state function. This condition is related to dissipation rates and guarantees that every storage function is a state function in the discrete time lossless case.

In our opinion, the above sufficient condition involving nonnegativities of storage functions can be connected with physical situations intuitively. Therefore, we conclude that every nonnegative storage function is a state function, and this finding is the most useful and interesting result of this paper. Since the conditions obtained in this paper are sufficient conditions, we must notice that a storage function satisfying the conditions neither in Theorem 5.1 nor in Theorem 5.2 may be a state function. However, the results of this paper will be useful for expanding the discrete time dissipation theory from both practical and theoretical viewpoints.

Further studies are as follows. First, we will derive not only sufficient conditions but also necessary conditions for a storage function to be a state function. Second, we will consider the physical meaning of the assumption on dissipation rates implied by (11). Third, we will study how Theorems 5.1 and 5.2 are related to each other. Finally, by using the results of the above investigations, we will develop various applications stated in section 1.

Appendix A. Notation. Let $\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{C}$ denote the set of integers, real numbers, and complex numbers, respectively. Let $\mathbb{R}^q$ ($\mathbb{C}^q$) denote the set of real (complex, respectively) vectors of size q. For $\lambda \in \mathbb{C}$, $\bar{\lambda}$ denotes the conjugate of λ. For $a \in \mathbb{R}^q$, $\|a\|^2 := a^T a$. For $a \in \mathbb{C}^q$, a^* denotes the conjugated transpose of a and $\|a\|^2 := a^* a$. Let $\mathbb{R}^{p\times q}$ denote the set of real matrices of size $p \times q$. Let $\mathbb{R}^{\bullet\times\bullet}$ ($\mathbb{R}^{\bullet}$) denote the set of matrices (vectors, respectively) whose size are suitable. For a constant matrix A, let "rank(A)" denote the rank of A. Let I_q and $0_{p\times q}$ denote the identity matrix of size $q \times q$ and the zero matrix of size $p \times q$, respectively. If the size of an identity matrix or a zero matrix is transparent, we omit its suffix. For an arbitrary subset $\mathbb{Z}'$ of $\mathbb{Z}$, let "min$\{\mathbb{Z}'\}$" denote the minimum integer of $\mathbb{Z}'$.

Let $\mathbb{R}[\xi]$ denote the set of polynomials in the indeterminate ξ with coefficients in $\mathbb{R}$. Similarly, let $\mathbb{R}[\zeta, \eta]$ denote the set of two-variable polynomials in the indeterminates ζ and η with coefficients in $\mathbb{R}$. In addition, let $\mathbb{R}[\xi^{-1}, \xi]$ denote the set of (one-variable) dipolynomials with coefficients in $\mathbb{R}$; i.e., an element of $\mathbb{R}[\xi^{-1}, \xi]$ consists of not only nonnegative but also negative powers of indeterminate ξ. The set of their matrix versions are written, respectively, by $\mathbb{R}^{p\times q}[\xi]$, $\mathbb{R}^{p\times q}[\zeta, \eta]$, and $\mathbb{R}^{p\times q}[\xi^{-1}, \xi]$ for real coefficient matrices of size $p \times q$. We use the symbol "λ" in order to denote an element of $\mathbb{C}$ and "ξ" in order to denote the indeterminate of one-variable polynomials and dipolynomials. For a nonsingular dipolynomial matrix $D(\xi^{-1}, \xi) \in \mathbb{R}^{\bullet\times\bullet}[\xi^{-1}, \xi]$, we call it a unimodular matrix on $\mathbb{R}^{\bullet\times\bullet}[\xi^{-1}, \xi]$ if there exist a nonzero $\alpha \in \mathbb{R}$ and $d \in \mathbb{Z}$ such that $\det(D(\xi^{-1}, \xi)) = \alpha\xi^d$. For a nonsingular polynomial matrix $D(\xi) \in \mathbb{R}^{\bullet\times\bullet}[\xi]$, we call it Hurwitz (anti-Hurwitz) if $\det(D(\lambda)) \neq 0$ for all $\lambda \in \mathbb{C}$ such that $|\lambda| \geq 1$ ($|\lambda| \leq 1$, respectively). For given matrices A_i $(i = 1, \ldots, n)$ having the same number of columns, col$[A_1, A_2, \ldots, A_n]$ denotes the matrix $[A_1^T, \quad A_2^T, \quad \ldots, \quad A_n^T\ ]^T$.

Let $(\mathbb{R}^q)^Z$ denote the set of real time series vectors of size q. For $w \in (\mathbb{R}^q)^Z$, the shift operator σ is defined by $(\sigma w)(t) := w(t + 1)$. The backward shift of w is also defined by $(\sigma^{-1}w)(t) := w(t - 1)$. Let l_2^q denote the set defined by $\{w \in (\mathbb{R}^q)^Z | \sum_{t=-\infty}^{t=\infty} \|w(t)\|^2 < \infty\}$. For $T \in \mathbb{Z}$, let $l_2^q|_T$ denote the set defined by $\{w \in (\mathbb{R}^q)^Z | \sum_{t=-\infty}^{t=T} \|w(t)\|^2 < \infty\}$. For $w_1, w_2 \in (\mathbb{R}^q)^Z$ and $T \in \mathbb{Z}$, let $w_1 \wedge_T w_2$ denote the

trajectory defined by $(w_1 \wedge_T w_2)(t) = w_1(t)$ for $t \leq T$ and $(w_1 \wedge_T w_2)(t) = w_2(t)$ for $t > T$. For $w \in (\mathbb{R}^q)^Z$ and $i, j \in \mathbb{Z}$ $(i \leq j)$, we use the notation $w|_{[i,j]} := \mathrm{col}[w(i), w(i+1), \ldots, w(j)] \in \mathbb{R}^{q(j-i+1)}$. For an arbitrary $R(\xi) \in \mathbb{R}^{p\times q}[\xi]$, let $\mathrm{Ker}(R(\sigma))$ denote the set defined by $\{w \in (\mathbb{R}^q)^Z | R(\sigma)w = 0\}$. Similarly, for an arbitrary $M(\xi) \in \mathbb{R}^{q\times d}[\xi]$, let $\mathrm{Im}(M(\sigma))$ denote the set defined by $\{w \in (\mathbb{R}^q)^Z | {}^{\exists}\ell \in (\mathbb{R}^d)^Z$ such that $w = M(\sigma)\ell\}$.

Appendix B. Behavioral system theory. In this section, we give a review of the behavioral system theory briefly. For more details, see these useful references: [8], [18], [19], [20], [21], and [22].

A discrete time dynamical system is defined as a triple $\Sigma = (\mathbb{Z}, \mathbb{W}, \mathfrak{B})$, where $\mathbb{Z}$ is the discrete time axis, $\mathbb{W}$ is the signal space (e.g., $\mathbb{W} = \mathbb{R}^q$), and $\mathfrak{B} \subseteq \mathbb{W}^Z$ is the (manifest) behavior. From Proposition 4.1 A in [21], a dynamical system $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathfrak{B})$ is linear, time-invariant, and complete if and only if Σ can be described by a kernel representation $R(\sigma^{-1}, \sigma)w = 0$ with $R(\xi^{-1}, \xi) \in \mathbb{R}^{\bullet\times q}[\xi^{-1}, \xi]$ for all $w \in \mathfrak{B}$. In addition to manifest variables w, there are many cases in which some auxiliary variables, say ℓ, are required to describe a dynamics. It is called a latent variable, and a dynamical system with latent variables is defined as a quadruple $\Sigma_a = (\mathbb{Z}, \mathbb{W}, \mathbb{A}, \mathfrak{B}_a)$, where $\mathbb{A}$ is the signal space of latent variables and $\mathfrak{B}_a \subseteq \mathbb{W}^Z \times \mathbb{A}^Z$ is the full behavior.

A dynamical system $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathfrak{B})$ is said to be controllable if for all $w_1, w_2 \in \mathfrak{B}$ there exist $w \in \mathfrak{B}$ and $T(\in \mathbb{Z}) \geq 0$ such that $w(t) = w_1(t)$ for $t < 0$ and $w(t) = (\sigma^{-T} w_2)(t)$ for $t \geq T$. As stated in Proposition 4.3 in [21], a linear time-invariant complete system Σ is controllable if and only if it can be described by an image representation $w = M(\sigma^{-1}, \sigma)\ell$ with $M(\xi^{-1}, \xi) \in \mathbb{R}^{q\times\bullet}[\xi^{-1}, \xi]$ for all $(w, \ell) \in \mathfrak{B}_a$. If $\{(w, \ell_1), (w, \ell_2) \in \mathfrak{B}_a\}$ implies $\ell_1 = \ell_2$, ℓ is said to be observable from w. If Σ has an image representation $w = M(\sigma^{-1}, \sigma)\ell$, ℓ is observable from w if and only if $M(\lambda^{-1}, \lambda)$ is full column rank for all nonzero $\lambda \in \mathbb{C}$. A controllable system has many observable image representations. If $M(\xi^{-1}, \xi)$ is full column rank, then there exists a nonsingular matrix $P \in \mathbb{R}^{q\times q}$ such that $PM(\xi^{-1}, \xi) = \mathrm{col}\left[\ U(\xi^{-1}, \xi), Y(\xi^{-1}, \xi)\ \right]$, where $\det(U(\xi^{-1}, \xi)) \neq 0$ and $Y(\xi^{-1}, \xi)U(\xi^{-1}, \xi)^{-1}$ is proper. We can regard $u := U(\sigma^{-1}, \sigma)\ell$ $(y := Y(\sigma^{-1}, \sigma)\ell)$ as inputs (outputs, respectively).

In $w = M(\sigma^{-1}, \sigma)\ell$ induced by $M(\xi^{-1}, \xi) \in \mathbb{R}^{q\times d}[\xi^{-1}, \xi]$, $\mathfrak{B} = \mathrm{Im}(M(\sigma))$ is invariant under postmultiplying $M(\xi^{-1}, \xi)$ by an arbitrary unimodular matrix of $\mathbb{R}^{d\times d}[\xi^{-1}, \xi]$ in the time axis $T = \mathbb{Z}$. In particular, if $M(\xi^{-1}, \xi)$ induces an observable image representation, it is easy to see that there exists a unimodular matrix $V(\xi^{-1}, \xi)$ such that $M(\xi^{-1}, \xi)V(\xi^{-1}, \xi)$ is a polynomial matrix and $M(\lambda^{-1}, \lambda)V(\lambda^{-1}, \lambda)$ is full column rank for all $\lambda \in \mathbb{C}$. Moreover, in the case where $M(\xi^{-1}, \xi)$ does not induce an observable image representation, the set of nonzero unobservable modes including multiplicities is also invariant. Thus, it is enough to focus on mathematical representations induced by polynomial matrices if we consider $\mathfrak{B}$. Of course, the same discussion holds for kernel representations. Hence, we treat representations induced by polynomial matrices in this paper.

Finally, we prepare the notions of state space systems. Let $\Sigma_s = (\mathbb{Z}, \mathbb{R}^q, \mathbb{R}^n, \mathfrak{B}_s)$ denote a system with latent variables. Then Σ_s is said to be a state space system if $\{(w_1, x_1), (w_2, x_2) \in \mathfrak{B}_s$ and $x_1(0) = x_2(0)\} \Longrightarrow \{(w_1, x_1) \wedge_0 (w_2, x_2) \in \mathfrak{B}_s\}$. Here, a latent variable x is said to be a state variable of Σ. A state variable x plays a role as a memory that stores the information of the past trajectory. It follows from Proposition 4.10 in [21] that Σ_s is a state space system if and only if there exist constant matrices $E \in \mathbb{R}^{\bullet\times n}$, $F \in \mathbb{R}^{\bullet\times n}$, and $G \in \mathbb{R}^{\bullet\times q}$ such that Σ_s can be represented by $E\sigma x + Fx + Gw = 0$ for all $(w, x) \in \mathfrak{B}_s$. It is referred to as a state representation of $\mathfrak{B}$. Of course, a state space representation of $\mathfrak{B}$ is not unique. Let

$\Sigma_s = (\mathbb{Z}, \mathbb{R}^q, \mathbb{R}^n, \mathfrak{B}_s)$ denote a state space system whose manifest behavior is $\mathfrak{B}$. For all state space systems $\Sigma'_s = (\mathbb{Z}, \mathbb{R}^q, \mathbb{R}^{n'}, \mathfrak{B}'_s)$ inducing the same manifest behavior $\mathfrak{B}$, Σ_s is said to be minimal if $n \leq n'$ holds.

Appendix C. Proofs.

Proof of Theorem 3.4. As is well known, it follows from $\Phi(e^{-j\omega}, e^{j\omega}) > 0$ for all $\omega \in [0, 2\pi)$ that there exist spectral factors such that $\Phi(\xi^{-1}, \xi) = A(\xi^{-1})^T A(\xi) = H(\xi^{-1})^T H(\xi)$, where $\det(A(\xi))$ is anti-Hurwitz and $\det(H(\xi))$ is Hurwitz. It is also clear that $A(\zeta)^T A(\eta) \geq 0$, so this is an element of $\mathcal{D}(\Phi)$. Define $\Psi^+(\zeta, \eta) := \Psi(\Phi, A(\zeta)^T A(\eta))$. From (7), we can write their relation as

$$Q_{\Psi^+}(w)(t+1) - Q_{\Psi^+}(w)(t) = Q_\Phi(w)(t) - \|A(\sigma)w(t)\|^2 \tag{12}$$

for all $t \in \mathbb{Z}$ and $w \in (\mathbb{R}^q)^Z$. Consider another $\Psi(\zeta, \eta) \in \mathcal{S}(\Phi)$ and $\Delta(\zeta, \eta) := \Delta(\Phi, \Psi)$. Define $\Psi_d(\zeta, \eta) := \Psi^+(\zeta, \eta) - \Psi(\zeta, \eta)$ and $h := N(\Psi_d)$. Subtracting the dissipation relation for $\Psi(\zeta, \eta)$ and $\Delta(\zeta, \eta)$ from (12) yields

$$w|^T_{[1,h+1]} \tilde{\Psi}_d w|_{[1,h+1]} - w|^T_{[0,h]} \tilde{\Psi}_d w|_{[0,h]} = Q_\Delta(w)(t) - \|(A(\sigma)w)(t)\|^2, \tag{13}$$

where $\tilde{\Psi}_d$ is the coefficient matrix of $\tilde{\Psi}_d(\zeta, \eta)$.

Here, suppose that $A(\xi)$ is described by $A(\xi) =: A_m\xi^m + \cdots + A_1\xi + A_0$. Since $\det(A(\xi))$ is an anti-Hurwitz polynomial, it has no poles in the unit disk. Thus, $\det(A(0)) = \det(A_0) \neq 0$. This enables us to observe that for arbitrary $w(m)$, ..., $w(1)$, there exists $w(0)$ such that $w(0) = -A_0^{-1}(A_m w(m) + \cdots + A_1 w(1))$. Similarly, we have $w(t) = -A_0^{-1}(A_m w(t+m) + \cdots + A_1 w(t+1))$ for all $t < 0$. That is, there exists a trajectory $w \in l_2^q|_m$ such that $A(\sigma)w(t) = 0$ in $t \leq 0$ for arbitrary $w(m)$, ..., $w(1) \in \mathbb{R}^q$. Define ν as the maximum integer of $h+1$, m, and $N(\Delta)$. Then we can see that there exists $w \in l_2^q|_\nu$ such that $w(\nu), \ldots, w(1) \in \mathbb{R}^q$ are arbitrary and $w(t)$ in $t \leq 0$ satisfies the difference equation $A(\sigma)w(t) = 0$. Substituting this w into (13) and summing it from $t = -\infty$ to 0 yields $w|^T_{[1,h+1]} \tilde{\Psi}_d w|_{[1,h+1]} = \sum_{t=-\infty}^{0} Q_\Delta(w)(t) \geq 0$. It follows from the arbitrariness of $w(h+1), \ldots, w(1) \in \mathbb{R}^q$ that this inequality means $a^T \tilde{\Psi}_d a \geq 0$ for all $a \in \mathbb{R}^{q(h+1)}$. Hence, we conclude that $\Psi_d(\zeta, \eta) \geq 0$, which implies $\Psi^+(\zeta, \eta) \geq \Psi(\zeta, \eta)$.

Regarding the minimum storage function, the proof is analogous and is left to the reader. □

Proof of Theorem 5.2. Theorem 5.2 is used to prove Theorem 5.1. Thus, we prove Theorem 5.2 before the proof of Theorem 5.1.

Without loss of generality, the observability of $\Phi(\zeta, \eta)$ enables us to split a canonical factor $M(\xi)$ into $M(\xi) = \mathrm{col}\left[\ U(\xi),\ \ Y(\xi)\ \right]$, where $U(\xi)$ is nonsingular and $Y(\xi)U(\xi)^{-1}$ is proper, after appropriate permutations of rows. Here, consider the following two additional assumptions:

(a) $U(\xi)$ is column reduced.

(b) $U(0)$ is nonsingular.

Under these assumptions, we show that Theorem 5.2 holds. Next, we eliminate these assumptions step by step.

1. *Proof under the assumptions.* First, we prepare the following lemma. It is easy to see that the statement is true, so we leave the proof to the reader.

LEMMA C.1. *Let* $M(\xi) \in \mathbb{R}^{\bullet \times q}[\xi]$ *and* $U(\xi) \in \mathbb{R}^{q \times q}[\xi]$. *Assume that* $\det(U(\xi)) \neq 0$, $\det(U(0)) \neq 0$, *and* $M(\xi)U(\xi)^{-1}$ *is proper. Then* $U(\xi^{-1})^{-T} M(\xi^{-1})^T$ *is proper.*

Let $\Phi(\zeta,\eta) = M(\zeta)^T\Sigma_\Phi M(\eta)$, $\Psi(\zeta,\eta) = F(\zeta)^T\Sigma_\Psi F(\eta)$, and $\Delta(\zeta,\eta) = D(\zeta)^T D(\eta)$ denote canonical factorizations. Consider a dissipation equation

$$(\zeta\eta - 1)F(\zeta)^T\Sigma_\Psi F(\eta) = M(\eta)^T\Sigma_\Phi M(\zeta) - D(\zeta)^T D(\eta). \tag{14}$$

By replacing ζ and η with ξ^{-1} and ξ in (14), respectively, we obtain $M(\xi^{-1})^T\Sigma_\Phi M(\xi) = D(\xi^{-1})^T D(\xi)$. Premultiplying this equation by $U(\xi^{-1})^{-T}$ and postmultiplying it by $U(\xi)^{-1}$ yields

$$U(\xi^{-1})^{-T}M(\xi^{-1})^T\Sigma_\Phi M(\xi)U(\xi)^{-1} = U(\xi^{-1})^{-T}D(\xi^{-1})^T D(\xi)U(\xi)^{-1}. \tag{15}$$

Since $M(\xi)U(\xi)^{-1}$ is proper and $U(0)$ is nonsingular, it follows from Lemma C.1 that $U(\xi^{-1})^{-T}M(\xi^{-1})^T$ is proper. Thus the two matrices on both sides of (15) are proper.

Next, we show that $D(\xi)U(\xi)^{-1}$ is proper. Here, assume that $D(\xi)U(\xi)^{-1}$ is nonproper; i.e., there exist $L(\xi) \in \mathbb{R}^{\bullet\times q}[\xi]$, $\pi(\xi) \in \mathbb{R}[\xi]$, and $S_j \in \mathbb{R}^{\bullet\times q}$ $(j = 0,\dots,p)$ satisfying

$$D(\xi)U(\xi)^{-1} = \frac{L(\xi)}{\pi(\xi)} + S_0 + S_1\xi + \cdots + S_{p-1}\xi^{p-1} + S_p\xi^p, \tag{16}$$

where $L(\xi)/\pi(\xi)$ is strictly proper. Notice that $\pi(0) \neq 0$ because of $\det(U(0)) \neq 0$. Substitute $L(\xi)$ and $\pi(\xi)I_q$ into $M(\xi)$ and $U(\xi)$, respectively, in Lemma C.1. It then follows from $\pi(0) \neq 0$ that $L(\xi^{-1})^T/\pi(\xi^{-1})$ is proper. By noting this fact and some algebraic manipulations, the right-hand side of (15) can be written by

$$\begin{aligned}
&U(\xi^{-1})^{-T}D(\xi^{-1})^T D(\xi)U(\xi)^{-1}\\
&= \left\{\frac{L(\xi^{-1})^T}{\pi(\xi^{-1})} + S_0^T + \frac{S_1^T}{\xi} + \cdots + \frac{S_p^T}{\xi^p}\right\}\left\{\frac{L(\xi)}{\pi(\xi)} + S_0 + S_1\xi + \cdots + S_p\xi^p\right\}\\
&= \left\{\left(\frac{L(\xi^{-1})^T}{\pi(\xi^{-1})} + S_0^T\right)S_p\right\}\xi^p + \left\{\left(\frac{L(\xi^{-1})^T}{\pi(\xi^{-1})} + S_0^T\right)S_{p-1} + S_1^T S_p\right\}\xi^{p-1}\\
&\qquad + \left\{\left(\frac{L(\xi^{-1})^T}{\pi(\xi^{-1})} + S_0^T\right)S_{p-2} + S_1^T S_{p-1} + S_2^T S_p\right\}\xi^{p-2}\\
&\cdots + \left\{\left(\frac{L(\xi^{-1})^T}{\pi(\xi^{-1})} + S_0^T\right)S_1 + S_1^T S_2 + S_2^T S_3 + \cdots + S_{p-1}^T S_p\right\}\xi + \text{proper part.}
\end{aligned} \tag{17}$$

It is easy to see that the constant part of $L(\xi^{-1})^T/\pi(\xi^{-1}) + S_0^T$ is equal to $L(0)^T/\pi(0) + S_0^T$. Let $D(\lambda)$ be described by $D(\xi) = \sum_{i=0}^h D_i\xi^i$. Then $\operatorname{rank}\left(\begin{bmatrix} D_0 & D_1 & \cdots & D_h \end{bmatrix}\right) = \operatorname{rank}(\tilde{\Delta})$ holds, because $D(\lambda)$ is obtained by a canonical factorization of $\Delta(\zeta,\eta)$. Moreover, $D_0^T D_0 = \Delta(0,0)$, and thus $\operatorname{rank}(\Delta(0,0)) = \operatorname{rank}(D_0)$ holds. By using these facts and (11), we can obtain

$$\operatorname{rank}\left(\begin{bmatrix} D_0 & D_1 & \cdots & D_h \end{bmatrix}\right) = \operatorname{rank}(\tilde{\Delta}) = \operatorname{rank}(\Delta(0,0)) = \operatorname{rank}(D_0). \tag{18}$$

Since $\Delta(\zeta,\eta) = D(\zeta)^T D(\eta)$ is a canonical factorization, the matrix on the left-hand side of (18) is full row rank. Thus (18) implies that D_0 $(= D(0))$ is also full row rank. In addition, $U(0)^{-T}D(0)^T$ is full column rank due to $\det(U(0)) \neq 0$. Substituting $\xi = 0$ to (16) and transposing it yields $U(0)^{-T}D(0)^T = L(0)^T/\pi(0) + S_0^T$, so $L(0)^T/\pi(0) + S_0^T$ is full column rank. At the same time, the left-hand side of (17) has no polynomial

part due to the properness of (15). This implies that the highest order term of the right-hand side of (17) vanishes; that is, $(L(0)^T/\pi(0)+S_0^T)S_p$ must be a zero matrix. Thus S_p is equal to $0_{\bullet\times q}$. Substituting $S_p = 0_{\bullet\times q}$ into (17), the highest order term is described by $\{(L(0)^T/\pi(0)+S_0^T)S_{p-1}\}\xi^{p-1}$. This term must also vanish, so it follows from the same reason that $S_{p-1} = 0_{\bullet\times q}$. In the same way, by repeating the above discussion up to the first order term in (17), we obtain $S_i = 0_{\bullet\times q}$, $i = 1,\dots,p$ in (16). Consequently, $D(\xi)U(\xi)^{-1}$ is proper.

Next, we show that $F(\xi)U(\xi)^{-1}$ is strictly proper. The technique of this part is similar to that of the continuous time case (cf. the proofs of Theorem 5.5 in [23] and Theorem 6.1 in [10]). Postmultiplying (14) by $U(\eta)^{-1}$ yields

$$(\zeta\eta-1)F(\zeta)^T\Sigma_\Psi F(\eta)U(\eta)^{-1} = M(\zeta)^T\Sigma M(\eta)U(\eta)^{-1} - D(\zeta)^T D(\eta)U(\eta)^{-1}. \tag{19}$$

From the above discussions, the right-hand side of (19) is proper with respect to η. Assume that $F(\eta)U(\eta)^{-1}$ is described by $F(\eta)U(\eta)^{-1} = R(\eta)+P_0+P_1\eta+\cdots+P_l\eta^l$, where $R(\xi)$ is the strictly proper rational function of column size q and $P_i \in \mathbb{R}^{\bullet\times q}$ $(i=0,\dots,l)$. Substituting this expression into (19) and equating powers of η in (19) enables us to obtain that $F(\zeta)^T\Sigma_\Psi P_i = 0_{q\times q}$, $i=0,\dots,l$. At the same time, the columns of $F(\zeta)^T \in \mathbb{R}^{q\times\bullet}[\zeta]$ are linearly independent over $\mathbb{R}$, so $\Sigma_\Psi P_i = 0_{\bullet\times q}$ for all i. This implies $P_i = 0_{\bullet\times q}$ for all i because of the nonsingularity of Σ_Ψ. Thus $F(\xi)U(\xi)^{-1}$ is strictly proper. It then follows from Proposition 2.1 that $\Psi(\zeta,\eta) \in \mathcal{F}_s(\Phi)$ under assumptions (a) and (b).

2. *Elimination of assumption* (b). If $U(0)$ is *not nonsingular*, then it is impossible to use the above discussion to show that $D(\xi)U(\xi)^{-1}$ is proper. In order to overcome this difficulty, we first modify the original $U(\xi)$ to $U(\xi)+\epsilon I_q$ for some $\epsilon \in \mathbb{R}$ so as to satisfy assumption (b). Next, by using assumption (a), we prove the properness of $(U(\xi)+\epsilon I_q)^{-T}D(\xi^{-1})^TD(\xi)(U(\xi)+\epsilon I_q)^{-1}$. Finally, we show that this properness implies that of $D(\xi)U(\xi)^{-1}$. The remaining proof is similar to that stated in the above discussion.

First, we modify the original $\Phi(\zeta,\eta)$ so as to satisfy assumption (b). Before going to this modification, we prepare some notation as a matter of convenience of the following proof. In a canonical factorization $\Phi(\zeta,\eta) = M(\zeta)^T\Sigma_\Phi M(\eta)$, let r denote the size of Σ_Φ. Moreover, Σ_Φ is a signature matrix, so it can be described by $\Sigma_\Phi = \mathrm{diag}\{\Gamma_{11},\Gamma_{22}\}$, where $\Gamma_{11} \in \mathbb{R}^{q\times q}$ and $\Gamma_{22} \in \mathbb{R}^{(r-q)\times(r-q)}$. Note that the observability of $\Phi(\zeta,\eta)$ guarantees $r \geq q$. Under these preparations, factorize $\Phi(\zeta,\eta) = M_\epsilon(\zeta)^T\Sigma_{\Phi_\epsilon}M_\epsilon(\eta)$ for an arbitrary $\epsilon \in \mathbb{R}$, where $M_\epsilon(\xi) := \mathrm{col}[U_\epsilon(\xi), Y(\xi), I_q, -\epsilon U(\xi)]$, $U_\epsilon(\xi) := U(\xi)+\epsilon I_q$, $\Sigma_{\Phi_\epsilon} := \begin{bmatrix} \Sigma_\Phi & 0_{r\times(r+q)} \\ 0_{(r+q)\times r} & \Sigma_\epsilon \end{bmatrix}$, and $\Sigma_\epsilon := \begin{bmatrix} -\epsilon^2\Gamma_{11} & \Gamma_{11} \\ \Gamma_{11} & 0_{q\times q} \end{bmatrix}$.

Next, we prove that $U_\epsilon(\xi^{-1})^{-T}D(\xi^{-1})^TD(\xi)U_\epsilon(\xi)^{-1}$ is proper. Note that $M_\epsilon(\xi)U_\epsilon(\xi)^{-1}$ can be written by

$$M_\epsilon(\xi)U_\epsilon(\xi)^{-1} = \begin{bmatrix} U(\xi)+\epsilon I_q \\ Y(\xi) \\ I_q \\ -\epsilon U(\xi) \end{bmatrix}(U(\xi)+\epsilon I_q)^{-1} = \begin{bmatrix} I_q \\ Y(\xi)U(\xi)^{-1}(I_q+\epsilon U(\xi)^{-1})^{-1} \\ U(\xi)^{-1}(I_q+\epsilon U(\xi)^{-1})^{-1} \\ -\epsilon(I_q+\epsilon U(\xi)^{-1})^{-1} \end{bmatrix}. \tag{20}$$

It follows from Lemma 6.3.11 in [4] and assumption (a) that $U(\xi)^{-1}$ is proper. Since $U(\xi)^{-1}$ is proper, $I_q+\epsilon U(\xi)^{-1}$ is biproper or, equivalently, $(I_q+\epsilon U(\xi)^{-1})^{-1}$ is biproper for almost every $\epsilon \in \mathbb{R}$. Thus, by (20), $M_\epsilon(\xi)U_\epsilon(\xi)^{-1}$ is also proper for almost every $\epsilon \in \mathbb{R}$. Furthermore, it is also possible to take this ϵ so that

$\det(U_\epsilon(0)) = \det(U(0) + \epsilon I_q) \neq 0$. Applying Lemma C.1 with the properness of $M_\epsilon(\xi)U_\epsilon(\xi)^{-1}$ and $\det(U_\epsilon(0)) \neq 0$ enables us to observe that $U_\epsilon(\xi^{-1})^{-T}M_\epsilon(\xi^{-1})^T$ is proper, which implies that $U_\epsilon(\xi^{-1})^{-T}\Phi(\xi^{-1},\xi)U_\epsilon(\xi)^{-1}$ is also proper. By premultiplying and postmultiplying $\Phi(\xi^{-1},\xi) = D(\xi^{-1})^T D(\xi)$ by $U_\epsilon(\xi^{-1})^{-T}$ and $U_\epsilon(\xi)^{-1}$, respectively, we can see that $U_\epsilon(\xi^{-1})^{-T}D(\xi^{-1})^T D(\xi)U_\epsilon(\xi)^{-1}$ is also proper.

Finally, we prove the properness of $D(\xi)U(\xi)^{-1}$. For $U_\epsilon(\xi^{-1})^{-T}D(\xi^{-1})^T D(\xi)U_\epsilon(\xi)^{-1}$, notice that $D(0)$ is full row rank and $\det(U_\epsilon(0)) \neq 0$ holds. In this point, repeating the same technique in the proof under the assumptions leads to the conclusion that $D(\xi)U_\epsilon(\xi)^{-1}$ is proper. Rewriting this proper rational function as $D(\xi)U_\epsilon(\xi)^{-1} = D(\xi)U(\xi)^{-1}(I_q + \epsilon U(\xi)^{-1})^{-1}$ and using the fact that $(I + \epsilon U(\xi)^{-1})^{-1}$is biproper, we can see that $D(\xi)U(\xi)^{-1}$ is proper. The strictly properness of $F(\xi)U(\xi)^{-1}$ can be shown by applying the same technique in the previous argument.

3. *Elimination of assumption* (a). We first prepare the following lemma. The proof is based on algebraic computations of polynomial matrices, so we leave the proof to the reader.

LEMMA C.2. *Let* $\Delta(\zeta,\eta) \in \mathbb{R}_s^{q\times q}[\zeta,\eta]$. *For an arbitrary unimodular matrix* $V(\xi) \in \mathbb{R}^{q\times q}[\xi]$, *define* $\Delta'(\zeta,\eta) := V(\zeta)^T\Delta(\zeta,\eta)V(\eta)$. *Then,*

$$\operatorname{rank}(\Delta(0,0)) = \operatorname{rank}(\Delta'(0,0)) \text{ and } \operatorname{rank}(\tilde{\Delta}) = \operatorname{rank}(\tilde{\Delta}'),$$

where $\tilde{\Delta}$ *and* $\tilde{\Delta}'$ *are the coefficient matrices of* $\Delta(\zeta,\eta)$ *and* $\Delta'(\zeta,\eta)$, *respectively.*

We suppose that $U(\xi)$ is *not column reduced*. In this case, there exists a unimodular matrix $V(\xi) \in \mathbb{R}^{q\times q}[\xi]$ such that $U(\xi)V(\xi)$ is column reduced (cf. p. 386 of [4]). By using this $V(\xi)$, consider $(\zeta\eta - 1)V(\zeta)^T\Psi(\zeta,\eta)V(\eta) = V(\zeta)^T\Phi(\zeta,\eta)V(\eta) - V(\zeta)^T\Delta(\zeta,\eta)V(\eta)$. Since $U(\xi)V(\xi)$ is column reduced, it follows from Lemma 6.3.11 in [4] that $(U(\xi)V(\xi))^{-1}$ is proper. In addition, it follows from Lemma C.2 and the unimodularity of $V(\xi)$ that $V(\zeta)^T\Delta(\zeta,\eta)V(\eta)$ satisfies (11). Hence, by repeating the previous discussion, we can observe that $(U(\zeta)V(\zeta))^{-T}(V(\zeta)^T\Psi(\zeta,\eta)V(\eta))(U(\eta)V(\eta))^{-1} = U(\zeta)^{-T}\Psi(\zeta,\eta)U(\eta)^{-1}$ is strictly proper with respect to ζ and η. This completes the proof of Theorem 5.2. □

Proof of Theorem 5.1. Similarly to Theorem 5.2, the observability of $\Phi(\zeta,\eta)$ enables us to split a canonical factor $M(\xi)$ into $M(\xi) = \operatorname{col}\begin{bmatrix} U(\xi), & Y(\xi) \end{bmatrix}$, where $U(\xi)$ is nonsingular and $Y(\xi)U(\xi)^{-1}$ is proper, after appropriate permutations of rows. Here, consider the following two additional assumptions:

(a') $U(\xi)$ is column reduced.

(b') $\Phi(e^{-j\omega}, e^{j\omega}) > 0$ for all $\omega \in [0, 2\pi)$.

As in the proof of Theorem 5.2, we show that the statement holds under the above assumptions. Next, we eliminate these assumptions. Note that the nonnegativity of $\Psi(\zeta,\eta)$ is invariant after postmultiplying and premultiplying it by $V(\zeta)^T$ and $V(\eta)$, respectively, where $V(\xi)$ is an arbitrary unimodular matrix of $\mathbb{R}^{q\times q}[\xi]$. Thus, the elimination of (a') is similar to that of (a) in the proof of Theorem 5.2. For this reason, we consider only the elimination of (b') here.

1. *Proof under the assumptions.* Using assumption (b'), we can obtain $\Phi(\xi^{-1},\xi) =: A(\xi^{-1})^T A(\xi)$, where $A(\xi) =: A_0 + \cdots + A_m\xi^m$ is an anti-Hurwitz spectral factor. It then follows from Theorem 3.4 that the maximum storage function is induced by $\Psi^+(\zeta,\eta) = \Psi(\Phi, A(\zeta)^T A(\eta))$. In addition, the existence of nonnegative storage functions yields $\Psi^+(\zeta,\eta) \geq 0$, so a canonical factorization of $\Psi^+(\zeta,\eta)$ can be described by $\Psi^+(\zeta,\eta) = F^+(\zeta)^T F^+(\eta)$, where $F^+(\xi) \in \mathbb{R}^{\bullet\times q}[\xi]$.

Consider an arbitrary nonnegative storage function induced by $\Psi(\zeta,\eta) \in \mathcal{S}(\Phi)$. A canonical factorization of $\Psi(\zeta,\eta)$ can also be described by $\Psi(\zeta,\eta) = F(\zeta)^T F(\eta)$, where

$F(\xi) \in \mathbb{R}^{\bullet \times q}[\xi]$. As in the proof of Theorem 3.4, define $\Psi_d(\zeta,\eta) = \Psi^+(\zeta,\eta) - \Psi(\zeta,\eta)$ and $\tilde{\Psi}_d$. It follows from $\Psi^+(\zeta,\eta) \geq \Psi(\zeta,\eta)$ that $\tilde{\Psi}_d \geq 0$. Here, for an arbitrary $\lambda \in \mathbb{C}$ and $\alpha \in \mathbb{C}^q$, define a complex vector $a := \mathrm{col}\left[\begin{array}{cccc} I_q, & \lambda I_q, & \ldots, & \lambda^h I_q \end{array}\right] \alpha \in \mathbb{C}^{q(h+1)}$, where $h := N(\Psi_d)$. The nonnegativity of $\tilde{\Psi}_d$ implies $\alpha^* \Psi_+(\bar{\lambda},\lambda)\alpha - \alpha^* \Psi(\bar{\lambda},\lambda)\alpha \geq 0$; i.e.,

$$\alpha^* F^+(\bar{\lambda})^T F^+(\lambda)\alpha \geq \alpha^* F(\bar{\lambda})^T F(\lambda)\alpha \geq 0 \tag{21}$$

for all $\lambda \in \mathbb{C}$ and $\alpha \in \mathbb{C}^q$. Next, by using an arbitrary $u \in \mathbb{C}^q$, define $\alpha' := U(\lambda)^{-1}u$. Since we can apply an arbitrary complex vector α in (21), substituting α' into α in (21) leads to

$$\|F^+(\lambda)U(\lambda)^{-1}u\|^2 \geq \|F(\lambda)U(\lambda)^{-1}u\|^2 \geq 0 \tag{22}$$

for all $u \in \mathbb{C}^q$ and almost every $\lambda \in \mathbb{C}$, i.e., λ satisfying $\det(U(\lambda)) \neq 0$.

At the same time, consider the anti-Hurwitz spectral factor $A(\xi)$ of $\Phi(\xi^{-1},\xi)$. Since $A(0)$ is nonsingular, it is easy to see that $A(\zeta)^T A(\eta)$ satisfies (11). Hence, it then follows from $F^+(\zeta)^T F^+(\eta) = \Psi(\Phi, A(\zeta)^T A(\eta))$ and Theorem 5.2 that $F^+(\xi)U(\xi)^{-1}$ is strictly proper. By using this fact, $|\lambda| \to \infty$ in (22) leads to

$$\lim_{|\lambda|\to\infty} \|F(\lambda)U(\lambda)^{-1}u\|^2 = 0$$

for all $u \in \mathbb{C}^q$. Due to the arbitrariness of u, $F(\xi)U(\xi)^{-1}$ is strictly proper. It then follows from Proposition 2.1 that $\Psi(\zeta,\eta) \in \mathcal{F}_s(\Phi)$ under the assumptions (a') and (b').

2. *Elimination of assumption* (b). We suppose that (b') does not hold. By using a canonical factor $M(\xi)$ of $\Phi(\zeta,\eta)$, we introduce $\Phi_\alpha(\zeta,\eta) := \left[\begin{array}{cc} U(\zeta)^T & Y(\zeta)^T \end{array}\right]\left[\begin{array}{c} U(\eta) \\ Y(\eta) \end{array}\right] + \alpha I_q$ for an arbitrary positive real α. Note that $\Phi_\alpha(\zeta,\eta) > 0$. Let $M_\alpha(\xi)$ denote a canonical factor of $\Phi_\alpha(\zeta,\eta)$. Suppose that $M_\alpha(\xi)$ is not full column rank; i.e., there exists a nonzero $a \in \mathbb{R}^q$ such that $M_\alpha(\xi)a = 0$. This contradicts the positivity of $\Phi_\alpha(\zeta,\eta) = M(\zeta)^T M(\eta) + \alpha I_q = M_\alpha(\zeta)^T M_\alpha(\eta)$. Thus, $M_\alpha(\xi)$ is full column rank, which implies that $M_\alpha(\xi)$ is described by $M_\alpha(\xi) = \mathrm{col}\left[\begin{array}{cc} U_\alpha(\xi), & Y_\alpha(\xi) \end{array}\right]$, where $\det(U_\alpha(\xi)) \neq 0$ and $Y_\alpha(\xi)U_\alpha(\xi)^{-1}$ is proper after appropriate permutations of rows.

It is clear that $\Phi_\alpha(e^{-j\omega}, e^{j\omega}) > 0$ for all $\omega \in [0, 2\pi)$, which implies that $\Phi_\alpha(\zeta,\eta)$ satisfies assumption (b'). It then follows from Theorem 3.4 that the maximum storage function for Q_{Φ_α} can be described by using an anti-Hurwitz spectral factor of $\Phi_\alpha(\xi^{-1},\xi)$, say $A_\alpha(\xi)$, as $\Psi^+_\alpha(\zeta,\eta) = (\Phi_\alpha(\zeta,\eta) - A_\alpha(\zeta)^T A_\alpha(\eta))/(\zeta\eta - 1)$, where $\Psi^+_\alpha(\zeta,\eta)$ induces the maximum storage function. Due to this fact, the same discussion in the proof under the assumptions enables us to observe that $U_\alpha(\zeta)^{-T}\Psi^+_\alpha(\zeta,\eta)U_\alpha(\eta)^{-1}$ is strictly proper with respect to ζ and η.

Here we consider

$$(\zeta\eta - 1)\Psi(\zeta,\eta) \leq \Phi(\zeta,\eta) \leq \Phi_\alpha(\zeta,\eta). \tag{23}$$

The first inequality is the dissipation inequality for the original supply rate induced by $\Phi(\zeta,\eta)$, and the second one is transparent from the definition of $\Phi_\alpha(\zeta,\eta)$. Equation (23) also means that $\mathcal{S}(\Phi) \subseteq \mathcal{S}(\Phi_\alpha)$. Thus, $\Psi(\zeta,\eta) \leq \Psi^+_\alpha(\zeta,\eta)$ holds for all $\Psi(\zeta,\eta) \in \mathcal{S}(\Phi)$. Assume that $\Psi(\zeta,\eta) \in \mathcal{S}(\Phi)$ is nonnegative. In addition, let $\Psi(\zeta,\eta) = F(\zeta)^T F(\eta)$ and $\Psi^+_\alpha(\zeta,\eta) = F^+_\alpha(\zeta)^T F^+_\alpha(\eta)$ denote canonical factorizations. Then, similarly to (22), $\|F^+_\alpha(\lambda)U_\alpha(\lambda)^{-1}u\|^2 \geq \|F(\lambda)U_\alpha(\lambda)^{-1}u\|^2 \geq 0$ holds for all

$u \in \mathbb{C}^q$ and almost every $\lambda \in \mathbb{C}$, i.e., λ satisfying $\det(U_\alpha(\lambda)) \neq 0$. Since $F_\alpha^+(\lambda)U_\alpha(\lambda)^{-1}$ is strictly proper, taking $|\lambda| \to \infty$ leads to $\lim_{|\lambda|\to\infty} \|F(\lambda)U_\alpha(\lambda)^{-1}u\|^2 = 0$. Due to the arbitrariness of u, $F(\xi)U_\alpha(\xi)^{-1}$ is strictly proper.

In the following, in order to show that $F(\xi)U(\xi)^{-1}$ is also strictly proper, we investigate the properness of $U_\alpha(\xi)U(\xi)^{-1}$. Since $M_\alpha(\xi)$ is a canonical factor and $\Phi_\alpha(\zeta,\eta) > 0$, we can see that

$$\Phi_\alpha(\zeta,\eta) = \begin{bmatrix} U(\zeta)^T & Y(\zeta)^T \end{bmatrix} \begin{bmatrix} U(\eta) \\ Y(\eta) \end{bmatrix} + \alpha I_q = \begin{bmatrix} U_\alpha(\zeta)^T & Y_\alpha(\zeta)^T \end{bmatrix} \begin{bmatrix} U_\alpha(\eta) \\ Y_\alpha(\eta) \end{bmatrix}. \tag{24}$$

By premultiplying and postmultiplying (24) by $U(\zeta)^{-T}$ and $U(\eta)^{-1}$, respectively, we also obtain

$$\begin{aligned} &U(\zeta)^{-T} \begin{bmatrix} U(\zeta)^T & Y(\zeta)^T \end{bmatrix} \begin{bmatrix} U(\eta) \\ Y(\eta) \end{bmatrix} U(\eta)^{-1} + \alpha U(\zeta)^{-T}U(\eta)^{-1} \\ &\quad = U(\zeta)^{-T}U_\alpha(\zeta)^T \left(I_q + U_\alpha(\zeta)^{-T}Y_\alpha(\zeta)^T Y_\alpha(\eta)U_\alpha(\eta)^{-1}\right) U_\alpha(\eta)U(\eta)^{-1}. \end{aligned} \tag{25}$$

In (25), it follows from assumption (b') that $U(\xi)^{-1}$ is proper, so the left-hand side is also proper with respect to the indeterminates of ζ and η. Moreover, it is easy to see that $(I_q + U_\alpha(\zeta)^{-T}Y_\alpha(\zeta)^T Y_\alpha(\eta)U_\alpha(\eta)^{-1})$ in the right-hand side is proper with respect to ζ and η and its constant term is a positive definite matrix, say V. Suppose that $U_\alpha(\xi)U(\xi)^{-1}$ is not proper. Let $P(\xi) =: P_1\xi + \cdots + P_k\xi^k \in \mathbb{R}^{q\times q}[\xi]$ denote the pure polynomial part of $U_\alpha(\xi)U(\xi)^{-1}$. This means that there exists a term described by $P_k^T V P_k(\zeta\eta)^k$ and $P_k^T V P_k \neq 0$, which contradicts the properness of the left-hand side. Thus, we obtain $P_k^T V P_k = 0$ and then $P_k = 0$ due to $V > 0$. Repeating the same technique, we can see that the matrices P_i $(i = k-1, \ldots, 1)$ are equal to zero. Thus, $U_\alpha(\xi)U(\xi)^{-1}$ is proper. By using this result, we conclude that $F(\xi)^{-1}U_\alpha(\xi)^{-1}U_\alpha(\xi)U(\xi)^{-1} = F(\xi)U(\xi)^{-1}$ is strictly proper. This completes the proof of Theorem 5.1.

Proof of Theorem 6.1. (Only if). Assume that there exists a nonnegative $\Psi(\zeta,\eta) \in \mathcal{S}(\Phi)$. Summing $Q_\Psi(w)(t+1) - Q_\Psi(w)(t) \leq Q_\Phi(w)(t)$ from $t = -\infty$ to an arbitrary $t = T$ along an arbitrary $w \in l_2^q$ yields $0 \leq Q_\Psi(w)(T+1) \leq \sum_{t=-\infty}^{t=T} Q_\Phi(w)(t)$.

(If). Since $\sum_{t=-\infty}^{T} Q_\Phi(w)(t) \geq 0 (\forall T \in \mathbb{Z})$ implies $\sum_{t=-\infty}^{\infty} Q_\Phi(w)(t) \geq 0$ for all $w \in l_2^q$, it also implies $\Phi(e^{-j\omega}, e^{j\omega}) \geq 0$ for all $\omega \in [0, 2\pi)$ (cf. Proposition 3.1 (1) in [5]).

First, suppose that $\Phi(e^{-j\omega}, e^{j\omega}) > 0$ for all $\omega \in [0, 2\pi)$. In this case, it is possible to calculate an anti-Hurwitz spectral factorization $\Phi(\xi^{-1}, \xi) = A(\xi^{-1})^T A(\xi)$. Moreover, it follows from the proof of Theorem 3.4 that $A(\zeta)^T A(\eta)$ is one dissipation rate in $\mathcal{D}(\Phi)$. Define $\Psi^+(\zeta,\eta) := \Psi(\Phi, A(\zeta)^T A(\eta))$. Then, we can write their relation as (12). Substituting an arbitrary $w \in l_2^q$ to (12) and summing it from $t = -\infty$ to 0 yields

$$w|_{[1,h+1]}^T \tilde{\Psi}^+ w|_{[1,h+1]} = \sum_{t=-\infty}^{0} \left\{ Q_\Phi(w)(t) - \|(A(\sigma)w)(t)\|^2 \right\}, \tag{26}$$

where $\tilde{\Psi}^+$ is the coefficient matrix of $\Psi^+(\zeta,\eta)$ and $h :=N(\Psi^+)$. Again, by using the proof of Theorem 3.4, we can see that there exists a trajectory $w \in l_2^q|_\nu$ such that $(A(\sigma)w)(t) = 0$ from 0 to $-\infty$ for arbitrary $w(1), \ldots, w(\nu)$, where ν is the maximum integer of $h+1$, $N(\Phi)$, and the degree of $A(\xi)$. Substituting this w into (26) allows us to obtain $\tilde{\Psi}^+ \geq 0$, which also implies $\Psi^+(\zeta,\eta) \geq 0$.

Next, we consider the general case; i.e., $\Phi(e^{-j\omega}, e^{j\omega}) \geq 0$ for all $\omega \in [0, 2\pi)$. Define $\Phi_\alpha(\zeta, \eta) := \Phi(\zeta, \eta) + \alpha I$ for an arbitrary real $\alpha > 0$. Then, $\Phi_\alpha(e^{-j\omega}, e^{j\omega}) > 0$ for all $\omega \in [0, \pi)$ holds, so the assumption in the statement of Theorem 3.4 is satisfied. Thus, there exist storage functions $\Psi^+_\alpha(\zeta, \eta)$ and $\Psi^-_\alpha(\zeta, \eta) \in \mathcal{S}(\Phi_\alpha)$ such that $\Psi^-_\alpha(\zeta, \eta) \leq \Psi_\alpha(\zeta, \eta) \leq \Psi^+_\alpha(\zeta, \eta)$ for any other $\Psi(\zeta, \eta) \in \mathcal{S}(\Phi_\alpha)$. At the same time, $\Psi^+_\alpha(\zeta, \eta) \geq 0$ is an immediate consequence of the previous discussion. Since our purpose is to show the existence of nonnegative storage functions for Q_Φ, it suffices to show that $\Psi^+_0(\zeta, \eta) := \lim_{\alpha \to +0} \Psi^+_\alpha(\zeta, \eta) \in \mathcal{S}(\Phi)$ and $\Psi^+_0(\zeta, \eta) \geq 0$.

We show the convergence of $\lim_{\alpha \to +0} \Psi^+_\alpha(\zeta, \eta)$. For arbitrary real numbers α_1 and α_2 such that $0 < \alpha_2 < \alpha_1$, we define $\Phi_{\alpha i}(\zeta, \eta) := \Phi(\zeta, \eta) + \alpha_i I$, where $i = 1, 2$. Since $\Phi_{\alpha_i}(\lambda^{-1}, \lambda)$ $(i = 1, 2)$ are positive on the unit circle, it follows from Theorem 3.4 that there exist maximum and minimum storage functions for $Q_{\Phi_{\alpha i}}$. For each $i = 1, 2$, let $\Psi^+_{\alpha i}(\zeta, \eta) \in \mathcal{S}(\Phi_{\alpha i})$ denote the maximum storage functions and $h_i := N(\Psi^+_{\alpha i})$. Then, the corresponding dissipation relation can be described by

$$Q_{\Psi^+_{\alpha i}}(w)(t+1) - Q_{\Psi^+_{\alpha i}}(w)(t) = Q_{\Phi_{\alpha i}}(w)(t) - \|A_{\alpha i}(\sigma) w(t)\|^2, \tag{27}$$

where $A_{\alpha i}(\xi) \in \mathbb{R}^{q \times q}[\xi]$ is an anti-Hurwitz spectral factor of $\Phi_{\alpha i}(\xi^{-1}, \xi)$ $(i = 1, 2)$. For each $i = 1, 2$, summing (27) from $-\infty$ to 0 along an arbitrary $w \in l^q_2$ and subtracting one from the other yields

$$Q_{\Psi^+_{\alpha 1}}(w)(1) - Q_{\Psi^+_{\alpha 2}}(w)(1) = \sum_{t=-\infty}^{0} \left\{ (\alpha_1 - \alpha_2)\|w(t)\|^2 + \|A_{\alpha 2}(\sigma) w(t)\|^2 - \|A_{\alpha 1}(\sigma) w(t)\|^2 \right\}. \tag{28}$$

Let m_i denote the degree of $A_{\alpha_i}(\xi)$ $(i = 1, 2)$. In the same way as the proof of Theorem 3.4, it follows from $|A_{\alpha 1}(0)| \neq 0$ that there exists an anti-Hurwitz trajectory w such that $A_{\alpha 1}(w)(t) = 0$ from $-\infty$ to 0 for an arbitrary $w(1), \ldots, w(m_1)$. Consider $w \in l^q_2|_\nu$ such that $w(\nu), \ldots, w(1)$ are arbitrary and $w(t)$ in $t \leq 0$ satisfies $A_{\alpha 1}(\sigma) w(t) = 0$, where ν is the maximum integer of $h_1 + 1$, $h_2 + 1$, m_1, and m_2. Then, substituting this $w \in l^q_2|_\nu$ into (28) yields

$$Q_{\Psi_{\alpha 1+}}(w)(1) - Q_{\Psi_{\alpha 2+}}(w)(1) = \sum_{t=-\infty}^{0} \left\{ (\alpha_1 - \alpha_2)\|w(t)\|^2 + \|A_{\alpha 2}(\sigma) w(t)\|^2 \right\} > 0$$

for arbitrary $w(1), \ldots, w(\nu)$. This implies $\Psi^+_{\alpha 1}(\zeta, \eta) > \Psi^+_{\alpha 2}(\zeta, \eta)$. Similarly, we obtain $\Psi^-_{\alpha_2}(\zeta, \eta) > \Psi^-_{\alpha_1}(\zeta, \eta)$. Hence, we can see that $\Psi^-_{\alpha_1}(\zeta, \eta) < \Psi^-_{\alpha_2}(\zeta, \eta) < \Psi^+_{\alpha_2}(\zeta, \eta) < \Psi^+_{\alpha_1}(\zeta, \eta)$ for arbitrary real numbers α_1 and α_2 such that $0 < \alpha_2 < \alpha_1$. This implies the convergence of $\lim_{\alpha \to +0} \Psi^+_\alpha(\zeta, \eta)$ and $\lim_{\alpha \to +0} \Psi^-_\alpha(\zeta, \eta)$.

Next, we show $\Psi^+_0(\zeta, \eta) := \lim_{\alpha \to +0} \Psi^+_\alpha(\zeta, \eta)$ and $\Psi^-_0(\zeta, \eta) := \lim_{\alpha \to +0} \Psi^-_\alpha(\zeta, \eta)$ are included in $\mathcal{S}(\Phi)$. For fixed $w \in (\mathbb{R}^q)^Z$ and $t \in \mathbb{Z}$, $Q_{\Psi^+_\alpha}(w)(t)$ and $Q_{\Psi^+_\alpha}(w)(t+1)$ converge to $Q_{\Psi^+_0}(w)(t)$ and $Q_{\Psi^+_0}(w)(t+1)$, respectively, as $\alpha \to 0$. By noting this fact, for $\alpha > 0$, define $g_{w(t)}(\alpha) := Q_\Phi(w)(t) + \alpha\|w(t)\|^2 - Q_{\Psi^+_\alpha}(w)(t+1) + Q_{\Psi^+_\alpha}(w)(t)$ and $g^0_{w(t)} := Q_\Phi(w)(t) - Q_{\Psi^+_0}(w)(t+1) + Q_{\Psi^+_0}(w)(t)$. It follows from the dissipation inequality that

$$g_{w(t)}(\alpha) \geq 0 \ \forall \alpha > 0. \tag{29}$$

Moreover, it is clear that $\lim_{\alpha \to +0} g_{w(t)}(\alpha) = g^0_{w(t)}$, so

$$\forall \epsilon > 0, \exists \delta(\epsilon) > 0 \text{ such that } 0 < \alpha < \delta(\epsilon) \Rightarrow g^0_{w(t)} - \epsilon < g_{w(t)}(\alpha) < g^0_{w(t)} + \epsilon. \tag{30}$$

By using these facts and notations, we show $\Psi_0^+(\zeta,\eta) \in \mathcal{S}(\Phi)$, which is equivalent to $g^0_{w(t)} \geq 0$. Suppose that $g^0_{w(t)} < 0$. Then, there exists a sufficiently small $\epsilon_1 > 0$ such that $g^0_{w(t)} + \epsilon_1 < 0$. By setting this ϵ_1 to ϵ in (30), we obtain $g_{w(t)}(\alpha) < g^0_{w(t)} + \epsilon_1 < 0$ for α satisfying $0 < \alpha < \delta(\epsilon_1)$. Thus, $g^0_{w(t)} < 0$ contradicts (29). Therefore, we conclude $\Psi_0^+(\zeta,\eta) \in \mathcal{S}(\Phi)$. Similarly, we can also obtain $\Psi_0^-(\zeta,\eta) \in \mathcal{S}(\Phi)$.

Finally, the remaining proof is to show $\Psi_0^+(\zeta,\eta) \geq 0$. Again, we prove it by a contradiction. Suppose that a canonical factorization of $\Psi_0^+(\zeta,\eta)$ is described by $\Psi_0^+(\zeta,\eta) = F(\zeta)^T \mathrm{diag}\{I_{r+}, -I_{r-}\}F(\eta)$, where $F(\xi) = \sum_{k=0}^{h_0} F^k \xi^k \in \mathbb{R}^{(r_- + r_+)\times q}[\xi]$ and r_+ (r_-) is the number of positive (negative, respectively) eigenvalues of the coefficient matrix of $\Psi_0^+(\zeta,\eta)$. Since $F(\xi)$ is a canonical factor, there exist $w_0, \ldots, w_{h_0}$ such that

$$\begin{bmatrix} 0 \\ f \end{bmatrix} = \begin{bmatrix} F^0 & F^1 & \cdots & F^{h_0} \end{bmatrix} \mathrm{col} \begin{bmatrix} w_0, & w_1, & \ldots, & w_{h_0} \end{bmatrix}$$

for an arbitrary nonzero $f \in \mathbb{R}^{r-}$. This implies that there exist $w \in (\mathbb{R}^q)^Z$ and $t \in \mathbb{Z}$ such that $Q_{\Psi_0^+}(w)(t) < 0$, where $w(t+k) = w_k$, $(k = 0, \ldots, h_0)$. This also means that there exists a sufficiently small $\epsilon_2 > 0$ such that $Q_{\Psi_0^+}(w)(t) + \epsilon_2 \|w(t)\|^2 < 0$ for this w and t. At the same time, $\Psi_\alpha^+(\zeta,\eta)$ is the decreasing function with respect to α and $\Psi_0^+(\zeta,\eta) = \lim_{\alpha\to 0} \Psi_\alpha^+(\zeta,\eta)$, so for all $\epsilon > 0$ there exists $\delta(\epsilon) > 0$ such that $0 < \alpha < \delta(\epsilon) \Rightarrow 0 < Q_{\Psi_\alpha^+}(w)(t) - Q_{\Psi_0^+}(w)(t) < \epsilon\|w(t)\|^2$. By setting ϵ_2 to ϵ, we obtain $Q_{\Psi_\alpha^+}(w)(t) < Q_{\Psi_0^+}(w)(t) + \epsilon_2\|w(t)\|^2 < 0$ for α satisfying $0 < \alpha < \delta(\epsilon_2)$. This contradicts $\Psi_\alpha^+(\zeta,\eta) \geq 0$ for $\alpha > 0$. By using the same technique again, we can also show that $\Psi_0^+(\zeta,\eta) < 0$ contradicts $\Psi_\alpha^+(\zeta,\eta) \geq 0$ for $\alpha > 0$. Therefore, $\Psi_0^+(\zeta,\eta) \geq 0$ holds. This completes the proof. □

Acknowledgment. The authors wish to express their sincere thanks to the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] D. Z. Arov, *Passive linear stationary dynamical systems*, Siberian Math. J., 20 (1979), pp. 149–162.

[2] D. Z. Arov, M. A. Kaashoek, and D. R. Pik, *Minimal and optimal linear discrete time-invariant dissipative scattering systems*, Integral Equations Operator Theory, 29 (1997), pp. 127–154.

[3] F. Fagnani and J. C. Willems, *Deterministic Kalman filtering in a behavioral framework*, Systems Control Lett., 32 (1997), pp. 301–312.

[4] T. Kailath, *Linear Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[5] O. Kaneko and T. Fujii, *Discrete time average positivity and spectral factorization in a behavioral framework*, Systems Control Lett., 39 (2000), pp. 31–44.

[6] V. M. Popov, *Hyperstability of Control Systems*, Springer-Verlag, Berlin, New York, 1973.

[7] C. Praagman, *Inputs, outputs, and the states in the representation of time series*, in Analysis and Optimization of Systems, Lecture Notes in Control and Inform. Sci. 111, A. Bensoussan and J. Lions, eds., Springer-Verlag, Berlin, 1988, pp. 1069–1078.

[8] P. Rapisarda and J. C. Willems, *State maps for linear systems*, SIAM J. Control Optim., 35 (1997), pp. 1053–1091.

[9] P. Rapisarda and J. C. Willems, *The subspace Nevenlinna interpolation problem and the most powerful unfalsified model*, Systems Control Lett., 32 (1997), pp. 291–300.

[10] H. L. Trentelman and J. C. Willems, *Every storage function is a state function*, Systems Control Lett., 32 (1997), pp. 249–259.

[11] H. L. Trentelman and J. C. Willems, *H_∞ control in a behavioral context: The full information case*, IEEE Trans. Automat. Control, 44 (1999), pp. 521–536.

[12] H. L. Trentelman, J. C. Willems and S. Shanker, *H_∞ filtering in a behavioral framework*, in Proceedings of the 14th International Symposium on Mathematical Theory of Networks and Systems, Perpignan, France, 2000, CD-ROM.

[13] H. L. Trentelman and P. Rapisarda, *New algorithms for polynomial J-spectral factorization*, Math. Control Signals Systems, 12 (1999), pp. 24–61.

[14] H. L. Trentelman and J. C. Willems, *Synthesis of dissipative systems using quadratic differential forms: Part* II, IEEE Trans. Automat. Control, 47 (2002), pp. 70–86.

[15] S. Weiland and J. C. Willems, *Dissipative dynamical systems in a behavioral context*, Math. Models Methods Appl. Sci., 1 (1991), pp. 1–25.

[16] J. C. Willems, *Dissipative dynamical systems. Part* I: *General theory*, Arch. Rational Mech. Anal., 45 (1972), pp. 321–351.

[17] J. C. Willems, *Dissipative dynamical systems. Part* II: *Linear systems with quadratic supply rates*, Arch. Rational Mech. Anal., 45 (1972), pp. 352–393.

[18] J. C. Willems, *From time series to linear system. Part* I: *Finite dimensional linear time invariant systems*, Automatica J. IFAC, 22 (1986), pp. 561–580.

[19] J. C. Willems, *From time series to linear system. Part* II: *Exact modeling*, Automatica J. IFAC, 22 (1986), pp. 675–694.

[20] J. C. Willems, *From time series to linear system. Part* III: *Approximate Modeling*, Automatica J. IFAC, 23 (1987), pp. 87–115.

[21] J. C. Willems, *Models for dynamics*, in Dynamics Reported, Vol. 2, Wiley, Chichester, 1989, pp. 171–269.

[22] J. C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.

[23] J. C. Willems and H. L. Trentelman, *On quadratic differential forms*, SIAM J. Control Optim., 36 (1998), pp. 1703–1749.

[24] J. C. Willems and H. L. Trentelman, *Synthesis of dissipative systems using quadratic differential forms: Part* I, IEEE Trans. Automat. Control, 47 (2002), pp. 53–69.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1395–1415

MINIMIZING INFINITE TIME HORIZON DISCOUNTED COST WITH MEAN, VARIANCE, AND BOUNDED VARIATION CONTROLS*

ANANDA WEERASINGHE†

Abstract. We consider an infinite time horizon discounted cost minimization problem for a class of Itô processes. The available controls are drift and diffusion coefficients and the added bounded variation process. We show that the value function is twice continuously differentiable and derive an optimal policy which has feedback-type drift and diffusion coefficients. When the absolute value of the optimal drift grows faster than the running cost function, the optimal bounded variation process is identically zero. When it grows weaker than the running cost function, optimal bounded variation process is a local time-type process. In this case, we relate the control problem with an optimal stopping problem. We also establish the Abelian limit relations between the value functions of the discounted cost problem and the stationary problem.

Key words. stochastic optimal control, principle of smooth fit, optimal stopping, diffusion processes with reflections

AMS subject classifications. 49A60, 93E20, 60H30

DOI. 10.1137/S0363012902403948

1. Introduction. We consider a class of discounted stochastic control problems where the controlled state process is a one-dimensional Itô process. The controller is allowed to control the drift and diffusion parameters as well as an added bounded variation process. The controller's objective is to minimize the overall cost which is an infinite time horizon integral with two components: a running cost due to the location of the state process, and a cost for using the bounded variation process which is proportional to the increase in total variation. The running cost is minimal near the origin, and it grows as the state moves away from the origin. The precise mathematical model will be described in section 2.

The problem considered here is related to the seminal work of [4], which initiated the derivation of explicit optimal policies for bounded variation control problems. In an interesting article [9], Karatzas considered three stochastic control problems related to the Brownian motion: the discounted control problem, the finite horizon control problem, and the stationary (Ergodic) control problem. Explicit optimal policies are derived there, and the Abelian limit relationships among the value functions are also established. In our work, the case of zero mean and the constant diffusion agrees with the results of [9]. In [15], Ma considered the discounted cost problem for a diffusion process with a linear drift term, and the available control is the bounded variation process. There, the optimal control is a "local time"-type process which enforces the optimal process to take values in a finite interval. Our results in section 5 are similar. But in [15], the running cost function $h(x)$ described below in (2.4) is assumed convex and the "discount factor" α in (2.4) below has a positive lower bound. Thus, the consideration of the Abelian limit $\lim_{\alpha\to 0} \alpha V_\alpha(x)$ is not possible, where V_α is the value function. In contrast, here $h(x)$ need not be convex, α is allowed to take any positive

*Received by the editors March 11, 2002; accepted for publication (in revised form) January 17, 2003; published electronically October 2, 2003.
http://www.siam.org/journals/sicon/42-4/40394.html
†Department of Mathematics, 400 Carver Hall, Iowa State University, Ames, IA 50011 (weera@math.iastate.edu).

value, and we compute the Abelian limit in section 6. In [14], one can find a novel approach of relating similar stochastic control problems with linear programming problems over the space of probability measures to obtain general existence theorems for optimal Markovian controls. The higher dimensional discounted control problem for Brownian motion is considered in the articles of [5], [6], [13], [21], and [23]. The optimal process, in general, is a reflected Brownian motion, and the reflecting boundary turned out to be the free boundary associated with the Hamilton–Jacobi–Bellman (HJB) equation.

Arisawa and Lions [3] consider an ergodic control problem (also known as the stationary control problem) in a compact state space. They systematically analyze the stationary problem and establish the Abelian limit relationships of value functions. They also point out that the existence of a solution to the HJB equation related to the stationary problem, in general, is not known. We refer to a recent article [17] for some results in this direction. Here we also provide a solution to the HJB equation of the stationary problem in section 6. In [22], the author has analyzed the stationary control problem with the same control parameters and established the optimal policy under quite general conditions. We use the results in [22] to establish Abelian relationships for the value function.

In a series of articles (see [1] and [2]), Alvarez considered bounded variation control problems for diffusions with an absorbing barrier at the origin and with a controlled increasing process in place of the bounded variation process. He used the connection between stochastic control and optimal stopping to derive optimal policies. This is a known theme in stochastic control; see [10] and [11]. This relationship is used in [18] to derive an optimal policy for a discounted control problem with degenerate variance controls without the presence of a bounded variation process. In [19], the authors provide a solution to the stationary control problem for degenerate variance controls and establish the Abelian limit relationships of the value functions of the related control problems.

In section 2 we develop the mathematical model and state the basic assumptions. Section 3 is devoted to a simple verification lemma which enables us to find optimal strategies. In many of the above mentioned articles (see [1], [2], [4], [9], [15]), the optimal policy involves a local time-type process for the bounded variation control. However, with the generality of the problem considered here, the running cost function $h(x)$ may grow much slower than $\mu_0(x)$ as $|x|$ tend to infinity, where $\mu_0(x)$ represents the maximum magnitude of the drift that can be enforced at x as described in (2.3). In this case, it is optimal not to use the bounded variation process at all. This result is proved in section 4.

In section 5, we consider the situation where $h(x)$ grows faster than $\mu_0(x)$. Here our optimal state process is a reflecting diffusion on a finite interval. To obtain this result, first we solve an auxiliary optimal stopping problem whose value function is indeed the derivative of the value function of the control problem. In section 6, we establish the Abelian relationship between the value functions of the discounted control problem and the stationary control problem. We also derive a solution to the HJB equation of the stationary problem.

In [8], Fujita and Morimoto address an ergodic control problem without the presence of a bounded variation control process $A(t)$. They control only the drift process $u_1(t)$. With our notation, $|u_1(t) \leq 1,\ u_2(t) \equiv 1$, and $A(t) \equiv 0$. Their optimal policy is similar to our unbounded optimal process in section 4. They also derive explicit expressions for the optimal value λ_0. In a simple example with $h(x) = x^2$ their opti-

mal value $\lambda_0 = \frac{1}{2}$. In contrast, with our model $\mu_0(x) \equiv 1$ and $h(x) = x^2$; the optimal process is a reflecting diffusion in a finite interval as described in section 6. By solving (6.16)–(6.18) explicitly, we can show that the optimal value λ_0 is the positive root of the equation $(x^2 - 6)e^x + 2x + 2 = 0$ and $\lambda_0 \approx 0.33590$ approximately. Thus, the efficient use of the control $A(t)$ is justified since it lowers the overall cost.

In our problem, if the position of the state process at time $t-$ is r, then the available choices for mean-variance pair at time t can be considered as belonging to the control set $C(r) = \{(\mu, \sigma) : |\mu| \leq \mu_0(r),\ |\sigma| \geq \sigma_0(r)\}$. In contrast with many articles in stochastic control literature, our control sets $C(r)$ are not compact. In fact, each $C(r)$ is unbounded.

2. Problem formulation. Consider a stochastic process $X_x(t)$ which can be considered as a weak solution to a one-dimensional stochastic differential equation

$$X_x(t) = x + \int_0^t u_1(s)ds + \int_0^t u_2(s)dW(s) + A(t), \tag{2.1}$$

where x belongs to $\mathbb{R}$, and $\{W(t) : t \geq 0\}$ is a standard Brownian motion adapted to a right-continuous filtration $\{F_t : t \geq 0\}$ on some probability space (Ω, F, P). F_0 contains all the P-null sets in F; each F_t is contained in F and is independent of the increments $\{W(t+s) - W(t) : s \geq 0\}$. The processes $u_1(t)$ and $u_2(t)$ are real valued, progressively measurable with respect to $\{F_t\}$, and satisfy condition (2.2) below. The process $A(t)$ is also $\{F_t\}$-adapted, right continuous with left limits, and is of bounded variation on finite intervals. Hence $A(t)$ is also progressively measurable. Let $|A|(t)$ be the total variation of the process A on the interval $[0, t]$.

The processes $u_1(t), u_2(t)$, and $A(t)$ are considered to be control processes. We assume that there is an increasing sequence of stopping times (τ_n) with respect to $\{F_t\}$ such that $\lim_{n\to\infty} \tau_n = +\infty$ and for each $T > 0$,

$$\begin{aligned} &\text{(i)} \quad E\left[|A|(T \wedge \tau_n) + \int_0^{T\wedge\tau_n} \left(|u_1(s)| + |u_2(s)|^2\right) ds\right] < \infty \ \text{ and} \\ &\text{(ii)} \quad \lim_{n\to\infty} E\left[|X_x(\tau_n)|e^{-\alpha\tau_n}\right] = 0, \end{aligned} \tag{2.2}$$

where $\alpha > 0$ is a constant associated with the cost function in (2.4) below. Part (i) of (2.2) is imposed to make sense of (2.1). Both parts of (2.2) will be effectively used in the proof of the verification lemma in section 3. Both conditions of (2.2) are satisfied if $E[\int_0^T (|u_1(s)| + u_2^2(s))ds] < \infty$ for each $T > 0$ and if the cost $J(x, \underline{u}, A)$ defined in (2.4) is finite. With our assumption A.2 below, the finiteness of $J(x, u, A)$ implies that $E[|A|(T)] < \infty$ for each $T > 0$ and guarantees the existence of a sequence (T_n) such that $\lim_{n\to\infty} T_n = +\infty$ and $\lim_{n\to\infty} e^{-\alpha T_n} E\left[|X_x(T_n)|\right] = 0$.

We make the following basic assumptions on the functions $\mu_0(y), \sigma_0(y)$, and $h(y)$ throughout this article. Additional assumptions will be made in each section appropriately.

A.1. $\mu_0(y)$ and $\sigma_0(y)$ are nonnegative even functions and are continuously differentiable on $\mathbb{R}$. Furthermore, $\mu_0(y)$ is decreasing on $(-\infty, 0)$, increasing on $(0, +\infty)$, and $\sigma_0(y)$ satisfies $\inf_{\mathbb{R}} \sigma_0(y) > 0$.

A.2. $h(y)$ is a continuously differentiable, nonnegative even function and is decreasing on $(-\infty, 0)$ and increasing on $(0, \infty)$. Also we let $h(0) = 0$ and $h(y) \geq c_0 + c_1|y|$ for all y, where c_0 and c_1 are constants and $c_1 > 0$.

At each time instant $t \geq 0$, the controller is required to choose $A(t)$ as well as the drift-diffusion pair $(u_1(t), u_2(t))$ satisfying (2.2) and

$$|u_1(t)| \leq \mu_0(X(t-)) \text{ and } |u_2(t)| \geq \sigma_0(X(t-)) \text{ for each } t \geq 0. \tag{2.3}$$

(We let $X(0-) \equiv X(0)$ when $t = 0$.)

For a given x in $\mathbb{R}$, we call $((\Omega, F, P), (F_t), W(t), X_x(t), \underline{u}(t), A(t))$ an admissible control system if (i) $X_x(t)$ is a weak solution to (2.1) with respect to $\{F_t\}$-Brownian motion $W(t)$, and the control processes $\underline{u}(t) \equiv (u_1(t), u_2(t))$ and $A(t)$ in the probability space (Ω, F, P), and (ii) the control processes $\underline{u}(t)$ and $A(t)$ satisfy the above described assumptions and (2.2) and (2.3).

This definition of admissible control systems is quite standard; for example, see [7]. Let $\sum(x)$ be the collection of all available control systems. For each admissible control system we define the associated cost function by

$$J(x, \underline{u}, A) = E\left[\int_0^\infty e^{-\alpha t}(h(X_x(t))dt + c.d|A|(t))\right], \tag{2.4}$$

where $\alpha > 0$ is a constant. For simplicity we assume $c = 1$. Our control problem is to minimize $J(x, \underline{u}, A)$ over all available admissible control systems in $\sum(x)$. We define the value function

$$V(x) = \inf_{\sum(x)} J(x, \underline{u}, A). \tag{2.5}$$

We intend to characterize a set of optimal controls $\underline{u}^*(t) = (u_1^*(t), u_2^*(t)), A^*(t)$ and derive an optimal state process $X^*(t)$ which yields $J(x, \underline{u}^*, A^*) = V(x)$ for each x in $\mathbb{R}$.

The formal HJB equation for the value function $V(x)$ is given by

$$\min\left\{\inf_{|u_1| \leq \mu_0(x)} \frac{1}{2}u_2^2 V''(x) + u_1 V'(x) - \alpha V(x) + h(x),\ 1 - |V'(x)|\right\} = 0. \tag{2.6}$$

In sections 4 and 5 we verify that the value function is a C^2-function and that it satisfies (2.6).

In our approach, first we consider the case where the variance coefficient is not available for control. Hence we take $u_2(s) \equiv \sigma_0(X_x(s-))$ and obtain the optimal strategy. In the next step, we approach the general case.

Therefore, for each x in $\mathbb{R}$, we introduce a subcollection $\sum_1(x)$ of $\sum(x)$ by taking $u_2(s)$ identically equal to $\sigma_0(X(s-))$ in (2.1) where $\sum(x)$ is defined prior to (2.4). In this case, each admissible control system $((\Omega, F, P), (F_t), W(t), X_x(t), \underline{u}(t), A(t))$ in $\sum_1(x)$ satisfies

$$\text{(i) } X_x(t) = x + \int_0^t u_1(s)ds + \int_0^t \sigma_0(X_x(s-))dW(s) + A(t) \tag{2.7}$$

and

$$\text{(ii) conditions (2.2) and (2.3) with } u_2(t) \equiv \sigma_0(X_x(t-)) \text{ for all } t. \tag{2.8}$$

For this collection $\sum_1(x)$, we define the analogous value function $U(x)$ by

$$U(x) = \inf_{\sum_1(x)} J(x, \underline{u}, A), \tag{2.9}$$

where $J(x, \underline{u}, A)$ is defined in (2.4).

Our optimal policies also solve the following control problem of a more theoretical nature. Consider a weak solution

$$X_x(t) = x + \int_0^t u_2(s)dW(s) + \Lambda(t), \tag{2.10}$$

where $W(t)$ is a Brownian motion, the control $u_2(t)$ satisfies the previous assumptions, and $|u_2(t)| \geq \sigma_0(X_x(t-))$. The control $\Lambda(t)$ is a bounded variation process which satisfies all the assumptions of $A(t), \Lambda(0) = 0$, and, in addition, we assume the following conditions: Let

$$\Lambda(t) = \Lambda_1(t) + \Lambda_2(t) \tag{2.11}$$

be the Lebesgue decomposition of $\Lambda(t)$ with $\Lambda_1(t)$ the absolutely continuous control, and let $\Lambda_2(t)$ be the singular control (possibly with the jumps) with respect to the Lebesgue measure. We assume Λ_1 and Λ_2 are adapted to $\{F_t\}$ and let $\frac{d\Lambda_1}{dt} = u_1(t)$. We consider $\Lambda_1(t)$ as a cost-free control process subject to the constraint $|u_1(t)| \leq \mu_0(X_x(t-))$, where μ_0 is described in A.1. The singular control process $\Lambda_2(t)$ incurs a cost when used, but it is not subject to any further constraints. Consider the cost function

$$\widehat{J}(x, u_2, \Lambda) = E\left[\int_0^\infty e^{-\alpha t}(h(X_x(t))dt + d|\Lambda_2|(t))\right], \tag{2.12}$$

where $|\Lambda_2|(t)$ denotes the total variation process of Λ_2 on $[0, t]$. The control problem is to minimize $\widehat{J}(x, u_2, \Lambda)$ over all available controls u_2 and Λ. Observe that $\widehat{J}(x, u_2, \Lambda) = J(x, \underline{u}, \Lambda_2)$, where J is given in (2.4), and since our optimal bounded variation control process $A(t)$ for (2.5) is either identically zero or is a singular process with respect to the Lebesgue measure, it follows that $\inf_{u_2,\Lambda} \widehat{J}(x, u_2, \Lambda)$ is also equal to $V(x)$ and the same optimal policies minimize $\widehat{J}(x, u_2, \Lambda)$. Our results obtain the conditions under which it is optimal to choose Λ_2 identically zero. In the case when a nonzero Λ_2 is optimal, it is identical to the optimal $A(t)$ process described in section 5.

3. A verification lemma. Here we formulate a simple verification lemma which is adequate for the purpose of this paper. This lemma is closely related to Theorem 2.1 of [15]. For results on very general verification theorems, we refer to Chapter 8 of [7].

LEMMA 3.1. *Let $Q(x)$ be a twice continuously differentiable even function satisfying the following conditions:*

$$\text{(i)} \quad -1 \leq Q'(x) \leq 0 \text{ for } x \leq 0 \text{ and } 0 \leq Q'(x) \leq 1 \text{ for } x \geq 0. \tag{3.1}$$

$$\text{(ii)} \quad \frac{\sigma_0^2(x)}{2}Q''(x) - \mu_0(x)sign(x)Q'(x) - \alpha Q(x) + h(x) \geq 0 \quad \text{for every } x. \tag{3.2}$$

Then the following conclusions hold.

(a) *$U(x) \geq Q(x)$ for all x, where $U(x)$ is defined as in* (2.9).

(b) *In addition to* (i) *and* (ii), *assume $Q''(x) \geq 0$ for all x; then $V(x) \geq Q(x)$, where $V(x)$ is defined in* (2.5).

Proof. It suffices to consider the processes $X_x(t)$ satisfying (2.1) and that the corresponding cost $J(x, \underline{u}, A)$ given in (2.4) is finite; otherwise $J(x, \underline{u}, A) \geq Q(x)$ is trivial. This together with (A.2) implies that $E\int_0^\infty e^{-\alpha t}|X_x(t)|dt$, $E\int_0^\infty e^{-\alpha t}d|A|(t)$, and $E[|A(T)|]$ are all finite. Consequently, $\liminf_{t\to\infty} E|X_x(t)|e^{-\alpha t} = 0$. We use these facts in the proof below. Let (τ_n) be as in (2.2).

To prove (a), we apply the generalized Itô formula (cf. Meyer [16, p. 285]).

$$
\begin{aligned}
&Q(X_x(T\wedge\tau_k))e^{-\alpha(T\wedge\tau_k)}\\
&= Q(x) + \int_0^{T\wedge\tau_k} e^{-\alpha s}Q'(X_x(s-))\sigma_0(X_x(s-)dW(s) + \int_0^{T\wedge\tau_k} e^{-\alpha s}Q'(X_x(s-))dA(s)\\
&\quad + \int_0^{T\wedge\tau_k} e^{-\alpha s}\left(\frac{1}{2}\sigma_0^2(X_x(s-))Q'' + u_1(s)Q' - \alpha Q\right)(X_x(s-))ds\\
&\quad + \sum_{0<s\le T\wedge\tau_k} e^{-\alpha s}[\Delta Q(X_x(s)) - Q'(X_x(s-))\Delta A(s)],
\end{aligned}
\tag{3.3}
$$

where $\Delta Q(X_x(s)) = Q(X_x(s)) - Q(X_x(s-))$.

Since $|Q'(x)\le 1$ for all x, the quantities $E[\sum_{0<s\le T\wedge\tau_k} e^{-\alpha s}|\Delta Q(X_x(s))|]$ and $E[\sum_{0<s\le T\wedge\tau_k} e^{-\alpha s}|Q'(X_x(s-))|.|\Delta A(s)|]$ are all bounded above by $E\int_0^\infty e^{-\alpha t}d|A|(t)$, and hence they are finite. Let $\{A^c(t)\}$ be the continuous part of $\{A(t)\}$. Then since $|Q'(x)|\le 1$ for all x, we have

$$
\begin{aligned}
&E\int_0^{T\Lambda\tau_k} e^{-\alpha s}Q'(X_x(s-))dA(s) + E\sum_{0<s\le T} e^{-\alpha s}[\Delta Q(X_x(s)) - Q'(X_x(s-))\Delta A(s)]\\
&= E\int_0^{T\Lambda\tau_k} e^{-\alpha s}Q'(X_x(s-))dA^c(s) + E\sum_{0<s\le T} e^{-\alpha s}\Delta Q(X_x(s))\\
&\ge -E\int_0^{T\Lambda\tau_k} e^{-\alpha s}d|A|(s).
\end{aligned}
\tag{3.4}
$$

By (2.2) and (2.8), $E[\int_0^{T\wedge\tau_k} e^{-\alpha s}Q'(X_x(s-))\sigma_0(X_x(s-))dW(s)] = 0$, and by a straightforward computation using (3.1), (3.2), and (3.4) in (3.3), we obtain

$$
E\left[e^{-\alpha(T\wedge\tau_k)}\left[|Q(0)| + |X_x(T\wedge\tau_k)|\right]\right] + J(x,u,A) \ge Q(x).
$$

By (2.2), $\lim_{k\to\infty} E\left[|X_x(\tau_k)|e^{-\alpha\tau_k}\right] = 0$ and also $\lim_{T\to\infty}\inf e^{-\alpha T}E|X_x(T)| = 0$ as we observed above. Hence by letting $\tau_k\to+\infty$ and then $T\to+\infty$ we obtain $J(x,\underline{u},A)\ge Q(x)$ for each x and (a) follows.

Proof of (b) is essentially the same once we obtain the inequality described below by employing (3.2) and the fact $Q''(x)\ge 0$ for all x. For each x and (u_1,u_2) which satisfy $|u_1|\le\mu_0(x)$ and $|u_2|\ge\sigma_0(x)$,

$$
\begin{aligned}
&\frac{1}{2}u_2^2Q''(x) + u_1Q'(x) - \alpha Q(x) + h(x)\\
&\qquad \ge \frac{\sigma_0^2(x)}{2}Q''(x) - \mu_0(x)\ \mathrm{sign}(x)Q'(x) - \alpha Q(x) + h(x) \ge 0.
\end{aligned}
$$

The last inequality follows from (3.2). Now the proof of (b) follows from the same argument as in (a). □

4. An unbounded optimal process. Consider a weak solution

$$
Z_x(t) = x - \int_0^t \mu_0(Z_x(s))\ \mathrm{sign}(Z_x(s))ds + \int_0^t \sigma_0(Z_x(s))dW(s) \tag{4.1}
$$

on some probability space (Ω, F, P) with respect to a Brownian motion $\{W(t) : t \geq 0\}$. $Z_x(t)$ is satisfying (2.1) with feedback controls $u_1(s) = -\mu_0(Z_x(s))\ \mathrm{sign}(Z_x(s))$, $u_2(s) = \sigma_0(Z_x(s))$, and $A(t)$ is identically zero. Throughout this section, we assume the following assumptions in addition to A.1 and A.2 given in section 2, and they will guarantee that $Z_x(t)$ is an admissible state process satisfying (2.2) and $Z_x(t)$ is finite for each t.

(4.2) (i) For each $x > 0$, $\quad \alpha + \mu_0'(x) > h'(x)$.

(ii) Let $\rho_0(x) = \dfrac{\mu_0(x)}{\sigma_0^2(x)}$ for $x > 0$, and we assume *either*

(4.3) $$\text{(a)} \int_1^\infty \rho_0(u)du = +\infty \quad \text{or} \quad \text{(b)} \int_1^\infty \frac{u}{\sigma_0^2(u)}\, du = +\infty.$$

The assumptions (4.2) and (4.3) are restricted only to this section.

We introduce a sequence of stopping times $\{\tau_n\}$ for each $n > |x|$ by

(4.4) $$\begin{aligned} \tau_n &= \inf\{t \geq 0 : |Z_x(t)| \geq n\} \\ &= +\infty \ \text{ if the above set is empty.} \end{aligned}$$

PROPOSITION 4.1. *Assume* (4.3) *and let* $\{Z_x(t)\}$ *be defined by* (4.1).

Then $\lim_{N\to\infty} E[|Z_x(\tau_N)|e^{-\alpha\tau_N} I_{[\tau_N<\infty]}] = 0$, *and hence* $Z_x(t)$ *is finite for each* $t > 0$ *and* $Z_x(t)$ *satisfies the admissibility condition* (2.2) *with respect to the sequence* $\{\tau_n\}$ *defined in* (4.4).

Proof. Let $H_0(x)$ be the solution to the differential equation

(4.5) $$\frac{\sigma_0^2(x)}{2} H_0''(x) - \mu_0(x)H_0'(x) - \alpha H_0(x) = 0 \text{ for } x > 0, \quad H_0(0) = 1, \qquad H_0'(0) = 0.$$

First we establish

(4.6) $$\lim_{x\to\infty} H_0'(x) = +\infty \ \text{ and thus } \lim_{x\to\infty} \frac{H_0(x)}{x} = +\infty.$$

Since $\mu_0(x) \geq 0$, by elementary analysis, $H_0(x)$ has no positive local maxima, and hence $H_0(x)$ is increasing and therefore $H_0''(x)$ is positive. Therefore, $H_0(x)$ is convex and $\lim_{x\to\infty} H_0(x) = +\infty$. Let $\rho_0(x)$ be given in (4.3). Introduce $A(x) = \int_0^x \rho_0(r)dr$. Then using (4.5) we obtain

(4.7) $$H_0'(x) = e^{2A(x)} \int_0^x \frac{2\alpha H_0(r)}{\sigma_0^2(r)}\, e^{-2A(r)} dr.$$

Assume (4.3)(a). Since $H_0(r) \geq 1$ it follows that $H_0'(x) \geq e^{2A(x)} \int_0^1 \frac{2\alpha}{\sigma_0^2(r)}\, e^{-2A(r)} dr$ for $x > 1$, and thus $\lim_{x\to\infty} H_0'(x) = +\infty$ as $\lim_{x\to\infty} A(x) = +\infty$.

(4.8) $$\text{Now assume } \int_1^\infty \frac{u}{\sigma_0^2(u)} du = +\infty \text{ as in (4.3)(b).}$$

It suffices to consider that the case $\lim_{x\to\infty} A(x) = L$ is finite. Since $H_0(x)$ is strictly convex, $H_0(x) > m_0 + m_1(x-1)$, where $m_0 = H_0(1)$ and $m_1 = H_0'(1) > 0$. By (4.7) we have

(4.9) $$H_0'(x) > \left[\int_0^1 \frac{2\alpha H_0(r)}{\sigma_0^2(r)}\, dr + \int_1^x \frac{2\alpha(m_0 + m_1(u-1))}{\sigma_0^2(u)}\, du\right] \ \text{ for } x > 1.$$

By (4.8), the right-hand side of (4.9) also tends to infinity. Hence $\lim_{x\to\infty} H_0'(x) = +\infty$ also follows in this case. Since $\lim_{x\to\infty} H_0'(x) = +\infty$, by L'Hopital's rule, $\lim_{x\to\infty} \frac{H_0(x)}{x} = +\infty$. This proves (4.6).

Next we extend $H_0(x)$ to $\mathbb{R}$ as an even function. So for $x < 0$, $H_0(x) = H_0(-x)$. For each $N > |x|$, introduce $H_N(x) = \frac{H_0(x)}{H_0(N)}$. Then H_N satisfies

$$\begin{aligned} &\frac{\sigma_0^2(x)}{2} H_N''(x) - \mu_0(x)\ \mathrm{sign}(x) H_N'(x) - \alpha H_N(x) = 0 \text{ for all } x, \\ &\qquad H_N'(0) = 0,\ H_N(\pm N) = 1. \end{aligned} \tag{4.10}$$

$H_N(x)$ is an even function and by Itô's lemma, we observe that the process $Y(t) = H_N(Z_x(t \wedge \tau_N))e^{-\alpha(t\wedge\tau_N)}$ is a positive martingale. Consequently,

$$E[e^{-a\tau_N} I_{[\tau_N<T]} + e^{-\alpha T} I_{[T<\tau_N]}] \le H_N(x). \tag{4.11}$$

By letting T tend to infinity, we have $E[e^{-\alpha\tau_N} I_{[\tau_N<\infty]}] \le \frac{H_0(x)}{H_0(N)}$. Therefore $E[|Z_x(\tau_N)|e^{-\alpha\tau_N} I_{[\tau_N<\infty]}] \le (\frac{N}{H_0(N)})H_0(x)$. Hence using (4.6), $\lim_{N\to\infty} \frac{N}{H_0(N)} = 0$, and we obtain

$$\lim_{N\to\infty} E[|Z_x(\tau_N)|e^{-\alpha\tau_N} I_{[\tau_N<\infty]}] = 0. \tag{4.12}$$

This clearly implies $\tau_\infty \equiv +\infty$ a.s., where τ_∞ is the explosion time for $Z_x(t)$.

To verify the admissibility condition (2.2) with respect to $\{\tau_N\}$ defined above notice that

$$E\int_0^{T\wedge\tau_N} [\mu_0(Z_x(s)) + \sigma_0^2(Z_x(s))]ds \le \left[\max_{[-N,N]} (\mu_0(x) + \sigma_0^2(x))\right] T < \infty. \tag{4.13}$$

Hence by (4.12) and (4.13), condition (2.2) follows. □

Remark 1. Using the assumption (4.3), one can directly verify the Khasminski's criteria for nonexplosion [20, p. 297] of a diffusion process for $(Z_x(t))$.

Remark 2. Using Itô's lemma for $H_0(Z_x(T \wedge \tau_n))$ and Gronwall's inequality we can derive $E[H_0(Z_x(T))] \le H_0(x)e^{\alpha T}$, and thus $E|Z_x(T)|$ is finite.

Next, we assume A.1, A.2, and (4.2) above. For each $n \ge 1$, let $W_n(x)$ be the solution of the differential equation

$$\frac{\sigma_0^2(x)}{2} W_n''(x) + (\sigma_0(x)\sigma_0'(x) - \mu_0(x))W_n'(x) - (\alpha + \mu_0'(x))W_n(x) + h'(x) = 0 \tag{4.14}$$

for $0 < x < n$ and with the boundary conditions $W_n(0) = 0$ and $W_n(n) = 1$.

Since μ_0', σ_0' and h' are continuous and $\mu_0'(x) \ge 0$, we extend each $W_n(x)$ to $[0, +\infty)$ so that W_n satisfies (4.14) on $[0, \infty)$. We prove the following lemma which leads to our main theorem.

LEMMA 4.2. *Assume* (4.2). *Let* $W_n(x)$ *be defined on* $[0, \infty)$ *satisfying* (4.14).

(i) *For each* $n, 0 < W_n(x) < 1$ *for* $0 < x < n$ *and* $W_n(x) > 1$ *for* $x > n$.

(ii) *For any* $x > 0$, $\{W_n(x)\}$ *is decreasing in* n, *and hence* $W_\infty(x) = \lim_{n\to\infty} W_n(x)$ *exists.*

(iii) *For each* $x > 0$, $0 < W_\infty(x) < 1$, $W_\infty(0) = 0$, *and* W_∞ *satisfies the same differential equation* (4.14) *on* $(0, \infty)$ *as* W_n.

Proof. Let W_n be a solution to (4.14) and introduce $\phi(x)$ by

$$\phi(x) = \frac{h'(x)}{\alpha + \mu_0'(x)} \text{ for } x \geq 0. \tag{4.15}$$

For each $c > 0$ such that $W_n'(c) = 0$ we have

$$\frac{\sigma_0^2(c)}{2} W_n''(c) = (\alpha + \mu_0'(c))(W_n(c) - \phi(c)). \tag{4.16}$$

By (4.2), $\phi(x) < 1$ for all $x > 0$. Thus, (4.16) implies that W_n cannot have nonpositive local minima on $(0, n)$. Furthermore, if $x = c$ is a local maxima with $0 < c < n$, then $0 < W_n(c) < 1$. Hence $0 < W_n < 1$ on $(0, n)$ and $W_n'(n) \geq 0$. Using (4.16) again, it follows that $W_n'(n) > 0$ and W_n cannot have local maxima for $x > n$. This proves (i). To prove (ii) and (iii), let $n > m \geq 1$. Since W_m and W_n both satisfy (4.14), $W_m(0) = W_n(0) = 0$, and $W_n(m) < W_m(m) = 1$, it follows that $W_m(x) > W_n(x)$ for all $x > 0$ and $\{W_n\}$ is a decreasing sequence. Thus $W_\infty(x) = \lim_{n\to\infty} W_n(x)$ exists and $0 \leq W_\infty(x) < 1$. Now let $\psi(x) = W_1(x) - W_2(x)$ for all $x > 0$. Then $\psi(0) = 0$ and $\psi(x) > 0$ for all $x > 0$. For each $n \geq 1$, since W_n satisfy (4.14), we can represent W_n by $W_n(x) = W_1(x) - t_n\psi(x)$, where $t_n = \frac{1-W_n(1)}{1-W_2(1)} > 0$. Since $\{W_n\}$ is decreasing in n, the sequence (t_n) is increasing and is bounded above by $\frac{1}{1-W_2(1)}$. Thus $\lim_{n\to\infty} t_n = t_\infty$ exists and consequently $W_\infty(x) = W_1(x) - t_\infty\psi(x)$ for all $x > 0$ and W_∞ satisfies (4.14). Thus (ii) and (iii) are complete. □

Next we introduce an even function $F : \mathbb{R} \to \mathbb{R}$ as follows:

$$F(x) = \frac{\sigma_0^2(0)}{2\alpha} W_\infty'(0) + \int_0^x W_\infty(r)dr \text{ for } x \geq 0 \tag{4.17}$$

and

$$F(x) = F(-x) \text{ for } x < 0. \tag{4.18}$$

Clearly, $F'(x) = W_\infty(x)$, $F'(0) = W_\infty(0) = 0$, and $\frac{\sigma_0^2(0)}{2} F''(0) = \alpha F(0)$. Now we claim that F satisfies

$$\frac{\sigma_0^2(x)}{2} F''(x) - \mu_0(x)\ \text{sign}(x) F'(x) - \alpha F(x) + h(x) = 0 \text{ on } \mathbb{R} \tag{4.19}$$

and

$$F(0) = \frac{\sigma_0^2(0)}{2\alpha} W_\infty'(0) \geq 0, \qquad F'(0) = 0. \tag{4.20}$$

$$\text{Also, } F \geq 0, \quad -1 \leq F'(x) \leq 0 \text{ for } x \leq 0, \text{ and } \quad 0 \leq F'(x) \leq 1 \text{ for } x \geq 0. \tag{4.21}$$

Since F, μ_0, σ_0 and h are even functions, it suffices to verify (4.19) on $[0, \infty)$. Let $R(x)$ be the left-hand side of (4.19) on $[0, \infty)$. Then $R(0) = \frac{\sigma_0^2(0)}{2} W_\infty'(0) - \alpha F(0) = 0$, and $R'(x)$ is identically zero on $(0, \infty)$ since W_∞ satisfies (4.14). Thus (4.19) follows. (4.20) is obvious and (4.21) is a consequence of Lemma 4.2(iii). Next we state and prove the main theorem in this section.

THEOREM 4.3. *Assume* (4.2) *and* (4.3). *Let* F *be defined by* (4.17) *and* (4.18). *Then*

(i) $F(x) = U(x)$ *for all* x, *where* U *is the value function of the control problem defined in* (2.9). *Moreover,* $Z_x(t)$ *defined in* (4.1) *is an optimal process.*

(ii) *In addition to* (4.2) *and* (4.3), *assume that* $\phi(x)$ *defined in* (4.15) *is increasing. Then* $F(x) = V(x)$, *where* V *is the value function for the control problem described in* (2.5). *Moreover,* $Z_x(t)$ *is an optimal process.*

Proof. Step 1. First we show that $F(x)$ is the cost function due to process $Z_x(t)$ in (4.1). Let τ_N be defined by (4.4). By a direct application of Itô's lemma to $F(e^{-\alpha(T\wedge\tau_N)}Z_x(T\wedge\tau_N))$ and employing Proposition 4.1, we obtain

$$F(x) = E\left[\int_0^\infty e^{-\alpha t}h(Z_x(t))dt\right]. \tag{4.22}$$

Step 2. Next we prove $F(x) = U(x)$, where U is given in (2.9) by using the verification lemma, Lemma 3.1. By (4.21), F clearly satisfies Lemma 3.1(i). F also satisfies (4.19), and hence it satisfies the assumption (3.2) there. Consequently, $U(x) \geq F(x)$, but by (4.22) it follows that $F(x) = U(x)$ and $Z_x(t)$ in (4.1) define an optimal process.

Step 3. To prove (ii), we intend to apply Lemma 3.1. All we have to do is to verify $F''(x) \geq 0$ for all x. It suffices to show $F''(x) \geq 0$ for $x \geq 0$. Notice $F''(x) = W'_\infty(x)$, where W_∞ is given in Lemma 4.2. By (4.22), $F(0) > 0$, and thus $F''(0) > 0$ by (4.19). Now suppose that $F''(x_1) < 0$ for some $x_1 > 0$. Since $W_\infty(x) \equiv F'(x)$ for $x \geq 0$ and also $W_\infty(x)$ is the decreasing limit of $\{W_n(x)\}$, where $W_n(x)$ satisfy (4.14) with $W_n(0) = 0$, and $W_n(n) = 1$, it follows that there exist large N and two points $0 < \xi_1 < \xi_2$ such that W_N has a local maxima at $x = \xi_1$ and local minima at $x = \xi_2$. Hence $W''_N(\xi_1) \leq 0 \leq W''_N(\xi_2)$ and using (4.14) together with the fact $W'_N(\xi_1) = W'_N(\xi_2) = 0$ we obtain $\frac{\sigma_0^2(\xi_1)}{2}W''_N(\xi_1) = (\alpha + \mu'_0(\xi_1))[W_N(\xi_1) - \phi(\xi_1)] \leq 0$ and $\frac{\sigma_0^2(\xi_2)}{2}W''_N(\xi_2) = (\alpha + \mu'_0(\xi_2))[W_N(\xi_2) - \phi(\xi_2)] \geq 0$, where ϕ is given by (4.15). Since $\mu'_0(x) \geq 0$, we have $\phi(\xi_1) \geq W_N(\xi_1) > W_N(\xi_2) \geq \phi(\xi_2)$ which contradicts the fact that ϕ is increasing. Hence $F''(x) \geq 0$ for all x and $F(x) \leq V(x)$ for all x, using Lemma 3.1. Moreover, by (4.22), $Z_x(t)$ is an optimal process and $F(x) = V(x)$ for all x. This completes the proof. □

5. A bounded optimal process.

5.1. Assumptions and reflecting diffusions. In this section we make the following assumption regarding the functions μ_0 and h in addition to A.1 and A.2 of section 2. Let $\phi(x) = \frac{h'(x)}{\alpha+\mu'_0(x)}$ be the function defined on $\mathbb{R}$ as in (4.15). Notice $\phi(0) = 0$. We assume the following:

(i) There exist positive constants $R > \beta > 0$ and $\delta_0 > 0$ such that $\phi(x) < 1$ for
$x < \beta,\ \phi(\beta) = 1,\ \phi(x) > 1$ for $x > \beta$, and $\phi(x) > 1 + \delta_0$ for every $x > R$. (5.1)

Our candidate for an optimal policy is derived from the class of diffusion processes with reflecting barriers at $-a$ and $+a$. Therefore for each $a > 0$, we consider a weak solution to the equation

$$X_x^a(t) = x - \int_0^t \mu_0(X_x^a(s))\ \mathrm{sign}(X_x^a(s))ds + \int_0^t \sigma_0(X_x^a(s))dW(s) + K_a(t), \tag{5.2}$$

where the bounded variation process K_a is given by

$$dK_a(t) = dL_{-a}(t) - dL_a(t) \quad \text{for } t > 0 \tag{5.3}$$

and

$$d|K_a|(t) = dL_{-a}(t) + dL_a(t), \tag{5.4}$$

and $\{W(t) : t \geq 0\}$ is a Brownian motion process in some probability space (Ω, F, P) with respect to a filtration $\{F_t\}$. $L_{-a}(t)$ and $L_a(t)$ are local time processes of $X_x^a(t)$ at $-a$ and $+a$, respectively. If $x \in [-a, a]$, then $X_x^a(t) \in [-a, +a]$ for all $t \geq 0$. If $x \notin [-a, a]$, there will be an initial jump to the nearest point of $\{-a, +a\}$ at $t = 0$, and then (5.2) and (5.3) follow. In this case $X(0+) \in \{-a, +a\}$ and $K_a(0)$ corresponds to the jump size at $t = 0$. Clearly, $\{X_x^a(t)\}$ satisfies the assumptions (2.2)–(2.3), and hence they are admissible processes.

Let

$$V_a(x) = E \int_0^\infty e^{-\alpha t}(h(X_x^a(t))dt + d|K_a|(t)). \tag{5.5}$$

Then by Itô's lemma, it is easy to verify that V_a satisfies the following differential equation:

$$\frac{\sigma_0^2(x)}{2} V_a''(x) - \mu_0(x)\ \mathrm{sign}(x) V_a'(x) - \alpha V_a(x) + h(x) = 0 \quad \text{for } -a < x < a, \tag{5.6}$$

$$V_a'(x) = -1 \text{ for } x \leq -a \text{ and } V_a'(x) = 1 \text{ for } x \geq a. \tag{5.7}$$

By symmetry, V_a is an even function. We expect the optimal process reflects at $-a^*$ and a^* and the corresponding V_{a^*}'' vanishes at $-a^*$ and a^* so that the principle of smooth fit holds (see [4], [15]). Hence we expect to find a point $a^* > 0$ such that V_{a^*} satisfies (5.6) and (5.7), V_{a^*}' is continuous on $\mathbb{R}$, and $|V_{a^*}'(x)| \leq 1$ for all x. It suffices to consider (5.6) and (5.7) on $[0, \infty)$. Thus we let $W_a(x) = V_a'(x)$ on $(0, a)$, and it satisfies

$$\frac{\sigma_0^2(x)}{2} W_a''(x) + (\sigma_0(x)\sigma_0'(x) - \mu_0(x)) W_a'(x) - (\alpha + \mu_0'(x)) W_a(x) + h'(x) = 0, \tag{5.8}$$

$$W_a(0) = 0 \text{ and } W_a(a) = 1. \tag{5.9}$$

We need to find the point $a = a^*$ so that $W_{a^*}(x)$ satisfies an additional condition

$$W_{a^*}'(a^*) = 0 \text{ and } 0 < W_{a^*}(x) < 1 \text{ for } 0 < x < a^*. \tag{5.10}$$

We establish the existence of a unique point $a^* > 0$ with the aid of an optimal stopping problem.

5.2. An auxiliary stopping problem. By considering (5.8), (5.9), and (5.10), we formulate the following optimal stopping problem. For each $x > 0$, consider the process

$$Y_x(t) = x + \int_0^t [\sigma_0(Y_x(s))\sigma_0'(Y_x(s)) - \mu_0(Y_x(s))]ds + \int_0^t \sigma_0(Y_x(s))dW(s) \tag{5.11}$$

which is a weak solution to the above equation with respect to a Brownian motion $\{W(t)\}$ adapted to a filtration $\{F_t\}$ in a probability space (Ω, F, P). We introduce the stopping time τ_0 by

$$\begin{aligned} \tau_0 &= \inf\{t > 0 : Y_x(t) = 0\} \\ &= +\infty \ \text{ if the above set is empty.} \end{aligned} \tag{5.12}$$

Introduce the set

$$\prod = \{\tau \geq 0 : \tau \text{ is a stopping time with respect to } \{F_t\}\} \tag{5.13}$$

and consider the stopping problem

$$S(x) = \inf_{\tau \in \prod} E\left[e^{-R(\tau)} I_{[\tau < \tau_0]} + \int_0^{\tau \wedge \tau_0} e^{-R(u)} h'(Y_x(u)) du\right], \tag{5.14}$$

where

$$R(t) = \int_0^t (\alpha + \mu_0'(Y_x(s))) ds. \tag{5.15}$$

We claim that the optimal stopping time is of the form τ_{a^*}, where τ_{a^*} is the first hitting time of the level a^* by the process $Y_x(t)$ for some $a^* > 0$. V_{a^*} characterized by (5.5)–(5.7) turned out to be the value function for the control problem and $V_{a^*}'(x) = S(x)$ for each $x > 0$. To verify these assertions, first we define the stopping time τ_a by

$$\begin{aligned} \tau_a &= \inf\{t > 0 : Y_x(t) = a\} \\ &= +\infty \text{ if the above set is empty.} \end{aligned} \tag{5.16}$$

We introduce the stochastic representation for the solution $W_a(x)$ of (5.8) and (5.9):

$$W_a(x) = E\left[e^{-R(\tau_a)} I_{[\tau_a < \tau_0]} + \int_0^{\tau_a \wedge \tau_0} e^{-R(u)} h'(Y_x(u)) du\right] \tag{5.17}$$

for each x in $[0, a]$, where $R(t)$ is given in (5.15). By Itô's lemma, it is easy to verify that W_a also satisfies (5.8) and (5.9). We extend each W_a to $\mathbb{R}^+$ so that it satisfies the differential equation (5.8) on the interval $(0, \infty)$. On $[0, a]$, W_a has the representation (5.17) and $V_a'(x) = W_a(x)$. We use these facts in our next lemma which derives an upper bound for optimal a^*.

LEMMA 5.1. *Assume* (5.1). *Then there exists a constant* $a_1 > R$ *such that for each* $a > a_1$, $\sup_{[0,a]} W_a(x) > 1$, *where* R *is given in* (5.1).

Proof. We assume the contrary. Suppose that there exists a strictly increasing sequence (b_n) so that $b_n > R$ for all n, $\lim_{n\to\infty} b_n = +\infty$, and $\sup_{[0,b_n]} W_{b_n}(x) \leq 1$ for each n. We claim $0 < W_{b_n}(x) < 1$ for $0 < x < b_n$. Suppose $W_{b_n}(x_n) = 1$ and $0 < x_n < b_n$; then W_{b_n} has a local maximum at x_n. Hence we use (5.8) to obtain $h'(x_n) \geq \alpha + \mu_0'(x_n)$ and therefore by (5.1), $x_n \geq \beta$. Since $W_{b_n}(x_n) = W_{b_n}(b_n) = 1$ and W_{b_n} satisfies (5.8), it has a local minimum at a point $x = z_n$, where $\beta \leq x_n < z_n < b_n$ and $W_{b_n}(z_n) < 1$. Again using (5.8) at $x = z_n$ we obtain $\alpha + \mu_0'(z_n) > h'(z_n)$ which contradicts (5.1). Thus our claim is true and furthermore $W_{b_n}'(b_n) \geq 0$. Next we claim $W_{b_n}'(b_n) > 0$ for all n. Suppose that $W_{b_n}'(b_n) = 0$. By (5.8) and (5.1), since $b_n > \beta$ we have $W_{b_n}''(b_n) < 0$. First, as in the above argument, we observe that W_{b_n} cannot have any local minima at a point $x = z$ satisfying $z > \beta$ and $W_{b_n}(z) \leq 1$. Hence W_{b_n} achieves its maximum at $x = b_n$ and decreasing on $[b_n, \infty)$. Next we compare W_{b_n} and $W_{b_{n+1}}$. $W_{b_{n+1}}(b_n) < W_{b_n}(b_n) = 1$, $W_{b_n}(b_{n+1}) < W_{b_{n+1}}(b_{n+1}) = 1$, and $W_{b_n}(0) = W_{b_{n+1}}(0) = 0$. Therefore W_{b_n} and $W_{b_{n+1}}$ both satisfy (5.8) and meet at $x = 0$ and at a point $x = q_n$, where $b_n < q_n < b_{n+1}$. Hence $W_{b_n} \equiv W_{b_{n+1}}$, and this leads to a contradiction as $W_{b_{n+1}}(b_n) < 1$. Thus $W_{b_n}'(b_n) > 0$ for all n.

Our next claim is $W'_{b_n}(0) \le W'_{b_1}(0)$ for all n. Suppose not. Assume $W'_{b_m}(0) > W'_{b_1}(0)$ for some m. Since $W_{b_m}(0) = W_{b_1}(0) = 0$, there exist $c_1 > 0$ such that $W_{b_m}(c_1) > W_{b_1}(c_1)$. Again W_{b_m} and W_{b_1} satisfy (5.8) and meet at $x = 0$ and at another point on (c_1, b_1), and hence $W_{b_m} \equiv W_{b_1}$. This contradicts the fact $W_{b_m}(b_1) < W_{b_1}(b_1) = 1$. Consequently, we conclude that $\sup_n W'_{b_n}(0) \le W'_{b_1}(0)$.

By (5.6) we observe that $\alpha V_{b_n}(0) = \frac{\sigma_0^2(0)}{2} W'_{b_n}(0) \le \frac{\sigma_0^2(0)}{2} W'_{b_1}(0)$, where V_{b_n} is given in (5.5) and $V'_{b_n}(x) \equiv W_{b_n}(x)$ on $[0, b_n]$. Hence by (5.6) we obtain

$$\frac{\sigma_0^2(b_n)}{2} W'_{b_n}(b_n) + h(b_n) - \mu_0(b_n) = \frac{\sigma_0^2(0)}{2} W'_{b_n}(0) + \alpha \int_0^{b_n} W_{b_n}(r)dr. \tag{5.18}$$

Since $W'_{b_n}(b_n) > 0$ and $\sup_n W'_{b_n}(0) \le W'_{b_1}(0)$, and $\sup_{[0,b_n]} W_{b_n}(x) \le 1$, we obtain

$$h(b_n) - \mu_0(b_n) < \frac{\sigma_0^2}{2} W'_{b_1}(0) + \alpha b_n \quad \text{for all} \quad n. \tag{5.19}$$

But (5.1) implies that

$$h(b_n) - \mu_0(b_n) > \alpha(1 + \delta_0)(b_n - R) + \delta_0 \mu_0(b_n) + [h(R) - (1 + \delta_0)\mu_0(R)] \tag{5.20}$$

and consequently by (5.19) and (5.20)

$$\alpha(1 + \delta_0)\left(1 - \frac{R}{b_n}\right) - \frac{(1 + \delta_0)\mu_0(R)}{b_n} < \alpha + \frac{\sigma_0^2(0)}{2b_n} W'_{b_1}(0). \tag{5.21}$$

By letting $b_n \to +\infty$ we obtain $\alpha(1 + \delta_0) \le \alpha$ which is a contradiction as $\alpha > 0$ and $\delta_0 > 0$. This proves the lemma. □

Now we derive the solution to the stopping problem described in (5.14).

THEOREM 5.2. *Assume* (5.1). *Then*

(i) *there exist* $a^* > 0$ *such that the function* W_{a^*} *defined by* (5.17) *satisfies* (5.8), (5.9), *and* (5.10);

(ii) *let* $S(x)$ *be the value function for the stopping problem as in* (5.14) *for* $x \ge 0$. *Then*

$$S(x) = \begin{cases} W_{a^*}(x) & \text{for } x \le a^*, \\ 1 & \text{for } x > a^*. \end{cases} \tag{5.22}$$

Thus $S'(x)$ *is continuous on* $(0, \infty)$, *and* $S''(x)$ *is continuous on* $(0, \infty)$ *except at* $x = a^*$. *But* $S''(a^*-)$ *and* $S''(a^*+)$ *both exist finitely.*

Moreover, an optimal stopping time τ^* *is defined by*

$$\tau^* = \begin{cases} \tau_{a^*} & \text{if } 0 < x < a^*, \\ 0 & \text{if } x \ge a^*. \end{cases} \tag{5.23}$$

Proof. Let $\beta > 0$ be as in assumption (5.1). Thus for each a in $(0, \beta)$ we consider the function $W_a(x)$ which satisfies (5.8) and (5.9) on $(0, +\infty)$. By Lemma 4.2, $0 < W_a(x) < 1$ for each x in $(0, a)$. Thus $W'_a(a) \ge 0$, but if $W'_a(a) = 0$, then by (5.7) and

(5.8), W_a has a strict local minima at $x = a$, and this contradicts $0 < W_a(x) < 1$ on $(0, a)$. Hence $W'_a(a) > 0$. We let

$$a^* = \sup\{b < 0 : \text{ for each } 0 < a < b,\ W_a(x) \in (0,1) \text{ for } x \text{ in } (0,a)\}. \tag{5.24}$$

Thus, $0 < \beta < a^* < a_1 < \infty$, where a_1 is given in Lemma 5.1, and hence a^* is finite. We claim that $0 < W_{a^*}(x) < 1$ and $W'_{a^*}(a^*) = 0$. If (a_n) is a sequence increasing to a^*, then $W_{a_n}(x)$ is decreasing to $W_{a^*}(x)$ as in Lemma 4.2. Hence $W_{a^*}(x) < 1$. But W_{a^*} satisfies (5.17); hence $W_{a^*}(x) > 0$ on $(0, a^*)$ and $W_{a^*}(a^*) = 1$. Thus $W'_{a^*}(a^*) \geq 0$. But if $W'_{a^*}(a^*) > 0$, we extend W_{a^*} to $[0, \infty)$ satisfying (5.8) and consider $W_{a^*}(x) - t\psi(x)$, where $\psi(x)$ is a positive solution to homogeneous equation related to (4.14), $\psi(0) = 0$, and t is a parameter. Thus elementary arguments show that for $0 < t < \varepsilon_1$, $W_{a^*} - t\psi$ also satisfies the conditions of the set in (5.24), and hence $a^* + \varepsilon$ also belong to the set in (5.24) leading to a contradiction. Consequently, $W'_{a^*}(a^*) = 0$.

To prove (ii), we let

$$\widehat{S}(x) = \begin{cases} W_{a^*}(x) & \text{for } 0 \leq x < a^*, \\ 1 & \text{for } x \geq a^*. \end{cases} \tag{5.25}$$

By (i), $\widehat{S}$ is C^1, and $\widehat{S}''$ is continuous everywhere except at $x = a^*$. Also $\widehat{S}''(a^*-)$ is finite and negative by (5.8) and $\widehat{S}(a^*+) = 0$. Notice that

$$\begin{aligned} &\frac{1}{2}\sigma_0^2(x)\widehat{S}''(x) + (\sigma_0(x)\sigma_0'(x) - \mu_0(x))\widehat{S}'(x) - (\alpha + \mu_0'(x))\widehat{S}(x) \\ &\qquad = \begin{cases} -h'(x) & \text{if } x < a^*, \\ -(\alpha + \mu_0'(x)) & \text{if } x > a^*. \end{cases} \end{aligned} \tag{5.26}$$

But $a^* > \beta > 0$; hence $-(\alpha + \mu_0'(x)) > -h'(x)$ for $x > a^*$. Consequently, the right-hand side of (5.26) is greater than or equal to $-h'(x)$ for all $x \neq a^*$. For any τ in $\prod$, we apply Itô's lemma [12, p. 219] to $\widehat{S}(Y_x(t \wedge \tau \wedge \tau_0))e^{-R(t\wedge\tau\wedge\tau_0)}$ to obtain

$$E\left[e^{-R(\tau\wedge\tau_0)}\widehat{S}(Y_x(\tau \wedge \tau_0)) + \int_0^{\tau\wedge\tau_0} e^{-R(s)}h'(Y_x(s))ds\right] \geq \widehat{S}(x). \tag{5.27}$$

To obtain (5.27), we used a standard localization argument, the fact $0 \leq \widehat{S}(x) \leq 1$ on $[0, \infty)$, the bounded convergence theorem, and the montone convergence theorem. Since $\widehat{S}(0) = 0$ and $0 \leq \widehat{S}(x) \leq 1$, now it follows that

$$E\left[e^{-R(\tau)}I_{[\tau<\tau_0]} + \int_0^{\tau\wedge\tau_0} e^{-R(s)}h'(Y_x(s))ds\right] \geq \widehat{S}(x).$$

Consequently, $S(x) \geq \widehat{S}(x)$. But using $\tau = \tau^*$, where τ^* is given in (5.23) and Itô's lemma, we obtain

$$E\left[e^{-R(\tau^*)}I_{[\tau^*<\tau_0]} + \int_0^{\tau^*\wedge\tau_0} e^{-R(s)}h'(Y_x(s))ds\right] = \widehat{S}(x). \tag{5.28}$$

Thus $S(x) \equiv \widehat{S}(x)$ and τ^* is optimal. This completes Theorem 5.2. □

Now we are ready to prove the main theorem in this section. For the point $a^* > 0$ derived in the previous theorem, we consider the reflecting diffusion process $\{X_x^{a^*}(t)\}$

defined by (5.2)–(5.4) with values in $[-a^*, a^*]$ for $t > 0$. If the initial point x is outside $[-a^*, a^*]$ we allow a jump to $\{-a^*, a^*\}$ at $t = 0$ as described below (5.7). To simplify the notation we denote $X_x^{a^*}(t)$ by $X_x^*(t)$ in the next theorem.

THEOREM 5.3. *Assume* (5.1). *Let* $\{X_x^*(t)\}$ *be the above described reflecting diffusion process which satisfies* (5.2)–(5.4) *with values in* $[-a^*, a^*]$ *for* $t > 0$, *and let* $V_{a^*}(x)$ *be the associated cost defined by* (5.5). *Then the following holds.*

(i) $\{X_x^*(t)\}$ *is an optimal process for* (2.9) *and* $V_{a^*}(x) = U(x)$ *for all* x.

(ii) *In addition to* (5.1), *assume that* $\phi(x)$ *is increasing on* $[0, \infty)$, *where* ϕ *is given by* (4.15). *Then* $\{X_x^*(t)\}$ *is also optimal for the control problem* (2.5) *and* $V_{a^*}(x) = V(x)$ *for all* x.

(iii) *In each case, the value functions of stochastic control problem and optimal stopping problem* (5.14) *are related by* $V'_{a^*}(x) = S(x)$ *for all* x, *where* S *is given in* (5.14).

Proof. Let $V_{a^*}(x)$ be defined by (5.5). Clearly, V_{a^*} is an even function which satisfies (5.6) and (5.7). For $0 \le x \le a^*$, $V'_{a^*}(x) \equiv W_{a^*}(x)$ and it satisfies (5.8), (5.9), and (5.17). Moreover, W_{a^*} satisfies (5.10) by Theorem 5.2. For $x \ge a^*$, $V'_{a^*}(x) = 1$. Hence, by Theorem 5.2, $V'_{a^*}(x) = S(x)$ for $x > 0$, where $S(x)$ is the value function of the stopping problem (5.14). In particular, $S(x)$ completely determines $V_{a^*}(x)$ by (5.6)–(5.10). Since $h(0) = 0$, by (5.6) we obtain $\alpha V_{a^*}(0) = \frac{\sigma_0^2(0)}{2} W'_{a^*}(0)$ and

$$V_{a^*}(x) = \frac{\sigma_0^2(0)}{2\alpha} S'(0) + \int_0^x S(r)dr \quad \text{for } x > 0. \tag{5.29}$$

V_{a^*} is a twice continuously differentiable even function on $\mathbb{R}$, since $S(x)$ is C^1 on $[0, \infty)$ by Theorem 5.2. To establish (i) and (ii) of the theorem, we use the verification lemma, Lemma 3.1, and hence first we verify the conditions (i) and (ii) of Lemma 3.1. The condition (i) follows from (5.29), and it suffices to verify condition (ii).

By (5.6), condition (ii) holds for all x in $[0, a^*]$. By (5.6) and (5.10) we have $h(a^*) = \alpha V_{a^*}(a^*) + \mu_0(a^*)$. Since $a^* > \beta$, by (5.1), we also have $h'(x) > \alpha + \mu_0'(x)$ for $x > \beta$. Consequently, $h(x) = h(a^*) + \int_{a^*}^x h'(r)dr > \alpha V_{a^*}(a^*) + \mu_0(a^*) + \int_{a^*}^x (\alpha + \mu_0'(r))dr$ for $x > a^*$. This yields $h(x) > \alpha V_{a^*}(x) + \mu_0(x)$ for $x > a^*$. Since $V'_{a^*}(x) = 1$ for $x > a^*$, we see that $\frac{\sigma_0^2(x)}{2} V''_{a^*}(x) - \mu_0(x) V'_{a^*}(x) - \alpha V_{a^*}(x) + h(x) = h(x) - (\alpha V_{a^*}(x) + \mu_0(x)) > 0$ for $x > a^*$. Thus condition (ii) in Lemma 3.1 is satisfied, and consequently Theorem 5.3(i) follows.

To obtain (ii) of the theorem, again we can use Lemma 3.1, once we verify $V''_{a^*}(x) \ge 0$ for all x. By (5.29) it is clear that $V''_{a^*}(x) = 0$ if $|x| \ge a^*$. Since V_{a^*} is even, it remains to show $V''_{a^*}(x) \ge 0$ on $[0, a^*]$. By (5.29), $V''_{a^*}(0) = S'(0) > 0$ since $V_{a^*}(0) > 0$ by (5.5). Now $W_{a^*}(x) = V'_{a^*}(x)$ on $[0, a^*]$, we have $W'_{a^*}(0+) > 0$, and thus W_{a^*} is strictly increasing on an interval $[0, \varepsilon]$ for some $\varepsilon > 0$, $W_{a^*}(0) = 0$, and $W_{\alpha^*}(a^*) = 1$. Suppose that $W'_{a^*}(x) < 0$ for some $0 < x < a^*$. Thus, it is clear that there exist $0 < c_1 < c_2 < a^*$, where W_{a^*} has a local maxima at $x = c_1$, W_{a^*} has a local minima at $x = c_2$, and W_{a^*} is decreasing on $[c_1, c_2]$. Then by an argument similar to Step 3 of the proof of Theorem 4.3 we obtain

$$\phi(c_1) \ge W_a(c_1) > W_a(c_2) \ge \phi(c_2) \quad \text{and} \quad c_1 < c_2. \tag{5.30}$$

This contradicts the added assumption that ϕ is increasing on $[0, \infty)$. Consequently, $V''_{a^*}(x) \equiv W'_{a^*}(x) \ge 0$ on $[0, a^*)$ and now by Lemma 3.2(b), we can conclude Theorem 5.3(ii).

Part (iii) is immediate from (5.29) and Theorem 5.2. This completes the proof of this theorem. □

6. Abelian relations.

6.1. Main results. We intend to compute the asymptotic behavior of the value function as α tends to zero. In this section we write V_α for V and U_α for U to signify the dependence on α, where the value functions V and U are introduced in (2.5) and (2.9). We analyze the behavior of U_α here, and thus the same results hold for V_α as well. With the usual notation as similar to (2.4) and (2.9), we introduce the value of the stationary control problem by

$$\lambda_0 = \inf_{\sum_1(x)} \limsup_{T\to\infty} \frac{1}{T} E[|A|(T) + \int_0^T h(X_x(t))dt], \tag{6.1}$$

where $X_x(t)$ satisfies (2.1). The value λ_0 is independent of the initial point x because for any y, the process can jump to y within a very short time interval $[0,\varepsilon]$ using the $A(t)$ process, but the cost of this jump disappears in the limit of (6.1) as T tend to infinity.

Our aim here is to establish $\lim_{\alpha\to 0^+} \alpha U_\alpha(x) = \lambda_0$ (uniformly on compact sets) under the assumptions in sections 4 and 5. For our optimal policy of the stationary control problem described in Theorem 6.1 below, the limit in (6.1) (instead of lim sup) exists and is equal to λ_0. A similar result in the Brownian motion case was proved in [9]. Throughout this section we assume there is a small interval $[0,\eta]$ with $\eta > 0$ so that the assumption (2.2) holds for every α in $(0,\eta]$. Notice also that the sequence (τ_n) in (2.2) may depend on $\alpha > 0$. Furthermore, in section 4, the assumption (4.2) depends on α, and here we replace it by a slightly stronger condition

$$\mu_0'(x) \geq h'(x) \text{ for each } x \geq 0. \tag{6.2}$$

Similarly with regard to assumption (5.1) in section 5, we replace it by (6.3) below so that (5.1) holds uniformly for all α in $(0,\eta]$.

There exists positive constants $\delta_0 > 0$, $\varepsilon > 0$, and $R > 0$ such that the function $h'(x) - \mu_0'(x)$ is monotone increasing on $[0,R]$ and

$$h'(x) - \mu_0'(x) > \delta_0 \mu_0'(x) + \varepsilon \quad \text{for all} \quad x \geq R. \tag{6.3}$$

Now we state our main theorem in this section.

THEOREM 6.1. *Under the assumptions* A.1, A.2, (6.2), *and* (4.3) *related to section* 4 *or* A.1, A.2, *and* (6.3) *related to section* 5, *we have*

$$\lim_{\alpha\to 0} \sup_{|x|\leq M} |\alpha U_\alpha(x) - \lambda_0| = 0 \quad \textit{for each} \quad M > 0, \tag{6.4}$$

where λ_0 *is given in* (6.1). *Furthermore, an optimal strategy for the stationary problem* (6.1) *will be obtained.*

Our first step is to prove the following technical lemma.

LEMMA 6.2. *Under the assumptions of Theorem* 6.1, *the following results hold.*

(i) *For every* $\alpha > 0$, $|U_\alpha(x) - U_\alpha(0)| \leq |x|$ *for all* x.

(ii) *For each* $M > 0$, $\lim_{\alpha\to 0^+} \sup_{|x|\leq M} |\alpha U_\alpha(x) - \alpha U_\alpha(0)| = 0$.

Proof. In both sections 4 and 5 we observed that U_α' is continuous and $|U_\alpha'(x)| \leq 1$ for each $\alpha > 0$. Hence (i) holds. Part (ii) is immediate from (i). □

Thus, to establish (6.4) it suffices to show $\lim_{\alpha\to 0^+} \alpha U_\alpha(0) = \lambda_0$. We verify this in the next subsections.

6.2. Case of the unbounded optimal diffusion. Throughout this subsection we make the assumptions A.1, A.2, (6.2), and (4.3) to use the results in section 4. The function W_∞ obtained in Lemma 4.2 depends on α, and henceforth we write W_α instead of W_∞.

Proof of Theorem 6.1 (under the assumptions of A.1, A.2, (6.2), and (4.3)). By (4.19), (4.20), and Theorem 4.3 we obtain $\alpha U_\alpha(0) = \frac{\sigma_0^2(0)}{2} W'_\alpha(0)$ and $U'_\alpha(x) = W_\alpha(x)$ for $x > 0$. W_α satisfies the differential equation (4.14) for $x > 0$, together with $W_\alpha(0) = 0$ and $0 < W_\alpha(x) < 1$ for all $x > 0$. An elementary application of Itô's lemma yields the following stochastic representation for $W_\alpha(x)$ for $x > 0$:

$$W_\alpha(x) = E\left[\int_0^{\tau_0} e^{-\int_0^r (\alpha + \mu_0'(Y_x(s)))ds} h'(Y_x(r))dr\right], \tag{6.5}$$

where Y_x is given in (5.11) and the stopping time τ_0 is given by (5.12). Now with the same notation, we define

$$W_0(x) = E\left[\int_0^{\tau_0} e^{-\int_0^r \mu_0'(Y_x(s))ds} h'(Y_x(r))dr\right] \quad \text{for} \quad x \geq 0. \tag{6.6}$$

Clearly, $W_0(0) = 0$ and by (4.2) we have $0 < W_0(x) \leq 1$ for $x > 0$. Since $U'_\alpha \equiv W_\alpha$, by (4.19) we obtain

(6.7)

$$\begin{aligned} &\alpha U_\alpha(0) \int_0^x \frac{2}{\sigma_0^2(r)}\, dr \\ &\quad = W_\alpha(x) + 2\int_0^x \frac{(h(r) - \mu_0(r)W_\alpha(r))}{\sigma_0^2(r)}\, dr - 2\alpha \int_0^x \frac{1}{\sigma_0^2(r)} \int_0^r W_\alpha(u)du\, dr. \end{aligned}$$

Since W_α is increasing to W_0 as α tends to 0, the right-hand side of (6.7) converges to a finite quantity as $\alpha \to 0^+$. Hence $\lim_{\alpha \to 0^+} \alpha U_\alpha(0)$ exists and is finite. Let

$$\lim_{\alpha \to 0^+} \alpha U_\alpha(0) = \wedge_0, \quad \text{where} \quad 0 \leq \wedge_0 < +\infty. \tag{6.8}$$

Thus by letting $\alpha \to 0^+$ in (6.7) we obtain

$$\wedge_0 \int_0^x \frac{2}{\sigma_0^2(r)}\, dr = W_0(x) + 2\int_0^x \frac{h(r) - \mu_0(r)W_0(r)}{\sigma_0^2(r)}\, dr \text{ for } x > 0. \tag{6.9}$$

Consequently, W_0 satisfies the first order equation $\frac{\sigma_0^2(x)}{2} W_0'(x) - \mu_0(x)W_0(x) + h(x) = \wedge_0$ for $x > 0$, $W_0(0) = 0$, and $0 < W_0(x) \leq 1$.

Introduce $Q_0(x)$ on $\mathbb{R}$ by

$$\begin{aligned} Q_0(x) &= \int_0^x W_0(r)dr \quad \text{for} \quad x \geq 0 \\ &= Q_0(-x) \quad \text{for} \quad x < 0. \end{aligned} \tag{6.10}$$

Then Q_0 satisfies

$$\frac{\sigma_0^2(x)}{2} Q_0''(x) - \mu_0(x)\ \text{sign}(x) Q_0'(x) + h(x) = \wedge_0 \text{ for all } x, \tag{6.11}$$

$Q_0(0) = 0,\ Q_0'(0) = 0,\ 0 < Q_0'(x) \le 1$ for $x > 0$, and $-1 \le Q_0'(x) < 0$ for $x < 0$. Hence $|Q_0(x)| \le |x|$ for all x.

First we intend to show $\lambda_0 \ge \wedge_0$, where λ_0 is given in (6.1). Let $X_0(t)$ be any available process satisfying (2.1) and (2.2) for each $\alpha > 0$ with $X_0(0) = 0$ and which has a finite value for $\limsup_{T\to\infty} \frac{1}{T} E[|A|(T) + \int_0^T h(X_0(t))dt]$. Thus there are constants $M > 0$ and $T_0 > 0$ such that for all $T > T_0$,

$$MT > E\left[|A|(T) + \int_0^T h(X_0(t))dt\right].$$

Since h satisfies A.2 as similar to the proof of Lemma 3.1, this implies that the quantities $\limsup_{T\to\infty} \frac{E|A|(T)}{T}$ and $\limsup_{T\to\infty} \frac{1}{T} E\int_0^T |X_0(t)|dt$ are finite and, as a consequence, $\liminf_{T\to\infty} \frac{E|X_0(T)|}{T} = 0$. Next we apply Itô's formula to $Q_0(X_0(t\wedge\tau_n))$ as similar to the proof of Lemma 3.1, where (τ_n) as in (2.2). Using the above properties of Q_0, we obtain

$$E|X_0(T\wedge\tau_n)| + E\left[|A|(T) + \int_0^T h(X_x(s))ds\right] \ge \wedge_0 E[T\wedge\tau_n]. \tag{6.12}$$

Notice that $E|X_0(T\wedge\tau_n)| \le E|X_0(T)| + e^{\alpha T} E[|X_0(\tau_n)|e^{-\alpha\tau_n}]$. Hence we obtain

$$\begin{aligned} e^{\alpha T} E\left[|X_0(\tau_n)|e^{-\alpha\tau_n}\right] + E|X_0(T)| + E\left[|A|(T) + \int_0^T h(X_x(s))ds\right] \\ \ge \wedge_0 E[T\wedge\tau_n]. \end{aligned} \tag{6.13}$$

Now keeping T fixed and by letting τ_n tend to infinity and using (2.2) we have

$$E|X_0(T)| + E\left[|A|(T) + \int_0^T h(X_x(s))ds\right] \ge \wedge_0 T. \tag{6.14}$$

Since $\liminf_{T\to\infty} \frac{E|X_0(T)|}{T} = 0$, by (6.14) we derive

$$\limsup_{T\to\infty} \frac{1}{T} E\left[|A|(T) + \int_0^T h(X_0(t))dt\right] \ge \wedge_0,$$

and thus $\lambda_0 \ge \wedge_0$.

Next we consider our candidate for the optimal process of the stationary problem. Let $Z_x(t)$ be as in (4.1). In this case $A(t)$ is identically zero. Let (τ_n) be as in (4.4). Apply Itô's lemma to $Q_0(Z_x(T\wedge\tau_n))$ and use properties of Q_0 and (6.11) to obtain $E[\int_0^T h(Z_x(s))ds] \le Q_0(x) + \wedge_0 T$. Hence we obtain

$$\limsup_{T\to\infty} \frac{1}{T} E\int_0^T h(Z_x(s))ds \le \wedge_0. \tag{6.15}$$

Thus $\lambda_0 \le \wedge_0$, where λ_0 as in (6.1). Consequently, $\lambda_0 = \wedge_0$ and the process $Z_x(t)$ defined in (4.1) is also optimal for the stationary problem (6.1). This completes the proof. □

The proof of the above lemma also yields the following important corollary, which can be compared with the results obtained in [3] and [17].

COROLLARY 6.3. *Let Q_0 be as in* (6.10). *Then*

(i) *Q_0 is a C^2-solution to the HJB equation of the stationary control problem;*

(ii) $\lim_{\alpha\to 0^+}(U_\alpha(x) - U_\alpha(0)) = Q_0(x)$ *uniformly on compact sets.*

Proof. Part (i) follows from (6.10), (6.11), and the discussion therein. The proof of (ii) follows from (6.5), (6.6), and (6.11) and the fact that $W_\alpha(x)$ is increasing to $W_0(x)$. □

6.3. Case of the bounded optimal process. Throughout this subsection we assume A.1, A.2, and (6.3). The condition (6.3) implies (5.1), and therefore the results we derived in section 5 remain valid here. In [22] a general stationary control problem for singular stochastic controls is resolved, and here we rely on the results developed in [22].

Proof of Theorem 6.1 (under the assumptions A.1, A.2, and (6.3)). A.1, A.2, and (6.3) imply the assumptions of Theorem 2.1 of [22] and, in particular, Proposition 4.4 of [22]. In our case $x^* = 0$ and $a^* = -b^*$ with the notation there. (See the remark on page 6 of [22].) By Proposition 4.4 of [21], there exist a point $b^* > 0$ and a function Q_0 (we use Q_0 instead of V there) defined on $\mathbb{R}$, related to the stationary problem (6.1) and satisfying the following conditions:

(6.16) (i) $Q_0(x)$is a twice differentiable even function satisfying $0 \le Q_0'(x) \le 1$ on $[0, b^*]$ and $Q_0'(x) \equiv 1$ for $x \ge b^*$.

(6.17) (ii) $\lambda_0 = h(b^*) - \mu_0(b^*) > 0$, where λ_0 as in (6.1).

(6.18) (iii) $\frac{1}{2}\sigma_0^2(x)Q_0''(x) - \mu_0(x)\,\mathrm{sign}(x)Q_0'(x) + h(x) = \lambda_0$ for $|x| < b^*$ and $\ge \lambda_0$ for $|x| > b^*$.

(6.19) (iv) Also by (6.16) and the assumption (A.2), there exist two positive constants K_1, K_2 so that $|Q_0(x)| \le K_1 + |x| \le K_2(1 + h(x))$ for all x.

Furthermore, from Theorem 2.1 of [22], an optimal state process for the stationary control problem (6.1) is given by a reflecting diffusion with state space $[-b^*, b^*]$ and described by (5.2), (5.3), and (5.4) with $a = b^*$.

For each $\alpha > 0$, let $\{X_\alpha^*(t)\}$ be the optimal reflecting diffusion process described by Theorem 5.3 with initial data $X_\alpha^*(0) = 0$ with the value for the discounted problem $U_\alpha(0)$. We apply Itô's lemma to $Q_0(X_\alpha^*(t))e^{-\alpha t}$ and obtain

$$
\begin{aligned}
E[Q_0(X_\alpha^*(T))e^{-\alpha T}] &= Q_0(0) + E\int_0^T e^{-\alpha t}Q_0'(X_\alpha^*(t))dA_\alpha^*(t) \\
&+ E\int_0^T e^{-\alpha t}\left[\frac{1}{2}\sigma_0^2(X_\alpha^*(t))Q_0'' - \mu_0(X_\alpha^*(t))\,\mathrm{sign}(X_\alpha^*(t))Q_0' - \alpha Q_0\right](X_\alpha^*(t))dt.
\end{aligned} \tag{6.20}
$$

Using (6.16), (6.17), and (6.18) we obtain

(6.21)

$$
\begin{aligned}
U_\alpha(0) &\ge E\left[\int_0^T e^{-\alpha t}(h(X_\alpha^*(t))dt + d|A_\alpha^*|(t))\right] \\
&\ge Q_0(0) + \lambda_0\int_0^T e^{-\alpha t}dt - \alpha E\left[\int_0^T e^{-\alpha t}Q_0(X_\alpha^*(t))dt\right] - E\left[Q_0(X_\alpha^*(T))e^{-\alpha T}\right].
\end{aligned}
$$

By Theorem 5.3, $\{X_\alpha^*(t)\}$ take values in a compact interval, and hence $|Q_0(X_\alpha^*(t))| < M_\alpha$ for some constant $M_\alpha > 0$ which may depend on α. By (6.19), we obtain

$E\int_0^T e^{-\alpha t}Q_0(X_\alpha^*(t))dt \le K_2\left(\frac{1}{\alpha}+U_\alpha(0)\right)$. Using this with (6.22) we obtain

$$(1+\alpha K_2)U_\alpha(0) \ge Q_0(0)+\lambda_0\int_0^T e^{-\alpha t}dt-\alpha K_2\int_0^T e^{-\alpha t}dt - M_\alpha e^{-\alpha T}. \tag{6.22}$$

Next, first we let T tend to infinity and then multiply the inequality by α, and we let α tend to zero to obtain

$$\liminf_{\alpha\to 0^+}\alpha U_\alpha(0) \ge \lambda_0. \tag{6.23}$$

It remains to verify $\limsup_{\alpha\to 0^+}\alpha U_\alpha(0) \le \lambda_0$. Let $\{X_0^*(t)\}$ be the optimal state process for the stationary control problem (6.1) with the associated bounded variation process $\{A_0^*(t)\}$ and the initial position $X_0^*(0)=0$ with $|X_0^*(t)|\le b^*$ as described in Theorem 2.1 of [22]. We apply Itô's lemma to $Q_0(X_0^*(t))e^{-\alpha t}$ and as similar to (6.20) we obtain

$$\begin{aligned} E\left[\int_0^T e^{-\alpha t}[h(X_0^*(t))dt+d|A_0^*|(t)]\right] &= \lambda_0\int_0^T e^{-\alpha t}dt \\ &\quad -\alpha E\int_0^T e^{-\alpha t}Q_0(X_0^*(t))dt+Q_0(0). \end{aligned} \tag{6.24}$$

By (6.16), $Q_0(x)\ge Q_0(0)$ for all x, and hence $E[\int_0^T e^{-\alpha t}(h(X_0^*(t))dt+d|A_0^*|(t))] \le \lambda_0\int_0^T e^{-\alpha t}dt + Q_0(0)e^{-\alpha T}$. By letting T tend to infinity, we have $U_\alpha(0) \le E[\int_0^\infty e^{-\alpha t}(h(X_0^*(t))dt+d|A_0^*|(t))] \le \frac{\lambda_0}{\alpha}$, and we conclude that $\limsup_{\alpha\to 0^+}\alpha U_\alpha(0)\le \lambda_0$. Hence, using (6.23), it follows that $\lim_{\alpha\to 0^+}\alpha U_\alpha(0)=\lambda_0$, where λ_0 is the value of the stationary problem (6.1). This completes the proof. □

We also have the following important corollary, as similar to Corollary 6.3.

COROLLARY 6.4. *Let Q_0 be the function satisfying* (6.16)–(6.18) *and $Q_0(0)=0$. Then*

(i) *Q_0 is a C^2-solution to the HJB equation corresponding to the stationary control problem;*

(ii) *$\lim_{\alpha\to 0^+}(U_\alpha(x)-U_\alpha(0))=Q_0(x)$ uniformly on compact sets.*

Proof. Part (i) follows from (6.16)–(6.19). To prove (ii), since $|U_\alpha'(x)|\le 1$ for all x, it suffices to show $\lim_{\alpha\to 0^+}U_\alpha'(x)=Q_0'(x)$ for all $x>0$. We sketch the proof of this fact. In Theorem 5.2, the point a^* may depend on α, and hence we relabel it a_α^*. Using (5.6), (5.7), and (5.10), we obtain $h(a_\alpha^*)-\mu_0(a_\alpha^*)=\alpha U_\alpha(a_\alpha^*)>0$. By (6.3) and since $h(0)=0$, there is a $\beta_0>0$ so that $\sup_{[0,\beta_0]}(h(x)-\mu_0(x))<0$. Hence $a_\alpha^*>\beta_0>0$ for every α. Let a_∞^* be any limit point ($a_\infty^*=+\infty$ is allowed) of $\{a_\alpha^*:\alpha>0\}$. Thus $a_\infty^*\ge\beta_0>0$. Let (a_{α_n}) be a sequence so that $\lim_{\alpha_n\to 0^+}a_{\alpha_n}=a_\infty^*$. We consider (U_{α_n}'). Each U_{α_n}' satisfies (5.8) on $(0,a_{\alpha_n}^*)$ with $U_{\alpha_n}'(0)=0$ and $U_{\alpha_n}''(0)=\frac{2\alpha_n}{\sigma_0^2(0)}U_{\alpha_n}(0)$. By Theorem 6.1, $\lim_{\alpha_n\to 0^+}U_{\alpha_n}''(0)=\frac{2\lambda_0}{\sigma_0^2(0)}$. Hence by elementary arguments, $\lim_{\alpha_n\to 0^+}U_{\alpha_n}'(x)\equiv U_\infty'(x)$ exists for all $x\ge 0$ and U_∞' satisfies (5.8) together with $U_\infty'(0)=0, U_\infty''(0)=\frac{2\lambda_0}{\sigma_0^2(0)}$, and $U_\infty'(x)\equiv 1$ for $x>a_\infty^*$. But Q_0' also satisfies (5.8) on $(0,b^*)$ together with the same boundary conditions as U_∞' at the origin and $Q_0'(x)\equiv 1$ for $x>b^*$. Hence $U_\infty'(x)\equiv Q_0'(x)$ for $0\le x\le a_\infty^*\wedge b^*$. Now suppose $a_\infty^*<b^*$, then $Q_0'(a_\infty^*)=Q_0'(b^*)=1$. Using (6.17) and (6.18) we conclude $Q_0'(x)\equiv 1$ on $[a_\infty^*,+\infty)$, and thus $Q_0'(x)\equiv U_\infty'(x)$. A similar result holds if $b^*<a_\infty^*$. Consequently, $\lim_{\alpha_n\to 0^+}U_{\alpha_n}'(x)=Q_0'(x)$ for all $x\ge 0$. Thus, $\lim_{\alpha\to 0^+}U_\alpha'(x)=Q_0'(x)$ follows. □

Acknowledgment. The author thanks the referees for their helpful comments.

REFERENCES

[1] L. H. R. Alvarez, *Singular stochastic control, linear diffusions, and optimal stopping: A class of solvable problems*, SIAM J. Control Optim., 39 (2001), pp. 1697–1710.
[2] L. R. Alvarez, *A class of solvable singular stochastic control problems*, Stochastics Stochastics Rep., 67 (1999), pp. 83–122.
[3] M. Arisawa and P.-L. Lions, *On ergodic stochastic control*, Comm. Partial Differential Equations, 23 (1998), pp. 2187–2217.
[4] V. E. Benes, L. A. Shepp, and H. S. Witsenhausen, *Some solvable stochastic control problems*, Stochastics, 4 (1980), pp. 39–82.
[5] M. B. Chiarolla and U. G. Haussmann, *The free boundary of the monotone follower*, SIAM J. Control Optim., 32 (1994), pp. 690–727.
[6] M. B. Chiarolla and U. G. Haussmann, *The optimal control of the cheap monotone follower*, Stochastic Stochastics Rep., 48 (1994), pp. 99–128.
[7] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[8] Y. Fujita and H. Morimoto, *Ergodic control of stochastic differential systems with controller constraints*, Stochastics Stochastics Rep., 58 (1996), pp. 245–257.
[9] I. Karatzas, *A class of singular stochastic control problems*, Adv. in Appl. Probab., 15 (1983), pp. 225–254.
[10] I. Karatzas and S. E. Shreve, *Connections between optimal stopping and singular stochastic control* I. *Monotone follower problems*, SIAM J. Control Optim., 22 (1984), pp. 856–877.
[11] I. Karatzas and S. E. Shreve, *Connections between optimal stopping and singular stochastic control* II. *Reflected follower problems*, SIAM J. Control Optim., 23 (1985), pp. 433–451.
[12] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.
[13] L. Kruk, *Optimal policies for n-dimensional singular stochastic control problems part* I: *The Skorokhod problem*, SIAM J. Control Optim., 38 (2000), pp. 1603–1622.
[14] T. G. Kurtz and R. H. Stockbridge, *Existence of Markov controls and characterization of optimal Markov controls*, SIAM J. Control Optim, 36 (1998), pp. 609–653.
[15] J. Ma, *On the principle of smooth fit for a class of singular stochastic control problems for diffusions*, SIAM J. Control Optim., 30 (1992), pp. 975–999.
[16] P. A. Meyer, *Un cours sur les integrales stochastiques*, in Seminaire de Probabilitiés X, Lecture Notes in Math. 511, Springer-Verlag, New York, 1974, pp. 245–400.
[17] H. Morimoto and M. Okada, *Some results on the Bellman equation of Ergodic control*, SIAM J. Control Optim., 38 (1999), pp. 159–174.
[18] D. Ocone and A. Weerasinghe, *Degenerate variance control of a one-dimensional diffusion*, SIAM J. Control Optim., 39 (2000), pp. 1–24.
[19] D. Ocone and A. Weerasinghe, *Degenerate variance control in the one dimensional stationary case*, Electron. J. Probab., submitted.
[20] L. C. G. Rogers and D. Williams, *Diffusions, Markov Processes, and Martingales*, in Itô Calculus, Wiley Ser. Probab. Statist. Probab. Statist. 2, John Wiley, New York, 1987.
[21] H. M. Soner and S. E. Shreve, *Regularity of the value function for a two dimensional singular stochastic control problem*, SIAM J. Control Optim., 27 (1989), pp. 876–907.
[22] A. Weerasinghe, *A stationary stochastic control for Itô Processes*, Adv. in Appl. Probab., 34 (2002), pp. 128–140.
[23] S. A. Williams, P. L. Chow, and J. L. Menaldi, *Regularity of the free boundary in singular stochastic control*, J. Differential Equations 111 (1994), pp. 175–201.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1416–1430

PARAMETER IDENTIFICATION BY REGULARIZATION FOR SURFACE REPRESENTATION VIA THE MOVING GRID APPROACH*

STEFAN KINDERMANN† AND ANDREAS NEUBAUER†

Abstract. We consider the identification of a diffusion parameter in a second order elliptic equation in two dimensions by interior measurements. The diffusion parameter is assumed to have discontinuities. For its reconstruction we propose regularization algorithms with an adaptive grid. The grid is adapted according to a measure of the smoothness of the regularized solution. For the numerical computation we compare several iterative methods such as the minimal error method, the steepest descent method, and an inexact iteratively regularized Gauss–Newton method. The computations show that these algorithms can effectively identify the discontinuities.

Key words. parameter identification, regularization via moving grids, ill-posed problems

AMS subject classification. 65J20

DOI. 10.1137/S0363012902408034

1. Introduction.

Our interest lies in the identification of a possibly discontinuous diffusion coefficient $\tilde{\gamma}$ defined on a domain $\Omega \subset \mathbb{R}^2$ in the equation

$$-\mathrm{div}(\tilde{\gamma}\nabla u) = f, \quad u_{|\partial\Omega} = 0, \tag{1.1}$$

where $f \in L^2(\Omega)$ is given, and $u(x), x \in \Omega$ is measured.

To ensure ellipticity, we assume that positive constants γ_1, γ_2 exist such that $\gamma_1 \leq \tilde{\gamma}(x) \leq \gamma_2$ almost everywhere. We want to find that part of $\tilde{\gamma}$ which differs from a constant background diffusivity which we assume to be 1, without loss of generality. Hence, our unknown is γ with $\tilde{\gamma} = 1 + \gamma$. In what follows, we denote the nonlinear operator that maps $\gamma = \tilde{\gamma} - 1$ to the solution of (1.1) by F. The problem we are faced with is an ill-posed equation

$$F(\gamma) = u, \tag{1.2}$$

where u are the given data.

It is well known that this parameter identification problem is ill-posed and that, for general data u, (1.2) does not necessarily have a solution. However, in the context of regularization theory often least squares solutions of (1.2) are considered.

For nonlinear problems, the question of whether a least squares solution is also a solution of (1.2) and if such a solution is unique is rather involved and is outside the scope of this work. For sufficient conditions for uniqueness of a solution see, e.g., [7, 18].

To avoid such problems with existence of a solution we will in the following always assume attainability; i.e., for exact data u there exists a $\gamma^\dagger$ for which (1.2) holds. Obviously in this case a least squares solution will also be a solution of (1.2).

*Received by the editors May 21, 2002; accepted for publication (in revised form) March 24, 2003; published electronically October 2, 2003. This work was supported by the Austrian Science Foundation Funds under grant SFB F013/F1317.

http://www.siam.org/journals/sicon/42-4/40803.html

†Institut für Industriemathematik, Johannes-Kepler-Universität, A-4040 Linz, Austria (kinderman@indmath.uni-linz.ac.at, neubauer@indmath.uni-linz.ac.at).

In many cases a standard regularization using Hilbert space norms (see, e.g., [5]) implemented, for example, by Tikhonov regularization does not give satisfactory results, since the discontinuities either are smeared out, if regularization is too strong, or oscillations occur when the regularization norm is too weak.

In particular for discontinuous solutions, the use of the BV-seminorm has been quite effective for this kind of problem (cf., e.g., [4, 13]).

However, the BV-functional suffers from the drawback that it is not differentiable, so usually a differentiable approximation to this functional is used. Moreover, BV-regularization shows the so-called staircase effect [19]; i.e., regularized solutions have the tendency to become piecewise constant.

Our motivation to use an adaptive grid for discontinuous solutions comes from the work on regularization by curve and surface representations (cf. [10, 11, 12]). Here the discontinuous functions are regarded as curves or surfaces. Regularization is applied to their parameterizations. In the discretized case this has the effect that the unknown γ is always defined on a grid, which changes with the iteration and is adapted to the regularized solution.

In [8, 9] it was shown that in one and two dimensions H^1-functions and a suitably adaptive grid may be used to approximate any BV-function in the sense of weak convergence.

The methods in [10, 11, 12, 16, 17] used a grid which is defined via optimization problems. Unfortunately, in higher dimensions, these optimization functionals are rather flat, leading to slow convergence. The resulting grid has the property that the mesh size is small wherever the solution exhibits discontinuities, as expected. So instead of computing the grid via optimization it seems more efficient to adapt the mesh size directly to the smoothness of γ. This adaption can be implemented efficiently by the deformation method.

We show that with this moving grid method we can use standard algorithms from regularization theory in Hilbert spaces and still obtain good results for discontinuous solutions with a small extra amount of recalculating the grid in each step.

2. Moving grid method.

2.1. General algorithm. For a numerical approach to problem (1.2), we have to set up a discrete approximation to γ by, for instance, finite element functions defined on a suitable grid.

In particular, if γ has discontinuities, and if the grid is kept fixed, a reasonable approximation of γ requires a small mesh size leading to a large number of unknowns. On the other hand, an adaptive grid allows a better resolution of discontinuities with a moderate number of variables. Thus, it seems advantageous to combine regularization with an adaptive grid.

Moreover, by our method we may use a discretization of γ that is smoother than γ itself, for example continuous ansatz functions even for discontinuous γ. Since the grid will be adaptive to the solution, the grid size will be small wherever γ has jumps, and this compensates the approximation error at the nonsmooth parts.

In our examples we work with weakly differentiable ansatz functions. The grid will be made finer wherever some measure of smoothness indicates a large gradient of γ which is regarded as possible lack of differentiability.

This idea of a moving grid has been used for numerical computations in PDEs; see, e.g., [20]. For instance, in the framework of hyperbolic equations the development of shock waves and the corresponding lack of smoothness of the solutions may not be

handled efficiently by a fixed grid. In the context of ill-posed problems, an adaptive grid approach has been successfully applied in [14] to linear integral equations.

We restrict ourselves to a moving grid; i.e., the adaptive grid is a transformation of a fixed, uniform one. Therefore, we need a transformation function ϕ which is one-to-one and onto on Ω

$$\phi : \quad \Omega \to \Omega. \tag{2.1}$$

A sufficient condition for a C^1-function ϕ to fulfill these conditions is that

$$\phi(\partial\Omega) = \partial\Omega \quad \text{and} \quad \det D\phi > 0 \quad \text{in } \Omega. \tag{2.2}$$

If c is defined on Ω on a fixed uniform grid, then we can find an approximation to γ on an adaptive grid by $c(\phi^{-1})$, where ϕ is an appropriate transformation function (again defined on a uniform grid).

We briefly describe the main ideas of our algorithm. In each step we find an approximation $c_n(\phi_n^{-1})$ of the solution, which is obtained by applying regularization to the equation

$$F(c(\phi_n^{-1})) = u \tag{2.3}$$

with fixed ϕ_n. In the next step we compute an error estimator, which measures the smoothness of $c_n(\phi_n^{-1})$. Then we calculate a new transformation function ϕ_{n+1} depending on this error estimation.

Thus, the general steps of the algorithm look as follows:

1. Start with a uniform grid and the identity as transformation function $\phi_0(\xi) = \xi \in \Omega \subset \mathbb{R}^2$, $n = 0$.
2. Compute c_n by regularization of the equation
$$F(c(\phi_n^{-1})) = u;$$
set $\gamma_n := c_n(\phi_n^{-1})$.
3. If a stopping criterion is satisfied, set $\gamma = \gamma_n$; otherwise
4. update the transformation function
$$\phi_{n+1} = T(\phi_n, c_n), \tag{2.4}$$
where T is the method of choice to define the moving grid; go to step 2.

For the stopping rule in step 3 we used Morozov's well-known discrepancy principle (cf., e.g., [5]). A detailed description of steps 2 and 4 is given below.

2.2. Regularization method. Note that in step 2 we apply regularization only to the function c and not to $c(\phi_n^{-1})$. This motivation comes from the previously mentioned idea of regularization for surface representations, where it has been observed that $c(\xi) = \gamma(\phi(\xi))$ can be chosen in H^1 even for functions γ being merely in BV.

Using the estimates in [12], it can be shown that F is a continuous and Fréchet-differentiable operator from H^1 to L^2. Hence, we may use, e.g., Tikhonov regularization with the H^1-seminorm as regularization term. In this case c_n is computed by the minimization problem

$$c_n = \operatorname*{argmin}_{c \in H^1(\Omega)} J(c, \phi_n), \tag{2.5}$$

$$J(c, \phi_n) := \|F(c(\phi_n^{-1})) - u_\delta\|^2 + \alpha \|\nabla c\|^2_{L^2(\Omega)}. \tag{2.6}$$

Here u_δ denotes the measured data, possibly contaminated with noise, where δ is the noise level, i.e.,

$$\|u - u_\delta\|_{L^2} \le \delta.$$

With fixed ϕ_n, (2.5) is a convergent regularization method for the considered problem (cf. [5]).

If ϕ_n satisfies (2.2), then the minimization problem (2.5) can also be written as

$$c_n(\phi_n^{-1}) = \operatorname*{argmin}_{w \in H^1} \|F(w) - u\|^2 + \alpha \|([D\phi_n]^{-T} \nabla w)(\det D\phi_n^{-1})^{\frac{1}{2}}\|^2_{L^2(\Omega)} \tag{2.7}$$

with $w = c(\phi_n^{-1})$. Here and in the following $D\phi$ denotes the Jacobian matrix of ϕ.

Equation (2.7) indicates that each regularization step can be seen as usual Tikhonov regularization with a weighted norm that is adapted to the grid.

Note that in (2.7) $\gamma_n = c_n(\phi_n^{-1})$ is always in H^1. This does not contradict our aim to approximate discontinuous coefficients $\gamma \notin H^1$, since the algorithm does not yield a uniform bound of $\|\gamma_n\|_{H^1}$ for all n. In fact, only $\|([D\phi_n]^{-T} \nabla \gamma_n)(\det D\phi_n^{-1})^{\frac{1}{2}}\|_{L^2(\Omega)}$ will be bounded.

Although our algorithm uses ϕ_n, where $\det D\phi_n(\xi) > 0$ always holds in Ω, after some iterations regions will occur, where $\det D\phi_n$ is numerically close to 0. These will be the parts of γ_n corresponding to the discontinuities of γ.

If we allowed generalized diffeomorphisms with $\det D\phi_n = 0$ on some part of $\Omega_0 \subset \Omega$ (i.e., if the grid were degenerate in this case), then γ_n could have discontinuities that approximate the discontinuities of the exact unknown γ. This idea has been exploited in [10, 11, 12] in one and two dimensions.

Of course, we are not restricted to Tikhonov regularization; any other convergent regularization method will be appropriate; in fact, for the numerical realization we prefer iterative regularization methods. We implemented three of them: the minimal error method, the steepest descent method, and an inexact iteratively regularized Gauss–Newton method.

To describe the ideas we introduce some notation. In step 2 we have to solve (2.3), where ϕ_n is kept fixed and c is the unknown. We use $F_{\phi_n}(c) := F(c(\phi_n^{-1}))$, and $F'_{\phi_n}(c)$ for the Fréchet derivative with respect to c, and $F'_{\phi_n}(c)^*$ for its adjoint (in the space H_0^1).

For exact data the iterate c_n in step 2 is computed by

$$c_n = \lim_{k \to \infty} c_{n,k}\,,$$

where $c_{n,k}$ is obtained out of one of the following iteration methods.

The minimal error and the steepest descent method use the iteration

$$c_{n,k+1} = c_{n,k} + \alpha_k s_k, \qquad s_k = -F'_{\phi_n}(c_{n,k})^*(F'_{\phi_n}(c_{n,k}) - u_\delta) \tag{2.8}$$

with

$$\alpha_k = \begin{cases} \dfrac{\|F_{\phi_n}(c_{n,k}) - u_\delta\|^2_{L^2}}{\|s_k\|^2_{H_0^1}} & \text{for the minimal error method,} \\[2ex] \dfrac{\|s_k\|^2_{H_0^1}}{\|F'_{\phi_n}(c_{n,k}) s_k\|^2_{L^2}} & \text{for the steepest descent method.} \end{cases} \tag{2.9}$$

Both iteration methods start with an appropriate initial value $c_{n,0}$, which we set $c_{n,0} = \gamma_{n-1}(\phi_n)$, $c_{0,0} = 0$. These iterations can be seen as Landweber iteration with an iteration dependent steplength α_k. Convergence and convergence rates for these

two methods have been investigated in [15]. If the data are exact, i.e., $\delta = 0$, then $c_{n,k}$ converges under suitable conditions on F_{ϕ_n} to a minimal norm solution of (2.3). In the noisy case the iteration is stopped according to Morozov's discrepancy principle, i.e., at the first iterate $c_{n,k}$ satisfying

$$\|F_{\phi_n}(c_{n,k}) - u_\delta\|_{L^2} \leq \tau\delta$$

with a suitable parameter $\tau > 1$.

The third iteration method we use is a variant of the iteratively regularized Gauss–Newton algorithm. For the exact algorithm a new update for $c_{n,k+1}$ is defined by the equation

$$\begin{aligned}\left(F'_{\phi_n}(c_{n,k})^* F'_{\phi_n}(c_{n,k}) + \alpha_k I\right)(c_{n,k+1} - c_{n,k}) \\ = -F'_{\phi_n}(c_{n,k})^*(F_{\phi_n}(c_{n,k}) - u_\delta) + \alpha_k c_{n,k}\,.\end{aligned} \tag{2.10}$$

The sequence α_k plays the role of a regularization parameter and can be chosen as geometrically decaying sequence $\alpha_k = \alpha_0 r^k$ with $0 < r < 1$. Again Morozov's discrepancy principle is used to stop the iteration in the presence of data noise. For the iteratively regularized Gauss–Newton iteration, convergence and convergence rates have been proven in [1].

In our computations we prefer an inexact Gauss–Newton method: instead of solving system (2.10) exactly, we use a conjugate gradient (CG) method to approximate the exact solution. That means that $c_{n,k+1} = c_{n,k} + \nu_l$, where ν_l denotes the lth step of a CG method applied to the equation

$$\left(F'_{\phi_n}(c_{n,k})^* F'_{\phi_n}(c_{n,k}) + \alpha_k I\right)\nu = -F'_{\phi_n}(c_{n,k})^*(F_{\phi_n}(c_{n,k}) - u_\delta) + \alpha_k c_{n,k}. \tag{2.11}$$

In the CG method the operator on the left-hand side has to be applied to functions ν. This means that we have only to calculate directional derivatives and hence it is not necessary in the discretized version to compute and store the matrix corresponding to the operator $F'_{\phi_n}(c_{n,k})^* F'_{\phi_n}(c_{n,k})$.

Since for the evaluation of $F, F'h, F'^* z$ we always have to solve a PDE, we intend to use a limited number of CG-iterations to save computation time. Thus we propose to keep the number of CG-iterations fixed ($l = 10$ suffices). Alternatively, we stop the iteration if (2.11) is solved with a precision to 10% of the initial error. These stopping criteria keep the computational effort moderate; however, since the equation is not solved exactly, the question of convergence of the algorithm arises. Practically, the method shows convergence, and we expect that it could also be verified theoretically from the following reasons: Note that the first step in a CG-iteration for (2.11) is identical to a steepest descent step for the Tikhonov functional. On the other hand, an exact solution (i.e., $\lim_{l\to\infty} \nu_l$) of (2.11) is identical to one Gauss–Newton step for this functional. Both iterations—steepest descent and iteratively regularized Gauss–Newton—yield convergence under reasonable conditions, and our inexact iteration is somewhere in between. Thus, we expect convergence also for the iteration with a fixed number of CG-steps l.

2.3. Deformation method. We now turn to step 4 in our algorithm above: the transformation T is defined via the deformation method. It provides direct control over the cell size. We briefly describe the main ideas following [20].

Given a positive monitoring function $m(\zeta, t) > 0$ depending on the space variable ζ and time t, we want to construct a deformation function $\phi(\xi, t)$ such that

$$\begin{aligned}\det D\phi(\xi,t) &= m(\phi(\xi,t),t)\,, \quad \xi \in \Omega, t > 0\,, \\ \phi(\xi,0) &= \phi_{\text{init}}(\xi), \qquad\quad \xi \in \Omega.\end{aligned} \tag{2.12}$$

In numerical computations for PDEs m usually describes the smoothness of the solutions. If m is small—indicating lack of smoothness—then the volume of a grid element will be small too. A necessary solvability condition for m is the normalization property:

$$\int_\Omega \left(\frac{1}{m(\zeta,t)} - 1 \right) d\zeta = 0. \tag{2.13}$$

Although (2.12) is a highly nonlinear PDE there is an elegant algorithm to solve it (cf. [2]).

First, we define a velocity field $v(\zeta,t)$ by

$$\begin{aligned} \operatorname{div} v(\zeta,t) &= -\frac{\partial}{\partial t}\frac{1}{m(\zeta,t)}, \qquad \zeta \in \Omega, t \geq 0, \\ \langle\, v(\zeta,t), n(\zeta,t)\,\rangle &= 0 \quad \zeta \in \partial\Omega, t > 0; \end{aligned}$$

here $n(\zeta,t)$ denotes the unit outward normal to $\partial\Omega$. v may be calculated by the gradient $v = \nabla w$ of the solution of the Neumann problem (for fixed $t \geq 0$)

$$\begin{aligned} \Delta w(\zeta,t) &= -\frac{\partial}{\partial t}\frac{1}{m(\zeta,t)}\,, \quad \zeta \in \Omega\,, \\ \frac{\partial}{\partial n} w &= 0 \qquad\qquad\quad \text{on } \partial\Omega\,. \end{aligned} \tag{2.14}$$

Note that the solvability condition of the Neumann problem is satisfied by the normalization property (2.13). A function $\phi(\xi,t)$ satisfying (2.12) is obtained as solution of the system of ODEs (for fixed $\xi \in \Omega$)

$$\begin{aligned} \frac{d}{dt}\phi(\xi,t) &= v(\phi(\xi,t),t)m(\phi(\xi,t),t)\,, \quad t > 0\,, \\ \phi(\xi,0) &= \phi_{\text{init}}(\xi)\,. \end{aligned} \tag{2.15}$$

In our case, t plays the role of a homotopy parameter connecting the initial grid at $t = 0$ with the final grid satisfying (2.12).

In each iteration in step 4 of our algorithm, we compute the deformation function ϕ_{n+1} by

$$\det D\phi_{n+1}(\xi) = m_n(\phi_{n+1}(\xi)). \tag{2.16}$$

In our numerical realization we choose

$$m_n(\zeta) := \frac{C_n}{(1+\beta|\nabla\gamma_n(\zeta)|^2)^{\frac{1}{2}}}, \quad \gamma_n(\zeta) = c_n(\phi_n^{-1}(\zeta)), \tag{2.17}$$

$\beta > 0$ being a fixed parameter, and C_n such that the normalization property (2.13) holds.

Since we are using the monitoring function (2.17), this would require the inversion of ϕ_n. However, in [14] a variant of the above algorithm is described to circumvent this inversion. The idea is to include the transformation function ϕ_n from the previous step by defining

$$\phi_{n+1}(\xi) := \phi_n(\sigma(\xi,1))$$

with an unknown function $\sigma(\xi,t), t \in [0,1]$, such that (2.16) holds:

$$\det D\phi_{n+1}(\xi) = \det D\phi_n(\sigma(\xi,1)) \det D\sigma(\xi,1) = m_n(\phi_n(\sigma(\xi,1))).$$

This yields an equation for σ:

$$\det D\sigma(\xi,1) = \tilde{m}_n(\sigma(\xi,1)) \tag{2.18}$$

with

$$\tilde{m}_n(\xi) = \frac{m_n(\phi_n(\xi))}{\det D\phi_n(\xi)}.$$

If we denote components of the deformation function in the nth step by a, b, i.e., $\phi_n(\xi) = (a(\xi), b(\xi))$ and $\gamma_n(\zeta) = c(\phi_n^{-1}(\zeta))$, the chain rule yields ($\xi = (\xi_1, \xi_2)$)

$$\begin{aligned}\tilde{m}_n(\xi) &= \frac{C_n}{\det D(a(\xi), b(\xi))(1+\beta|\nabla\gamma_n(\phi_n(\xi))|^2)^{\frac{1}{2}}} \\ &= \frac{C_n}{(a_{\xi_1}b_{\xi_2} - a_{\xi_2}b_{\xi_1})^2 + \beta((b_{\xi_2}c_{\xi_1} - b_{\xi_1}c_{\xi_2})^2 + (a_{\xi_1}c_{\xi_2} - a_{\xi_2}c_{\xi_1})^2)^{\frac{1}{2}}}.\end{aligned}$$

We start with $\sigma(\xi,0) = \mathrm{id}(\xi) = \xi$ and use the parameter $t \in [0,1]$ to connect $\sigma(\xi,0)$ with $\sigma(\xi,1)$. The function $\sigma(\xi,t)$ is chosen to solve

$$\det \sigma(\xi,t) = \frac{1}{(1-t) + t\frac{1}{\tilde{m}(\sigma(\xi,t))}}, \quad t \in [0,1]. \tag{2.19}$$

This equation has the form (2.12) and can be solved as above (see (2.14), (2.15)). Note that by the choice of how the right-hand side in (2.19) depends on t, $w(x,t)$ in (2.14) will not depend on t and has to be solved only once in step 4.

3. Numerical realization.

3.1. Approximation of the direct problem. For the numerical computations we restrict ourselves to the unit square in $\mathbb{R}^2$, $\Omega = [0,1]^2$. Our algorithm requires solving (1.1) on Ω with $\tilde{\gamma} = 1+\gamma$. In the weak formulation we have to find $u \in H_0^1(\Omega)$ such that

$$\langle\, (1+\gamma)\nabla u, \nabla\psi \,\rangle = \langle\, f, \psi \,\rangle \qquad \forall \psi \in H_0^1(\Omega), \tag{3.1}$$

where $\langle\, \cdot, \cdot \,\rangle$ denotes the inner product on $L^2(\Omega)$. Since $\gamma = c(\phi^{-1})$ and since we want to avoid inverting ϕ, we transform the differential equation. In fact, if $u = F_\phi(c)$ solves (3.1), then $\tilde{u} := u \circ \phi$ solves

$$\langle\, (1+c)M\nabla\tilde{u}, \nabla\psi \,\rangle = \langle\, (f\circ\phi)\det D\phi, \psi \,\rangle \qquad \forall \psi \in H_0^1(\Omega), \tag{3.2}$$

with the matrix-valued function

$$M(\xi) = \det D\phi(\xi)[D\phi(\xi)]^{-1}[D\phi(\xi)]^{-T}.$$

For a numerical computation we use Courant (i.e., piecewise linear and continuous) elements for c and for the transformation function ϕ on the same uniform grid. The nodal points of the grid are given by

$$\xi_{i,j} = (\tfrac{i}{N}, \tfrac{j}{N}), \quad i,j = 0,\ldots,N. \tag{3.3}$$

This yields N^2 small squares, which are again subdivided into two triangles by the diagonal running from the lower left to the upper right corner. The $2N^2$ triangle elements are thus the triangles with corners and

$$\xi_{i,j}, \xi_{i,j+1}, \xi_{i+1,j+1}\,, \quad i,j = 0,\dots,N-1\,.$$

As usual the linear finite element basis functions $\psi_{i,j}(\xi)$ are first order polynomials on each triangle and satisfy $\psi_{i,j}(\xi_{k,l}) = \delta_{i,k}\delta_{j,l}$. $\psi_{i,j}$ is supported on at most 6 triangles which have $\xi_{i,j}$ as corner.

Given c, ϕ, then a numerical approximation $\tilde{u}_N$ of $\tilde{u}$ solving (3.2) is given as solution of

$$\begin{aligned} \langle\,(1+c)M\nabla\tilde{u}_N, \nabla\psi_{i,j}\,\rangle &= \langle\,(f\circ\phi)\det D\phi, \psi_{i,j}\,\rangle \quad \forall i,j = 1,\dots,N-1\,, \\ \tilde{u}_N &= \sum_{i,j=1}^{N-1} \tilde{u}_{i,j}\psi_{i,j}\,. \end{aligned} \tag{3.4}$$

This system is equivalent to a numerical approximation of the original weak form (3.1) on a transformed grid; i.e., the operator $F_\phi(c)$ is approximated by

$$F_\phi^N(c) := u_N = \sum_{i,j=1}^{N-1} \tilde{u}_{i,j}(\psi_{i,j}\circ\phi^{-1})\,. \tag{3.5}$$

Since we control only the cell size of the transformed grid and not the shape, the discretized equation (3.4) can be badly conditioned if the shape of the triangles on the transformed grid become nearly degenerate. This can be avoided by subdividing these degenerate triangles appropriately.

By our choice of ϕ and c, the Jacobian matrix $D\phi$ and $\nabla\psi_{i,j}$ are constant on each triangle, and c is a first order polynomial on each element; thus the integrals on the left-hand side of (3.4) can be evaluated exactly. For the right-hand side we calculate the values $f\circ\phi$ at the nodal points and then interpolate this function linearly on each triangle to evaluate the integral. By this procedure we obtain a $(N-1)^2 \times (N-1)^2$ linear system

$$A(c,\phi)\vec{u} = \vec{f}, \tag{3.6}$$

where $\vec{u}$ is a vector with entries $\tilde{u}_{i,j}$, $i,j = 1,\dots,N-1$, $A(c,\phi)$ is a sparse, band limited matrix with six nonzero off-diagonals, and $\vec{f}$ is a load vector with entries $\langle\,(f\circ\phi)\det D\phi, \psi_{i,j}\,\rangle$, $i,j = 1,\dots,N-1$. Equation (3.6) is solved using MATLAB's PCG method.

3.2. Realization of the regularization method. For the computation of the regularized solutions (cf. (2.8), (2.9), (2.11)) we have to replace the operator $F_{\phi_n}(c_{n,k})$ by $F_{\phi_n}^N(c_{n,k})$ (see (3.5)).

The effort for computing the derivative and its adjoint into one direction is of the same order as one evaluation of $F_{\phi_n}^N(c_{n,k})$. In fact, for a calculation of $u_N = F_{\phi_n}^N(c_{n,k})$ we have to solve the equation (see (3.5), (3.6))

$$A(c_{n,k},\phi_n)\vec{u} = \vec{f}$$

once, and for the derivative $w_N := (F_{\phi_n}^N)'(c_{n,k})h = \sum_{i,j=1}^{N-1} \tilde{w}_{i,j}(\psi_{i,j}\circ\phi^{-1})$ we additionally have to solve

$$A(c_{n,k},\phi_n)\vec{w} = -A(h-1,\phi_n)\vec{u}\,.$$

Note that we are interested only in elements h in the discretized space

$$X_N := \left\{ \sum_{i,j=1}^{N-1} h_{i,j}\psi_{i,j} \,:\, h_{i,j} \in \mathbb{R} \right\} \subset H_0^1 \,.$$

The adjoint of the operator $(F_{\phi_n}^N)'(c_{n,k}) : X_N \to L^2$ applied to some element $z = \sum_{i,j=1}^{N-1} z_{i,j}(\psi_{i,j} \circ \phi^{-1})$ is given by the solution of

$$\langle\, \nabla((F_{\phi_n}^N)'(c_{n,k})^* z), \nabla\psi_{i,j} \,\rangle = -\langle\, \nabla\tilde{u}^T M \nabla\tilde{\eta}, \psi_{i,j} \,\rangle \qquad \forall i,j = 1,\dots,N-1\,,$$

in X_N, where $\eta = \sum_{i,j=1}^{N-1} \eta_{i,j}\psi_{i,j}$ solves

$$A(c_{n,k},\phi)\vec{\eta} = G\vec{z}$$

with the Gramian matrix

$$G_{(i,j);(l,k)} = \langle\, \psi_{i,j} \det D\phi, \psi_{l,k} \,\rangle\,. \tag{3.7}$$

Note that $c_{n,k}$ has to satisfy $\gamma_1 \le 1 + c_{n,k} \le \gamma_2$. Therefore, we project $c_{n,k}$ onto this convex set, in each iteration step if necessary.

In our regularization methods we have two iteration loops: the index n corresponds to the update of the grid and the iteration indexed by k corresponds to the regularization step. An obvious improvement of the algorithm can be expected by including the information of the regularization iteration into the regridding step and combining these two iterations into one.

So instead of finishing the iteration with respect to k until the stopping rule is satisfied, we perform at most k_0 steps, where k_0 is a small fixed number. This means that even if the stopping rule is not yet satisfied after k_0 steps, we perform a regridding step and choose $c_{n+1,0}$ such that $\gamma_{n+1,0}$ is approximately equal to γ_{n,k_0}; i.e., $c_{n+1,0}$ equals the linear interpolant of $c_{n,k_0}(\phi_n^{-1} \circ \phi_{n+1})$. The numerical results indicate that this mixed iteration converges.

3.2.1. Realization of the deformation method. For the computation of the grid update $\phi_{n+1}(\xi) = \phi_n(\sigma(\xi,1))$ we have to solve a Neumann problem for the Poisson equation (2.14) once and the system of ODE (2.15). To do this we first compute the monitoring function $\tilde{m}_n$, which is piecewise constant on each triangle element. Note that the Poisson equation is defined on a uniform grid, and it is again solved by MATLAB's PCG algorithm. $\tilde{m}_n$ is scaled by taking C_n as $\int_\Omega \tilde{m}_n^{-1}(\xi)d\xi$ such that the solvability condition (2.13) holds. This integral can be evaluated exactly, because $\tilde{m}_n$ is piecewise constant.

The system of ODEs is solved by the classical fourth order Runge–Kutta method. Since the right-hand side of (2.15) is piecewise constant, we use bilinear interpolation on the triangles to obtain a continuous function.

Note that ϕ_n has to map the boundary of Ω onto itself. This is achieved by keeping ϕ_n fixed at the corner points of the unit square. Moreover, the first component of ϕ is not changed at the lines $y = 0$ and $y = 1$, and vice versa for the second component at $x = 0$ and $x = 1$.

Since, due to discretization, it may happen that the function σ has a negative determinant, we also include a smoothing step then by setting the values of σ at the corner points $\xi_{i,j}$ to the mean value of neighboring nodal points. However, it turned

out in our computations that such a smoothing step is rarely necessary as long as β is not too large. For large values of β a smaller step size in the Runge–Kutta method was sufficient to guarantee the positivity of the determinant.

If the grid is updated, i.e., a new transformation function ϕ_{n+1} is computed, we also have to recalculate several vectors and matrices depending on the grid. In fact, we have to compute the Gramian matrix (3.7) and the load vector $\vec{f}$ (3.6). The stiffness matrix $A(c_{n,k}, \phi_n)$ has to be updated in every step, too.

4. Numerical results. For the numerical experiments we used $f(x, y) = \sin(2\pi x)\sin(2\pi y)$ as right-hand side in (1.1).

The following examples were considered.

Example 4.1. Circle: $\gamma = 1 + 2\chi_{B_{0.55,0.45}(0.3)}$.

Example 4.2. Ramp: $\gamma(x, y) = 1 + 2\frac{x-0.25}{0.35}\chi_{\{(x,y)|0.25\le x\le 0.6, 0.2\le y\le 0.8\}}$.

Example 4.3. Moon: $\gamma = 1 + 2(\chi_{B_{0.55,0.5}(0.3)}(1 - \chi_{B_{0.4,0.5}(0.25)}))$.

Example 4.4. Rectangle (chosen from [6]): $\gamma = 1 + \chi_{\{(x,y)|0.3\le x\le 0.5, 0.3\le x+y\le 0.6\}}$.

Example 4.5. Circle and rectangle:

$$\gamma = 1 + 2(\chi_{B_{0.35,0.65}(0.15)} + \chi_{\{(x,y)|0.1\le x-y\le 0.6, 0.7\le x+y\le 1.1\}}.$$

($B_{x_0,y_0}(r)$ denotes the circle with midpoint at (x_0, y_0) and radius r.)

For all examples the data points were first computed using a fine uniform grid with $N = 120$ and then contaminated by random noise. This grid is much finer than the one used to calculate the regularized solutions (usually $N = 40$). By this we avoid so-called inverse crimes, namely, to use the same setup for the calculation of the simulated data and the regularization itself.

Since all matrices in our above algorithm are sparse we need a storage effort of order $\mathcal{O}(N^2)$. This shows the advantage of using a CG method for the iteratively regularized Gauss–Newton iteration, since the full matrix in (2.11) has $\mathcal{O}(N^4)$ entries.

Viewing computational time, the most expensive step is to solve (3.6). Basically, the PCG algorithm needs $\mathrm{cond}^{\frac{1}{2}} \log \epsilon^{-1}$ iterations to obtain an error reduction of ϵ. Here cond is the condition number of the preconditioned stiffness matrix $A(c, \phi)$. Each iteration needs a complexity of $\mathcal{O}(N^2)$ flops. Note that it is not necessary to solve the equation to a precision which is below the data noise.

We emphasize that the regridding step is not very expensive at all. In fact, it requires solving one PDE and the Runge–Kutta step. After all, its complexity is about of the same order as one function evaluation of $F^N_{\phi_n}(c_{n,k})$.

We first report about the differences of the iteration methods we used. Figures 4.1 and 4.2 show the relative L^2-error $\frac{\|\gamma_{n,k}-\gamma^\dagger\|}{\|\gamma_0-\gamma^\dagger\|}$ versus the iteration number for the minimal error method and the steepest descent method for Example 4.1 with discretization $N = 40$ and unperturbed data. Here every step of the form (2.8) is counted as one iteration. Hence if the grid is updated after k_0 steps of (2.8), then the total iteration number is $n_{tot} = (n-1)k_0 + k$ in the previous notation.

We chose different intervals for the grid update step. The dotted line indicates a grid update in every second step, i.e., $c_{n+1,0} \approx c_{n,2}$. The full lines correspond to a grid update in every 5th step and the dashed lines to one in every 10th step.

The two iteration methods show a different behavior: the error reduction for the minimal error method is not as smooth as for the steepest descent method. Moreover, the former exhibits a stronger dependence on the choice of the grid update intervals k_0, whereas the latter is quite insensitive to it. In fact, for the minimal error method we found the best results for $k_0 = 10$. However, the minimal error method performs better with respect to the required CPU-time. For an error reduction of 70% the

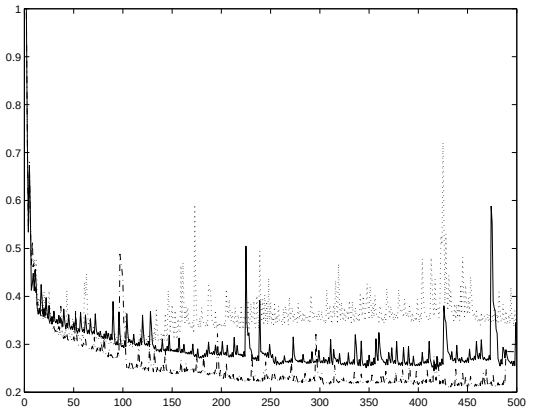


FIG. 4.1. *Error reduction vs. iteration: minimal error method.*

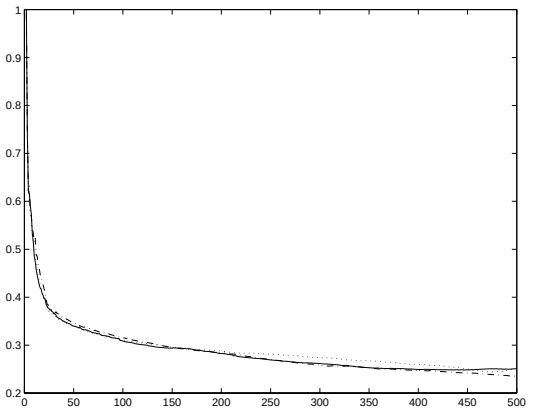


FIG. 4.2. *Error reduction vs. iteration: steepest descent method.*

steepest descent method needed more than three times the CPU-time for the minimal error method.

The inexact iteratively regularized Gauss–Newton iteration yields the best results. It is quite insensitive to the choice of k_0 compared to the steepest descent method and performs better with respect to the CPU-time than the minimal error method.

Figures 4.3–4.7 show the results for our examples with exact and noisy data (5% noise). We used $N = 40$, $\beta = 10$, and the inexact iteratively regularized Gauss–Newton method (2.11) with $\alpha_{k+1} = 0.9\alpha_k$. A grid update was done in every second step (i.e., $k_0 = 2$). For exact data the iteration was stopped at $\alpha_k = 10^{-10}$ and for noisy data we used the discrepancy principle as stopping rule.

The results show that we can identify the location of discontinuities quite well. Obviously, for noisy data the resolution is not as sharp as for exact ones. Note that even for exact data we have noise due to discretization. The second example exhibits that our algorithm does not suffer from the staircasing effect of several BV-regularizations. We observed that it is difficult to identify γ in regions where the gradient of u vanishes or is small. This effect can be expected, since γ is not identifiable at points where $\nabla u = 0$.

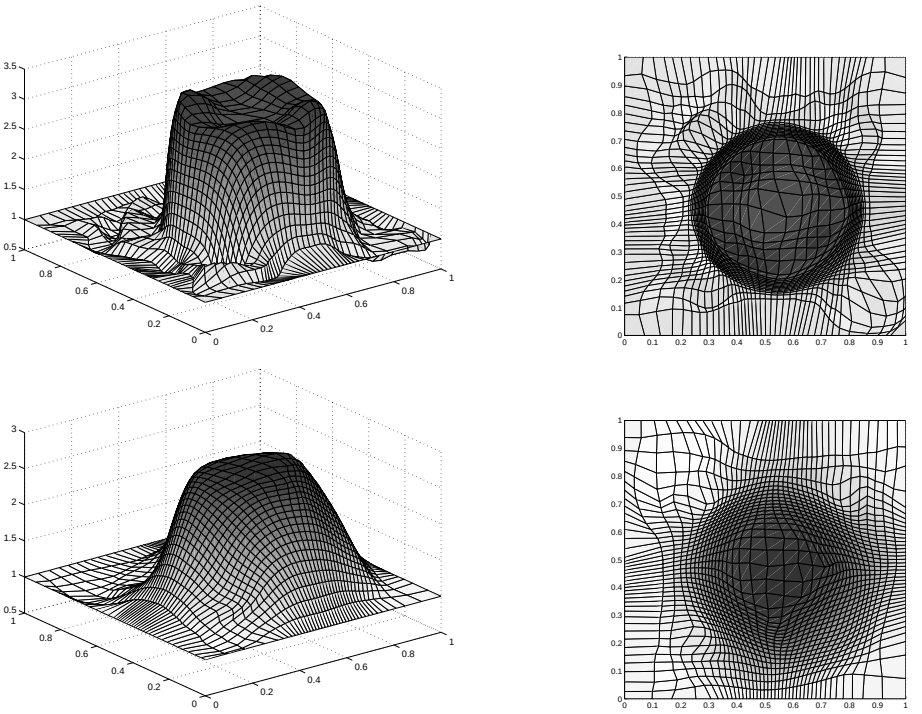


FIG. 4.3. *Reconstruction for Example* 4.1 *for exact data (above) and for* 5% *noise (below).*

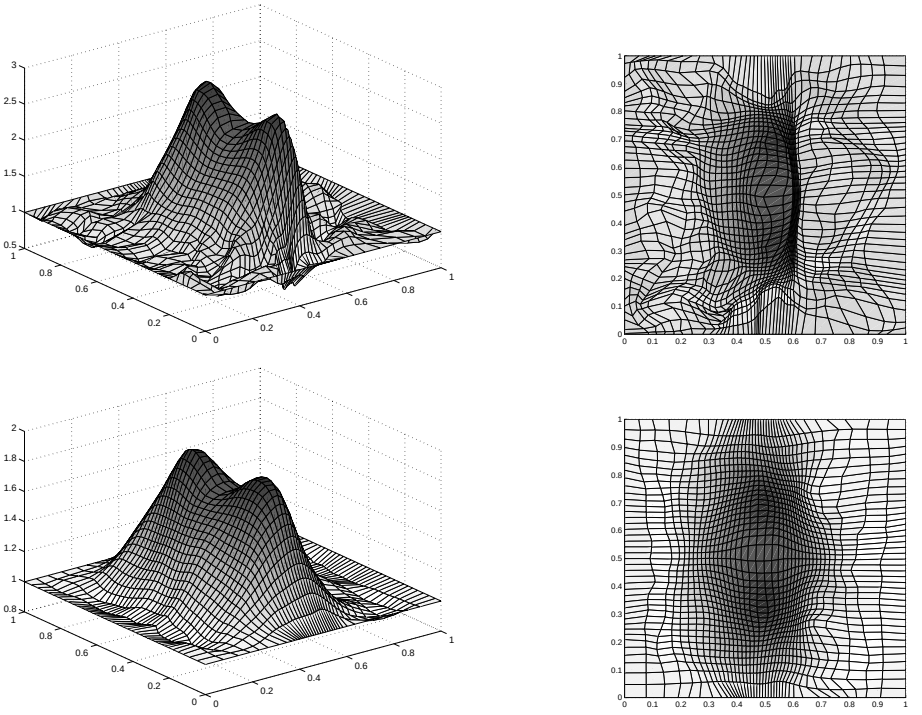


FIG. 4.4. *Reconstruction for Example* 4.2 *for exact data (above) and for* 5% *noise (below).*

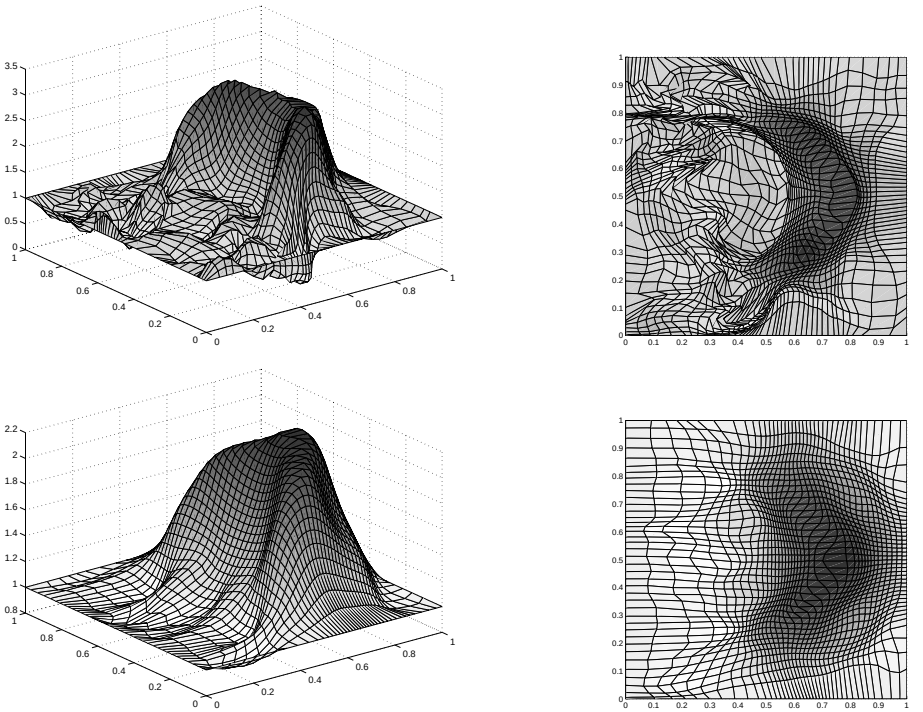


FIG. 4.5. *Reconstruction for Example* 4.3 *for exact data (above) and for* 5% *noise (below).*

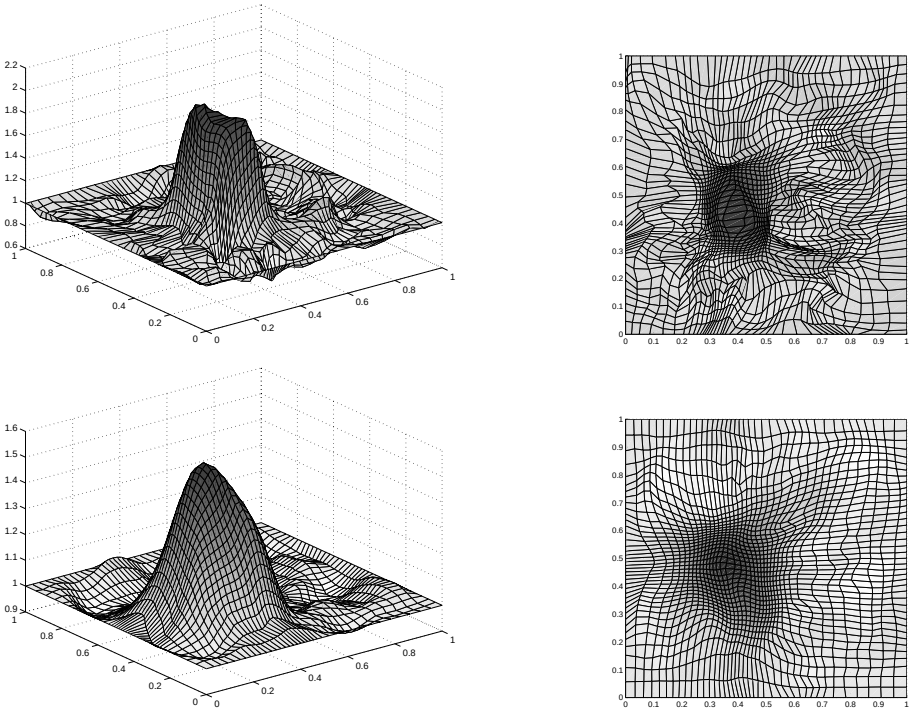


FIG. 4.6. *Reconstruction for Example* 4.4 *for exact data (above) and for* 5% *noise (below).*

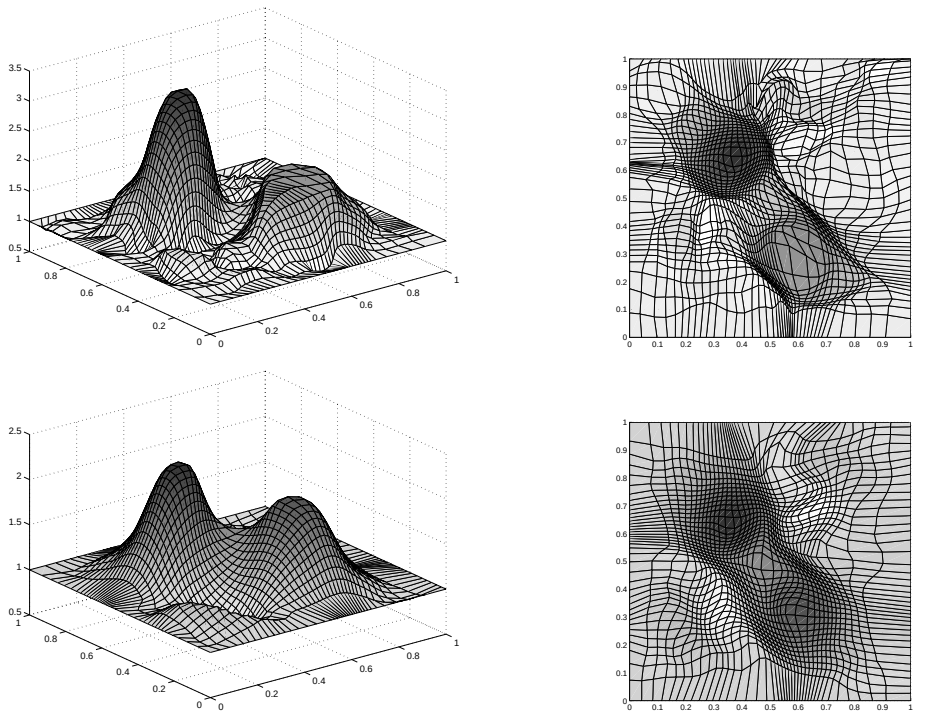


FIG. 4.7. *Reconstruction for Example* 4.5 *for exact data (above) and for* 5% *noise (below).*

Moreover, for exact data, the results for γ show a rather low dependence on the choice of the parameter β and the number of regridding steps. There seems to be a broad region of values of β yielding similar results. For noisy data we obtained slightly better results with respect to the resolution of discontinuities with larger values of β and more regridding steps.

The choice of the method for solving the ODE (2.15) is quite important. We additionally tried to use an explicit Euler method instead of the Runge–Kutta method, with unsatisfactory results. Using a method with lower order often yields negative determinants and requires smoothing steps of the grid, which makes the resolution of discontinuities not so sharp.

We want to compare our results of Example 4.4 with the ones in [6]: In the noise-free case we need about 40 iterations of Gauss–Newton steps. A rough comparison of the number of iterations to the augmented Lagrangian algorithm used in [6] shows that the latter method might be faster. However, a fair comparison seems to be hard, since the performance strongly depends on the regularization parameter and also on the regularization norm which was not BV-like in [6]. The quality of the results of both approaches is comparable. However, we think that the advantage of our approach is that good results will also be obtained if the discontinuity lines are not chosen parallel to the grid lines (see Example 4.5).

Of course the moving grid idea does not depend on any specific regularization method; hence a combination of the augmented Lagrangian method or SQP-like methods (cf. [3]) with a moving grid seems to be quite promising.

Finally, we want to mention that our regularization algorithm is less dependent on the choice of the size of the initial grid, since we are using a variable grid that is

adapted to the solution. A comparison of our results with numerical computations where the grid size was chosen as $N = 100$ instead of $N = 40$ showed practically no improvement for Examples 4.1–4.3. It is obvious that the possibility of choosing a coarser grid saves a lot of computation time. In Example 4.5 two disjoint regions of discontinuity have to be identified. For a good resolution more grid lines are needed for the gap between the two regions. Hence, the result was slightly better for the finer grid.

REFERENCES

[1] B. Blaschke, A. Neubauer, and O. Scherzer, *On convergence rates for the iteratively regularized Gauss-Newton method*, IMA J. Numer. Anal., 17 (1997), pp. 421–436.

[2] P. Bochev, G. Liao, and G. dela Pena, *Analysis and computation for adaptive moving grids by deformation*, Numer. Methods Partial Differential Equations, 12 (1996), pp. 489–506.

[3] M. Burger and W. Mühlhuber, *Iterative regularization of parameter identification problems by sequential quadratic programming methods*, Inverse Problems, 18 (2002), pp. 943–969.

[4] Z. Chen and J. Zou, *An augmented Lagrangian method for identifying discontinuous parameters in elliptic systems*, SIAM J. Control. Optim, 37 (1999), pp. 892–910.

[5] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, The Netherlands, 1996.

[6] K. Ito, M. Kroller, and K. Kunisch, *A numerical study of an augmented Lagrangian method for the estimation of parameters in elliptic systems*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 884–910.

[7] K. Ito and K. Kunisch, *On the injectivity and linearization of the coefficient-to-solution mapping for elliptic boundary value problems*, J. Math. Anal. Appl., 188 (1994), pp. 1040–1066.

[8] S. Kindermann, *Regularization of Ill-Posed Problems with Discontinuous Solutions by Curve and Surface Representations*, Ph.D. thesis, University of Linz, Austria, 2001.

[9] S. Kindermann and A. Neubauer, *Each $\mathcal{BV}$-Function is Representable by an $\mathcal{H}^1$-Curve*, Technical report 1/1999, Industrial Mathematics Institute, University of Linz, Austria, 1999.

[10] S. Kindermann and A. Neubauer, *Identification of discontinuous parameters by regularization for curve representations*, Inverse Problems, 15 (1999), pp. 1559–1572.

[11] S. Kindermann and A. Neubauer, *Estimation of discontinuous parameters of elliptic partial differential equations by regularization for surface representations*, Inverse Problems, 17 (2001), pp. 789–803.

[12] S. Kindermann and A. Neubauer, *Regularization for surface representations of discontinuous solutions of linear ill-posed problems*, Numer. Funct. Anal. Optim., 22 (2001), pp. 79–105.

[13] R. Luce and S. Perez, *Parameter identification for an elliptic partial differential equation with distributed noisy data*, Inverse Problems, 15 (1999), pp. 291–307.

[14] A. Neubauer, *Estimation of discontinuous solutions of ill-posed problems by regularization for surface representations: Numerical realization via moving grids*, in Recent Developments in Theories and Numerics, International Conference on Inverse Problems, Y. C. Hon, M. Yamamoto, J. Cheng, and J. Y. Lee, eds., World Scientific, Singapore, 2003, pp. 67–83.

[15] A. Neubauer and O. Scherzer, *A convergent rate result for a steepest descent method and a minimal error method for the solution of nonlinear ill-posed problems*, Z. Anal. Anwendungen, 14 (1995), pp. 369–377.

[16] A. Neubauer and O. Scherzer, *Reconstruction of discontinuous solutions from blurred data*, in Computational, Experimental, and Numerical Methods for Solving Ill-Posed Inverse Imaging Problems: Medical and Nonmedical Applications, Proc. SPIE 3171, R. L. Barbour, M. J. Carvlin, and M. A. Fiddy, eds., SPIE, Bellingham, WA, 1997, pp. 34–41.

[17] A. Neubauer and O. Scherzer, *Regularization for curve representations: Uniform convergence for discontinuous solutions of ill-posed problems*, SIAM J. Appl. Math., 58 (1998), pp. 1891–1900.

[18] G. R. Richter, *An inverse problem for the steady state diffusion equation*, SIAM J. Appl. Math., 41 (1981), pp. 210–221.

[19] W. Ring, *Structural properties of solutions of total variation regularization problems*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 799–810.

[20] B. Semper and G. Liao, *A moving grid finite element method using grid deformation*, Numer. Methods Partial Differential Equations, 11 (1995), pp. 603–615.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1431–1450

UNCONTROLLABILITY FOR LINEAR AUTONOMOUS MULTI-INPUT DYNAMICAL SYSTEMS DEPENDING ON PARAMETERS*

ALEXEI A. MAILYBAEV†

Abstract. Linear multi-input dynamical systems smoothly depending on parameters are considered. A set of parameter values corresponding to uncontrollable systems (an uncontrollability set) is studied. The typical (generic) structure of the uncontrollability set is described. A constructive method of perturbation analysis of the uncontrollability set is developed. Formulae of first-order approximations for the uncontrollability set and generalized eigenvalues (uncontrollable modes) are derived and used for numerical construction of the uncontrollability set. The method is based on the versal deformation theory for matrix pairs under feedback equivalence. As an example, the uncontrollability set is found for a three-parameter two-degree-of-freedom mechanical system.

Key words. uncontrollability, feedback equivalence, uncontrollable mode, perturbation, matrix pair, versal deformation

AMS subject classification. 93B

DOI. 10.1137/S0363012902405339

1. Introduction.

The concept of controllability is very important in the study of control problems. It describes the possibility of transferring a system to a required state using a given set of input (control) variables. Uncontrollability makes operation of a system in the whole state space impossible and signals the fundamental trouble with a control problem or underlying physical system. Similar difficulties appear for nearly uncontrollable systems, which require big control resources for control performance and which are strongly affected by imperfections and disturbances of the system.

The well-developed control theory exists for linear dynamical systems [3, 13]. Nevertheless, there are essential problems in using classical controllability criteria for numerical implementation. This is related to the structural instability of an uncontrollable system, which becomes controllable under an arbitrarily small perturbation. In this respect, it is important to know how far our system is from the nearest uncontrollable one. This problem was studied by many authors; see [2, 10, 11, 12] and references therein.

Design of a particular control system requires checking the controllability condition for different values of parameters. In this analysis, the knowledge on the structure of the uncontrollability set (a set of parameter values corresponding to uncontrollable systems) is very useful and helps in avoiding the dangerous nearness to uncontrollability. In this paper, basic qualitative properties of the uncontrollability set for a generic (typical) multi-input linear dynamical system depending on several parameters are investigated. This includes a description of a regular part of the uncontrollability set and its basic singularities. Then the quantitative perturbation method for local analysis of the uncontrollability set near its regular points is developed. Application

*Received by the editors April 11, 2002; accepted for publication (in revised form) April 1, 2003; published electronically October 2, 2003. This work was supported by Russian Foundation of Basic Research grant 02-01-39004.

http://www.siam.org/journals/sicon/42-4/40533.html

†Institute of Mechanics, Moscow State Lomonosov University, Michurinsky pr. 1, 119192 Moscow, Russia (mailybaev@imec.msu.ru).

of this method to numerical calculation of a regular part of the uncontrollability set in the parameter space is proposed. As an example, an elastic mechanical system controlled by a force and dependent on three design parameters is studied. The results of the paper are based on the versal deformation theory for matrix pairs under state feedback equivalence [6, 8, 14].

The paper is organized as follows. Section 2 describes basic concepts and results of the mathematical control theory. Section 3 studies the qualitative structure of the uncontrollability set. Quantitative perturbation method for local analysis of the uncontrollability set is developed in section 4. In section 5, a numerical method for computation of a regular part of the uncontrollability set is constructed and applied to the analysis of a specific mechanical system. Section 6 is devoted to singularities of the uncontrollability set. The conclusion summarizes the contribution.

2. Controllability, feedback equivalence, and versal deformation. Let us consider a dynamical system described by the system of linear ordinary differential equations

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \tag{2.1}$$

where $\mathbf{x} \in \mathbb{R}^n$ is a state vector, $\mathbf{u} \in \mathbb{R}^m$ is an input vector, and $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ and $n \times m$ real time-invariant matrices, respectively; the dot denotes differentiation with respect to time t. System (2.1) is called (state) *controllable* if it is possible to construct a control signal $\mathbf{u}(t)$ that will transfer an initial state to any final state in finite time [3, 13]. Otherwise, the system is said to be *uncontrollable*. The classical criterion of controllability says that the system is controllable if and only if the $n \times nm$ *controllability matrix* $\mathbf{C} = [\mathbf{B}, \mathbf{AB}, \dots, \mathbf{A}^{n-1}\mathbf{B}]$ has full rank [3, 13]

$$\operatorname{rank}[\mathbf{B}, \mathbf{AB}, \dots, \mathbf{A}^{n-1}\mathbf{B}] = n. \tag{2.2}$$

System (2.1) is determined by a matrix pair $(\mathbf{A}, \mathbf{B})$. Let us denote the set of all matrix pairs by $\mathcal{M} = \{(\mathbf{A}, \mathbf{B}) \mid \mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{B} \in \mathbb{R}^{n\times m}\}$. We will use the short notation $\alpha = (\mathbf{A}, \mathbf{B})$ for a matrix pair. The matrix pair α is called controllable (or uncontrollable) if the corresponding system (2.1) is controllable (or uncontrollable).

2.1. Feedback equivalence. Let us apply a linear feedback transformation of the input vector and perform a change of basis in the state and input spaces. The new state vector $\bar{\mathbf{x}}$ and input vector $\bar{\mathbf{u}}$ are related to $\mathbf{x}$ and $\mathbf{u}$ by the expressions

$$\mathbf{x} = \mathbf{P}\bar{\mathbf{x}}, \quad \mathbf{u} = \mathbf{Q}\bar{\mathbf{u}} + \mathbf{R}\bar{\mathbf{x}}, \tag{2.3}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are $n \times n$ and $m \times m$ nonsingular real matrices; $\mathbf{R}$ is an $m \times n$ real matrix. The substitution of (2.3) into (2.1) yields

$$\dot{\bar{\mathbf{x}}}(t) = \bar{\mathbf{A}}\bar{\mathbf{x}}(t) + \bar{\mathbf{B}}\bar{\mathbf{u}}(t), \tag{2.4}$$

where

$$\bar{\mathbf{A}} = \mathbf{P}^{-1}(\mathbf{AP} + \mathbf{BR}), \quad \bar{\mathbf{B}} = \mathbf{P}^{-1}\mathbf{BQ}. \tag{2.5}$$

Systems (2.1) and (2.4) (as well as the corresponding matrix pairs $\alpha = (\mathbf{A}, \mathbf{B})$ and $\bar{\alpha} = (\bar{\mathbf{A}}, \bar{\mathbf{B}})$) are called *feedback equivalent* (also called *block equivalent*) [9, 13]. The controllability property is invariant under the feedback equivalence transformation.

Let us introduce the short notation $\bar{\alpha} = \gamma \circ \alpha$ for the feedback equivalence transformation (2.5) applied to a matrix pair $\alpha = (\mathbf{A}, \mathbf{B})$, where γ denotes a triple $\gamma = (\mathbf{P}, \mathbf{Q}, \mathbf{R})$. We denote the set of all feedback equivalence transformations by $\mathcal{G} = \{(\mathbf{P}, \mathbf{Q}, \mathbf{R}) \mid \mathbf{P} \in \mathbb{R}^{n\times n}, \mathbf{Q} \in \mathbb{R}^{m\times m}, \mathbf{R} \in \mathbb{R}^{m\times n}, \det \mathbf{P} \neq 0, \det \mathbf{Q} \neq 0\}$. Note that relations (2.5) determine the Lie group structure in $\mathcal{G}$ with the multiplication and inversion of elements given by

$$\begin{gathered} \gamma_1\gamma_2 = (\mathbf{P}_1\mathbf{P}_2,\, \mathbf{Q}_1\mathbf{Q}_2,\, \mathbf{R}_1\mathbf{P}_2 + \mathbf{Q}_1\mathbf{R}_2) \in \mathcal{G}, \quad \gamma_i = (\mathbf{P}_i, \mathbf{Q}_i, \mathbf{R}_i) \in \mathcal{G}, \\ \gamma^{-1} = (\mathbf{P}^{-1},\, \mathbf{Q}^{-1},\, -\mathbf{Q}^{-1}\mathbf{R}\mathbf{P}^{-1}) \in \mathcal{G}, \quad \gamma = (\mathbf{P}, \mathbf{Q}, \mathbf{R}) \in \mathcal{G} \end{gathered} \tag{2.6}$$

such that $\gamma_1\gamma_2 \circ \alpha = \gamma_2 \circ (\gamma_1 \circ \alpha)$ and $\gamma\gamma^{-1} \circ \alpha = \alpha$ for any $\alpha \in \mathcal{M}$. The unit element of $\mathcal{G}$ is $e = (\mathbf{I}_n, \mathbf{I}_m, 0)$, where $\mathbf{I}_n$ and $\mathbf{I}_m$ are $n \times n$ and $m \times m$ identity matrices, respectively. This triple has the property $\alpha = e \circ \alpha$ for any pair $\alpha \in \mathcal{M}$.

2.2. Equivalence classes and their local structure. Let us consider a fixed matrix pair $\alpha_0 = (\mathbf{A}_0, \mathbf{B}_0) \in \mathcal{M}$. A set of all pairs α feedback equivalent to α_0 is called the *orbit* of α_0 and denoted by

$$\mathcal{O}(\alpha_0) = \{\alpha \in \mathcal{M} \mid \alpha = \gamma \circ \alpha_0,\ \gamma \in \mathcal{G}\}. \tag{2.7}$$

The orbit is a smooth submanifold of $\mathcal{M}$.

The orbit $\mathcal{O}(\alpha_0)$ can be represented by its arbitrary member $\alpha \in \mathcal{O}(\alpha_0)$. Therefore, to describe the orbit it is convenient to choose a pair α having, in some sense, the simplest form. One such form, called a *Brunovsky canonical form*, is represented by the matrix pair [9, 13]

$$\alpha_b = (\mathbf{A}_b, \mathbf{B}_b), \quad \mathbf{A}_b = \begin{pmatrix} \mathbf{N} & 0 \\ 0 & \mathbf{J} \end{pmatrix}, \quad \mathbf{B}_b = \begin{pmatrix} \mathbf{E} & 0 \\ 0 & 0 \end{pmatrix}, \tag{2.8}$$

where $\mathbf{J}$ is the real Jordan canonical form (real counterpart of the Jordan form); $\mathbf{N} = \operatorname{diag}(\mathbf{N}_1, \dots, \mathbf{N}_r)$; $\mathbf{E} = \operatorname{diag}(\mathbf{E}_1, \dots, \mathbf{E}_r)$; $\mathbf{N}_i$ and $\mathbf{E}_i$ are $k_i \times k_i$ and $k_i \times 1$ matrices, respectively, having the form

$$\mathbf{N}_i = \begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}, \quad \mathbf{E}_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}. \tag{2.9}$$

The numbers $k_1 \geq \cdots \geq k_r > 0$ are called *controllability indices* or *Kronecker indices* of the system; $\mathbf{N}$ is called a *Kronecker part* and $\mathbf{J}$ is called a *Jordan part* of the Brunovsky form. Any matrix pair $\alpha \in \mathcal{M}$ is feedback equivalent to the corresponding Brunovsky form, which is unique up to the permutation of blocks in the Jordan matrix $\mathbf{J}$. Using condition (2.2), one can check that the matrix pair α is controllable if and only if its Brunovsky form has no Jordan part. If the pair is uncontrollable, then eigenvalues of the Jordan part $\mathbf{J}$ are called *generalized eigenvalues* or *uncontrollable modes*.

Let us introduce the function $f_{\alpha_0}(\gamma) = \gamma \circ \alpha_0$, which is a smooth function from $\mathcal{G}$ to $\mathcal{M}$. Then the orbit $\mathcal{O}(\alpha_0)$ can be seen as the range of the mapping f_{α_0}, i.e.,

$$\mathcal{O}(\alpha_0) = \operatorname{Im} f_{\alpha_0}. \tag{2.10}$$

Let us denote by $T_e\mathcal{G}$ the tangent space to $\mathcal{G}$ at the unit element e. Since $\mathcal{G}$ is an open set, we have $T_e\mathcal{G} = \{(\mathbf{U}, \mathbf{V}, \mathbf{W}) \mid \mathbf{U} \in \mathbb{R}^{n\times n}, \mathbf{V} \in \mathbb{R}^{m\times m}, \mathbf{W} \in \mathbb{R}^{m\times n}\}$. The linear spaces $\mathcal{M}$ and $T_e\mathcal{G}$ are equipped with the Euclidean scalar products

$$\begin{aligned} &\langle \alpha_1, \alpha_2\rangle_1 = \operatorname{trace}(\mathbf{A}_1\mathbf{A}_2^T) + \operatorname{trace}(\mathbf{B}_1\mathbf{B}_2^T),\\ &\langle \xi_1, \xi_2\rangle_2 = \operatorname{trace}(\mathbf{U}_1\mathbf{U}_2^T) + \operatorname{trace}(\mathbf{V}_1\mathbf{V}_2^T) + \operatorname{trace}(\mathbf{W}_1\mathbf{W}_2^T), \end{aligned} \tag{2.11}$$

where $\alpha_i = (\mathbf{A}_i, \mathbf{B}_i) \in \mathcal{M}$, $\xi_i = (\mathbf{U}_i, \mathbf{V}_i, \mathbf{W}_i) \in T_e\mathcal{G}$, and $\mathbf{A}^T$ denotes the transposed matrix. Let us introduce linear mappings $df_{\alpha_0} : T_e\mathcal{G} \longrightarrow \mathcal{M}$ and $df^*_{\alpha_0} : \mathcal{M} \longrightarrow T_e\mathcal{G}$ as follows:

$$\begin{aligned} &df_{\alpha_0}(\xi) = (\mathbf{A}_0\mathbf{U} - \mathbf{U}\mathbf{A}_0 + \mathbf{B}_0\mathbf{W},\ \mathbf{B}_0\mathbf{V} - \mathbf{U}\mathbf{B}_0), \quad \xi = (\mathbf{U}, \mathbf{V}, \mathbf{W}),\\ &df^*_{\alpha_0}(\alpha) = (\mathbf{A}_0^T\mathbf{A} - \mathbf{A}\mathbf{A}_0^T - \mathbf{B}\mathbf{B}_0^T,\ \mathbf{B}_0^T\mathbf{B},\ \mathbf{B}_0^T\mathbf{A}), \quad \alpha = (\mathbf{A}, \mathbf{B}). \end{aligned} \tag{2.12}$$

It is straightforward to check that df_{α_0} is the differential of the function f_{α_0} at the unit element $e = (\mathbf{I}_n, \mathbf{I}_m, 0)$ [6], and $df^*_{\alpha_0}$ is the adjoint function defined by the relation [7]

$$\langle df_{\alpha_0}(\xi), \alpha\rangle_1 = \langle \xi, df^*_{\alpha_0}(\alpha)\rangle_2, \quad \alpha \in \mathcal{M}, \quad \xi \in T_e\mathcal{G}. \tag{2.13}$$

A local structure of the orbit $\mathcal{O}(\alpha_0)$ near the point α_0 is determined by the range of the mapping df_{α_0} and null-space of $df^*_{\alpha_0}$ as follows [8]:

$$T_{\alpha_0}\mathcal{O}(\alpha_0) = \operatorname{Im} df_{\alpha_0}, \tag{2.14}$$

$$(T_{\alpha_0}\mathcal{O}(\alpha_0))^{\perp} = \operatorname{Ker} df^*_{\alpha_0}, \tag{2.15}$$

where $T_{\alpha_0}\mathcal{O}(\alpha_0)$ is the tangent space to $\mathcal{O}(\alpha_0)$ at the point α_0; $(T_{\alpha_0}\mathcal{O}(\alpha_0))^{\perp}$ denotes the normal complimentary subspace to $T_{\alpha_0}\mathcal{O}(\alpha_0)$ in $\mathcal{M}$. In addition, we denote by $(T_{\alpha_0}\mathcal{O}(\alpha_0))^c$ an arbitrary complimentary subspace to $T_{\alpha_0}\mathcal{O}(\alpha_0)$ in $\mathcal{M}$.

2.3. Versal deformation. A multiparameter dynamical system (2.1) is described by a matrix pair $\alpha(\mathbf{p})$ smoothly dependent on a vector of parameters $\mathbf{p} = (p_1, \dots, p_k) \in \mathbb{R}^k$. The function $\alpha(\mathbf{p})$ is called the *family* of matrix pairs. The family $\alpha(\mathbf{p})$ determined in the neighborhood of a point $\mathbf{p}_0$ is called a *deformation* of the matrix pair $\alpha_0 = \alpha(\mathbf{p}_0)$. Using feedback equivalence transformation $\gamma(\mathbf{p}) \circ \alpha(\mathbf{p})$, the family $\alpha(\mathbf{p})$ can be reduced to a more simple form. Nevertheless, a reduction to the Brunovsky form generally cannot be achieved by the feedback equivalence transformation $\gamma(\mathbf{p})$ smoothly dependent on parameters. The following theorem proved in [6, 8] provides another form called a versal deformation that can be used for multiparameter families of matrix pairs. Note that the concept of versal deformation was first introduced by Arnold [1] for families of square complex matrices; see also [14] for the generalization to the case of a Lie group acting on a complex manifold.

THEOREM 2.1. *Let $\alpha(\mathbf{p})$ be a family of matrix pairs. Then in the neighborhood of a point $\mathbf{p}_0$, the family $\alpha(\mathbf{p})$ can be represented in the form*

$$\alpha(\mathbf{p}) = \gamma(\mathbf{p}) \circ \left(\alpha_0 + \sum_{i=1}^{\ell} q_i(\mathbf{p})\, \alpha_i^c\right). \tag{2.16}$$

In this formula $\{\alpha_1^c, \dots, \alpha_\ell^c\}$, $\ell = \dim (T_{\alpha_0}\mathcal{O}(\alpha_0))^c$, is a basis of $(T_{\alpha_0}\mathcal{O}(\alpha_0))^c$; $\gamma(\mathbf{p})$ is a feedback equivalence transformation smoothly dependent on $\mathbf{p}$ such that $\gamma(\mathbf{p}_0) = e$; $q_1(\mathbf{p}), \dots, q_\ell(\mathbf{p})$ are smooth functions, whose values and derivatives at $\mathbf{p}_0$ are

$$q_1(\mathbf{p}_0) = \dots = q_\ell(\mathbf{p}_0) = 0, \tag{2.17}$$

$$\begin{pmatrix} \dfrac{\partial q_1}{\partial p_j} \\ \vdots \\ \dfrac{\partial q_\ell}{\partial p_j} \end{pmatrix} = \mathbf{Z}^{-1} \begin{pmatrix} \left\langle \dfrac{\partial \alpha}{\partial p_j}, \alpha_1^n \right\rangle_1 \\ \vdots \\ \left\langle \dfrac{\partial \alpha}{\partial p_j}, \alpha_\ell^n \right\rangle_1 \end{pmatrix}, \tag{2.18}$$

where $\{\alpha_1^n, \dots, \alpha_\ell^n\}$ *is a basis of the linear space* $(T_{\alpha_0}\mathcal{O}(\alpha_0))^\perp$*;* $\mathbf{Z}$ *is a nonsingular* $\ell \times \ell$ *matrix with the elements* $z_{ij} = \langle \alpha_j^c, \alpha_i^n \rangle_1$.

Formulae for derivatives of the functions $q_1(\mathbf{p}), \dots, q_\ell(\mathbf{p})$ and $\gamma(\mathbf{p})$ of any order were derived in [8]. The family of matrix pairs

$$\beta(\mathbf{q}) = \alpha_0 + \sum_{i=1}^{\ell} q_i \, \alpha_i^c, \quad \mathbf{q} = (q_1, \dots, q_\ell), \tag{2.19}$$

is called a *versal deformation* of the matrix pair α_0; $\mathbf{q}$ is a parameter vector of the versal deformation [1]. The main idea of the above theorem is that any matrix family $\alpha(\mathbf{p})$ with a given pair $\alpha_0 = \alpha(\mathbf{p}_0)$ can be transformed locally to the versal deformation $\beta(\mathbf{q})$, which has an explicit and simple form, by the feedback equivalence transformation $\gamma(\mathbf{p})$ smoothly dependent on $\mathbf{p}$ and smooth change of parameters $\mathbf{q} = \mathbf{q}(\mathbf{p})$. Note that the bases $\{\alpha_1^c, \dots, \alpha_\ell^c\}$ and $\{\alpha_1^n, \dots, \alpha_\ell^n\}$ have been found explicitly in [6] for matrix pairs reduced to the Brunovsky canonical form.

Example 2.1. Let us consider a one-parameter family of matrix pairs $\alpha(p) = (\mathbf{A}(p), \mathbf{B}(p))$, where

$$\mathbf{A}(p) = \begin{pmatrix} p & 0 & p^2 \\ 2p & p & -p \\ 3p & p & 2+p^3 \end{pmatrix}, \quad \mathbf{B}(p) = \begin{pmatrix} 1 & p \\ p & 1 \\ 0 & -p \end{pmatrix}. \tag{2.20}$$

Family (2.20) determines a one-parameter dynamical system (2.1) with three-dimensional state space and two-dimensional input vector. The pair $\alpha_0 = \alpha(p_0)$ for $p_0 = 0$ has the form

$$\alpha_0 = \left(\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \right), \tag{2.21}$$

which is the Brunovsky canonical form (2.8), (2.9) with $r = 2$, $k_1 = k_2 = 1$, and $\mathbf{J} = (2)$. Solving the linear equation $df^*_{\alpha_0}(\alpha) = 0$ with respect to α and using relation (2.15), we find that the space $(T_{\alpha_0}\mathcal{O}(\alpha_0))^\perp$ has dimension $\ell = 3$ and consists of the matrix pairs

$$\left(\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ q_1 & q_2 & q_3 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 2q_1 & 2q_2 \end{pmatrix} \right) \in (T_{\alpha_0}\mathcal{O}(\alpha_0))^\perp, \quad q_1, q_2, q_3 \in \mathbb{R}. \tag{2.22}$$

The complimentary subspace $(T_{\alpha_0}\mathcal{O}(\alpha_0))^c$ can be chosen in a more simple form as follows:

$$\left(\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ q_1 & q_2 & q_3 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \right) \in (T_{\alpha_0}\mathcal{O}(\alpha_0))^c, \quad q_1, q_2, q_3 \in \mathbb{R}. \tag{2.23}$$

Expressions (2.22) and (2.23) generate the bases $\{\alpha_1^n, \alpha_2^n, \alpha_3^n\}$ and $\{\alpha_1^c, \alpha_2^c, \alpha_3^c\}$ as coefficients corresponding to q_1, q_2, q_3.

By Theorem 2.1, family (2.20) can be represented in the form $\alpha(p) = \gamma(p) \circ \beta(\mathbf{q}(p))$, where $\beta(\mathbf{q})$ is the versal deformation

$$\beta(\mathbf{q}) = \left(\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ q_1 & q_2 & 2+q_3 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \right), \quad \mathbf{q} = (q_1, q_2, q_3), \tag{2.24}$$

$\gamma(p)$ is a feedback equivalence transformation smoothly dependent on p, and $\mathbf{q} = \mathbf{q}(p)$ is a smooth change of parameters. Derivatives of the functions $q_1(p)$, $q_2(p)$, $q_3(p)$ at $p_0 = 0$ can be calculated by expression (2.18) as follows:

$$\frac{dq_1}{dp} = 3, \quad \frac{dq_2}{dp} = -1, \quad \frac{dq_3}{dp} = 0. \tag{2.25}$$

3. Structure of the uncontrollability set. Let us consider a multiparameter dynamical system described by a family of matrix pairs $\alpha(\mathbf{p})$. A set of values of the parameter vector $\mathbf{p}$ such that $\alpha(\mathbf{p})$ is uncontrollable is called the *uncontrollability set* and denoted by

$$\mathcal{N} = \{\mathbf{p} \mid \alpha(\mathbf{p}) \text{ is uncontrollable}\}. \tag{3.1}$$

It is known that $\mathcal{N}$ is typically a set of zero measure [13]. In particular, any uncontrollable system can be made controllable by an arbitrarily small perturbation of the matrices $\mathbf{A}$ and $\mathbf{B}$. Nevertheless, when the parameter vector is close to the uncontrollability set $\mathcal{N}$, the system becomes practically uncontrollable due to small perturbations and uncertainties existing in every physical system, and the necessity of using big control resources (large $\|\mathbf{u}\|$) or long time for control operation. This makes analysis and construction of the uncontrollability set in the parameter space important for the design of control systems.

As was mentioned in the previous section, for a fixed value of $\mathbf{p}$ the matrix pair $\alpha(\mathbf{p})$ can be transformed to the Brunovsky canonical form (2.8), and the uncontrollability of $\alpha(\mathbf{p})$ is equivalent to the existence of the Jordan part in this form. Let us consider two specific types of the Brunovsky form that will be important for further analysis. These two forms are represented by the 1×1 and 2×2 Jordan parts

$$\mathbf{J}_\sigma = (\sigma), \quad \mathbf{J}_{\sigma \pm i\omega} = \begin{pmatrix} \sigma & \omega \\ -\omega & \sigma \end{pmatrix}, \quad \sigma, \omega \in \mathbb{R},\ \omega > 0, \tag{3.2}$$

corresponding to a real simple eigenvalue σ and a pair of complex conjugate simple eigenvalues $\sigma \pm i\omega$, respectively. Structure of the Kronecker part can be arbitrary. We refer to matrix pairs having the described structures of the Brunovsky form as pairs of $\mathbf{J}_\sigma$ and $\mathbf{J}_{\sigma \pm i\omega}$ types.

To describe the qualitative structure of the uncontrollability set $\mathcal{N}$, it is reasonable to restrict our attention to the *generic* (typical) situation. This corresponds to a typical form of the set $\mathcal{N}$ such that small perturbations of the family $\alpha(\mathbf{p})$ do not lead to qualitative changes in the geometry and structure of $\mathcal{N}$ but result only in its small shift in the parameter space. For more precise mathematical formulation of the concept "generic," see [1]. Consideration of the generic case allows extracting the most typical and interesting information on the structure of the uncontrollability

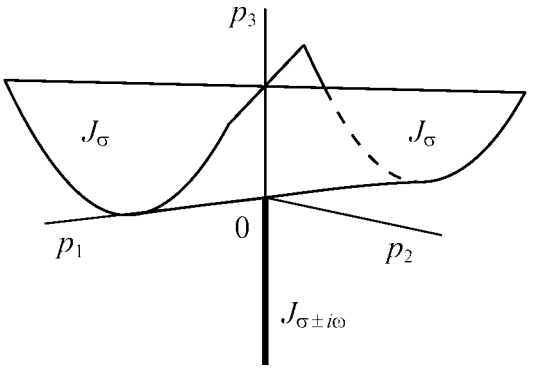


FIG. 3.1. *Structure of the uncontrollability set.*

set, which is valid for almost all multi-input linear dynamical systems dependent on parameters. The following theorem provides the basic qualitative description of this set.

THEOREM 3.1. *In the generic case the uncontrollability set $\mathcal{N}$ of the family of matrix pairs $\alpha(\mathbf{p})$ has a regular part, which consists of smooth surfaces of codimensions m and $2m$ corresponding to matrix pairs of $\mathbf{J}_\sigma$ and $\mathbf{J}_{\sigma\pm i\omega}$ types, respectively. Points $\mathbf{p} \in \mathcal{N}$ such that the matrix pair $\alpha(\mathbf{p})$ has a different type of the Brunovsky form belong to the boundary of these surfaces.*

Example 3.1. Let us consider the family of matrix pairs

$$\alpha(\mathbf{p}) = \left(\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ p_1 & 0 & 0 & 1 \\ p_2 & 0 & p_3 & 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \right), \quad \mathbf{p} = (p_1, p_2, p_3). \tag{3.3}$$

Using the controllability condition (2.2), we find the uncontrollability set in the form

$$\mathcal{N} = \{\mathbf{p} \in \mathbb{R}^3 \mid p_1^2 p_3 - p_2^2 = 0\}. \tag{3.4}$$

This set is shown in Figure 3.1. It represents the well-known Whitney–Cayley umbrella [1]. The structure of $\mathcal{N}$ agrees with Theorem 3.1. Indeed, the set $\mathcal{N}$ has a "handle" (the ray $p_1 = p_2 = 0$, $p_3 < 0$), which has codimension $2m = 2$ and corresponds to pairs of $\mathbf{J}_{\sigma\pm i\omega}$ type with the generalized eigenvalues $\sigma \pm i\omega = \pm i\sqrt{-p_3}$. There are two smooth surfaces of codimension $m = 1$ determined by the equation $p_1^2 p_3 - p_2^2 = 0$ for $p_3 \geq 0$, $p_1 < 0$ and $p_3 \geq 0$, $p_1 > 0$, which correspond to the pairs of $\mathbf{J}_\sigma$ type; the generalized eigenvalue is $\sigma = -p_2/p_1$. The "handle" and two surfaces form a regular part of $\mathcal{N}$. There are also different types of uncontrollable pairs: at the point $\mathbf{p} = 0$ we have a pair with the Jordan part $\mathbf{J} = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$, and points of the ray $p_1 = p_2 = 0$, $p_3 > 0$ correspond to pairs having two different real generalized eigenvalues $\sigma_{1,2} = \pm\sqrt{p_3}$. These points determine singularities of $\mathcal{N}$ and belong to the boundary of the regular part.

According to Theorem 3.1, we can expect to find a similar structure of the uncontrollability set for almost all families of matrix pairs. This means that for the construction of the uncontrollability set it is sufficient to find its regular part, i.e., points of $\mathbf{J}_\sigma$ and $\mathbf{J}_{\sigma\pm i\omega}$ types forming smooth surfaces of codimensions m and $2m$, respectively. Then the whole set $\mathcal{N}$ will be the closure of these surfaces.

Proof. The proof is based on general results of the singularity theory and uses versal deformations. In order to avoid special mathematical language, we describe general steps and ideas of the proof, while the details can be easily reconstructed by the reader using the cited literature.

The decomposition of the parameter space into subsets (*strata*) according to the Brunovsky structure of the corresponding matrix pair $\alpha(\mathbf{p})$ is called the *bifurcation diagram*. The singularity theory says that the qualitative local structure of the bifurcation diagram (and, hence, of the uncontrollability set) for a generic family $\alpha(\mathbf{p})$ is the same as for the versal deformation $\beta(\mathbf{q})$ [1]. In particular, codimensions of strata in the corresponding parameter spaces coincide. This follows from the relation $\alpha(\mathbf{p}) = \gamma(\mathbf{p}) \circ \beta(\mathbf{q}(\mathbf{p}))$ and the property that the Brunovsky canonical form is invariant under the feedback equivalence transformation $\gamma(\mathbf{p})$.

Let us consider a matrix pair $\alpha_0 = \alpha(\mathbf{p}_0)$. Without loss of generality, we can assume that α_0 is transformed to the Brunovsky canonical form (2.8). Then its versal deformation can be taken in the form

$$\beta(\mathbf{q}) = \left(\begin{pmatrix} \mathbf{N} & 0 \\ \mathbf{X}^c_{21} & \mathbf{J} + \mathbf{X}^c_{22} \end{pmatrix}, \begin{pmatrix} \mathbf{E} + \mathbf{Y}^c_{11} & \mathbf{Y}^c_{12} \\ 0 & \mathbf{Y}^c_{22} \end{pmatrix} \right), \tag{3.5}$$

where $\mathbf{X}^c_{ij}(\mathbf{q})$ and $\mathbf{Y}^c_{ij}(\mathbf{q})$ are matrices depending on a vector of parameters $\mathbf{q} \in \mathbb{R}^\ell$ such that every matrix is zero at $\mathbf{q} = 0$. Explicit form of these matrices depends on the structure of $\mathbf{N}$ and $\mathbf{J}$. In the case $\mathbf{J} = \mathbf{J}_\sigma$ we can take [6]

$$\mathbf{X}^c_{21} = \big(\mathbf{L}^c_1(q_1), \dots, \mathbf{L}^c_r(q_r)\big),\ \mathbf{Y}^c_{22} = (q_{r+1}, \dots, q_m),\ \mathbf{J} + \mathbf{X}^c_{22} = (\sigma + q_{m+1}), \tag{3.6}$$

where $\mathbf{L}^c_i(q_i) = (q_i, 0, \dots, 0)$ is a $1 \times k_i$ matrix; $k_1, \dots, k_r$ are the controllability indices of the Kronecker part $\mathbf{N}$. The matrices $\mathbf{Y}^c_{11}$ and $\mathbf{Y}^c_{12}$ depend on $q_{m+2}, \dots, q_\ell$. Using the controllability condition (2.2), one can show that in the vicinity of $\mathbf{q} = 0$ versal deformation (3.5), (3.6) is controllable if and only if at least one of the parameters $q_1, \dots, q_m$ is nonzero. Hence, the uncontrollability set for the versal deformation is given locally by the equalities

$$q_1 = \dots = q_m = 0. \tag{3.7}$$

Taking a point $\mathbf{q}$ such that (3.7) holds, it is easy to see that the pair $\beta(\mathbf{q})$ has $\mathbf{J}_\sigma$ type with the generalized eigenvalue $\sigma + q_{m+1}$. Since the uncontrollability set of a generic family of matrix pairs $\alpha(\mathbf{p})$ has the same local structure, we conclude that the set $\mathcal{N}$ in the neighborhood of a point $\mathbf{p}_0$, corresponding to a pair of $\mathbf{J}_\sigma$ type, is a smooth surface of codimension m, whose points correspond to pairs of $\mathbf{J}_\sigma$ type.

Analogously, in the case $\mathbf{J} = \mathbf{J}_{\sigma \pm i\omega}$ we can take [6]

$$\begin{gathered} \mathbf{X}^c_{21} = \big(\mathbf{L}^c_1(q_1, q_2), \dots, \mathbf{L}^c_r(q_{2r-1}, q_{2r})\big), \\ \mathbf{Y}^c_{22} = \begin{pmatrix} q_{2r+1} & \cdots & q_{2m-1} \\ q_{2r+2} & \cdots & q_{2m} \end{pmatrix}, \\ \mathbf{J} + \mathbf{X}^c_{22} = \begin{pmatrix} \sigma + q_{2m+1} & \omega + q_{2m+2} \\ -\omega - q_{2m+2} & \sigma + q_{2m+1} \end{pmatrix}, \end{gathered} \tag{3.8}$$

where $\mathbf{L}^c_i(q_{2i-1}, q_{2i})$ is a $2 \times k_i$ matrix of the form

$$\mathbf{L}^c_i(q_{2i-1}, q_{2i}) = \begin{pmatrix} q_{2i-1} & 0 & \cdots & 0 \\ q_{2i} & 0 & \cdots & 0 \end{pmatrix}. \tag{3.9}$$

The matrices $\mathbf{Y}_{11}^c$ and $\mathbf{Y}_{12}^c$ depend on $q_{2m+3}, \dots, q_\ell$. The uncontrollability set of this versal deformation in the vicinity of $\mathbf{q} = 0$ is given by the equations

$$q_1 = \cdots = q_{2m} = 0. \tag{3.10}$$

At the points $\mathbf{q}$ satisfying (3.10), the pair $\beta(\mathbf{q})$ has $\mathbf{J}_{\sigma\pm i\omega}$ type, and the complex conjugate generalized eigenvalues are equal to $\sigma + q_{2m+1} \pm i(\omega + q_{2m+2})$. Hence, the uncontrollability set of a generic family $\alpha(\mathbf{p})$ in the neighborhood of a point $\mathbf{p}_0$, corresponding to a pair of $\mathbf{J}_{\sigma\pm i\omega}$ type, is a smooth surface of codimension $2m$. Points of this surface correspond to matrix pairs of $\mathbf{J}_{\sigma\pm i\omega}$ type.

Now, let us consider a matrix pair α_0 having a different Jordan part $\mathbf{J}$. In this case the block $\mathbf{J} + \mathbf{X}_{22}^c$ is identical to the versal deformation of a square matrix $\mathbf{J}$ under the similarity equivalence in the space of square matrices [1, 6]. From [1] we know that, for any $\mathbf{J}$, taking arbitrarily small parameters $\mathbf{q}$, we can obtain the matrix $\mathbf{J} + \mathbf{X}_{22}^c(\mathbf{q})$ having only simple eigenvalues. Then, taking arbitrarily small parameters in the block $\mathbf{X}_{21}^c(\mathbf{q})$, we can destroy the Jordan structure corresponding to every simple eigenvalue, leaving only one real generalized eigenvalue σ or one pair of complex conjugate generalized eigenvalues $\sigma \pm i\omega$. Hence, an uncontrollable system of $\mathbf{J}_\sigma$ or $\mathbf{J}_{\sigma\pm i\omega}$ type can be found in any neighborhood of $\mathbf{q} = 0$. The same holds for a generic family of matrix pairs $\alpha(\mathbf{p})$; i.e., a point $\mathbf{p}_0 \in \mathcal{N}$ either lies on a surface represented by matrix pairs of $\mathbf{J}_\sigma$ or $\mathbf{J}_{\sigma\pm i\omega}$ types (regular points of $\mathcal{N}$) or belongs to a boundary of these surfaces. □

Remark. Let us consider matrix pairs α with a fixed Jordan part $\mathbf{J}$ in the Brunovsky form. In the generic case, a Kronecker part $\mathbf{N}$ in the Brunovsky form of almost all matrix pairs α has maximal possible number of blocks r, and the sizes $k_1, \dots, k_r$ of these blocks are different by no more than one, i.e., $k_1 = \cdots = k_{r'} = k_{r'+1} + 1 = \cdots = k_r + 1$ [6]. In this case $\mathbf{Y}_{11}^c = 0$ and $\mathbf{Y}_{12}^c = 0$ in versal deformation (3.5). Such a Kronecker part cannot be changed by a small perturbation of the parameters if we do not change the structure of the Jordan part. On the contrary, if the matrix pair has a different Kronecker part, then we can always find a nearby matrix pair with the generic Kronecker part, keeping the structure of the Jordan part [6].

4. Perturbation analysis of the uncontrollability set. Let us assume that we are given a point $\mathbf{p}_0 \in \mathcal{N}$ corresponding to a pair $\alpha(\mathbf{p}_0)$ of $\mathbf{J}_\sigma$ or $\mathbf{J}_{\sigma\pm i\omega}$ type. From Theorem 3.1 we know the generic structure of $\mathcal{N}$ in the neighborhood of $\mathbf{p}_0$. Nevertheless, some symmetry or degeneracy of the family $\alpha(\mathbf{p})$ may cause the appearance of a nongeneric structure. For example, the pair

$$\alpha(\mathbf{p}) = \left(\begin{pmatrix} 0 & 0 \\ 0 & p_1 + p_2 + p_3 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right) \tag{4.1}$$

is uncontrollable for all $\mathbf{p} \in \mathbb{R}^3$, and $\alpha(\mathbf{p})$ has $\mathbf{J}_\sigma$ type for any $\mathbf{p}$. Clearly, an arbitrarily small perturbation of the family can result in the generic structure of $\mathcal{N}$. For example, taking the $(2,1)$th element of the first matrix to be ϵp_1 for an arbitrarily small $\epsilon > 0$, the set $\mathcal{N}$ becomes the plane $p_1 = 0$ of codimension 1, which is the generic case. Therefore, it would be useful to have a constructive criterion guaranteeing that the structure of $\mathcal{N}$ is generic for a given family $\alpha(\mathbf{p})$. For applications and numerical analysis of the uncontrollability set it is also important to have quantitative local information on $\mathcal{N}$, i.e., its tangent plane and perturbations of the generalized eigenvalues. The solution of these problems is given in the following theorems.

Let $\mathbf{p}_0$ be a point of the uncontrollability set $\mathcal{N}$ for a family of matrix pairs $\alpha(\mathbf{p})$. Let $\gamma_b = (\mathbf{P}, \mathbf{Q}, \mathbf{R}) \in \mathcal{G}$ be a triple determining the feedback equivalence transformation of the pair $\alpha_0 = \alpha(\mathbf{p}_0)$ to its Brunovsky canonical form $\alpha_b = \gamma_b \circ \alpha_0$.

For the pair α_0 of $\mathbf{J}_\sigma$ type with the generalized eigenvalue σ, we define real vectors $\mathbf{f}_i = (f_i^1, \dots, f_i^k)$, $i = 1, \dots, m$, and $\mathbf{f}_\sigma = (f_\sigma^1, \dots, f_\sigma^k)$ with the components

$$
\begin{aligned}
f_i^j &= \mathbf{P}^{-1}(n,:)\left[\frac{\partial \mathbf{A}}{\partial p_j}\sum_{s=1}^{k_i}\sigma^{s-1}\mathbf{P}(:,K_i+s)\right. \\
&\quad \left. + \frac{\partial \mathbf{B}}{\partial p_j}\left(\sum_{s=1}^{k_i}\sigma^{s-1}\mathbf{R}(:,K_i+s) + \sigma^{k_i}\mathbf{Q}(:,i)\right)\right], \quad i = 1, \dots, r; \\
f_i^j &= \mathbf{P}^{-1}(n,:)\frac{\partial \mathbf{B}}{\partial p_j}\mathbf{Q}(:,i), \quad i = r+1, \dots, m; \\
f_\sigma^j &= \mathbf{P}^{-1}(n,:)\left[\frac{\partial \mathbf{A}}{\partial p_j}\mathbf{P}(:,n) + \frac{\partial \mathbf{B}}{\partial p_j}\mathbf{R}(:,n)\right];
\end{aligned}
\tag{4.2}
$$

where $K_1 = 0$, $K_i = k_1 + \dots + k_{i-1}$; $\mathbf{P}^{-1}(n,:)$, $\mathbf{P}(:,i)$, $\mathbf{Q}(:,i)$, and $\mathbf{R}(:,i)$ denote the nth row of $\mathbf{P}^{-1}$ and the ith columns of $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{R}$, respectively.

THEOREM 4.1. *Let us assume that $\alpha_0 = \alpha(\mathbf{p}_0)$ is a matrix pair of $\mathbf{J}_\sigma$ type, and the vectors $\mathbf{f}_1, \dots, \mathbf{f}_m$ are linearly independent. Then in the vicinity of $\mathbf{p}_0$ the uncontrollability set $\mathcal{N}$ is a smooth surface of codimension m corresponding to matrix pairs of $\mathbf{J}_\sigma$ type. Its tangent plane at $\mathbf{p}_0$ is given by the equations*

$$
(\mathbf{f}_1, \mathbf{p} - \mathbf{p}_0) = \dots = (\mathbf{f}_m, \mathbf{p} - \mathbf{p}_0) = 0,
\tag{4.3}
$$

where $(\mathbf{f}_i, \mathbf{p} - \mathbf{p}_0) = \sum_{j=1}^k f_i^j (p_j - p_{0j})$ is a scalar product in $\mathbb{R}^k$. The generalized eigenvalue for $\mathbf{p} \in \mathcal{N}$ in the neighborhood of the point $\mathbf{p}_0$ is given by the expression

$$
\sigma + (\mathbf{f}_\sigma, \mathbf{p} - \mathbf{p}_0) + o(\|\mathbf{p} - \mathbf{p}_0\|).
\tag{4.4}
$$

For the pair α_0 of $\mathbf{J}_{\sigma \pm i\omega}$ type with the generalized eigenvalues $\sigma \pm i\omega$, we define real vectors $\mathbf{f}_s = (f_s^1, \dots, f_s^k)$, $s = 1, \dots, 2m$, $\mathbf{f}_\sigma = (f_\sigma^1, \dots, f_\sigma^k)$, and $\mathbf{f}_\omega = (f_\omega^1, \dots, f_\omega^k)$ with the components

$$
\begin{aligned}
f_{2s-1}^j + i f_{2s}^j &= \sum_{z=0}^{1} i^{1-z}\mathbf{P}^{-1}(n-z,:)\left[\frac{\partial \mathbf{A}}{\partial p_j}\sum_{v=1}^{k_s}(\sigma - i\omega)^{v-1}\mathbf{P}(:,K_s+v)\right. \\
&\quad \left. + \frac{\partial \mathbf{B}}{\partial p_j}\left(\sum_{v=1}^{k_s}(\sigma - i\omega)^{v-1}\mathbf{R}(:,K_s+v) + (\sigma - i\omega)^{k_s}\mathbf{Q}(:,s)\right)\right], \quad s = 1, \dots, r;
\end{aligned}
\tag{4.5}
$$

$$
f_{2s-1}^j + i f_{2s}^j = \sum_{z=0}^{1} i^{1-z}\mathbf{P}^{-1}(n-z,:)\frac{\partial \mathbf{B}}{\partial p_j}\mathbf{Q}(:,s), \quad s = r+1, \dots, m;
\tag{4.6}
$$

$$
f_\sigma^j + i f_\omega^j = \frac{1}{2}\sum_{z=0}^{1}\sum_{v=0}^{1} i^{z-v}\mathbf{P}^{-1}(n-z,:)\left[\frac{\partial \mathbf{A}}{\partial p_j}\mathbf{P}(:,n-v) + \frac{\partial \mathbf{B}}{\partial p_j}\mathbf{R}(:,n-v)\right];
\tag{4.7}
$$

where i is the imaginary unit.

THEOREM 4.2. *Let us assume that $\alpha_0 = \alpha(\mathbf{p}_0)$ is a matrix pair of $\mathbf{J}_{\sigma\pm i\omega}$ type, and the vectors $\mathbf{f}_1, \dots, \mathbf{f}_{2m}$ are linearly independent. Then in the vicinity of $\mathbf{p}_0$ the uncontrollability set $\mathcal{N}$ is a smooth surface of codimension $2m$ corresponding to matrix pairs of $\mathbf{J}_{\sigma\pm i\omega}$ type. Its tangent plane at $\mathbf{p}_0$ is given by the equations*

$$(\mathbf{f}_1, \mathbf{p} - \mathbf{p}_0) = \cdots = (\mathbf{f}_{2m}, \mathbf{p} - \mathbf{p}_0) = 0. \tag{4.8}$$

The generalized eigenvalues for $\mathbf{p} \in \mathcal{N}$ in the neighborhood of $\mathbf{p}_0$ are given by the expression

$$\sigma + (\mathbf{f}_\sigma, \mathbf{p} - \mathbf{p}_0) \pm i\big(\omega + (\mathbf{f}_\omega, \mathbf{p} - \mathbf{p}_0)\big) + o(\|\mathbf{p} - \mathbf{p}_0\|). \tag{4.9}$$

The important consequence of Theorems 4.1 and 4.2 is that to determine the local structure of the uncontrollability set we need only derivatives of the family $\alpha(\mathbf{p})$ with respect to the parameters at $\mathbf{p}_0$ and the triple $\gamma_b = (\mathbf{P}, \mathbf{Q}, \mathbf{R})$ transforming the pair α_0 to the Brunovsky canonical form. The triple γ_b can be found using the software developed in [4, 5], which provides the Kronecker canonical form of the matrix pencil $(\mathbf{A}_0, \mathbf{B}_0) - \lambda(\mathbf{I}_n, 0)$. The Brunovsky form can be obtained from the Kronecker canonical form by permutation of columns [9].

Example 4.1. Let us consider a three-parameter two-input system (2.1) with the matrices $\mathbf{A}$ and $\mathbf{B}$ given by the relations

$$\mathbf{A}(\mathbf{p}) = \begin{pmatrix} -p_3 & p_1p_2 & p_1p_2 \\ 2 - p_3 & 3 + p_1p_2 & 1 + p_1p_2 \\ -2 - 2p_2^2 & p_1 - 1 & 1 - p_2 \end{pmatrix}, \quad \mathbf{B}(\mathbf{p}) = \begin{pmatrix} 1 & -p_2 \\ 1 & 1 - p_2 \\ -1 & p_2 - 1 \end{pmatrix}.$$

The matrix pair $\alpha_0 = (\mathbf{A}(\mathbf{p}_0), \mathbf{B}(\mathbf{p}_0))$ at $\mathbf{p}_0 = (0, 0, 0)$ is uncontrollable and has $\mathbf{J}_\sigma$ type. Its Brunovsky form consists of two 1×1 blocks $\mathbf{N}_1 = \mathbf{N}_2 = (0)$ in the Kronecker part ($r = 2$, $k_1 = k_2 = 1$) and the Jordan part $\mathbf{J} = (\sigma)$ with the generalized eigenvalue $\sigma = 2$. The transformation of α_0 to the Brunovsky form is performed by the triple $\gamma_b = (\mathbf{P}, \mathbf{Q}, \mathbf{R})$ with the matrices

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & -1 & 1 \end{pmatrix}, \ \mathbf{Q} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \ \mathbf{R} = \begin{pmatrix} 0 & 0 & 0 \\ -4 & -2 & -1 \end{pmatrix}. \tag{4.10}$$

Using formulae (4.2), we compute the vectors

$$\mathbf{f}_1 = (1, 1, -1), \quad \mathbf{f}_2 = (1, 1, 0), \quad \mathbf{f}_\sigma = (0, -1, 0). \tag{4.11}$$

Since the vectors $\mathbf{f}_1$ and $\mathbf{f}_2$ are linearly independent, by Theorem 4.1, the uncontrollability set $\mathcal{N}$ is a smooth curve in the neighborhood of $\mathbf{p}_0 = (0, 0, 0)$ with the tangent

$$(\mathbf{f}_1, \mathbf{p} - \mathbf{p}_0) = p_1 + p_2 - p_3 = 0, \quad (\mathbf{f}_2, \mathbf{p} - \mathbf{p}_0) = p_1 + p_2 = 0. \tag{4.12}$$

The generalized eigenvalue of the matrix pair $\alpha(\mathbf{p})$ for $\mathbf{p} \in \mathcal{N}$ is approximated by

$$\sigma + (\mathbf{f}_\sigma, \mathbf{p} - \mathbf{p}_0) + o(\|\mathbf{p} - \mathbf{p}_0\|) = 2 - p_2 + o(\|\mathbf{p}\|). \tag{4.13}$$

It is straightforward to check that $\alpha(\mathbf{p}) = \gamma(\mathbf{p}) \circ \beta(\mathbf{q}(\mathbf{p}))$, where

$$\beta(\mathbf{q}) = \left(\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ q_1 & q_2 & 2 + q_3 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \right), \qquad \begin{aligned} q_1(\mathbf{p}) &= p_1 + p_2 - p_3 - 2p_2^2, \\ q_2(\mathbf{p}) &= p_1 + p_2, \\ q_3(\mathbf{p}) &= -p_2 + p_1p_2, \end{aligned}$$

$$\gamma(\mathbf{p}) = \left(\begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & -p_2 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -p_3 & p_1p_2 & p_1p_2 \\ 2 & 3 & 1 \end{pmatrix} \right).$$

Hence, we find exact expressions for the uncontrollability set and the generalized eigenvalue in the form

$$\begin{gathered} \mathcal{N} = \{\mathbf{p} \in \mathbb{R}^3 \mid p_1 + p_2 - p_3 - 2p_2^2 = 0,\ p_1 + p_2 = 0\}, \\ \sigma(\mathbf{p}) = 2 - p_2 + p_1p_2. \end{gathered} \tag{4.14}$$

This agrees with the results obtained by Theorem 4.1.

Proof of Theorem 4.1. Let us consider a family of matrix pairs $\widetilde{\alpha}(\mathbf{p}) = \gamma_b \circ \alpha(\mathbf{p})$. This family is a deformation of the Brunovsky canonical form $\widetilde{\alpha}(\mathbf{p}_0) = \alpha_b$. By Theorem 2.1, the family $\widetilde{\alpha}(\mathbf{p})$ can be represented in the form $\widetilde{\alpha}(\mathbf{p}) = \widetilde{\gamma}(\mathbf{p}) \circ \beta(\mathbf{q}(\mathbf{p}))$, where $\beta(\mathbf{q})$ is the versal deformation of α_b having the form (3.5), (3.6). Since the controllability property is invariant under the feedback equivalence transformation, the uncontrollability sets for the families $\alpha(\mathbf{p})$ and $\beta(\mathbf{q}(\mathbf{p}))$ coincide in the neighborhood of $\mathbf{p}_0$ and, according to (3.7), have the form

$$q_1(\mathbf{p}) = \cdots = q_m(\mathbf{p}) = 0, \tag{4.15}$$

where $q_1(\mathbf{p}_0) = \cdots = q_m(\mathbf{p}_0) = 0$ by construction. Derivatives of the functions $q_i(\mathbf{p})$ at $\mathbf{p}_0$ are given by expression (2.18) of Theorem 2.1, where we take $\widetilde{\alpha}(\mathbf{p})$ and α_b instead of $\alpha(\mathbf{p})$ and α_0, respectively. Formula (2.18) requires the basis $\{\alpha_1^n, \dots, \alpha_\ell^n\}$ of the subspace $(T_{\alpha_b}\mathcal{O}(\alpha_b))^\perp$. This subspace for a matrix pair in the Brunovsky canonical form was found explicitly in [6]. It is convenient to represent the basis of $(T_{\alpha_b}\mathcal{O}(\alpha_b))^\perp$ in the form of the family $\alpha_1^n q_1 + \cdots + \alpha_\ell^n q_\ell$, which, for the matrix pair α_b under consideration, takes the form

$$\left(\begin{pmatrix} \mathbf{X}_{11}^n & 0 \\ \mathbf{X}_{21}^n & \mathbf{X}_{22}^n \end{pmatrix}, \begin{pmatrix} \mathbf{Y}_{11}^n & \mathbf{Y}_{12}^n \\ \mathbf{Y}_{21}^n & \mathbf{Y}_{22}^n \end{pmatrix} \right), \tag{4.16}$$

where

$$\begin{aligned} &\mathbf{X}_{21}^n = \big(\mathbf{L}_1^n(q_1),\ \dots,\ \mathbf{L}_r^n(q_r)\big), \quad \mathbf{X}_{22}^n = (q_{m+1}), \\ &\mathbf{Y}_{21}^n = (\sigma^{k_1}q_1,\ \dots,\ \sigma^{k_r}q_r), \quad \mathbf{Y}_{22}^n = (q_{r+1},\ \dots,\ q_m); \end{aligned} \tag{4.17}$$

$\mathbf{L}_i^n(q_i) = (q_i,\ \sigma q_i,\ \dots,\ \sigma^{k_i-1}q_i)$ is a $1 \times k_i$ matrix. The blocks $\mathbf{X}_{11}^n(\mathbf{q})$, $\mathbf{Y}_{11}^n(\mathbf{q})$, and $\mathbf{Y}_{12}^n(\mathbf{q})$ depend on $q_{m+2}, \dots, q_\ell$. Comparing this basis with the basis $\{\alpha_1^c, \dots, \alpha_\ell^c\}$ defined by (3.5), (3.6), we see that $z_{ij} = \langle \alpha_j^c, \alpha_i^n \rangle_1 = \delta_{ij}$ if $i \le m+1$ or $j \le m+1$, where δ_{ij} is the Kronecker delta. Hence, using (2.18), we find the derivatives of $q_1(\mathbf{p}), \dots, q_{m+1}(\mathbf{p})$ at $\mathbf{p}_0$ in the form

$$\frac{\partial q_i}{\partial p_j} = \Big\langle \frac{\partial \widetilde{\alpha}}{\partial p_j}, \alpha_i^n \Big\rangle_1, \quad i = 1, \dots, m+1. \tag{4.18}$$

Using $\widetilde{\alpha}(\mathbf{p})$ in its original form $\gamma_b \circ \alpha(\mathbf{p}) = (\mathbf{P}^{-1}(\mathbf{A}(\mathbf{p})\mathbf{P} + \mathbf{B}(\mathbf{p})\mathbf{R}),\ \mathbf{P}^{-1}\mathbf{B}(\mathbf{p})\mathbf{Q})$ and the explicit form of α_i^n, we find

$$\frac{\partial q_i}{\partial p_j} = f_i^j, \ \ i = 1, \dots, m; \quad \frac{\partial q_{m+1}}{\partial p_j} = f_\sigma^j \tag{4.19}$$

with f_i^j and f_σ^j defined in (4.2). Therefore, $\mathbf{f}_1, \dots, \mathbf{f}_m$, and $\mathbf{f}_\sigma$ are the gradient vectors of the functions $q_1(\mathbf{p}), \dots, q_{m+1}(\mathbf{p})$ at $\mathbf{p}_0$:

$$\nabla q_i = \mathbf{f}_i, \quad i = 1, \dots, m; \quad \nabla q_{m+1} = \mathbf{f}_\sigma; \quad \nabla = \left(\frac{\partial}{\partial p_1}, \dots, \frac{\partial}{\partial p_k} \right). \tag{4.20}$$

If the gradient vectors $\nabla q_1, \dots, \nabla q_m$ are linearly independent, then, by the implicit function theorem, the set $\mathcal{N}$ determined by (4.15) is a smooth surface of codimension m with the tangent plane (4.3). If $q_1(\mathbf{p}) = \cdots = q_m(\mathbf{p}) = 0$, then we see from (3.5), (3.6) that $\beta(\mathbf{q})$ and, hence, $\alpha(\mathbf{p})$ is a matrix pair of $\mathbf{J}_\sigma$ type with the generalized eigenvalue $\sigma + q_{m+1}(\mathbf{p}) = \sigma + (\mathbf{f}_\sigma, \mathbf{p} - \mathbf{p}_0) + o(\|\mathbf{p} - \mathbf{p}_0\|)$. □

Proof of Theorem 4.2. The case when the pair α_0 has $\mathbf{J}_{\sigma \pm i\omega}$ type is studied analogously. In this case the versal deformation $\beta(\mathbf{q})$ takes the form (3.5), (3.8), (3.9), and the basis $\{\alpha_1^n, \dots, \alpha_\ell^n\}$ is represented by family (4.16) with

$$\begin{aligned}
\mathbf{X}_{21}^n &= \left[\mathbf{L}_1^n(q_1, q_2), \ \dots, \ \mathbf{L}_r^n(q_{2r-1}, q_{2r}) \right], \\
\mathbf{X}_{22}^n &= \begin{pmatrix} q_{2m+1}/2 & q_{2m+2}/2 \\ -q_{2m+2}/2 & q_{2m+1}/2 \end{pmatrix}, \\
\mathbf{Y}_{21}^n &= \begin{pmatrix} \operatorname{Re}(q_1 + iq_2)(\sigma + i\omega)^{k_1} & \cdots & \operatorname{Re}(q_{2r-1} + iq_{2r})(\sigma + i\omega)^{k_r} \\ \operatorname{Im}(q_1 + iq_2)(\sigma + i\omega)^{k_1} & \cdots & \operatorname{Im}(q_{2r-1} + iq_{2r})(\sigma + i\omega)^{k_r} \end{pmatrix}, \\
\mathbf{Y}_{22}^n &= \begin{pmatrix} q_{2r+1} & \cdots & q_{2m-1} \\ q_{2r+2} & \cdots & q_{2m} \end{pmatrix},
\end{aligned} \tag{4.21}$$

where $\mathbf{L}_s^n(q_{2s-1}, q_{2s})$ is a $2 \times k_s$ matrix of the form

$$\mathbf{L}_s^n = \begin{pmatrix} q_{2s-1} & \sigma q_{2s-1} - \omega q_{2s} & \cdots & \operatorname{Re}(q_{2s-1} + iq_{2s})(\sigma + i\omega)^{k_s - 1} \\ q_{2s} & \omega q_{2s-1} + \sigma q_{2s} & \cdots & \operatorname{Im}(q_{2s-1} + iq_{2s})(\sigma + i\omega)^{k_s - 1} \end{pmatrix}, \tag{4.22}$$

and the blocks $\mathbf{X}_{11}^n$, $\mathbf{Y}_{11}^n$, and $\mathbf{Y}_{12}^n$ depend on $q_{2m+3}, \dots, q_\ell$. The uncontrollability set is given by the equations

$$q_1(\mathbf{p}) = \cdots = q_{2m}(\mathbf{p}) = 0, \tag{4.23}$$

where the gradients $\nabla q_1, \dots, \nabla q_{2m+2}$ are equal to the vectors $\mathbf{f}_1, \dots, \mathbf{f}_{2m}$, $\mathbf{f}_\sigma$, $\mathbf{f}_\omega$ defined in (4.5)–(4.7). The generalized eigenvalues on the surface (4.23) are determined by the expression

$$\begin{aligned}
\sigma + q_{2m+1}(\mathbf{p}) \pm i(\omega + q_{2m+2}(\mathbf{p})) = \sigma &+ (\mathbf{f}_\sigma, \mathbf{p} - \mathbf{p}_0) \pm i\big(\omega + (\mathbf{f}_\omega, \mathbf{p} - \mathbf{p}_0)\big) \\
&+ o(\|\mathbf{p} - \mathbf{p}_0\|).
\end{aligned}$$ □

5. Numerical construction of the uncontrollability set. Perturbation analysis developed in the previous section can be applied to numerical construction of the uncontrollability set by continuation if one point of this set is known.

Let us illustrate the implementation of this procedure for a specific case of a three-parameter system with one-input variable ($k = 3$ and $m = 1$); dimension of the state space is arbitrary. Let us assume that we are given a point $\mathbf{p}_0 \in \mathcal{N}$ corresponding to

a matrix pair $\alpha(\mathbf{p}_0)$ of $\mathbf{J}_{\sigma\pm i\omega}$ type. By Theorem 3.1, in the generic case the set $\mathcal{N}$ is a smooth curve in a vicinity of the point $\mathbf{p}_0$. Let us introduce the length parameter y along the curve $\mathcal{N}$. Then the curve $\mathcal{N}$ is given by a smooth function $\mathbf{p}(y)$ such that $\|d\mathbf{p}/dy\| = 1$. By Theorem 4.2, the vectors $\mathbf{f}_1, \mathbf{f}_2 \in \mathbb{R}^3$ evaluated by expressions (4.5), (4.6) at $\mathbf{p}(y)$ are normal vectors to the curve $\mathcal{N}$ at $\mathbf{p}(y)$. The vector product $\mathbf{f}_1 \times \mathbf{f}_2$ is a tangent vector to $\mathcal{N}$. Hence, we find

$$\frac{d\mathbf{p}}{dy} = \mathbf{g}, \quad \mathbf{g} = \frac{\mathbf{f}_1 \times \mathbf{f}_2}{\|\mathbf{f}_1 \times \mathbf{f}_2\|}. \tag{5.1}$$

Using expressions (4.9) and (5.1), we find derivatives of the functions $\sigma(y)$ and $\omega(y)$, which determine the generalized eigenvalues $\sigma(y) \pm i\omega(y)$ of the matrix pair $\alpha(\mathbf{p}(y))$, as follows:

$$\frac{d\sigma}{dy} = (\mathbf{f}_\sigma, \mathbf{g}), \quad \frac{d\omega}{dy} = (\mathbf{f}_\omega, \mathbf{g}). \tag{5.2}$$

Equations (5.1) and (5.2) represent a system of ordinary differential equations with respect to $\mathbf{p}$, σ, and ω. Initial conditions at $y = 0$ are given by

$$\mathbf{p}(0) = \mathbf{p}_0, \quad \sigma(0) = \sigma_0, \quad \omega(0) = \omega_0, \tag{5.3}$$

where $\sigma_0 \pm i\omega_0$ are generalized eigenvalues of the matrix pair $\alpha(\mathbf{p}_0)$ at the given point $\mathbf{p}_0 \in \mathcal{N}$. Integrating this system forwards and backwards, we find a regular part of the uncontrollability set of $\mathbf{J}_{\sigma\pm i\omega}$ type. Since we use the local information on the uncontrollability set, there is no guarantee that we have found the whole set $\mathcal{N}$ in the parameter space. Nevertheless, we obtain a finite piece of the uncontrollability set containing the given point $\mathbf{p}_0$. The integration can be continued until we reach a physical boundary of the parameter space or arrive at a singularity (boundary of the regular part of $\mathcal{N}$). A singularity causes numerical instability of the integration procedure and can be recognized by the appearance of a matrix pair with a different Jordan structure.

To determine right-hand sides of system (5.1), (5.2), we need to find the transformation to the Brunovsky canonical form

$$\alpha_b(y) = \gamma_b(y) \circ \alpha(\mathbf{p}(y)) \tag{5.4}$$

at each y. This can be done using the software [4, 5] providing the Kronecker canonical form of the matrix pencil $\big(\mathbf{A}(\mathbf{p}(y)), \mathbf{B}(\mathbf{p}(y))\big) - \lambda(\mathbf{I}_n, 0)$. Then permutation of columns provides the Brunovsky canonical form [9]. In these calculations, the information on the generalized eigenvalues $\sigma(y) \pm i\omega(y)$ can be used.

Alternatively, we can calculate $\alpha_b(y)$ and $\gamma_b(y)$ by taking advantage of the remark in section 3. In the generic case, the Brunovsky form $\alpha_b(y)$ has one $(n-2) \times (n-2)$ block $\mathbf{N}_1$ in the Kronecker part:

$$\alpha_b(y) = \left(\begin{pmatrix} \mathbf{N}_1 & 0 \\ 0 & \mathbf{J}_{\sigma\pm i\omega} \end{pmatrix}, \begin{pmatrix} \mathbf{E}_1 \\ 0 \end{pmatrix} \right). \tag{5.5}$$

Taking the derivative of (5.4) with respect to y, we find

$$\frac{d\alpha_b}{dy} = df_{\alpha_b}\left(\gamma_b^{-1}\frac{d\gamma_b}{dy}\right) + \gamma_b \circ \frac{d\alpha(\mathbf{p}(y))}{dy}, \tag{5.6}$$

where $df_{\alpha_b}(\xi)$ is the differential of the function $f_{\alpha_b}(\gamma)$ at the unit element e defined by expression (2.12). Introducing the matrix triple $\xi = \gamma_b^{-1} d\gamma_b/dy$ and using expression (5.1), we find a linear algebraic equation

$$df_{\alpha_b}(\xi) = \frac{d\alpha_b}{dy} - \gamma_b \circ \left(\sum_{i=1}^{k} \frac{\partial \alpha}{\partial p_i} g_i \right). \tag{5.7}$$

Then the derivative of the function $\gamma_b(y)$ is given by the relation

$$\frac{d\gamma_b}{dy} = \gamma_b(y)\xi. \tag{5.8}$$

Equation (5.7) does not determine the triple ξ uniquely (transformation to the Brunovsky form is not unique). It is convenient to choose a particular solution satisfying the condition

$$\xi \in \left(\operatorname{Ker} df_{\alpha_b}\right)^{\perp}, \tag{5.9}$$

which determines a unique ξ. A numerical method for finding this solution is given in the appendix.

Differential equation (5.8) can be integrated together with system (5.1), (5.2). As a result, we find the Brunovsky canonical form $\alpha_b(y)$ and the feedback equivalence transformation $\gamma_b(y)$ at each point of the uncontrollability set represented by the curve $\mathbf{p}(y)$. The initial conditions are given by

$$\gamma_b(0) = \gamma_b^0, \tag{5.10}$$

where γ_b^0 transforms the matrix pair $\alpha(\mathbf{p}_0)$ to the Brunovsky form. Note that matrices of the triple $\gamma_b(y) = (\mathbf{P}_b(y), \mathbf{Q}_b(y), \mathbf{R}_b(y))$ may become ill-conditioned when approaching a singularity of $\mathcal{N}$. To control the accuracy of calculations and to detect a singularity it is convenient to use the norm $\|\alpha_b(y) - \gamma_b(y) \circ \alpha(\mathbf{p}(y))\|$, which describes the error in equality (5.4).

Similarly, we can calculate a regular part of the uncontrollability set, corresponding to matrix pairs of $\mathbf{J}_\sigma$ type, which is a smooth surface for a three-parameter one-input dynamical system. Recall that, by Theorem 3.1, the surfaces corresponding to matrix pairs of $\mathbf{J}_\sigma$ and $\mathbf{J}_{\sigma \pm i\omega}$ types, together with their boundaries, form the whole uncontrollability set of a generic multi-input linear dynamical system.

Example 5.1. Let us consider a mechanical system shown in Figure 5.1. The system consists of a thin uniform platform of mass m and length $2l$ supported at both ends by springs having elastic coefficients k_1, k_2 and viscous damping coefficients d_1, d_2. There is a vertical force F applied to the platform at the distance ξl from the middle. As generalized coordinates, we take a vertical coordinate z of the center of the platform and an angle φ between the platform and horizontal axis. The equilibrium of the system for zero external force $F = 0$ is assumed to be $z = 0$, $\varphi = 0$.

Equations of motion of the system linearized near the equilibrium take the form

$$\begin{aligned} &m\ddot{z} + d_1(\dot{z} + l\dot{\varphi}) + d_2(\dot{z} - l\dot{\varphi}) + k_1(z + l\varphi) + k_2(z - l\varphi) = F, \\ &I_m\ddot{\varphi} + d_1 l(\dot{z} + l\dot{\varphi}) - d_2 l(\dot{z} - l\dot{\varphi}) + k_1 l(z + l\varphi) - k_2 l(z - l\varphi) = -\xi l F, \end{aligned} \tag{5.11}$$

where $I_m = ml^2/3$ is the moment of inertia of the platform with respect to the center of mass; the dot denotes differentiation with respect to time t. If $F = 0$, then the

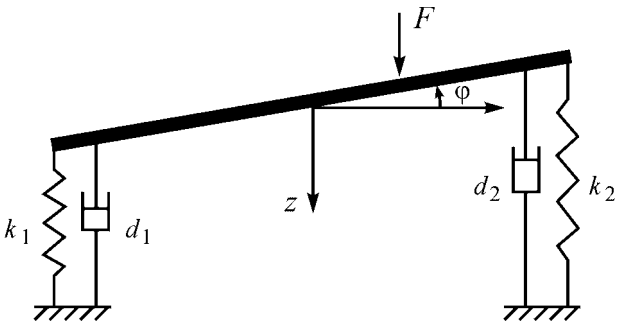


FIG. 5.1. *Mechanical system controlled by a vertical force.*

system oscillates with a decaying amplitude and comes to the equilibrium in infinite time $t \longrightarrow +\infty$. Let us consider the force F as a control parameter. If the system is controllable, then it can be damped (put into the equilibrium) in finite time. This task becomes difficult or impossible if the system is close to the uncontrollable system.

Let us introduce nondimensional variables

$$\begin{aligned} &\tau = t/\alpha, \quad f_1 = \frac{(d_1+d_2)\alpha}{m}, \quad f_2 = \frac{(d_1-d_2)\alpha}{m}, \\ &c_1 = \frac{(k_1+k_2)\alpha^2}{m}, \quad c_2 = \frac{(k_1-k_2)\alpha^2}{m}, \quad u = \frac{\alpha^2}{ml}\,F, \end{aligned} \tag{5.12}$$

where α is a time scale, and choose a state vector $\mathbf{x} \in \mathbb{R}^4$ in the form

$$x_1 = \frac{z}{l}, \quad x_2 = \varphi, \quad x_3 = \frac{\alpha \dot{z}}{l}, \quad x_4 = \alpha\dot{\varphi}. \tag{5.13}$$

Then system (5.11) can be written in the form $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u$, where derivative is taken with respect to nondimensional time τ, and the matrices $\mathbf{A}$ and $\mathbf{B}$ are

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -c_1 & -c_2 & -f_1 & -f_2 \\ -3c_2 & -3c_1 & -3f_2 & -3f_1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -3\xi \end{pmatrix}. \tag{5.14}$$

Let us fix the parameters $c_1 = 25/12$ and $f_1 = 1$, characterizing the joint stiffness and damping of supports, and consider the parameter vector $\mathbf{p} = (c_2, f_2, \xi)$. Let us consider the point $\mathbf{p}_0 = (0, 0, 0)$ corresponding to equal supports and the force applied at the center of the platform. The matrix pair $\alpha(\mathbf{p}_0)$ is uncontrollable and has the Brunovsky canonical form

$$\mathbf{A}_b = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_0 & \omega_0 \\ 0 & 0 & -\omega_0 & \sigma_0 \end{pmatrix}, \quad \mathbf{B}_b = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \tag{5.15}$$

with the uncontrollable modes $\sigma_0 \pm i\omega_0 = -1.5 \pm i2$. The triple $\gamma_b^0 = (\mathbf{P}_b, \mathbf{Q}_b, \mathbf{R}_b)$

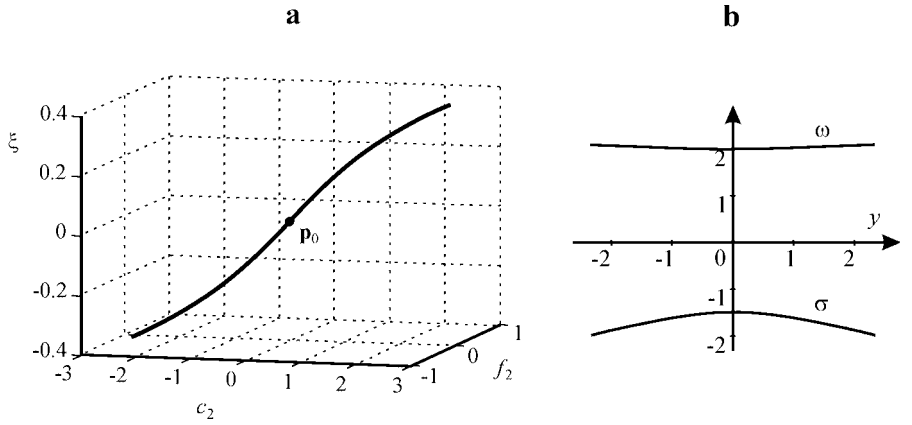


FIG. 5.2. (a) *Uncontrollability set in the parameter space;* (b) *uncontrollable modes* $\sigma(y) \pm i\omega(y)$.

transforming $\alpha(\mathbf{p}_0)$ to the Brunovsky form is

$$\mathbf{P}_b = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1.5 & 2 \end{pmatrix}, \quad \mathbf{Q}_b = (1), \quad \mathbf{R}_b = \begin{pmatrix} 25/12 & 1 & 0 & 0 \end{pmatrix}. \tag{5.16}$$

The point $\mathbf{p}_0$ belongs to the uncontrollability set and has $\mathbf{J}_{\sigma\pm i\omega}$ type. Integrating equations (5.1), (5.2), (5.8) with initial conditions (5.3), (5.10) numerically, we find the uncontrollability set $\mathcal{N}$ in the physical range of parameters $-c_1 < c_2 < c_1$ and $-f_1 < f_2 < f_1$ corresponding to positive characteristics of the supports. The result is shown in Figure 5.2(a), where the set $\mathcal{N}$ is represented by a solid line. Figure 5.2(b) shows graphs of the real and imaginary parts of the uncontrollable modes $\sigma(y) \pm i\omega(y)$. The maximal error $\|\alpha_b(y) - \gamma_b(y) \circ \alpha(\mathbf{p}(y))\|$ along the curve $\mathcal{N}$ is about $4 \cdot 10^{-7}$, which is less than the accuracy of the ordinary differential equation solver (the calculations were carried out in MATLAB using the standard ode45 solver).

Using the computed data, it can be shown that one mode of free vibrations of the system with the parameter vector $\mathbf{p}(y) = (c_2(y), f_2(y), \xi(y))$ has a node at the distance $\xi(y)l$ from the center of the platform; i.e., this mode represents rotation of the platform around a point. The force F applied at this point has no influence on the rotational mode, which leads to uncontrollability of the system. From Figure 5.2(a) we see that uncontrollability occurs when one of the supports has smaller stiffness and damping coefficients, and the force is applied at the side of a softer support. The obtained results are useful for the design of the system to avoid uncontrollability effects.

6. Singularities of the uncontrollability set. In this section we consider points of the uncontrollability set corresponding to matrix pairs whose types are different from $\mathbf{J}_\sigma$ and $\mathbf{J}_{\sigma\pm i\omega}$. By Theorem 3.1, these points belong to a boundary of the regular part of $\mathcal{N}$, represented by matrix pairs of $\mathbf{J}_\sigma$ and $\mathbf{J}_{\sigma\pm i\omega}$ types, and determine singularities (nonsmooth points) of the uncontrollability set.

In order to understand the role of singular points in the structure of the uncontrollability set, let us consider a specific case when the Jordan part of the Brunovsky

canonical form α_b of a pair $\alpha(\mathbf{p})$, $\mathbf{p} \in \mathcal{N}$, is a 2×2 Jordan block

$$\mathbf{J}_{\sigma^2} = \begin{pmatrix} \sigma & 1 \\ 0 & \sigma \end{pmatrix} \tag{6.1}$$

with a double real generalized eigenvalue σ. In the generic case, points $\mathbf{p} \in \mathcal{N}$ corresponding to matrix pairs of this type form a smooth surface of codimension $2m+1$ in the parameter space [6]. If the number of parameters is less than $2m+1$, then matrix pairs of $\mathbf{J}_{\sigma^2}$ type do not appear in generic families $\alpha(\mathbf{p})$.

In the case under consideration, a versal deformation of the pair α_b has the form (3.5), where [6]

$$\begin{gathered} \mathbf{X}^c_{21} = (\mathbf{L}^c_1(q_1, q_2), \dots, \mathbf{L}^c_r(q_{2r-1}, q_{2r})), \\ \mathbf{Y}^c_{22} = \begin{pmatrix} q_{2r+1} & \cdots & q_{2m-1} \\ q_{2r+2} & \cdots & q_{2m} \end{pmatrix}, \quad \mathbf{X}^c_{22} = \begin{pmatrix} q_{2m+2} & 0 \\ q_{2m+1} & q_{2m+2} \end{pmatrix}, \end{gathered} \tag{6.2}$$

and $\mathbf{L}^c_i(q_{2i-1}, q_{2i})$ is a $2 \times k_i$ matrix of the form

$$\mathbf{L}^c_i(q_{2i-1}, q_{2i}) = \begin{pmatrix} q_{2i-1} & 0 & \cdots & 0 \\ q_{2i} & 0 & \cdots & 0 \end{pmatrix}; \tag{6.3}$$

the blocks $\mathbf{Y}^c_{11}$ and $\mathbf{Y}^c_{12}$ depend on $q_{2m+3}, \dots, q_\ell$. Using controllability condition (2.2), we find that the uncontrollability set of the versal deformation in the neighborhood of $\mathbf{q} = 0$ is determined by the equations

$$q^2_{2i-1} q_{2m+1} - q^2_{2i} = 0, \quad i = 1, \dots, m. \tag{6.4}$$

Every equation in (6.4) determines in the space $(q_{2i-1}, q_{2i}, q_{2m+1})$ a surface shown in Figure 3.1 and discussed in Example 3.1. Hence, the regular part of the uncontrollability set consists of one smooth surface of $\mathbf{J}_{\sigma \pm i\omega}$ type and codimension $2m$, determined by the equations

$$q_i = 0, \quad i = 1, \dots, 2m, \quad q_{2m+1} < 0, \tag{6.5}$$

and smooth surfaces of $\mathbf{J}_\sigma$ type determined by the equations

$$q^2_{2i-1} q_{2m+1} - q^2_{2i} = 0, \quad q_{2i-1} \neq 0, \quad i = 1, \dots, m, \quad q_{2m+1} \geq 0. \tag{6.6}$$

There are 2^m separate surfaces of $\mathbf{J}_\sigma$ type corresponding to different combinations of signs of the parameters q_{2i-1}, $i = 1, \dots, m$, in (6.6). The singular part of $\mathcal{N}$ is a boundary of the regular part. It consists of several smooth surfaces, which are parts of the set (6.4) with the additional condition

$$q_{2j-1} = q_{2j} = 0, \quad q_{2m+1} \geq 0 \tag{6.7}$$

for some $j \in \{1, \dots, m\}$.

The structure of the uncontrollability set for a generic family $\alpha(\mathbf{p})$ in the neighborhood of a point $\mathbf{p}_0 \in \mathcal{N}$ of $\mathbf{J}_{\sigma^2}$ type is the same as for the versal deformation. These sets are related by a smooth change of parameters $\mathbf{q} = \mathbf{q}(\mathbf{p})$. Analogously to the method of Theorem 4.1, we can use formulae of Theorem 2.1 to calculate the gradients ∇q_i of the functions $q_i(\mathbf{p})$ at the singular point $\mathbf{p}_0$. Then expressions (6.4)–(6.7), where $q_i(\mathbf{p})$ is substituted by its linear approximation $(\nabla q_i, \mathbf{p} - \mathbf{p}_0)$, provide first-order approximations of the regular and singular parts of $\mathcal{N}$.

Singular points lead to a more rich and complicated structure of the uncontrollability set. This affects the behavior of the underlying dynamical system and causes numerical difficulties in the analysis of $\mathcal{N}$. The information on the local form of $\mathcal{N}$ at its regular or singular point is useful for the analysis and construction of the uncontrollability set. In particular, this information allows choosing the locally optimal change of design parameters in order to get a controllable system.

7. Conclusion. In this paper, fundamental properties of the uncontrollability set for a multi-input linear dynamical system dependent on parameters are investigated. It is shown that the uncontrollability set has a regular part, which consists of smooth surfaces corresponding to one real uncontrollable mode or a complex conjugate pair of uncontrollable modes. Explicit formulae for local quantitative analysis of the uncontrollability set and perturbation of the uncontrollable modes are derived and used for numerical construction of the uncontrollability set in the parameter space.

A constructive method for qualitative and quantitative analysis of the uncontrollability set based on the versal deformation theory is developed. The idea of regularization of the parameter space by transformation to a versal deformation, proposed in the paper, provides a powerful tool of multiparameter perturbation theory for control systems.

Using the duality theorem [3, 13], all the results of this paper can be applied to the analysis of an unobservability set for a multioutput linear dynamical system dependent on parameters.

Appendix. Let us consider the linear algebraic equation with respect to $\xi \in T_e\mathcal{G}$

$$df_\alpha(\xi) = \alpha', \quad \xi \in \bigl(\operatorname{Ker} df_\alpha\bigr)^\perp, \tag{7.1}$$

assuming that a solution exists, i.e., $\alpha' \in \operatorname{Im} df_\alpha$. We denote by $\operatorname{vec}(\mathbf{A})$ a column vector, which is an ordered stack of columns of $\mathbf{A}$ from left to right (its dimension is equal to the number of elements of $\mathbf{A}$). Analogously, we introduce the vectorization of the matrix pairs $\alpha = (\mathbf{A}, \mathbf{B})$, $\alpha' = (\mathbf{A}', \mathbf{B}')$, and matrix triple $\xi = (\mathbf{U}, \mathbf{V}, \mathbf{W})$ as follows:

$$\operatorname{vec}(\alpha) = \begin{pmatrix} \operatorname{vec}(\mathbf{A}) \\ \operatorname{vec}(\mathbf{B}) \end{pmatrix}, \ \operatorname{vec}(\alpha') = \begin{pmatrix} \operatorname{vec}(\mathbf{A}') \\ \operatorname{vec}(\mathbf{B}') \end{pmatrix}, \ \operatorname{vec}(\xi) = \begin{pmatrix} \operatorname{vec}(\mathbf{U}) \\ \operatorname{vec}(\mathbf{V}) \\ \operatorname{vec}(\mathbf{W}) \end{pmatrix}. \tag{7.2}$$

Then (7.1) for the mapping df_α defined in (2.12) can be written in the form

$$\mathbf{G}_\alpha \operatorname{vec}(\xi) = \operatorname{vec}(\alpha'), \quad \operatorname{vec}(\xi) \in \bigl(\operatorname{null} \mathbf{G}_\alpha\bigr)^\perp, \tag{7.3}$$

where $\mathbf{G}_\alpha$ is an $(n^2 + nm) \times (n^2 + m^2 + nm)$ real matrix of the form

$$\mathbf{G}_\alpha = \begin{pmatrix} \mathbf{I}_n \otimes \mathbf{A} - \mathbf{A}^T \otimes \mathbf{I}_n & 0 & \mathbf{I}_n \otimes \mathbf{B} \\ -\mathbf{B}^T \otimes \mathbf{I}_n & \mathbf{I}_m \otimes \mathbf{B} & 0 \end{pmatrix}, \tag{7.4}$$

where $\mathbf{I}_n \otimes \mathbf{A}$ denotes the Kronecker product of matrices. The vector $\operatorname{vec}(\xi)$ satisfying the condition $\operatorname{vec}(\xi) \in \bigl(\operatorname{null} \mathbf{G}_\alpha\bigr)^\perp$ can be expressed as follows:

$$\operatorname{vec}(\xi) = \mathbf{G}_\alpha^T \mathbf{y}, \tag{7.5}$$

where $\mathbf{y}$ is a column vector of dimension $n^2 + nm$. Substituting (7.5) into (7.3), we find

$$\mathbf{G}_\alpha \mathbf{G}_\alpha^T \mathbf{y} = \mathrm{vec}(\alpha'). \tag{7.6}$$

The linear mapping df_α considered in this paper has a nontrivial null-space. Hence, the matrix $\mathbf{G}_\alpha \mathbf{G}_\alpha^T$ is singular. Recall that $\{\alpha_1^n, \dots, \alpha_\ell^n\}$ is a basis of $\mathrm{Ker}\, df_\alpha^*$. Then $\{\mathrm{vec}(\alpha_1^n), \dots, \mathrm{vec}(\alpha_\ell^n)\}$ is a basis of the null-space of the symmetric matrix $\mathbf{G}_\alpha \mathbf{G}_\alpha^T$. Using the method described in [15], we construct the equation

$$\left(\mathbf{G}_\alpha \mathbf{G}_\alpha^T + \sum_{i=1}^{\ell} \mathrm{vec}(\alpha_i^n) \big(\mathrm{vec}(\alpha_i^n)\big)^T \right) \mathbf{y} = \mathrm{vec}(\alpha'), \tag{7.7}$$

where the matrix in the left-hand side is nonsingular. Solution $\mathbf{y}$ of (7.7) can be found numerically using the standard codes. The obtained vector $\mathbf{y}$ is a particular solution of (7.6), which determines the unique solution (7.5) of (7.1).

For a matrix pair $\alpha = \alpha_b$, where α_b is the Brunovsky canonical form (5.5) considered in section 5, we have $\ell = 4$, and the pairs $\alpha_1^n q_1 + \dots + \alpha_4^n q_4$ are given by expressions (4.16), (4.21), (4.22), where $r = 1$, $k_1 = n - 2$, $\mathbf{X}_{11}^n = 0$, $\mathbf{Y}_{11}^n = 0$, and $\mathbf{Y}_{12}^n = 0$.

REFERENCES

[1] V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1983.

[2] D. L. Boley and W.-S. Lu, *Measuring how far a controllable system is from an uncontrollable one*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 249–251.

[3] C.-T. Chen, *Linear System Theory and Design*, Holt, Rinehart, and Winston, New York, 1984.

[4] J. Demmel and B. Kågström, *The generalized Schur decomposition of an arbitrary pencil $A - \lambda B$: Robust software with error bounds and applications.* I, ACM Trans. Math. Software, 19 (1993), pp. 160–174.

[5] J. Demmel and B. Kågström, *The generalized Schur decomposition of an arbitrary pencil $A - \lambda B$: Robust software with error bounds and applications.* II, ACM Trans. Math. Software, 19 (1993), pp. 175–201.

[6] J. Ferrer, M. I. García-Planas, and F. Puerta, *Brunowsky local form of a holomorphic family of pairs of matrices*, Linear Algebra Appl., 253 (1997), pp. 175–198.

[7] F. R. Gantmacher, *The Theory of Matrices*, Vol. 1, Chelsea, New York, 1959.

[8] M. I. García-Planas and A. A. Mailybaev, *Reduction to versal deformations of matrix pencils and matrix pairs with application to control theory*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 943–962.

[9] I. Gohberg, P. Lancaster, and L. Rodman, *Invariant Subspaces of Matrices with Applications*, Wiley, New York, 1986.

[10] M. Gu, *New methods for estimating the distance to uncontrollability*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 989–1003.

[11] A. J. Laub, *Survey of computational methods in control theory*, in Electric Power Problems: The Mathematical Challenge, A. M. Erisman, K. W. Neves, and M. H. Dwarakanath, eds., SIAM, Philadelphia, 1980, pp. 231–260.

[12] R. V. Patel, A. J. Laub, and P. M. Van Dooren, *Numerical Linear Algebra Techniques for Systems and Control*, IEEE Press, New York, 1994.

[13] E. D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Texts Appl. Math. 6, Springer-Verlag, New York, 1990.

[14] A. Tannenbaum, *Invariance and System Theory: Algebraic and Geometric Aspects*, Lecture Notes in Math. 845, Springer-Verlag, New York, 1981.

[15] V. A. Yakubovich and V. M. Starzhinskii, *Parametric Resonance in Linear Systems*, Nauka, Moscow, 1987 (in Russian).

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1451–1465

OPTIMAL HARVESTING OF A JUMP DIFFUSION POPULATION AND THE EFFECT OF JUMP UNCERTAINTY*

NILS CHR. FRAMSTAD†

Abstract. The problem of irreversibly harvesting from a general one-dimensional (Wiener–Poisson) jump diffusion population model is studied. For a wide class of stochastic models, the optimal strategy has a downwards local time reflection at a trigger level x^*, which is typically known to be larger than in the corresponding deterministic problem if the uncertainty is Brownian. This paper shows that the presence of zero-mean jump uncertainty may or may not have the opposite effect on x^*. The property of uncertainty increasing x^* is related to the applicability of a comparison theorem.

Key words. optimal harvesting, singular stochastic control, reflected jump diffusion model, behavior towards risk

AMS subject classifications. 92D25, 91B32, 93E20, 49K45, 60G51, 60H10, 60J75, 60K37

DOI. 10.1137/S0363012902385910

1. Introduction.

The problem of optimally harvesting a population has been widely studied. The canonical example is asking how to get the most out of a logistic growth model. This paper is strongly inspired by two papers studying stochastic versions of the logistic growth models. Lungu and Øksendal [10] assume the process to follow the Itô (nonanticipative) stochastic differential equation

$$\mathrm{d}X_t = X_t(1 - X_t/K) \cdot (\beta_0 \,\mathrm{d}t + \sigma_{\mathrm{a}} \,\mathrm{d}B_t) - \mathrm{d}H_t, \tag{1.1a}$$

where, throughout this paper, H_t is the *cumulative harvested amount* up to and including time t. This is maybe the most straightforward generalization to a stochastic model, merely adding (a constant factor σ_{a} times) white noise to the (constant, positive) growth term β_0; here, B is standard Brownian motion and the positive constant K is the carrying capacity of the environment. In this model, volatility is largest when growth rate is. An alternative model, studied by, e.g., Alvarez and Shepp [3] and later by Myhre [11], generalizes the logistic model by adding a noise term where *relative* uncertainty is constant:

$$\mathrm{d}X_t = \beta_0 X_t(1 - X_t/K) \,\mathrm{d}t + \sigma_{\mathrm{b}} X_t \,\mathrm{d}B_t - X_{t^-} \cdot \mathrm{d}N_t - \mathrm{d}H_t \,. \tag{1.1b}$$

In their model there is also an implicit killing time, independent of everything else, where the population or the opportunity to harvest it disappears once and for all; the Poisson process N_t in (1.1b) makes this term explicit. Alvarez [2] has also analyzed a class of models described by

$$\mathrm{d}X_t = X_t\beta(X_t) \,\mathrm{d}t + \sigma_{\mathrm{c}}(X_t) \,\mathrm{d}B_t - \mathrm{d}H_t \,. \tag{1.1c}$$

All the aforementioned works are, as this paper, concerned with the problem of maximizing expected discounted total harvest. They find that the optimal strategy is a local time downwards reflection at a trigger level x^*, which turns out to be

*Received by the editors January 11, 2002; accepted for publication (in revised form) March 9, 2003; published electronically October 2, 2003. An earlier version of this paper appeared in the author's doctoral dissertation, University of Oslo, 2002. This work was supported by the Research Council of Norway.

http://www.siam.org/journals/sicon/42-4/38591.html

†Skogvollveien 40, NO-0580 Oslo, Norway (ncf@math.uio.no).

greater than in the corresponding deterministic case obtained by replacing σ by 0. Alvarez's paper [2], being the most general treatment of the above, concludes that "*the sign of the relationship between stochasticity and the optimal policy is unambiguously negative.*" A main object of this paper is to show that Alvarez's conclusions rely crucially upon modeling uncertainty with Brownian motion; it will turn out that if the Gaussian noise is replaced by another (infinitely divisible) zero-mean probability law, then the trigger level x^* may be lower than in the corresponding deterministic case. A careful note is appropriate: it is obvious that an *uncompensated* jump term (with nonzero mean) may lower x^*—like the simple "catastrophe intensity" in the model (1.1b) above, where the jump to zero intensity has the same effect as an increased discounting rate. We emphasize that we will instead be introducing a pure jump *martingale* to the model. Pure jump martingales may be regarded as a modeling alternative to the Brownian motion, at least if the jumps are small, and we shall see that the phenomena might have qualitatively different implications.

In the above cited works, the optimal strategy is described completely by the single value x^* at which one reflects the process downwards. However, if growth at zero does not exceed the interest rate, it may be that a profit maximizer will deplete a low population; that in particular goes with models with depensation, also called Allee effect, namely negative (expected) growth rate at sufficiently small populations due to, for example, the difficulty of finding a mate. If a population is doomed to decline until extinction, an economic agent will then harvest it right away. Optimality of immediate depletion also occurs in stochastic models: a case is treated by Alvarez in [1], where it is shown that the optimal strategy is described by two trigger values: downwards reflection at some x^*, and immediate total depletion whenever X falls below some $x_* < x^*$. Lande, Engen, and Sæther [9] have found by numerical methods that introducing a Gaussian noise may *increase* the optimal x_* in such a model. Therefore, even in the Brownian motion case, there are cases where the presence of uncertainty may lead to earlier total depletion of the population. We shall see why the arguments leading to our conclusions concerning the upper trigger level x^* do not apply to the lower trigger value x_* given that the latter exists (is strictly positive).

2. The model. Assume to be given a filtered probability space $(\Omega, \mathfrak{S}, \{\mathcal{F}_t\}_{t\geq t_0}, \mathsf{P})$ satisfying the usual conditions. The population is assumed to be an adapted process X following a stochastic differential equation given in (2.1) below. X will be driven by a standard Brownian motion B and by a centered integer-valued random measure governing jumps. It will be convenient to allow for state-dependent jump intensity, as well as the usual representation with Poisson random jumps, so let us give both forms. In the latter, $\tilde{N}$ is a centered Poisson random measure with Lévy measure λ, while in the former setting, we denote the centered random measure by $\tilde{M}$ while the jump intensity is q. Thus, if population at time t^- is at level x, then the intensity of a jump by a factor belonging to a Borel set Z is $q(x, Z)$, with $q(x, \{0\}) = 0$. Consider now two different representations for the dynamics:

$$\mathrm{d}X_t = X_{t^-} \cdot \Big(\beta(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}B_t + \int z\,\tilde{M}(\mathrm{d}t, \mathrm{d}z)\Big) - \mathrm{d}H_t \tag{2.1a}$$

$$= X_{t^-} \cdot \Big(\beta(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}B_t + \int \eta(X_{t^-}, z)\,\tilde{N}(\mathrm{d}t, \mathrm{d}z)\Big) - \mathrm{d}H_t, \tag{2.1b}$$

$$\text{both whenever } X_{t^-} > 0\text{; elsewhere, assume } \mathrm{d}X = 0. \tag{2.1c}$$

H_t is the total amount harvested up to and including time t, subject to our choice under the following restrictions.

DEFINITION 2.1. *Let* $\mathsf{P}^{t_0,x}$ *be the probability law of the time-space process* (t, X) *starting at* $(t_0, X_{t_0}) = (t_0, x)$ *with* $x \geq 0$*, and define* the class $\mathcal{H}$ of admissible controls *to be the* $\mathcal{F}_t$*-predictable, left-continuous nondecreasing stochastic processes* H *such that* $\mathsf{P}^{t_0,x}$*-a.s. we have* $H_{t_0} = 0$ *and* $X_t \geq 0$ *for all* $t \geq t_0$.

In particular, we will have to stop harvesting (i.e., $\mathrm{d}H = 0$) from the moment X hits zero. We emphasize that this can happen both by itself, with no harvesting, or by depleting the population by choosing $H_{t^+} = H_t + X_t$. We note that there always exists an admissible strategy, namely immediate depletion. As we do not allow negative values for X, we shall assume that for all x,

$$q(x, (-\infty, -1)) = 0 \leq 1 + \eta(x, \cdot) \quad \lambda\text{-almost everywhere.} \tag{2.2}$$

It will be useful to decompose q into mutually singular measures $\bar{q}, \check{q}, \hat{q}$ as follows:

$$\bar{q}(\cdot, \mathrm{d}z) := q(\cdot, \mathrm{d}z) \cdot \chi_{\{-1\}}(z) \qquad \text{(annihilation)}, \tag{2.3a}$$

$$\check{q}(\cdot, \mathrm{d}z) := q(\cdot, \mathrm{d}z) \cdot \chi_{(-1,0)}(z) \qquad \text{(other negative jumps)}, \tag{2.3b}$$

$$\hat{q}(\cdot, \mathrm{d}z) := q(\cdot, \mathrm{d}z) \cdot \chi_{(0,\infty)}(z) \qquad \text{(positive jumps)}, \tag{2.3c}$$

and, by abuse of notation, to identify the former with the function $\bar{q}(x) := q(x, \{-1\})$.

Existence and uniqueness will now be granted by familiar assumptions (i.e., sublinear growth and Lipschitz conditions on $x\beta$, $x\sigma$, $x\eta$; see, e.g., [4], [8], or [12] for more information about stochastic differential equations with jumps); we shall instead make the appropriate regularity assumptions ad hoc.

Assumption 2.2. The coefficients β, σ^2, and q are assumed to be one-sided continuous at each point, and admit existence and uniqueness of a weak solution to (2.1). We assume that $x\beta$ is locally upper bounded, and that q satisfies (2.2) and $\int z \wedge z^2 \, q(x, \mathrm{d}z) < \infty$ for all x.

DEFINITION 2.3. *Let* $\mathsf{E} = \mathsf{E}^{t_0,x}$ *be expectation w.r.t.* $\mathsf{P}^{t_0,x}$*,* $x \geq 0$*. Assume that* $\delta > 0$ *is a constant discount rate and that one wants to maximize* total discounted expected harvest *defined as*

$$J^H := \mathsf{E}\left[\int_{[t_0,\infty)} \mathrm{e}^{-\delta t} \, \mathrm{d}H_t\right] \tag{2.4}$$

and, if it exists, find an optimal strategy $H^* \in \mathcal{H}$ *such that*

$$J^{H^*} = \sup_{H \in \mathcal{H}} J^H =: \Phi(t_0, x) =: \mathrm{e}^{-\delta t_0} \Psi(x). \tag{2.5}$$

We note that by the Markov property, Ψ will indeed be a function of x only, not depending on t_0. Therefore, we can and will without loss of generality assume

$$t_0 = 0. \tag{2.6}$$

We also note that we need not really assume global existence of solution, but should the process possibly explode to infinity—which is very unrealistic from a biological point of view—we would be able to harvest an arbitrarily high amount in finite time. The same consideration lies behind the assumed local upper boundedness on $x\beta$; were it not, then one could in the deterministic case gain arbitrarily high discounted harvest by keeping the population at a level where $x\beta$ were sufficiently high. Clearly, there are stochastic cases with the same property: take $\sigma = 0$ and $q(\cdot, [-1, \infty))$ locally bounded in x, and consider only what happens before first negative jump.

3. Sufficient conditions and properties of the value function. Because it is always admissible to harvest any given amount not exceeding the present population, we have $\Psi(x) \geq \Psi(y)+x-y$ whenever $x \geq y \geq 0$, with equality if it is optimal to harvest at least $(x-y)$. Thus $\Psi'(x) \geq 1$ whenever it exists, with equality on intervals where harvesting is optimal. In the cases treated by the works [2], [3], [10], and [11] they find a unique threshold x^* such that $\Psi' > 1$ iff $x < x^*$; hence the *continuation region* D is of the form $D = [0, x^*)$. In the case of depensation, or if growth rate at zero is small, one may also want to deplete the population at small rates, and D may become an interval bounded away from zero, as found in [1]. Heuristically it is reasonable that D is at worst a countable union of intervals, and we shall restrict ourselves to this case. On D, we shall characterize Ψ by the Hamilton–Jacobi–Bellman (HJB) equation. Define

$$\mathsf{Q}\psi(x) := \int \big(\psi(x+zx) - \psi(x) - zx\psi'(x)\big)\, q(x, \mathrm{d}z) \tag{3.1a}$$

and

$$\mathsf{A}\psi(x) := -\delta\psi(x) + x\beta(x)\psi'(x) + \tfrac{1}{2}x^2\sigma^2(x)\psi''(x) + \mathsf{Q}\psi(x). \tag{3.1b}$$

For functions $\psi \in \mathsf{C}^2$ vanishing at 0 and with sublinear growth (cf. Theorem 3.3), $(\mathsf{A}+\delta)$ coincides with the generator of the process X when $\mathrm{d}H = 0$. It turns out—see Proposition 3.5 and Remark 3.6—that everywhere twice continuous differentiability frequently is too strong of a requirement: A function ψ is called *stochastically* C^2 in the sense of [5] if the Dynkin formula

$$\mathsf{E}[\mathrm{e}^{-\delta\tau}\psi(X_\tau)] = \psi(x) + \mathsf{E}\left[\int_0^\tau \mathsf{A}\psi(X_t)\,\mathrm{d}t\right]$$

holds for any bounded stopping time τ if there is no harvesting on $[0, \tau)$. It is possible to formulate conditions for (super)optimality in terms of viscosity solutions to the HJB quasi-variational inequality

$$\max\{\mathsf{A}\psi,\ 1-\psi'\} = 0$$

—see, e.g., [7, proof of Theorem 3] for a similar case. However, in section 5 below, we shall use the second derivative (see Proposition 3.5 for regularity at the reflection threshold). We therefore state a verification theorem which is tailored to the case we want to study.

THEOREM 3.1 (sufficient conditions). *Suppose we can find a* $\psi : [0,\infty) \mapsto [0,\infty)$ *which is stochastically* C^2 *and such that for all* $x > 0$,

$$\max\{\mathsf{A}\psi(x^-),\ 1-\psi'(x^-)\} \leq 0. \tag{3.2}$$

Then

$$\psi \geq \Psi. \tag{3.3}$$

For optimality, define the nonintervention region D *as*

$$D := \{x \geq 0;\ \psi'(x) > 1\}$$

and assume that D is a union of disjoint intervals $(\check{x}^{(k)}, \hat{x}^{(k)})$ and point 0. Suppose in addition that (3.2) holds with equality, $\psi \in \mathsf{C}^2((0,\infty) \setminus \cup_k\{\check{x}^{(k)}\})$, and $\psi(0) = 0$. Define the control $\hat{H}$ by

$$\hat{H}_{t^+} = \sum_k L_t^{\hat{x}^{(k)}} + \sum_{s\in[0,t]} \inf\{h \geq 0; X_s - h \in D\}, \tag{3.4}$$

where L^y is the local time of X at y. Suppose that using this $\hat{H}$ we have a sequence of stopping times τ_n such that

$$\mathsf{E}[\mathrm{e}^{-\delta\tau_n}\psi(X_{\tau_n^+})] \to 0. \tag{3.5}$$

Then $\psi = \Psi$ and $H^* := \hat{H}$ is optimal.

Proof. Using an arbitrary control H and expanding $\mathsf{E}[\mathrm{e}^{-\delta\tau}\psi(X_\tau)]$ using the Itô formula, we have

$$\psi(x) \geq \mathsf{E}\left[\int_{[0,\tau)} \mathrm{e}^{-\delta t}\,\mathrm{d}H_t\right] + \mathsf{E}[\mathrm{e}^{-\delta\tau}\psi(X_\tau)] \tag{3.6}$$

for all bounded stopping times τ, and superoptimality follows from the nonnegativity of ψ. To prove optimality we use H^*. This strategy involves action only on the at most countable points $\{\hat{x}^{(k)}\}$, $\{\check{x}^{(k)}\}$. At the former, we have reflection at a point where $\psi' = 1$, and ψ is C^2. At the latter, when X hits $\check{x}^{(k)}$ it is immediately harvested down to $\hat{x}^{(k-1)}$ (note that $\psi' = 1$ on $(\hat{x}^{(k-1)}, \check{x}^{(k)})$), so we may apply the Itô formula piecewise and disregard the possible discontinuity in $\mathsf{A}\psi$. In conclusion, equality in (3.2) implies equality in (3.6), and condition (3.5) grants that $\psi = \Psi$. □

Remark 3.2. The assumption (3.5) is ad hoc, and though it trivially holds if D is bounded, one may want an alternative condition. Note that if we have $\beta \leq \delta$ for all large enough x, then $\Psi(x) \leq x + F$, where $F := \sup(x(\beta/\delta - 1))$ is finite by the assumed local boundedness of $x\beta$. This can be verified by (3.2) but does not depend on the Markovian nature of the problem. In this case, it is no restriction to impose $\psi \leq x + F$ for our candidate ψ to be checked for optimality. This bound, together with a slightly stronger assumption on β, namely the condition that there exists some $\epsilon \in (0, \delta)$ such that $\beta < \delta - \epsilon$ for large enough x, will imply (3.5). Since $F\mathrm{e}^{-\delta T} \to 0$, it suffices to check that $\mathrm{e}^{-\delta T}\mathsf{E}[X_T]$ tends to zero. By Itô's formula,

$$\mathsf{E}[X_T] \leq x + \int_0^T \mathsf{E}[X_t\beta(X_t)]\,\mathrm{d}t \leq x + \int_0^T \big(\delta F + (\delta - \epsilon)\mathsf{E}[X_t]\big)\,\mathrm{d}t := Y_T \geq 0.$$

By a Gronwall argument, it easily follows that

$$\mathrm{e}^{-\delta T}\mathsf{E}[X_T] \leq \mathrm{e}^{-\delta T}Y_T \leq \mathrm{e}^{-\epsilon T}\left(x + \frac{\delta F}{\delta - \epsilon}\right) \xrightarrow[T\to\infty]{} 0.$$

Let us introduce the notation

$$x^* := \sup D, \qquad x_* := \inf D \setminus \{0\} \tag{3.7}$$

and also agree to say that some $x = \hat{x}$ if it is optimal to reflect downwards at x; i.e., $y \in D$ for all sufficiently large $y < x$ but no sufficiently small $y > x$. Similarly, we write $x = \check{x}$ if $y \in D$ for all sufficiently small $y > x$ but for no sufficiently large $y < x$. We then have some properties of the optimal harvesting problem.

THEOREM 3.3.

(a) *If there is an $\epsilon > 0$ such that $\beta - \delta > \epsilon$ everywhere, then $\Psi(x) = \infty$ for all $x > 0$.*

(b) *If $\Psi(x) = x$, then on $(0, x)$ we have $\beta(y) \le \delta$.*

(c) *Assume either both $x = x^*$ and $\Psi'(x^{*-}) = 1$, or both Ψ concave and $x < x^*$. Then $\beta(x^-) \ge \delta$ with equality iff Ψ linear on $[0, x]$.*

Proof.

(a) Consider the (admissible) strategy of harvesting a constant fraction $h \in (0, \epsilon]$ of the population, i.e., $\mathrm{d}H_t = hX_t \mathrm{d}t$. Expanding $\mathrm{e}^{-\delta t} X_t$ using the Itô formula and assuming $X_0 = x > 0$, we get

$$\begin{aligned} J^H &= h \int_0^\infty \mathsf{E}[\mathrm{e}^{-\delta t} X_t] \, \mathrm{d}t \\ &= h \int_0^\infty \int_0^t \left(x + \mathrm{e}^{-\delta s} \mathsf{E}[(\beta(X_s) - \delta - h) X_s] \right) \mathrm{d}s \, \mathrm{d}t \\ &\ge hx \int_0^\infty \int_0^t 1 \, \mathrm{d}s \, \mathrm{d}t = \infty, \end{aligned}$$

since $h \le \epsilon$ implies $\beta(x) - \delta - h \ge 0$ by assumption.

(b) For $y \in (0, x)$, we must have $\Psi(y) = y$ and

$$0 \ge \mathsf{A}\Psi(y) = -\delta y + y\beta(y) + \mathsf{Q}\Psi(y) \ge -\delta y + y\beta(y)$$

since $\Psi(y + zy) \ge y + zy$ always. So $0 \ge -\delta y + y\beta(y)$.

(c) Consider $\Psi(x + zx) - \Psi(x) - zx\Psi'(x)$. It is easy to see that it is negative if Ψ is concave, which also applies at x^{*-} if $z \ge 0$. On the other hand, for $z < 0$ we have

$$\Psi(x + zx) \le \Psi(x) - zx = \Psi(x) - zx\Psi'(x^{*-})$$

by assumption. So the hypothesis implies that $-\delta\Psi(x) + x\beta(x^-)\Psi'(x^-)$ is nonnegative. In the concave case, we know that $x\Psi'(x^-) \le \Psi(x)$ which proves the claim. At $x = x^{*-}$, it follows by $\Psi'(x^{*-}) = 1$ and $\Psi(x) \ge x$. □

Remark 3.4. Note that we cannot allow $\epsilon = 0$ in part (a) above. A counterexample is given: Let $\theta \in (0, 1)$ and $A > 0$, and assume $\beta = \frac{x + Ax^\theta}{x + \theta Ax^\theta} \delta$ and for simplicity $\sigma = 0 = q$. Then it is easy to show that no optimal strategy exists, as waiting is always better, and that $\psi(x) = x + Ax^\theta$ is superoptimal. However, in a real world population model it is hardly a restriction to assume that $\beta < \delta$ when x is large enough—and property (c) above then upper bounds x^*. Property (b) shows that if $\delta < \beta(0^+)$, it is never optimal to harvest the population to extinction. This is a well-known feature from the deterministic and Brownian motion settings and is also proven under more general concave or linear Hindy–Huang preferences in [6]. This is not the same as saying that the optimal harvesting strategy will not indirectly lead to extinction; let, for example, $y > 0$ be a trap of the process if uncontrolled, while for $x < y$ the annihilation intensity $\bar{q}$ is positive. Then if we harvest any amount, the population could become extinct in finite time (typically even a.s.).

Even if the viscosity solution concept enables us to consider also nonsmooth candidates for value function, smoothness is certainly a valuable property; indeed, we should expect the value function to be smooth (if finite) also at points $x = \hat{x}$ at which reflection is optimal.

PROPOSITION 3.5 (regularity at $\hat{x}$). *Assume that* $\mathsf{A}\Psi(\hat{x}^-) = 0$ *(in the classical sense), that* $\Psi'(x^{*-})$ *exists and is* 1, *and that* β *and* $x \mapsto q$ *are continuous at* $\hat{x}$. *If* $\lim_{x \nearrow \hat{x}} \sigma^2(x) = 0$, *then assume that* $\beta(\hat{x}) > 0$ *(which holds if* $\hat{x} = x^*$ *or if* $\hat{q}(\hat{x}, \cdot) = 0$*), and furthermore that* $\sigma^2(\hat{x}^-)\Psi'''(\hat{x}^-) = 0$ *and that the coefficients are differentiable at* $\hat{x}$. *Then* Ψ *is* C^2 *at* $\hat{x}$.

Proof. For any sequence $x_n \nearrow \hat{x}$, we have

$$0 \geq \mathsf{A}\Psi(\hat{x}^+) = \lim \left(\mathsf{A}\Psi(x_n) - \tfrac{1}{2}x_n^2\sigma^2(x_n)\Psi''(x_n)\right) \geq 0, \tag{3.8}$$

since Ψ' is decreasing at $\hat{x}$. Therefore, $\mathsf{A}\Psi$ is continuous at $\hat{x}$, and we must have Ψ'' continuous at $\hat{x}$ unless possibly if $\sigma^2(\hat{x}^-) = 0$. In that case, we first note that with $\hat{x} = x^*$ or $\hat{q}(\hat{x}, \cdot) = 0$, then $\mathsf{A}\Psi(\hat{x})$ will be strictly negative unless $\hat{x}\beta(\hat{x})\Psi'(\hat{x}) > 0$. Now $x \mapsto \mathsf{A}\Psi$ is decreasing at $\hat{x}$, while its left-hand derivative is zero. For $x < \hat{x}$,

$$\begin{aligned} 0 &= ((x\beta)' - \delta)\Psi' + (x\beta + \tfrac{1}{2}(x^2\sigma^2)')\Psi'' + \tfrac{1}{2}x^2\sigma^2\Psi''' + \tfrac{\mathrm{d}}{\mathrm{d}x}\mathsf{Q}\Psi(x) \\ &\geq \left[((x\beta)' - \delta)\Psi' + \tfrac{\mathrm{d}}{\mathrm{d}x}\mathsf{Q}\Psi(x)\right]_{x=\hat{x}}. \end{aligned}$$

Since $\hat{x}$ is a minimum for $x^2\sigma^2$, then $0 \leq \hat{x}\beta(\hat{x})\Psi''(\hat{x}^-)$. □

Remark 3.6. While Proposition 3.5 gives good reason to expect Ψ to be C^2 at points where it is optimal to reflect downwards, there is no reason to expect that the value function behaves equally nice at $\check{x}$: there, the value function is convex, which invalidates the positivity inequality in (3.8), and so we cannot expect $\mathsf{A}\Psi$ to be continuous at $\hat{x}$. Furthermore, at that point we will perform a proper impulse, not a reflection, which makes the problem akin to optimal stopping—indeed, the problem of optimizing w.r.t. x_* is a proper optimal stopping problem; those problems are known to frequently exhibit exact once continuous differentiability of the value function, and it is not difficult to construct examples where the value function cannot be C^2. Briefly, continuous second derivative would imply that $\beta^{-1}(\delta)$ determines indifference on whether or not to deplete the population, but it is easy to see that waiting for a sufficiently small time τ and then keeping the population constant from then on yields a higher performance in the deterministic case. The expected nonsmoothness at $\check{x}$ is the reason for the ad hoc formulation of Theorem 3.1 above. The nonsmoothness also helps to explain why the introduction of uncertainty may increase both x^* (as we will see in section 5 below) and x_* (as found in [9]) and thus, informally, postpone population at high levels but hasten it at low levels.

4. Finding an optimal solution: The concave case. This section concerns the search for a function satisfying the sufficient conditions of Theorem 3.1. We shall restrict ourselves to the case where

$$D = [0, x^*). \tag{4.1}$$

Consider, informally, the continuous case; then it is natural to proceed by finding a function f that vanishes at 0, solves the HJB equation $\mathsf{A}f = 0$, and has some inflection point $\tilde{x}$ for which $f'(\tilde{x}) > 0$. We then construct a candidate ψ by

$$\psi(x) = \frac{f(\min(\tilde{x}, x))}{f'(\tilde{x})} + \max\{0, x - \tilde{x}\}. \tag{4.2}$$

Then ψ will be C^2 and satisfy $(1 - \psi') \cdot (\mathsf{A}\psi) = 0$. We want both factors nonpositive, and if f is concave on $(0, \tilde{x})$, then we have $\psi' \geq 1$ and it suffices to check that $\mathsf{A}\psi \leq 0$

for $x \geq \tilde{x}$. Concavity is the reason for assuming the form (4.1)—if that does not hold, we will have $\psi' = 1$ on disjoint intervals. We also note that by Theorem 3.3 (b), we can hope for concavity if

$$\beta(x) > \delta \quad \text{for all small enough } x > 0. \tag{4.3}$$

Because of (4.1), the above approach will work just as well in the case with negative jumps, as we can then safely paste at high levels—we must, however, exclude positive jumps through part of the analysis.

So we proceed to show that there is a quite wide class of cases where we can find an f which vanishes at 0, is concave at 0, and solves the HJB equation. It turns out that if coefficients are real analytic (at least near 0), then we may, as in [10], adapt the Frobenius theory (at least near 0) if $\hat{q} = 0$. So assume that within some positive convergence radius, the coefficients may be represented as

$$\beta(x) = \sum_{j=0}^{\infty} \beta_j x^j, \quad \sigma^2(x) = \sum_{j=0}^{\infty} \zeta_j x^j, \quad \text{and} \quad q(x, \mathrm{d}z) = \sum_{j=0}^{\infty} q_j(\mathrm{d}z) x^j. \tag{4.4}$$

Now try to insert a nonzero function of the form $f(x) = x^\theta \sum_{i=0}^{\infty} a_i x^i$ into the HJB equation to get

$$\begin{aligned} 0 = \sum_{i=0}^{\infty} \Bigg(-\delta + \sum_{j=0}^{\infty} \Bigg[& \theta(\beta_j + \tfrac{1}{2}(\theta - 1)\zeta_j) \\ & + \int \left((1+z)^{\theta+i} - 1 - (\theta + i)z \right) q_j(\mathrm{d}z) \\ & + i(\beta_j + \theta\zeta_j) + \tfrac{1}{2} i(i-1)\zeta_j \Bigg] x^j \Bigg) a_i x^i. \end{aligned} \tag{4.5}$$

If $\beta_0 > \delta$, the constant term determines $\theta \in (0, 1)$:

$$0 = -\delta + \theta\beta_0 + \tfrac{1}{2}\theta(\theta - 1)\zeta_0 + \int \Big((1+z)^\theta - 1 - \theta z \Big) q_0(\mathrm{d}z). \tag{4.6}$$

Choosing $a_0 > 0$, we then have $f'(0^+) = +\infty = -f''(0^+)$ and thus f increasing and concave at zero. Assuming (4.3) the only remaining case is $\beta_0 = \delta$; rather than choosing $\theta = 1$ from (4.6), we equivalently pick $\theta = a_0 = 0$. It turns out that all solutions must then have $a_1 \neq 0$, and the index of the next nonzero coefficient a_{i_1} will be determined by

$$i_1 - 1 = j_1 := \min\{j \geq 1; \beta_j \neq 0\}. \tag{4.7}$$

To see this, first note that $\beta_{j_1} > 0$ by (4.3), while $\beta_0 = \delta$. Matching the i_1th coefficients we get

$$a_{i_1} \Big((i_1 - 1)(\delta + \tfrac{1}{2} i_1 \zeta_0) + \int \left((1+z)^{i_1} - 1 - i_1 z \right) q_0(\mathrm{d}z) \Big) = -a_1 \beta_{i_1 - 1}, \tag{4.8}$$

which shows both that $a_1 \neq 0$ and that i_1 is determined by (4.7), and that $a_{i_1} < 0$ and thus f concave at zero: it is easy to see that the integrand is nonnegative, and

so is $\zeta_0 = \sigma^2(0)$, while $\delta > 0$. So if the right-hand side were zero, then so would a_{i_1} and f. We have proved the following proposition.

PROPOSITION 4.1. *Assume* (4.3) *and that on some interval* $(0, R) \neq \varnothing$ *the coefficients are real analytic with* $\hat{q} = 0$. *Then on* $(0, R)$, *the HJB equation* $\mathsf{A}f = 0$ *has a solution* f *which vanishes at* 0, *increases at* 0, *and is concave at* 0.

Remark 4.2. The reader who might feel uncomfortable with the real analyticity of the measure-valued function $x \mapsto q$, may verify that a similar condition may be imposed equally well on the usual Lévy representation, which yields

$$\int \Big((1 + \eta(x,z))^{\theta+i} - 1 - (\theta + i)\eta(x,z) \Big) \lambda(\mathrm{d}z). \tag{4.9}$$

If $x \mapsto \eta$ is real analytic for each z, then for each i the integrand is analytic, being a composition of analytic functions (we may separately treat the z-values which yield $\eta = -1$). So the expression in (4.9) is analytic in x also, and we may proceed as above.

At this point, we do not know whether this concave f will have any inflection point $\tilde{x}$, or if $f'(\tilde{x}) > 0$, which is essential for the construction (4.2). This is addressed in Proposition 4.3, which is written to suit a ψ constructed by (4.2) but for the sake of generality has a slightly ad hoc formulation which does not exclude cases with positive jumps.

PROPOSITION 4.3. *Assume* (4.3). *Let* ψ *be nonlinear and concave, and assume it vanishes at* 0 *and increases at* 0. *Assume that there exists some* $\tilde{x} \in (0, \infty]$ *such that* ψ *is affine for* $x \geq \tilde{x}$, C^1 *at* $\tilde{x}$ *and solves the HJB equation* $\mathsf{A}\psi = 0$ *for* $x < \tilde{x}$. *Then* $\beta > \delta$ *on* $(0, \tilde{x})$ *and also at* $\tilde{x}^-$ *iff finite. Furthermore,* $\psi' > 0$ *everywhere and* $\mathsf{A}\psi \leq 0$ *whenever* $\beta \leq \delta$.

Proof. Concavity implies $-\delta\psi(x) \leq -\delta x(\psi'(x) + \psi(0))$ (with equality iff ψ linear on $(0, x)$), and also that $\mathsf{Q}\psi(x) \leq 0$. We therefore have

$$\mathsf{A}\psi(x) \leq x(\beta(x) - \delta)\psi'(x) + \frac{1}{2}x^2\sigma^2(x)\psi''(x) + \mathsf{Q}\psi(x) \leq x(\beta(x) - \delta)\psi'(x) \tag{4.10}$$

for all $x \in (0, \tilde{x})$. In fact, the right-hand side is strictly positive; if not, we would have $\psi(x) = kx$ on $(0, x)$, with $k > 0$ because ψ is assumed to be increasing at zero. But for sufficiently small $x > 0$ we have $\beta > \delta$ by (4.3). With the right-hand side strictly positive, ψ' cannot hit zero, and by continuity whenever $\mathsf{A}\psi$ is defined it is strictly positive, and so $\beta > \delta$. This holds for all $x \in (0, \tilde{x})$ and at $\tilde{x}^-$ if finite, in which case $\psi'(y) = \psi'(\tilde{x}) > 0$ for all $y \geq \tilde{x}$ as well. So $\psi' > 0$ everywhere. It now follows from (4.10) that $\beta \leq \delta$ implies $\mathsf{A}\psi \leq 0$. □

As long as Proposition 4.3 applies, we can by scaling with a constant assume $\psi'(\tilde{x}) = 1$, and the problem is solved if we can show that $\mathsf{A}\psi \leq 0$ on $(\tilde{x}, \bar{x})$, where $\bar{x} \leq \infty$ is defined as

$$\bar{x} := \sup\{x \geq \tilde{x};\ \beta(x) > \delta\}. \tag{4.11}$$

Note that the potentially troublesome interval $(\tilde{x}, \bar{x})$ is nonempty if β is continuous at $\tilde{x}$ and $\tilde{x} < \infty$.

Before we give the main result of this section, let us point out a few useful properties of the function $Q(x)$ defined for a fixed ψ as

$$Q(x) := \mathsf{Q}\psi(x). \tag{4.12a}$$

If $\psi \in \mathsf{C}^2$, we have

$$Q(x) = \int \int_{x+zx}^{x} \int_{y}^{x} \psi''(w)\,\mathrm{d}w\,\mathrm{d}y\,q(x,\mathrm{d}z). \tag{4.12b}$$

Assuming sufficient regularity, the first two derivatives of Q are then

$$\begin{aligned} Q'(x) &= \int \int_{x+zx}^{x} \int_{y}^{x} \psi''(w)\,\mathrm{d}w\,\mathrm{d}y\,q'(x,\mathrm{d}z) \\ &\quad + \int \Big((1+z)\big(\psi'(x+zx)-\psi'(x)\big) - zx\psi''(x)\Big)\,q(x,\mathrm{d}z) \end{aligned} \tag{4.12c}$$

and

$$\begin{aligned} Q''(x) &= \int \int_{x+zx}^{x} \int_{y}^{x} \psi''(w)\,\mathrm{d}w\,\mathrm{d}y\,q''(x,\mathrm{d}z) \\ &\quad + 2\int \Big((1+z)\big(\psi'(x+zx)-\psi'(x)\big) - zx\psi''(x)\Big)\,q'(x,\mathrm{d}z) \\ &\quad + \int \Big((1+z)\big((1+z)\psi''(x+zx)-\psi''(x)\big) - z(x\psi''(x))'\Big)\,q(x,\mathrm{d}z), \end{aligned} \tag{4.12d}$$

where $q'(x,\mathrm{d}z) := (\partial/\partial x)q(x,\mathrm{d}z)$ and $q''(x,\mathrm{d}z) := (\partial/\partial x)^2 q(x,\mathrm{d}z)$ are signed measures. We then have the following theorem.

THEOREM 4.4. *Assume that $\psi''(\tilde{x}) = 0$, with $0 < \tilde{x} < \infty$.*

(a) Optimality, $x(\beta-\delta)+Q$ eventually nonincreasing: *If Proposition* 4.3 *applies and*

$$\tilde{x}(\beta(\tilde{x})-\delta) + Q(\tilde{x}) \geq x(\beta(x)-\delta) + Q(x) \quad \forall x \geq \tilde{x}, \tag{4.13}$$

then we have $\Psi = \psi$ and the optimal continuation region is $D = [0,x^*) = [0,\tilde{x})$.

(b) Optimality, $x(\beta-\delta)+Q$ eventually concave: *If Proposition* 4.3 *applies and the coefficients are continuous at $\tilde{x}$, and $x\beta+Q$ is concave on $[\tilde{x},\bar{x}]$, then $\mathsf{A}\psi \leq 0$ and thus $\Psi = \psi$ and the optimal continuation region is $D = [0,x^*) = [0,\tilde{x})$. Note that if at some point $y \geq \tilde{x}$ we have $x \mapsto \check{q}$ nonincreasing ($\Rightarrow Q$ nondecreasing!) and convex and $\bar{q}$ is convex, then Q is concave at y.*

(c) Optimality w.r.t. a possibly modified problem: *Assume that Proposition* 4.3 *applies and that ψ is not C^∞ at $\tilde{x}$. Let the nth derivative $\psi^{(n)}$ be discontinuous at $\tilde{x}$ and assume in addition that we either have $\sigma(\tilde{x}) \neq 0$ and C^{n-2} coefficients, or $\beta(\tilde{x}) - \int z\,q(\tilde{x},\mathrm{d}z)$ exists and is > 0 and C^{n-1} coefficients. Then $\mathsf{A}\psi \leq 0$ for all small enough $x \geq \tilde{x}$. Therefore, if $\hat{q} = 0$, then we can construct a* new *problem with value function ψ and optimal continuation region $[0,\tilde{x})$ by leaving the coefficients unchanged on some right-open interval containing $[0,\tilde{x}]$ and changing β from there on.*

Proof.

(a) Since $\mathsf{A}\psi(\tilde{x}) = \psi''(\tilde{x}) = 0$, then for $x \geq \tilde{x}$,

$$\begin{aligned} \mathsf{A}\psi(x) &= \mathsf{A}\psi(x) - \mathsf{A}\psi(\tilde{x}) \\ &= -\delta\psi(x) + \delta\psi(\tilde{x}) + x\beta(x) - \tilde{x}\beta(\tilde{x}) + \mathsf{Q}\psi(x) - \mathsf{Q}\psi(\tilde{x}) \\ &= x(\beta(x)-\delta) - \tilde{x}(\beta(\tilde{x})-\delta) + \mathsf{Q}\psi(x) - \mathsf{Q}\psi(\tilde{x}). \end{aligned} \tag{4.14}$$

(b) Consider

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathsf{A}\psi(x) = (x(\beta-\delta))'\psi'(x) + \frac{\mathrm{d}}{\mathrm{d}x}\mathsf{Q}\psi(x) + \left(x\beta + \frac{1}{2}(x^2\sigma^2)'\right)\psi''(x) + \frac{1}{2}x^2\sigma^2\psi'''(x). \tag{4.15}$$

By the nonnegativity of the latter term at $\tilde{x}^-$, we know that $\mathsf{A}\psi$ has nonpositive derivative at $\tilde{x}$. Relaxing differentiability, we still have $\mathsf{A}\psi$ nonincreasing at $\tilde{x}$. Fix $x \in (\tilde{x}, \bar{x})$. Since $\psi' = 1$ around x, the derivative of (4.15) reduces to $(x\beta + Q)''$. For the last claim, consider (4.12d): Q'' reduces to

$$\int\int_{x+zx}^{x}\int_{y}^{x} \psi''(w)\,\mathrm{d}w\,\mathrm{d}y\,(\check{q}''(x,\mathrm{d}z) + \bar{q}''(x,\mathrm{d}z)) + 2\int (1+z)\big(\psi'(x+zx)-1\big)\,\check{q}'(x,\mathrm{d}z), \tag{4.16}$$

which is nonpositive if $\check{q}''$ and $\bar{q}''$ are nonnegative and $\check{q}'$ is nonpositive, since $\psi' \geq 1$ and decreasing.

(c) To prove the last assertion, let n be the smallest number such that the nth derivative $\psi^{(n)}$ is discontinuous at $\tilde{x}$. Then necessarily n is odd and $\psi^{(n)}(\tilde{x}^-) > 0 = \psi^{(n)}(\tilde{x}^+)$. If $\sigma(\tilde{x}) \neq 0$, differentiate $n-2$ times to get

$$\Big(\frac{\mathrm{d}}{\mathrm{d}x}\Big)^{n-2}\mathsf{A}\psi(\tilde{x}^+) - \Big(\frac{\mathrm{d}}{\mathrm{d}x}\Big)^{n-2}\mathsf{A}\psi(\tilde{x}^-) = -\frac{1}{2}\tilde{x}^2\sigma^2(\tilde{x})\cdot\psi^{(n)}(\tilde{x}^-), \tag{4.17a}$$

which is < 0 by assumption. If $\sigma(\tilde{x}) = 0$, differentiate instead $n-1$ times to get

$$\Big(\frac{\mathrm{d}}{\mathrm{d}x}\Big)^{n-1}\mathsf{A}\psi(\tilde{x}^+) - \Big(\frac{\mathrm{d}}{\mathrm{d}x}\Big)^{n-1}\mathsf{A}\psi(\tilde{x}^-) = -\tilde{x}\left(\beta(\tilde{x}) - \int z\,q(x,\mathrm{d}z)\right)\cdot\psi^{(n)}(\tilde{x}^-), \tag{4.17b}$$

which is < 0 by assumption. Finally, if $\hat{q} = 0$, then we can change β at x without affecting the quasi-variational inequality at values to the left of x. □

Note that if we remove the assumption $\tilde{x} < \infty$, we can conclude only that ψ is superoptimal; cf. Remark 3.4.

COROLLARY 4.5. *Assume that Proposition* 4.1 *applies and that f'' has a (finite) zero. Let $\tilde{x}$ be the leftmost, and define ψ by* (4.2). *Then Theorem* 4.4(c) applies.

Proof. By (4.10) we have that $f'(\tilde{x}) > 0$, so ψ is also concave. Now analytic functions are determined by their derivatives, so while the coefficients are C^∞ at x^*, ψ is not, unless identically equal to x, which is impossible by (4.3). □

5. The effect of uncertainty. Having found (under suitable conditions) the value function in section 4, we shall throughout this section assume some regularity.

Assumption 5.1. It is optimal to reflect the process downwards at $\hat{x}$. The value function Ψ is C^2 at $\hat{x}$.

We know from Proposition 3.5 and section 4 that this ad hoc assumption does cover a wide range of control problems, just as obtained in [2] in the nonjump case.

For a given stochastic problem denote by x_0^* the x^* obtained in the corresponding deterministic problems, i.e., the one that arises when the uncertainty terms σ and q

are replaced by zero (recall x^* defined by (3.7)). In the continuous cases studied in [3] and [10] they find that $x^* \geq x_0^*$ (equal to $\operatorname{argmax} x(\beta - \delta)$); Myhre [11] improves the bound to $x^* \geq x_0^* + \sigma_{\mathrm{b}}/\beta_0$ for the latter case (1.1b). The relation $x^* \geq x_0^*$ may have the interpretation of one being more careful under uncertainty. It turns out that jump uncertainty may violate this property (at least apparently; see the closing remarks for an interpretation). However, it holds if the jump intensity is nonincreasing in x, which leads us to the well-known comparison theorem. Consider for a moment the Lévy representation (2.1b). To extend the comparison theorem of the continuous case, it is sufficient that state after jump is nondecreasing as a function of the prejump state for almost every jump index z, i.e., that $x(1+\eta)$ is nondecreasing. This is also almost necessary—comparison fails if X is expected to spend positive time where $x(1+\eta)$ is strictly decreasing. Now in terms of the Lévy representation, we have

$$Q'(x) = \int \big((\Psi'(x + x\eta(x,z)) - \Psi'(x)) \cdot (x + x\eta)' - x\eta\Psi''(x)\big)\, \lambda(\mathrm{d}z). \tag{5.1}$$

At $\hat{x}$, Ψ' attains its minimum value and $\Psi'' = 0$. Therefore, comparison—locally at $\hat{x}$—is sufficient to grant $Q'(\hat{x}) \geq 0$, a property which turns out to be critical for the behavior towards risk.

PROPOSITION 5.2. *Assume $\Psi'''(\hat{x}^-)$ exists and assume (for simplicity, admits generalizations) coefficients differentiable at $\hat{x}$. If $Q'(\hat{x}) \geq 0$ (resp., > 0), then*

$$0 \geq \ \textit{(resp., >)}\ (x(\beta - \delta))'|_{x=\hat{x}}.$$

Hence for $x\beta$ concave, $\hat{x}$ is no smaller than (resp., strictly greater than) in the deterministic case. If, on the other hand, $\sigma(\hat{x}) = 0$ and $Q'(\hat{x}) \leq 0$ (resp., < 0), then $(x(\beta - \delta))'|_{x=\hat{x}} \geq 0$ (resp., > 0); hence for $x\beta$ concave, $\hat{x}$ is no larger than (resp., strictly smaller than) in the deterministic case. Furthermore, x^* is unaffected by $\hat{q}$ as long as $\sigma(x^*) = 0$.

Proof. Differentiate the equation $\mathsf{A}\Psi = 0$ and insert $\hat{x}^-$:

$$-\big(x(\beta - \delta)\big)'|_{x=\hat{x}} = \frac{1}{2}\hat{x}^2\sigma^2(\hat{x})\Psi'''(\hat{x}^-) + Q'(\hat{x}) \geq Q'(\hat{x}) \tag{5.2}$$

with equality if $\sigma(\hat{x}) = 0$; in that case, Ψ''' vanishes from (5.2), which is therefore not affected by $\hat{q}$—to see this, do the construction (4.12c) with $\hat{q}$ instead of q to see that the integrands become zero. □

Proposition 5.2 combined with Theorem 4.4(c) now yields a main point of this paper.

META-THEOREM 5.3 (behavior towards risk). *There is a wide class of problems for which the optimal solution is to reflect downwards at the same or a lower level than the corresponding deterministic problem, contrary to the case covered in* [2].

While this may appear a bit counterintuitive at first glance, it certainly makes sense: jump intensity decreasing in x may lead to one keeping the population at a higher level to reduce the probability of "disasters," i.e., negative jumps; on the other hand, the presence of the jump terms may lead us to harvest and reduce X before the jumps do. This is, however, a priori not a valid argument, since jumps are compensated. The case with only annihilation risk is illustrative, but it will be convenient (and no more complicated, as $\hat{q}$ does not affect $Q'(x^*)$) to include positive jumps if $\hat{x} = x^*$ in the following theorem.

THEOREM 5.4. *Assume $\hat{q} = 0$ or $\hat{x} = x^*$, and furthermore $\breve{q} = \sigma = 0$ and continuous and piecewise differentiable coefficients. Then $\hat{x}$ is a stationary point of*

$x(\beta-\delta)/(\delta+\bar{q})$. Furthermore, $x(\beta-\delta)$ is (strictly) increasing (resp., decreasing) at $\hat{x}$ iff $\bar{q}$ is. In particular, if $x(\beta-\delta)/\delta$ and $x(\beta-\delta)/(\delta+\bar{q})$ have unique maximum points $\hat{x}_0$, $\hat{x}$, respectively, their respective control problems have optimal continuation regions $D_0=[0,x_0^*)=[0,\hat{x}_0)$ and $D=[0,x^*)=[0,\hat{x})$ and $x^*\le x_0^*$ ($<x_0^*$) iff $\bar{q}$ is (strictly) increasing at x^*.

Proof. The HJB equation $\mathsf{A}\Psi=0$ is now

$$0=-(\delta+\bar{q})\Psi(x)+x(\beta+\bar{q})\Psi'(x)+Q(x). \tag{5.3a}$$

Inserting $x=\hat{x}$, then by our assumptions we have

$$0=-(\delta+\bar{q}(\hat{x}))\Psi(\hat{x})+\hat{x}(\beta(\hat{x})+\bar{q}(\hat{x})). \tag{5.3b}$$

Differentiate (5.3a), evaluate at $\hat{x}$, and insert for $\Psi(\hat{x})$ from (5.3b). Then we get that $\hat{x}$ must be a stationary point for $x(\beta-\delta)/(\delta+\bar{q})$, i.e., where

$$\frac{(x(\beta-\delta))'}{x(\beta-\delta)}=\frac{\bar{q}'}{(\delta+\bar{q})} \tag{5.4}$$

since $\beta(\hat{x})>\delta$ (if not, (5.3a) would yield $\Psi(\hat{x})\le\hat{x}$). □

It is known that Brownian uncertainty has the effect of reducing the optimal *value* (or leaving it unchanged)—see [2, Theorem 6]. We see from the above calculation that this property is not necessarily connected to the postponed harvesting associated with the Brownian uncertainty, as the introduction of a $\bar{q}$ not affecting x^* will reduce $\Psi(x^*)=x^*(1+(\beta(x^*)-\delta)/(\delta+\bar{q}(x^*)))$. But in Theorem 5.4, positive jumps do not affect the value function for $x\ge x^*$; nevertheless we have the following theorem.

THEOREM 5.5 (compensated positive jumps' effect on optimal value). *Assume that Theorem 5.4 applies and $\hat{q}$ is nonzero for all sufficiently large $x<x^*$. Then $\hat{q}$ reduces the optimal value for all sufficiently large $x<x_0^*$.*

Proof. By Theorem 5.4, we know that x^* and $\Psi(x^*)$ are not affected. Now since Ψ' is decreasing to 1, we have that Ψ is concave on (y,∞) for all sufficiently large $y<x^*$. Fix such a y; since $\beta(x^*)>0$, we can and will assume $\beta>0$ on (y,x^*). Let $\hat{\Psi}$ be the value function with $\hat{q}$ as assumed, and let $\hat{\Psi}_0$ be the value function with $\hat{q}$ replaced by the zero measure; then we have $\mathsf{Q}\hat{\Psi}<0=\mathsf{Q}\hat{\Psi}_0$ on (y,x^*), and by the HJB equation we have

$$-\delta\hat{\Psi}_0+x\beta\hat{\Psi}_0'=0<-\delta\hat{\Psi}+x\beta\hat{\Psi}'.$$

Let $\Delta:=\hat{\Psi}-\hat{\Psi}_0$; then we have $\Delta(x^*)=0$ and

$$\delta\Delta<x\beta\Delta' \quad \text{on } (y,x^*). \tag{5.5}$$

Assume for contradiction that $\Delta(x)\ge 0$ for some $x\in(y,x^*)$; then $\Delta'(x)>0$, and so Δ becomes positive and thus by (5.5) continues to increase and will ultimately violate $\Delta(x^*)=0$. We conclude that $\Delta<0$, i.e., $\hat{\Psi}<\hat{\Psi}_0$, on (y,x^*). □

We end our analysis with an example illustrating Theorem 5.4.

Example 5.6. If $x\beta$ is concave and $\bar{q}$ is convex, we can solve the problem completely if $x(\beta-\delta)$ is increasing at 0 and has some stationary point, and that $\beta(0)>\delta$ and $\beta'(0^+)$ and $\bar{q}(0^+)$ are both finite. Then it is easy to verify that $x(\beta-\delta)/(\delta+\bar{q})$ also is increasing at 0 and has some stationary point; let x^* be the smallest one. Then for $x\le x^*$,

$$\psi(x)=x^*\cdot\frac{\beta(x^*)+\bar{q}(x^*)}{\delta+\bar{q}(x^*)}\exp\left\{\int_{x^*}^{x}\frac{\delta+\bar{q}(y)}{y(\beta(y)+\bar{q}(y))}\,dy\right\}. \tag{5.6}$$

It is easy to verify that the HJB equation holds, that $\psi'' \le 0$, and hence that $\psi' \ge \psi'(x^*) = 1$ and that Theorem 4.4 (b) applies. By (5.4), we can merely check sign $\bar{q}'(x^*)$ to find whether the optimal trigger is higher than in the deterministic case or not. As a special case, consider the logistic growth model

$$\beta(x) = \beta_0(1 - x/K) \tag{5.7}$$

(with $\beta_0 > \delta$) modified with a compensated annihilation term with intensity $\bar{q}(x)$ of the form $(q_0 + q_1 x)^+$ (convex); assume $\sigma = \hat{q} = \check{q} = 0$. Then $x_0^* = (\beta_0 - \delta)K/2\beta_0$. Assume that $\bar{q}(x^*) > 0$; one may verify that

$$x^* = \frac{\delta + q_0}{q_1}\left(-1 + \sqrt{1 + 2x_0^* \frac{q_1}{\delta + q_0}}\right). \tag{5.8}$$

We see that x^* is strictly decreasing in $(\delta + q_0)/q_1$; therefore, increasing both q_0 and q_1 simultaneously gives no information on whether x^* increases or decreases. We may also find the value function: For $x < x^*$,

$$\ln \Psi(x) = \begin{cases} \frac{\delta+q_0}{\beta_0+q_0} \ln x + \left(\frac{Kq_1}{Kq_1-\beta_0} - \frac{\delta+q_0}{\beta_0+q_0}\right) \ln(\beta_0 + q_0 + (q_1 - \frac{\beta_0}{K})x) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } q_1 \ne \frac{\beta_0}{K}, \text{ and} \\ \frac{1}{\beta_0+q_0}\left((\delta + q_0)\ln x + \frac{\beta_0}{K}x\right) \qquad \text{if } q_1 = \frac{\beta_0}{K}. \end{cases} \tag{5.9}$$

We omit the details.

6. Closing remarks. We have seen that even in this simple model, the optimal strategy may adapt qualitatively differently to the introduction of a jump martingale compared to the introduction of Brownian noise to the model. The result suggests that one should be careful about how one models uncertainty. We have also seen that the property of increasing trigger value $\hat{x}$ is related not solely to the Brownian motion assumption but to the applicability of a comparison theorem. Comparison means, roughly speaking, that by saving rather than harvesting, we do not risk losing more than what we save, while a general jump term may cost us more if we allow the population to grow to a state where the jumps are "worse." We therefore interpret Proposition 5.2 as *a tradeoff between noise level on one hand, and on the other hand exposure to risk of falling to a lower level.* Speaking heuristically, the assertion that "risk leads to higher trigger level" is now modified by adding "as long as it does not become risky to increase the trigger level"—a reservation which is redundant under continuity or comparison.

From [2], we know that if we consider problems indexed by the Brownian volatility, as $\{\ell\sigma\}_{\ell \ge 0}$, then under suitable assumptions we have x^* monotonically increasing in ℓ. We have now established that introducing a pure jump Markov martingale to a deterministic model may either reduce or increase the optimal trigger x^*, and one may want to ask, If we introduce jump intensities $\{\ell q\}_{\ell \ge 0}$, for what q is x^* monotone with respect to ℓ? And, ultimately, for what $\{(\ell\sigma, \ell q)\}$ is x^* monotone in ℓ? These are topics for future research.

Acknowledgments. The author acknowledges the hospitality of the University of Kansas where part of the work was carried out. This paper has benefited from anonymous referee comments.

REFERENCES

[1] L. H. R. Alvarez, *Optimal harvesting under stochastic fluctuations and critical depensation*, Math. Biosci., 152 (1998), pp. 63–85.

[2] L. H. R. Alvarez, *Singular stochastic control, linear diffusions, and optimal stopping: A class of solvable problems*, SIAM J. Control Optim., 39 (2001), pp. 1697–1710.

[3] L. H. R. Alvarez and L. A. Shepp, *Optimal harvesting of stochastically fluctuating populations*, J. Math. Biol., 37 (1998), pp. 155–177.

[4] A. Bensoussan and J.-L. Lions, *Impulse Control and Quasivariational Inequalities*, Gauthier–Villars, Montrouge, France, 1984.

[5] K. A. Brekke and B. Øksendal, *The high contact principle as a sufficiency condition for optimal stopping*, in Stochastic Models and Option Values (Loen, 1989), D. Lund and B. Øksendal, eds., North–Holland, Amsterdam, 1991, pp. 187–208.

[6] N. C. Framstad, *A Remark on Non-Depletion of a Natural Resource under Intertemporal Preferences*, Preprint 16, Pure Mathematics, University of Oslo, Oslo, Norway, 2001; http://www.math.uio.no/eprint/pure_math/2001/16-01.html.

[7] N. C. Framstad, B. Øksendal, and A. Sulem, *Optimal consumption and portfolio in a jump diffusion market with proportional transaction costs*, J. Math. Econom., 35 (2001), pp. 233–257.

[8] J. Jacod and A. N. Shiryaev, *Limit Theorems for Stochastic Processes*, Springer-Verlag, Berlin, 1987.

[9] R. Lande, S. Engen, and B.-E. Sæther, *Optimal harvesting, economic discounting and extinction risk in fluctuating population*, Nature, 372 (1994), pp. 88–90.

[10] E. M. Lungu and B. Øksendal, *Optimal harvesting from a population in a stochastic crowded environment*, Math. Biosci., 145 (1997), pp. 47–75.

[11] T. Myhre, *A Connection Between Singular Stochastic Control and Optimal Stopping*, Cand.scient. thesis, University of Oslo, Oslo, Norway, 1997.

[12] P. Protter, *Stochastic Integration and Differential Equations*, Springer-Verlag, Berlin, 1990.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1466–1482

MARKOWITZ'S MEAN-VARIANCE PORTFOLIO SELECTION WITH REGIME SWITCHING: A CONTINUOUS-TIME MODEL*

XUN YU ZHOU† AND G. YIN‡

Abstract. A continuous-time version of the Markowitz mean-variance portfolio selection model is proposed and analyzed for a market consisting of one bank account and multiple stocks. The market parameters, including the bank interest rate and the appreciation and volatility rates of the stocks, depend on the market mode that switches among a finite number of states. The random regime switching is assumed to be independent of the underlying Brownian motion. This essentially renders the underlying market *incomplete*. A Markov chain modulated diffusion formulation is employed to model the problem. Using techniques of stochastic linear-quadratic control, mean-variance efficient portfolios and efficient frontiers are derived explicitly in *closed forms*, based on solutions of two systems of linear ordinary differential equations. Related issues such as a minimum-variance portfolio and a mutual fund theorem are also addressed. All the results are markedly different from those for the case when there is no regime switching. An interesting observation is, however, that if the interest rate is deterministic, then the results exhibit (rather unexpected) similarity to their no-regime-switching counterparts, even if the stock appreciation and volatility rates are Markov-modulated.

Key words. continuous time, regime switching, Markov chain, mean-variance, portfolio selection, efficient frontier, linear-quadratic control

AMS subject classifications. Primary, 90A09; Secondary, 93E20

DOI. 10.1137/S0363012902405583

1. Introduction.

Recently there has been an increasing interest in financial market models whose key parameters, such as the bank interest rate, stocks appreciation rates, and volatility rates, are modulated by some Markov processes. This is motivated by the need of more realistic models that better reflect random market environment. A factor that dominates the movement of a stock is the trend of the market. To reflect the market trend, it is necessary to allow the key parameters to respond to the general market movements. One such formulation is the regime switching model, where the market parameters depend on the market mode that switches among a finite number of states. The market mode could reflect the state of the underlying economy, the general mood of investors in the market, and other economic factors. For example, the market can be roughly divided as "bullish" and "bearish," while the market parameters can be quite different in the two modes. One could certainly introduce more intermediate states between the two extremes. A regime switching model can be formulated mathematically as a stochastic differential equation (SDE) whose coefficients are modulated by a continuous-time Markov chain. Such models have been mainly employed in the literature to deal with options; see Barone-Adesi and Whaley [1], Di Masi, Kabanov, and Runggaldier [6], Guo [10], Buffington and

*Received by the editors April 16, 2002; accepted for publication (in revised form) May 5, 2003; published electronically October 28, 2003.

http://www.siam.org/journals/sicon/42-4/40558.html

†Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (xyzhou@se.cuhk.edu.hk). The research of this author was supported in part by the RGC Earmarked Grants CUHK 4435/99E and CUHK 4175/00E.

‡Department of Mathematics, Wayne State University, Detroit, MI 48202 (gyin@math.wayne.edu). The research of this author was supported in part by the National Science Foundation under grants DMS-9877090 and DMS-0304928.

Elliott [3], and Yao, Zhang, and Zhou [26]. In addition, recently Zhang [30] studied an optimal stock selling rule for a Markov-modulated Black–Scholes model (see also [27] for a stochastic optimization approach).

In this paper, we develop a continuous-time version of the Nobel prize winning mean-variance portfolio selection model with regime switching and attempt to derive closed-form solutions for efficient portfolios and efficient frontier. The mean-variance model was originally proposed by Markowitz [20, 21] for portfolio construction in a single period. One of the salient features of his model is as follows: It enables an investor to seek the highest return after specifying his/her acceptable risk level that is quantified by the variance of the return. The mean-variance approach has become the foundation of modern finance theory and has inspired numerous extensions and applications. One natural extension is to investigate *dynamic* mean-variance models. Along this line, multiperiod mean-variance portfolio selection was studied in, for example, Samuelson [23], Hakansson [11], and Pliska [22] among others. On the other hand, continuous-time mean-variance hedging problems were attacked by Duffie and Richardson [7] and Schweizer [24], where optimal dynamic strategies were derived, based on the projection theorem, to hedge contingent claims in incomplete markets. In [7], the result was derived under the assumption that all the coefficients (interest rate, volatility rate, etc.) are deterministic, time-invariant constants. The model considered in [24] is mathematically general; however, the solution is based on an abstract martingale measure and is thus not easily decipherable.

It should be noted that the research works on dynamic portfolio selections have been dominated by those of maximizing expected utility functions of the terminal wealth. In the utility model, besides the difficulty in eliciting utility functions from the investors, tradeoff between the risk and return is implicit, making an investment decision much less intuitive. In this sense, Markowitz's mean-variance approach has not been fully utilized in the utility approach.

Using the recently developed stochastic linear-quadratic (LQ) control framework [4, 5, 29], Zhou and Li [31] studied the mean-variance problem for a continuous-time model from another angel. By embedding the original (not readily solvable) problem into a tractable auxiliary problem, following a similar embedding technique introduced in Li and Ng [19] for the multiperiod model, it was shown that this auxiliary problem in fact is a stochastic optimal LQ problem and can be solved explicitly by LQ theory. Such an approach establishes a natural connection of the portfolio selection problems and standard stochastic control models. The theory of stochastic control is rich, and many mathematical machineries are available; see Fleming and Soner [9] and Yong and Zhou [29], which provides an opportunity for treating more complicated situations. For example, a portfolio selection problem with random coefficients was solved in [16] using LQ theory and backward SDEs, a problem with short sell prohibition was studied in [15] via LQ and viscosity solution theories, and a mean-variance hedging problem was treated in [12] within the LQ framework.

In this work, we focus on a continuous-time mean-variance model modulated by a Markov chain representing the regime switching. The random switching of the market modes is assumed to be independent of the Brownian motion in defining the stock prices. Therefore, the underlying market is essentially incomplete, as the regime switching constitutes an additional dimension of uncertainty that cannot be perfectly hedged by any combination of the stocks and the bank account. We formulate the problem as a Markov-modulated stochastic LQ control model with a terminal

constraint representing the expected payoff of the investor. The feasibility due to the constraint is first addressed under a very mild condition. Then, using Lagrange multiplier techniques, the problem is converted to an unconstrained problem. We proceed with the solution of the unconstrained problem based on two systems of ordinary differential equations (ODEs). This leads to the analytic expressions of the efficient portfolios in a feedback form as well as the efficient frontier. In addition, the minimum variance is explicitly derived. In fact, one needs only to solve two systems of linear ODEs in order to completely determine the efficient portfolios/frontier of the underlying mean-variance problem. It is interesting, though rather expected, that the efficient frontier is no longer a straight line in the mean-standard deviation diagram. However, if the interest rate is independent of the Markov chain, then the efficient frontier becomes a straight line again, and the one-fund theorem is preserved, even if the appreciation and volatility rates of the stocks are random (i.e., Markov-modulated).

It should be noted that in our model the wealth process is allowed to take negative values, representing the bankruptcy situation. This is due to our definition of admissible portfolios. (A portfolio is defined to be a vector consisting of the dollar values of different stocks.) In most of the literature, a portfolio contains the *fractions* of wealth in stocks, which *automatically* ensures the positivity of the wealth process (see Remark 1). In our model, requiring a nonnegative wealth process imposes an *additional state constraint*, which is a very difficult problem from the stochastic control point of view. We are not able to treat such a case in this paper and will defer it to later consideration. We remark that portfolio selection problems with constraints on wealth have been studied by many researchers, mostly in the realm of utility optimization. In particular, Korn and Trautmann considered in [14], for the first time, a mean-variance problem with nonnegative terminal constraint and without regime switching; see also Korn [13, Chapter 4]. The basic idea presented in these references is to reduce the problem to one finding an optimal attainable terminal wealth, the latter being a quadratic optimization problem. Then the efficient portfolio is the one that duplicates the optimal attainable terminal wealth.

The rest of the paper is arranged as follows. Section 2 begins with the precise problem formulation. Section 3 is concerned with the feasibility issue of the underlying model. Section 4 proceeds with the solution of the unconstrained optimization problem. The efficient frontier is obtained in section 5. Section 6 specializes in the case when the interest rate is independent of the modulating Markov chain. Finally, concluding remarks are made in the last section.

2. Problem formulation. Throughout the paper, let $(\Omega, \mathcal{F}, P)$ be a fixed complete probability space on which are defined a standard d-dimensional Brownian motion $W(t) \equiv (W_1(t), \dots, W_d(t))'$ and a continuous-time stationary Markov chain $\alpha(t)$ taking value in a finite state space $\mathcal{M} = \{1, 2, \dots, l\}$ such that $W(t)$ and $\alpha(t)$ are independent of each other. The Markov chain has a generator $Q = (q_{ij})_{l\times l}$ and stationary transition probabilities

$$p_{ij}(t) = P(\alpha(t) = j | \alpha(0) = i), \quad t \geq 0, \quad i, j = 1, 2, \dots, l. \tag{2.1}$$

Define $\mathcal{F}_t = \sigma\{W(s), \alpha(s) : 0 \leq s \leq t\}$. We denote by $L^2_{\mathcal{F}}(0, T; \mathbb{R}^m)$ the set of all $\mathbb{R}^m$-valued, measurable stochastic processes $f(t)$ adapted to $\{\mathcal{F}_t\}_{t\geq 0}$, such that $E\int_0^T |f(t)|^2 dt < +\infty$. We will also use the following notation.

Notation.

M':	the transpose of any vector or matrix M;
m_j:	the jth component of any vector M;
$\|M\|$:	$= \sqrt{\sum_{i,j} m_{ij}^2}$ for any matrix vector $M = (m_{ij})$ or $= \sqrt{\sum_j m_j^2}$ for any vector $M = (m_j)$;
$\mathrm{tr}(M)$:	the trace of a square matrix M;
$C([0,T];X)$:	the Banach space of X-valued continuous functions on $[0,T]$ endowed with the maximum norm $\|\cdot\|$ for a given Hilbert space X;
$C^2([0,T]\times\mathbb{R}^n)$:	the space of all twice continuously differentiable functions on $[0,T]\times\mathbb{R}^n$;
$L^2(0,T;X)$:	the Hilbert space of X-valued integrable functions on $[0,T]$ endowed with the norm $(\int_0^T \| f(t) \|_X^2 \, dt)^{\frac{1}{2}}$ for a given Hilbert space X.

Consider a market in which $d+1$ assets are traded continuously. One of the assets is a bank account whose price $P_0(t)$ is subject to the stochastic ODE

$$\begin{cases} dP_0(t) = r(t,\alpha(t))P_0(t)dt, \quad t \in [0,T], \\ P_0(0) = p_0 > 0, \end{cases} \tag{2.2}$$

where $r(t,i) \geq 0,\ i = 1,2,\ldots,l$, are given as the interest rate processes corresponding to different market modes. The other d assets are stocks whose price processes $P_m(t)$, $m = 1,2,\ldots,d$, satisfy the system of SDEs

$$\begin{cases} dP_m(t) = P_m(t)\left\{ b_m(t,\alpha(t))dt + \sum_{n=1}^{d} \sigma_{mn}(t,\alpha(t))dW_n(t) \right\}, \quad t \in [0,T], \\ P_m(0) = p_m > 0, \end{cases} \tag{2.3}$$

where for each $i = 1,2,\ldots,l$, $b_m(t,i)$ is the appreciation rate process and $\sigma_m(t,i) := (\sigma_{m1}(t,i),\ldots,\sigma_{md}(t,i))$ is the volatility or the dispersion rate process of the mth stock, corresponding to $\alpha(t) = i$.

Define the volatility matrix

$$\sigma(t,i) := \begin{pmatrix} \sigma_1(t,i) \\ \vdots \\ \sigma_d(t,i) \end{pmatrix} \equiv (\sigma_{mn}(t,i))_{d\times d} \quad \text{for each } i = 1,\ldots,l. \tag{2.4}$$

We assume throughout this paper that the nondegeneracy condition

$$\sigma(t,i)\sigma(t,i)' \geq \delta I \quad \forall t \in [0,T] \ \text{ and } \ i = 1,2,\ldots,l \tag{2.5}$$

is satisfied for some $\delta > 0$. We also assume that all the functions $r(t,i), b_m(t,i), \sigma_{mn}(t,i)$ are measurable and uniformly bounded in t.

Suppose that the initial market mode $\alpha(0) = i_0$. Consider an agent with an initial wealth $x_0 > 0$. These initial conditions are fixed throughout the paper. Denote by $x(t)$ the total wealth of the agent at time $t \geq 0$. Assuming that the trading of

shares takes place continuously and that transaction cost and consumptions are not considered, then one has (see, e.g., [29, p. 57])

$$\begin{cases} dx(t) &= \left\{ r(t,\alpha(t))x(t) + \sum_{m=1}^{d} \big[b_m(t,\alpha(t)) - r(t,\alpha(t))\big] u_m(t) \right\} dt \\ &\quad + \sum_{n=1}^{d}\sum_{m=1}^{d} \sigma_{mn}(t,\alpha(t)) u_m(t) dW_n(t), \\ x(0) &= x_0 > 0, \quad \alpha(0) = i_0, \end{cases} \tag{2.6}$$

where $u_m(t)$ is the total market value of the agent's wealth in the mth asset, $m = 0, 1, \ldots, d$, at time t. We call $u(\cdot) = (u_1(\cdot), \ldots, u_d(\cdot))'$ a *portfolio* of the agent. Note that once $u(\cdot)$ is determined, $u_0(\cdot)$, the asset in the bank account is completely specified since $u_0(t) = x(t) - \sum_{i=1}^{d} u_i(t)$. Thus, in our analysis to follow, only $u(\cdot)$ is considered.

Setting

$$B(t,i) := (b_1(t,i) - r(t,i), \ldots, b_d(t,i) - r(t,i)), \ i = 1, 2, \ldots, l, \tag{2.7}$$

we can rewrite the wealth equation (2.6) as

$$\begin{cases} dx(t) = [r(t,\alpha(t))x(t) + B(t,\alpha(t))u(t)]dt + u(t)'\sigma(t,\alpha(t))dW(t), \\ x(0) = x_0, \quad \alpha(0) = i_0. \end{cases} \tag{2.8}$$

Definition 2.1. *A portfolio $u(\cdot)$ is said to be admissible if $u(\cdot) \in L^2_{\mathcal{F}}(0,T;\mathbb{R}^d)$ and the SDE* (2.8) *has a unique solution $x(\cdot)$ corresponding to $u(\cdot)$. In this case, we refer to $(x(\cdot), u(\cdot))$ as an* admissible (wealth, portfolio) pair.

Remark 1. Most works in the literature define a portfolio, say, $\pi(\cdot)$, as the fractions of wealth allocated to different stocks. That is,

$$\pi(t) = \frac{u(t)}{x(t)}, \quad t \in [0,T]. \tag{2.9}$$

With this definition, (2.8) can be rewritten as

$$\begin{cases} dx(t) = x(t)[r(t,\alpha(t)) + B(t,\alpha(t))\pi(t)]dt + x(t)\pi(t)'\sigma(t,\alpha(t))dW(t), \\ x(0) = x_0, \quad \alpha(0) = i_0. \end{cases} \tag{2.10}$$

It is well known that this equation has a solution that can be expressed explicitly as an exponential of certain process, which therefore must be *automatically* positive if the initial wealth x_0 is positive. The reason for this guaranteed positivity of wealth is because the very definition of the portfolio, (2.9), has implicitly assumed that $x(t) \neq 0$; hence $x = 0$ becomes a natural barrier of the wealth process. It is our view, however, that a wealth process with possible zero or negative values is theoretically and practically sensible at least for some circumstances. Hence the nonnegativity of the wealth is better imposed as an *additional constraint*, rather than as a built-in feature, of the model. In our formulation, a portfolio is well defined even if the wealth is zero or negative, and the nonnegativity of the wealth, if so required, would be a constraint.

The agent's objective is to find an admissible portfolio $u(\cdot)$, among all the admissible portfolios whose expected terminal wealth is $Ex(T) = z$ for some given $z \in \mathbb{R}^1$, so that the risk measured by the variance of the terminal wealth

$$\text{Var } x(T) \equiv E[x(T) - Ex(T)]^2 = E[x(T) - z]^2 \tag{2.11}$$

is minimized. Finding such a portfolio $u(\cdot)$ is referred to as the *mean-variance portfolio selection problem.* Specifically, we have the following formulation.

DEFINITION 2.2. *The mean-variance portfolio selection is a constrained stochastic optimization problem, parameterized by* $z \in \mathbb{R}^1$*:*

$$\left\{ \begin{array}{ll} \textit{minimize} & J_{\mathrm{MV}}(x_0, i_0, u(\cdot)) := E[x(T) - z]^2, \\ \textit{subject to} & \left\{ \begin{array}{l} Ex(T) = z, \\ (x(\cdot), u(\cdot)) \quad \textit{admissible.} \end{array} \right. \end{array} \right. \tag{2.12}$$

Moreover, the problem is called feasible *if there is at least one portfolio satisfying all the constraints. The problem is called* finite *if it is feasible and the infimum of* $J_{\mathrm{MV}}(x_0, i_0, u(\cdot))$ *is finite. Finally, an optimal portfolio to the above problem, if it ever exists, is called an* efficient portfolio *corresponding to* z*, and the corresponding* $(\mathrm{Var}\ x(T), z) \in \mathbb{R}^2$ *and* $(\sigma_{x(T)}, z) \in \mathbb{R}^2$ *are interchangeably called an* efficient point*, where* $\sigma_{x(T)}$ *denotes the standard deviation of* $x(T)$*. The set of all the efficient points is called the* efficient frontier*.*

Remark 2. While in the above definition, an efficient portfolio is broadly defined for any given $z \in \mathbb{R}^1$; in the subsequent context we will see that it is practically sensible only for z greater than or equal to certain value. Also, the shape of the efficient frontier depends on whether it is plotted in the mean-variance plane or mean-standard deviation plane. In what follows, we will specify which one we are referring to only when ambiguity might arise.

Remark 3. The mean-variance portfolio selection problem may be defined in some different, albeit equivalent, ways. For example, in [31] the problem is formulated as a multiobjective optimization problem. It should be noted that the model in this paper is a faithful replication in form of the original Markowitz single-period model.

3. Feasibility. Since the problem (2.12) involves a terminal constraint $Ex(T) = z$, in this section, we derive conditions under which the problem is at least feasible. First, the following generalized Itô lemma [2] for Markov-modulated processes is useful.

LEMMA 3.1. *Given an* n*-dimensional process* $x(\cdot)$ *satisfying*

$$dx(t) = b(t, x(t), \alpha(t))dt + \sigma(t, x(t), \alpha(t))dW(t)$$

and a number of functions $\varphi(\cdot, \cdot, i) \in C^2([0, T] \times \mathbb{R}^n)$*,* $i = 1, 2, \ldots, l$*, we have*

$$d\varphi(t, x(t), \alpha(t)) = \Gamma\varphi(t, x(t), \alpha(t))dt + \varphi_x(t, x(t), \alpha(t))'\sigma(t, x(t), \alpha(t))dW(t),$$

where

$$\begin{aligned} \Gamma\varphi(t, x, i) := \quad & \varphi_t(t, x, i) + \varphi_x(t, x, i)'b(t, x, i) \\ & + \frac{1}{2}\mathrm{tr}[\sigma(t, x, i)'\varphi_{xx}(t, x, i)\sigma(t, x, i)] + \sum_{j=1}^{l} q_{ij}\varphi(t, x, j). \end{aligned}$$

Consider a portfolio $u^0(t) \equiv 0$, corresponding to the one that puts all the money in the bank account. The associated wealth process $x^0(\cdot)$ satisfies

$$\left\{ \begin{array}{l} dx^0(t) = r(t, \alpha(t))x^0(t)dt, \\ x^0(0) = x_0, \quad \alpha(0) = i_0, \end{array} \right. \tag{3.1}$$

with its expected terminal wealth

$$z^0 := Ex^0(T) = Ee^{\int_0^T r(s, \alpha(s))ds}x_0. \tag{3.2}$$

LEMMA 3.2. *Let $\psi(\cdot,i)$, $i=1,2,\dots,l$, be the solutions to the following system of linear ODEs:*

$$\begin{cases} \dot{\psi}(t,i) = -r(t,i)\psi(t,i) - \sum_{j=1}^{l} q_{ij}\psi(t,j), \\ \psi(T,i) = 1, \qquad i=1,2,\dots,l. \end{cases} \tag{3.3}$$

Then the mean-variance problem (2.12) *is feasible for every $z \in \mathbb{R}^1$ if and only if*

$$\gamma := E\int_0^T |\psi(t,\alpha(t))B(t,\alpha(t))|^2 dt > 0. \tag{3.4}$$

Proof. To prove the "if" part, construct a family of admissible portfolios $u^\beta(\cdot) = \beta u(\cdot)$ for $\beta \in \mathbb{R}^1$, where

$$u(t) = B(t,\alpha(t))'\psi(t,\alpha(t)). \tag{3.5}$$

Let $x^\beta(\cdot)$ be the wealth process corresponding to $u^\beta(\cdot)$. By linearity of the wealth equation, we have $x^\beta(t) = x^0(t) + \beta y(t)$, where $x^0(\cdot)$ satisfies (3.1) and $y(\cdot)$ is the solution to the following equation:

$$\begin{cases} dy(t) = [r(t,\alpha(t))y(t) + B(t,\alpha(t))u(t)]dt + u(t)'\sigma(t,\alpha(t))dW(t), \\ y(0) = 0, \quad \alpha(0) = i_0. \end{cases} \tag{3.6}$$

Therefore, problem (2.12) is feasible for every $z \in \mathbb{R}^1$ if there exists $\beta \in \mathbb{R}^1$ such that $z = Ex^\beta(T) \equiv Ex^0(T) + \beta Ey(T)$. Equivalently, (2.12) is feasible for every $z \in \mathbb{R}^1$ if $Ey(T) \neq 0$. However, applying the generalized Itô formula (Lemma 3.1) to $\varphi(t,x,i) = \psi(t,i)x$, we have

$$\begin{aligned} & d[\psi(t,\alpha(t))y(t)] \\ &= \Bigg\{ \psi(t,\alpha(t))[r(t,\alpha(t))y(t) + B(t,\alpha(t))u(t)]dt - r(t,\alpha(t))\psi(t,\alpha(t))y(t) \\ &\qquad - \sum_{j=1}^{l} q_{\alpha(t)j}\psi(t,j)y(t) \Bigg\} dt + \sum_{j=1}^{l} q_{\alpha(t)j}\psi(t,j)y(t)dt + \{\cdots\}dW(t) \\ &= \psi(t,\alpha(t))B(t,\alpha(t))u(t)dt + \{\cdots\}dW(t). \end{aligned}$$

Integrating from 0 to T, taking expectation, and using (3.5), we obtain

$$Ey(T) = E\int_0^T \psi(t,\alpha(t))B(t,\alpha(t))u(t)dt = E\int_0^T |\psi(t,\alpha(t))B(t,\alpha(t))|^2 dt. \tag{3.7}$$

Consequently, $Ey(T) \neq 0$ if (3.4) holds.

Conversely, suppose that problem (2.12) is feasible for every $z \in \mathbb{R}^1$. Then for each $z \in \mathbb{R}^1$, there is an admissible portfolio $u(\cdot)$ so that $Ex(T) = z$. However, we can always decompose $x(t) = x^0(t) + y(t)$, where $y(\cdot)$ satisfies (3.6). This leads to $Ex^0(T) + Ey(T) = z$. However, $Ex^0(T) \equiv z_0$ is independent of $u(\cdot)$; thus it is necessary that there is a $u(\cdot)$ with $Ey(T) \neq 0$. It follows then from (3.7) that (3.4) is valid. ∎

THEOREM 3.3. *The mean-variance problem* (2.12) *is feasible for every* $z \in \mathbb{R}^1$ *if and only if*

$$E\int_0^T |B(t,\alpha(t))|^2 dt > 0. \tag{3.8}$$

Proof. By virtue of Lemma 3.2, it suffices to prove that $\psi(t,i) > 0$ for all $t \in [0,T]$, $i = 1,2,\dots,l$. To this end, note that (3.3) can be rewritten as

$$\begin{cases} \dot{\psi}(t,i) = [-r(t,i) - q_{ii}]\,\psi(t,i) - \displaystyle\sum_{j\neq i}^{l} q_{ij}\psi(t,j), \\ \psi(T,i) = 1, \quad i = 1,2,\dots,l. \end{cases} \tag{3.9}$$

Treating this as a system of terminal-valued ODEs, a variation-of-constant formula yields

(3.10)

$$\psi(t,i) = e^{-\int_t^T [-r(s,i)-q_{ii}]ds} + \int_t^T e^{-\int_t^s [-r(\tau,i)-q_{ii}]d\tau} \sum_{j\neq i}^{l} q_{ij}\psi(s,j)ds,\ i = 1,2,\dots,l.$$

Construct a sequence $\{\psi^{(k)}(\cdot,i)\}$ (known as the Picard sequence) as follows:

$$\begin{aligned} &\psi^{(0)}(t,i) = 1,\ t \in [0,T],\ i = 1,2,\dots,l, \\ &\psi^{(k+1)}(t,i) = e^{-\int_t^T [-r(s,i)-q_{ii}]ds} + \int_t^T e^{-\int_t^s [-r(\tau,i)-q_{ii}]d\tau} \sum_{j\neq i}^{l} q_{ij}\psi^{(k)}(s,j)ds,\ t \in [0,T], \\ &\qquad\qquad i = 1,2,\dots,l,\ k = 0,1,\dots. \end{aligned}$$

Noting that $q_{ij} \geq 0$ for all $j \neq i$, we have

$$\psi^{(k)}(t,i) \geq e^{-\int_t^T [-r(s,i)-q_{ii}]ds} > 0,\ k = 0,1,\dots.$$

On the other hand, it is well known that $\psi(t,i)$ is the limit of the Picard sequence $\{\psi^{(k)}(t,i)\}$ as $k \to \infty$. Thus $\psi(t,i) > 0$. This proves the desired result. □

COROLLARY 3.4. *If* (3.8) *holds, then for any* $z \in \mathbb{R}^1$, *an admissible portfolio that satisfies* $Ex(T) = z$ *is given by*

$$u(t) = \frac{z - z^0}{\gamma} B(t,\alpha(t))'\psi(t,\alpha(t)), \tag{3.11}$$

where z^0 *and* γ *are given by* (3.2) *and* (3.4), *respectively.*

Proof. This is immediate from the proof of the "if" part of Lemma 3.2. □

COROLLARY 3.5. *If* $E\int_0^T |B(t,\alpha(t))|^2 dt = 0$, *then any admissible portfolio* $u(\cdot)$ *results in* $Ex(T) = z^0$.

Proof. This is seen from the proof of the "only if" part of Lemma 3.2. □

Remark 4. The condition (3.8) is very mild. For example, (3.8) holds as long as there is one stock whose appreciation-rate process is different from the interest-rate process at any market mode, which is obviously a practically reasonable assumption. On the other hand, if (3.8) fails, then Corollary 3.5 implies that the mean-variance problem (2.12) is feasible only if $z = z^0$. This is a pathological and trivial case that

does not warrant further consideration. Therefore, from this point on we shall assume that (3.8) holds or, equivalently, that the mean-variance problem (2.12) is feasible for any z.

Having addressed the issue of feasibility, we proceed with the study of optimality. The mean-variance problem (2.12) under consideration is a dynamic optimization problem with a constraint $Ex(T) = z$. To handle this constraint, we apply the Lagrange multiplier technique. Define

$$\begin{aligned} J(x_0, i_0, u(\cdot), \lambda) :=& \quad E\big\{|x(T) - z|^2 + 2\lambda[x(T) - z]\big\} \\ =& \quad E[x(T) + \lambda - z]^2 - \lambda^2, \qquad \lambda \in \mathbb{R}^1. \end{aligned} \tag{3.12}$$

Our first goal is to solve the following unconstrained problem parameterized by the Lagrange multiplier λ:

$$\begin{cases} \text{minimize} & J(x_0, i_0, u(\cdot), \lambda) = E\big[x(T) + \lambda - z\big]^2 - \lambda^2 \\ \text{subject to} & (x(\cdot), u(\cdot)) \ \text{ admissible.} \end{cases} \tag{3.13}$$

This turns out to be a Markov-modulated stochastic LQ optimal control problem, which will be solved in the next section.

4. Solution to the unconstrained problem. In this section we solve the unconstrained problem (3.13). First define

$$\rho(t,i) := B(t,i)[\sigma(t,i)\sigma(t,i)']^{-1}B(t,i)', \ i = 1, 2, \ldots, l. \tag{4.1}$$

Consider the following two systems of ODEs:

$$\begin{cases} \dot{P}(t,i) = [\rho(t,i) - 2r(t,i)]\, P(t,i) - \displaystyle\sum_{j=1}^{l} q_{ij} P(t,j), \\ P(T,i) = 1, \quad i = 1, 2, \ldots, l, \end{cases} \tag{4.2}$$

and

$$\begin{cases} \dot{H}(t,i) = r(t,i)H(t,i) - \dfrac{1}{P(t,i)} \displaystyle\sum_{j=1}^{l} q_{ij} P(t,j)[H(t,j) - H(t,i)], \\ H(T,i) = 1, \quad i = 1, 2, \ldots, l. \end{cases} \tag{4.3}$$

The existence and uniqueness of solutions to the above two systems of equations are evident as both are linear with uniformly bounded coefficients.

PROPOSITION 4.1. *The solutions of* (4.2) *and* (4.3) *must satisfy* $P(t,i) > 0$ *and* $0 < H(t,i) \le 1$ *for all* $t \in [0,T]$, $i = 1, 2, \ldots, l$. *Moreover, if for a fixed* i, $r(t,i) > 0$ *a.e.* $t \in [0,T]$, *then* $H(t,i) < 1$ *for all* $t \in [0,T)$.

Proof. The assertion $P(t,i) > 0$ can be proved in exactly the same way as that of $\psi(t,i) > 0$; see the proof of Theorem 3.3. Having proved the positivity of $P(t,i)$, one can then show $H(t,i) > 0$ using the same argument because now $P(t,j)/P(t,i) > 0$.

To prove that $H(t,i) \le 1$, first note that the system of ODEs

$$\begin{cases} \dfrac{d}{dt}\widetilde{H}(t,i) = -\dfrac{1}{P(t,i)} \displaystyle\sum_{j=1}^{l} q_{ij} P(t,j)[\widetilde{H}(t,j) - \widetilde{H}(t,i)], \\ \widetilde{H}(T,i) = 1, \quad i = 1, 2, \ldots, l, \end{cases} \tag{4.4}$$

has the only solutions $\widetilde{H}(t,i) \equiv 1$, $i = 1, 2, \ldots, l$, due to the uniqueness of solutions. Set

$$\widehat{H}(t,i) := \widetilde{H}(t,i) - H(t,i) \equiv 1 - H(t,i),$$

which solves the following equations:

(4.5)
$$\begin{cases} \dfrac{d}{dt}\widehat{H}(t,i) &= r(t,i)\widehat{H}(t,i) - r(t,i) - \dfrac{1}{P(t,i)}\displaystyle\sum_{j=1}^{l} q_{ij}P(t,j)[\widehat{H}(t,j) - \widehat{H}(t,i)] \\ &= \left[r(t,i) + \displaystyle\sum_{j\neq i} q_{ij}\right]\widehat{H}(t,i) - r(t,i) - \dfrac{1}{P(t,i)}\displaystyle\sum_{j\neq i} q_{ij}P(t,j)\widehat{H}(t,j), \\ \widehat{H}(T,i) &= 0, \quad i = 1, 2, \ldots, l. \end{cases}$$

A variation-of-constant formula leads to

$$\widehat{H}(t,i) = \int_t^T e^{-\int_t^s [r(\tau,i) + \sum_{j\neq i} q_{ij}]d\tau} \left[r(s,i) + \frac{1}{P(s,i)}\sum_{j\neq i} q_{ij}P(s,j)\widehat{H}(s,j)\right]. \tag{4.6}$$

A similar trick using the construction of Picard's sequence yields that $\widehat{H}(t,i) \geq 0$. In addition, $\widehat{H}(t,i) > 0$ for all $t \in [0,T)$ if $r(t,i) > 0$ a.e. $t \in [0,T]$. The desired result then follows from the fact that $\widehat{H}(t,i) = 1 - H(t,i)$. □

Remark 5. Equation (4.2) is a Riccati-type equation that arises naturally in studying the stochastic LQ control problem (3.13), whereas (4.3) is used to handle the nonhomogeneous terms involved in (3.13); see the proof of Theorem 4.2 below. On the other hand, $H(t,i)$ has a financial interpretation: for fixed (t,i), $H(t,i)$ is a *deterministic* quantity representing the *risk-adjusted discount factor* at time t when the market mode is i (note that the interest rate itself is random); see also Remark 11 in what follows.

THEOREM 4.2. *Problem* (3.13) *has an optimal feedback control*

$$u^*(t,x,i) = -[\sigma(t,i)\sigma(t,i)']^{-1}B(t,i)'[x + (\lambda - z)H(t,i)]. \tag{4.7}$$

Moreover, the corresponding optimal value is

$$\begin{aligned} &\inf_{u(\cdot)\text{admissible}} J(x_0, i_0, u(\cdot), \lambda) \\ &\quad = [P(0,i_0)H(0,i_0)^2 + \theta - 1](\lambda - z)^2 \\ &\qquad +2[P(0,i_0)H(0,i_0)x_0 - z](\lambda - z) + P(0,i_0)x_0^2 - z^2, \end{aligned} \tag{4.8}$$

where

$$\begin{aligned} \theta &:= E\int_0^T \sum_{j=1}^{l} q_{\alpha(t)j}P(t,j)[H(t,j) - H(t,\alpha(t))]^2 dt \\ &= \sum_{i=1}^{l}\sum_{j=1}^{l}\int_0^T P(t,j)p_{i_0 i}(t)q_{ij}[H(t,j) - H(t,i)]^2 dt \geq 0, \end{aligned} \tag{4.9}$$

with the transition probabilities $p_{i_0 i}(t)$ *given by* (2.1).

Proof. Let $u(\cdot)$ be any admissible control and $x(\cdot)$ be the corresponding state trajectory of (2.8). Applying the generalized Itô formula (Lemma 3.1) to

$$\varphi(t,x,i) = P(t,i)[x+(\lambda-z)H(t,i)]^2,$$

we obtain

(4.10)

$$\begin{aligned}
&d\Big\{P(t,\alpha(t))[x(t)+(\lambda-z)H(t,\alpha(t))]^2\Big\}\\
&\quad= P(t,\alpha(t))\Big\{u(t)'[\sigma(t,\alpha(t))\sigma(t,\alpha(t))']u(t)+2u(t)'B(t,\alpha(t))'[x(t)+(\lambda-z)H(t,\alpha(t))]\\
&\qquad+2r(t,\alpha(t))[x(t)+(\lambda-z)H(t,\alpha(t))]^2\Big\}dt\\
&\qquad-2(\lambda-z)[x(t)+(\lambda-z)H(t,\alpha(t))]\sum_{j=1}^{l}q_{\alpha(t)j}P(t,j)[H(t,j)-H(t,\alpha(t))]dt\\
&\qquad+[\rho(t,\alpha(t))-2r(t,\alpha(t))]P(t,\alpha(t))[x(t)+(\lambda-z)H(t,\alpha(t))]^2dt\\
&\qquad-\sum_{j=1}^{l}q_{\alpha(t)j}P(t,j)[x(t)+(\lambda-z)H(t,\alpha(t))]^2dt\\
&\qquad+\sum_{j=1}^{l}q_{\alpha(t)j}P(t,j)[x(t)+(\lambda-z)H(t,j)]^2dt+\{\cdots\}dW(t)\\
&\quad= P(t,\alpha(t))\Big\{u(t)'[\sigma(t,\alpha(t))\sigma(t,\alpha(t))']u(t)+2u(t)'B(t,\alpha(t))'[x(t)+(\lambda-z)H(t,\alpha(t))]\\
&\qquad+\rho(t,\alpha(t))[x(t)+(\lambda-z)H(t,\alpha(t))]^2\Big\}dt\\
&\qquad+(\lambda-z)^2\sum_{j=1}^{l}q_{\alpha(t)j}P(t,j)[H(t,j)-H(t,\alpha(t))]^2dt+\{\cdots\}dW(t)\\
&\quad= P(t,\alpha(t))[u(t)-u^*(t,x(t),\alpha(t))]'[\sigma(t,\alpha(t))\sigma(t,\alpha(t))'][u(t)-u^*(t,x(t),\alpha(t))]dt\\
&\qquad+(\lambda-z)^2\sum_{j=1}^{l}q_{\alpha(t)j}P(t,j)[H(t,j)-H(t,\alpha(t))]^2dt+\{\cdots\}dW(t),
\end{aligned}$$

where $u^*(t,x,i)$ is defined as the right-hand side of (4.7). Integrating the above from 0 to T and taking expectations, we obtain

$$\begin{aligned}
&E[x(T)+\lambda-z]^2\\
&\quad= P(0,i_0)[x_0+(\lambda-z)H(0,i_0)]^2+\theta(\lambda-z)^2\\
&\qquad+E\int_0^T P(t,\alpha(t))[u(t)-u^*(t,x(t),\alpha(t))]'[\sigma(t,\alpha(t))\sigma(t,\alpha(t))']\\
&\qquad\quad\times[u(t)-u^*(t,x(t),\alpha(t))]dt.
\end{aligned}\tag{4.11}$$

Consequently,

$$\begin{aligned}
&J(x_0,i_0,u(\cdot),\lambda)\\
&\quad= E[x(T)+\lambda-z]^2-\lambda^2\\
&\quad= [P(0,i_0)H(0,i_0)^2+\theta-1](\lambda-z)^2+2[P(0,i_0)H(0,i_0)x_0-z](\lambda-z)\\
&\qquad+P(0,i_0)x_0^2-z^2\\
&\qquad+E\int_0^T P(t,\alpha(t))[u(t)-u^*(t,x(t),\alpha(t))]'[\sigma(t,\alpha(t))\sigma(t,\alpha(t))']\\
&\qquad\qquad\cdot[u(t)-u^*(t,x(t),\alpha(t))]dt.
\end{aligned}\tag{4.12}$$

Since $P(t,\alpha(t)) > 0$ by Proposition 3.1, it follows immediately that the optimal feedback control is given by (4.7) and the optimal value is given by (4.8), provided that the corresponding equation (2.8) under the feedback control (4.7) has a solution. However, under (4.7), the system (2.8) is a nonhomogeneous linear SDE with coefficients modulated by $\alpha(t)$. Since all the coefficients of this linear equation are uniformly bounded and $\alpha(t)$ is independent of $W(t)$, the existence and uniqueness of the solution to the equation are straightforward based on a standard successive approximation scheme.

Finally, since

$$\theta = \sum_{i \neq j} \int_0^T P(t,j) p_{i_0 i}(t) q_{ij} [H(t,j) - H(t,i)]^2 dt$$

and $q_{ij} \geq 0$ for all $i \neq j$, we must have $\theta \geq 0$. This completes the proof. □

5. Efficient frontier. In this section we proceed to derive the efficient frontier for the original mean-variance problem (2.12).

THEOREM 5.1 (efficient portfolios and efficient frontier). *Assume that* (3.8) *holds. Then we have*

$$P(0,i_0)H(0,i_0)^2 + \theta - 1 < 0. \tag{5.1}$$

Moreover, the efficient portfolio corresponding to z, as a function of the time t, the wealth level x, and the market mode i, is

$$u^*(t,x,i) = -[\sigma(t,i)\sigma(t,i)']^{-1} B(t,i)'[x + (\lambda^* - z)H(t,i)], \tag{5.2}$$

where

$$\lambda^* - z = \frac{z - P(0,i_0)H(0,i_0)x_0}{P(0,i_0)H(0,i_0)^2 + \theta - 1}. \tag{5.3}$$

Furthermore, the optimal value of Var $x(T)$*, among all the wealth processes $x(\cdot)$ satisfying $Ex(T) = z$, is*

$$\begin{aligned} &\text{Var } x^*(T) \\ &\quad = \frac{P(0,i_0)H(0,i_0)^2 + \theta}{1 - \theta - P(0,i_0)H(0,i_0)^2} \left[z - \frac{P(0,i_0)H(0,i_0)}{P(0,i_0)H(0,i_0)^2 + \theta} x_0 \right]^2 \\ &\qquad + \frac{P(0,i_0)\theta}{P(0,i_0)H(0,i_0)^2 + \theta} x_0^2. \end{aligned} \tag{5.4}$$

Proof. By assumption (3.8) and Theorem 3.3, the mean-variance problem (2.12) is feasible for any $z \in \mathbb{R}^1$. Moreover, using exactly the same approach as in the proof of Theorem 4.2, one can show that problem (2.12) *without* the constraint $Ex(T) = z$ must have a finite optimal value; hence so does the problem (2.12). Therefore, (2.12) is finite for any $z \in \mathbb{R}^1$. Since $J_{\text{MV}}(x_0, i_0, \pi(\cdot))$ is strictly convex in $u(\cdot)$ and the constraint function $Ex(T) - z$ is affine in $u(\cdot)$, we can apply the well-known duality theorem (see, e.g., [17, p. 224, Theorem 1][1]) to conclude that for any $z \in \mathbb{R}^1$, the

[1]To be precise, one should apply [17, p. 236, Problem 7] together with the proof of [17, p. 224, Theorem 1] in our case, as there is an *equality* constraint, $Ex(T) = z$, in (2.12). To be able to use the result there, one needs to check a condition posed in [17, p. 236, Problem 7]; namely, 0 is an interior point of the set $\mathcal{T} := \{Ex(T) - z | x(\cdot)$ is the wealth process of an admissible portfolio $u(\cdot)\}$. In the present case this condition is implied by Theorem 3.3, which essentially yields that $\mathcal{T} = \mathbb{R}^1$.

optimal value of (2.12) is

$$J^*_{\mathrm{MV}}(x_0,i_0)=\sup_{\lambda\in\mathbb{R}^1}\ \inf_{u(\cdot)\ \text{admissible}} J(x_0,i_0,u(\cdot),\lambda)>-\infty. \tag{5.5}$$

By Theorem 4.2, $\inf_{u(\cdot)\ \text{admissible}} J(x_0,i_0,u(\cdot),\lambda)$ is a quadratic function (4.8) in $\lambda-z$. It follows from the finiteness of the supremum value of this quadratic function (see (5.5)) that

$$P(0,i_0)H(0,i_0)^2+\theta-1\le 0.$$

Now, if

$$P(0,i_0)H(0,i_0)^2+\theta-1=0,$$

then again by Theorem 4.2 and (5.5) we must have

$$P(0,i_0)H(0,i_0)x_0-z=0$$

for every $z\in\mathbb{R}^1$, which is a contradiction. This proves (5.1). On the other hand, in view of (5.5), we maximize the quadratic function (4.8) over $\lambda-z$ and conclude that the maximizer is given by (5.3), whereas the maximum value is given by the right-hand side of (5.4). Finally, the optimal control (5.2) is obtained by (4.7) with $\lambda=\lambda^*$. □

The efficient frontier (5.4) reveals explicitly the tradeoff between the mean (return) and variance (risk) at the terminal. Quite contrary to the case without Markovian jumps [31], the efficient frontier in the present case is no longer a perfect square (or, equivalently, the efficient frontier in the mean-standard deviation diagram is no longer a straight line). As a consequence, one is not able to achieve a risk-free investment. This, certainly, is expected since now the interest rate process is modulated by the Markov chain, and the interest rate risk cannot be perfectly hedged by any portfolio consisting of the bank account and stocks (as with the case studied in [16]) because the Markov chain is independent of the Brownian motion.

Nevertheless, the expression (5.4) does disclose the *minimum variance*, namely, the minimum possible terminal variance achievable by an admissible portfolio, along with the portfolio that attains this minimum variance.

THEOREM 5.2 (minimum variance). *The minimum terminal variance is*

$$\mathrm{Var}\ x^*_{\min}(T)=\frac{P(0,i_0)\theta}{P(0,i_0)H(0,i_0)^2+\theta}x_0^2\ge 0 \tag{5.6}$$

with the corresponding expected terminal wealth

$$z_{\min}:=\frac{P(0,i_0)H(0,i_0)}{P(0,i_0)H(0,i_0)^2+\theta}x_0 \tag{5.7}$$

and the corresponding Lagrange multiplier $\lambda^*_{\min}=0$. *Moreover, the portfolio that achieves the above minimum variance, as a function of the time* t, *the wealth level* x, *and the market mode* i, *is*

$$u^*_{\min}(t,x,i)=-[\sigma(t,i)\sigma(t,i)']^{-1}B(t,i)'[x-z_{\min}H(t,i)]. \tag{5.8}$$

Proof. The conclusions regarding (5.6) and (5.7) are evident in view of the efficient frontier (5.4). The assertion $\lambda^*_{\min}=0$ can be verified via (5.3) and (5.7). Finally, (5.8) follows from (5.2). □

Remark 6. As a consequence of the above theorem, the parameter z can be restricted to $z \geq z_{\min}$ when one defines the efficient frontier for the mean-variance problem (2.12).

THEOREM 5.3 (mutual fund theorem). *Suppose an efficient portfolio $u_1^*(\cdot)$ is given by* (5.2) *corresponding to $z = z_1 > z_{\min}$. Then a portfolio $u^*(\cdot)$ is efficient if and only if there is a $\mu \geq 0$ such that*

$$u^*(t) = (1-\mu)u^*_{\min}(t) + \mu u_1^*(t), \quad t \in [0,T], \tag{5.9}$$

where $u^*_{\min}(\cdot)$ is the minimum variance portfolio defined in Theorem 5.2.

Proof. We first prove the "if" part. Since both $u^*_{\min}(\cdot)$ and $u_1^*(\cdot)$ are efficient, by the explicit expression of any efficient portfolio given by (5.2), $u^*(t) = (1-\mu)u_0^*(\cdot) + \mu u_1^*(t)$ must be in the form of (5.2) corresponding to $z = (1-\mu)z_{\min} + \mu z_1$ (also noting that $x^*(\cdot)$ is linear in $u^*(\cdot)$). Hence $u^*(\cdot)$ must be efficient.

Conversely, suppose $u^*(\cdot)$ is efficient corresponding to a certain $z \geq z_{\min}$. Write $z = (1-\mu)z_{\min} + \mu z_1$ with some $\mu \geq 0$. Multiplying

$$u^*_{\min}(t) = -[\sigma(t,\alpha(t))\sigma(t,\alpha(t))']^{-1}B(t,\alpha(t))'[x^*_{\min}(t) - z_{\min}H(t,\alpha(t))]$$

by $(1-\mu)$, multiplying

$$u_1^*(t) = -[\sigma(t,\alpha(t))\sigma(t,\alpha(t))']^{-1}B(t,\alpha(t))'[x_1^*(t) + (\lambda_1^* - z_1)H(t,\alpha(t))]$$

by μ, and summing them up, we obtain that $(1-\mu)u^*_{\min}(t) + \mu u_1^*(t)$ is represented by (5.2) with $x^*(t) = (1-\mu)x^*_{\min}(t) + \mu x_1^*(t)$ and $z = (1-\mu)z_{\min} + \mu z_1$. This leads to (5.9). □

Remark 7. The above mutual fund theorem implies that any investor needs only to invest in the minimum variance portfolio and another prespecified efficient portfolio in order to achieve the efficiency. Note that in the case where all the market parameters are deterministic [31], the corresponding mutual fund theorem becomes the *one-fund theorem*, which yields that any efficient portfolio is a combination of the bank account and a given efficient risky portfolio (known as the *tangent fund*). This is equivalent to the fact that the fractions of wealth among the stocks are the same among all efficient portfolios. However, in the present Markov-modulated case, this feature is no longer available.

6. A special case: Interest rate unaffected by the Markov chain. In this section we consider a special case where the interest-rate process does not respond to the change in the market mode, namely, $r(t,i) = r(t)$ for any $i = 1,2,\ldots,l$, whereas the appreciation-rate and volatility-rate processes still do. This stems from the situations where substantial changes in the interest-rate process are much less frequent than those in the other processes. For example, the interest rate may typically change on a bimonthly, or even less often, basis, whereas the stock market mode may switch on a weekly, or more frequent, basis. It turns out that the results obtained in the previous sections can be substantially simplified in this case.

The key to the simplification is that when $r(t,i) = r(t)$, the only solutions to (4.3) are

$$H(t,i) = e^{-\int_t^T r(s)ds} \quad \forall i = 1,2,\ldots,l \tag{6.1}$$

due to the uniqueness of solutions to (4.3). It follows then from (4.9) that

$$\theta = 0. \tag{6.2}$$

As a result, Theorem 5.1 reduces to the following result.

THEOREM 6.1. *Assume that* (3.8) *holds and that* $r(t,i) = r(t)$ *for all* $i = 1,2,\ldots,l$. *Then we must have*

$$P(0,i_0) < e^{2\int_0^T r(s)ds}. \tag{6.3}$$

Moreover, the efficient portfolio corresponding to z, *as a function of the time* t, *the wealth level* x, *and the market mode* i, *is*

$$u^*(t,x,i) = -[\sigma(t,i)\sigma(t,i)']^{-1}B(t,i)'[x + (\lambda^* - z)e^{-\int_t^T r(s)ds}], \tag{6.4}$$

where

$$\lambda^* - z = \frac{z - P(0,i_0)e^{-\int_0^T r(s)ds}x_0}{P(0,i_0)e^{-2\int_0^T r(s)ds} - 1}. \tag{6.5}$$

Furthermore, the optimal value of $\operatorname{Var} x(T)$, *among all the wealth processes* $x(\cdot)$ *satisfying* $Ex(T) = z$, *is*

$$\operatorname{Var} x^*(T) = \frac{P(0,i_0)e^{-2\int_0^T r(s)ds}}{1 - P(0,i_0)e^{-2\int_0^T r(s)ds}}\left[z - e^{\int_0^T r(s)ds}x_0\right]^2. \tag{6.6}$$

Proof. This is straightforward by Theorem 5.1, together with (6.1) and (6.2). □

Remark 8. Note that in this case the efficient frontier involves a perfect square, even if the market parameters of the stocks are all random. The *capital market line* (see, e.g., [18]) in the mean-standard deviation diagram is

$$Ex^*(T) = e^{\int_0^T r(t)dt}x_0 + \sqrt{\frac{1 - P(0,i_0)e^{-2\int_0^T r(s)ds}}{P(0,i_0)e^{-2\int_0^T r(s)ds}}}\sigma_{x^*(T)}. \tag{6.7}$$

Therefore, the price of risk is given by

$$p = \sqrt{\frac{1 - P(0,i_0)e^{-2\int_0^T r(s)ds}}{P(0,i_0)e^{-2\int_0^T r(s)ds}}},$$

which depends only on the initial market mode i_0.

Remark 9. Clearly the minimum terminal variance in this case is zero, corresponding to putting all the money in the bank account. Moreover, $z_{\min} = e^{\int_0^T r(t)dt}x_0$. Consequently, the mutual fund theorem (Theorem 5.3) specifies that any efficient portfolio is a combination of the bank account and a given efficient portfolio. In other words, the one-fund theorem is valid in this case. In particular, the proportions of the stocks in all the efficient portfolios are the same under a particular market mode, irrespective of the wealth level and risk preference of the investors. This, in turn, will lead to the so-called market portfolio and *capital asset pricing model* (CAPM); see [18].

Remark 10. If we further assume that all the appreciation-rate and volatility-rate processes are independent of the market mode i, then $P(t,i) = e^{-\int_0^T[\rho(s)-2r(s)]ds}$ for each $i = 1,2,\ldots,l$ are the only solutions to (4.2). In this case, all the results reduce to those of [31].

Remark 11. We see from (6.1) that the functions $H(t,i)$, which are keys in our main results Theorems 5.1–5.3, are nothing else than a generalization of the discount factor between the present time to the terminal time under different market modes. Note that Proposition 4.1 stipulates that if the interest rate $r(t,i) > 0$ for a mode i, then the corresponding $H(t,i) < 1$, representing a genuine discount.

7. Concluding remarks. We have developed mean-variance optimal portfolio selection for a market with regime switching. The formulation allows the market to have random switching among a finite number of possible configurations that are modulated by a continuous-time Markov chain. Such a setup takes into consideration the discrete changes in a regime across which the behavior of the corresponding market could be markedly different. Our main effort has been devoted to obtaining efficient portfolios and an efficient frontier. It is interesting to note that for the Markov-modulated model, the efficient frontier is no longer a perfect square, except in the case when the interest rate is independent of the Markov chain.

There are several interesting problems that deserve further investigation. One is a model with nonnegativity constraints on the terminal wealth. As discussed earlier, this would render a stochastic LQ control problem with a sample-wise state constraint, which is a very challenging problem. Another problem is one with transaction costs. Although with the rapidly growing use of on-line trading, transaction costs nowadays represent a very small, if not at all negligible, portion of the total transacted values, the problem with transaction costs is theoretically interesting as it leads to a singular stochastic control problem whose solution would normally exhibit very different behavior than its no-transaction counterpart. In particular, with transaction costs optimal strategies would no longer be continuously trading strategies as opposed to the no-transaction case. In some sense, one motivation of introducing the transaction costs is to limit the changes in the optimal strategy. Indeed, Soner and Touzi [25] considered a market, in the absence of transaction costs, with the so-called gamma constraints in order to restrict the unbounded variation of the portfolios under consideration. Yet another problem is to remove the assumption that the Markov chain is independent of the underlying Brownian motion. Note that the mean-variance portfolio selection with the Brownian motion adapted random market coefficients has been completely solved in [16]. The model of this paper represents another "extreme" where the random coefficients are entirely independent of the Brownian motion. A more general model where the randomness in the coefficients is neither adapted to nor independent of the Brownian motion may be tackled by decomposing the problem into the two extremes that have been solved. On the other hand, the Markov chain describing the regime switching is assumed to be *completely* observable in this paper. A more realistic and theoretically interesting model is that the Markov chain is "hidden" and only partially observable through the stock prices. In this case, one needs to first perform filtering in order to estimate the state of the current regime before making efficient investment strategies. In [8], Elliott, Malcolm, and Tsoi developed schemes to estimate the appreciation rate, the volatility, and the generator of the underlying Markov chain. Estimates of the generator were also obtained via the stochastic approximation method in [28]. These estimation techniques may be used in conjunction with the portfolio selection approach presented in this work. Finally, a corresponding discrete-time model will be useful in the actual computing. In addition, to take into consideration that the Markov chain may have a large state space, an interesting problem is to reduce complexity via a singular perturbation approach.

Acknowledgments. We thank the editors and three reviewers for their helpful comments and suggestions on an earlier version of the paper.

REFERENCES

[1] G. Barone-Adesi and R. Whaley, *Efficient analytic approximation of American option values*, J. Fin., 42 (1987), pp. 301–320.

[2] T. Björk, *Finite dimensional optimal filters for a class of Itô-processes with jumping parameters*, Stochastics, 4 (1980/1981), pp. 167–183.
[3] J. Buffington and R. Elliott, *American options with regime switching*, Int. J. Theor. Appl. Finance, 5 (2002), pp. 497–514.
[4] S. Chen, X. Li, and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs*, SIAM J. Control Optim., 36 (1998), pp. 1685–1702.
[5] S. Chen and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs* II, SIAM J. Control Optim., 39 (2000), pp. 1065–1081.
[6] G. B. Di Masi, Y. M. Kabanov, and W. J. Runggaldier, *Mean variance hedging of options on stocks with Markov volatility*, Theory Probab. Appl., 39 (1994), pp. 173–181.
[7] D. Duffie and H. Richardson, *Mean-variance hedging in continuous time*, Ann. Appl. Probab., 1 (1991), pp. 1–15.
[8] R. J. Elliott, W. P. Malcolm, and A. Tsoi, *Robust parameter estimation for asset price models with Markov modulated volatilities*, J. Econom. Dynam. Control, 27 (2003), pp. 1391–1409.
[9] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[10] X. Guo, *Insider Information and Stock Fluctuations*, Ph.D. thesis, Rutgers University, New Brunswick, NJ, 1999.
[11] N. H. Hakansson, *Multi-period mean-variance analysis: Toward a general theory of portfolio choice*, J. Fin., 26 (1971), pp. 857–884.
[12] M. Kohlmann and X. Y. Zhou, *Relationship between backward stochastic differential equations and stochastic controls: A linear-quadratic approach*, SIAM J. Control Optim., 38 (2000), pp. 1392–1407.
[13] R. Korn, *Optimal Portfolios—Stochastic Models for Optimal Investment and Risk Management in Continuous Time*, World Scientific, Singapore, 1997.
[14] R. Korn and S. Trautmann, *Continuous-time portfolio optimization under terminal wealth constraints*, ZOR—Math. Methods Oper. Res., 42 (1995), pp. 69–92.
[15] X. Li, X. Y. Zhou, and A. E. B. Lim, *Dynamic mean-variance portfolio selection with no-shorting constraints*, SIAM J. Control Optim., 40 (2002), pp. 1540–1555.
[16] A. E. B. Lim and X. Y. Zhou, *Mean-variance portfolio selection with random coefficients in a complete market*, Math. Oper. Res., 27 (2002), pp. 101–120.
[17] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York, 1968.
[18] D. G. Luenberger, *Investment Science*, Oxford University Press, New York, 1998.
[19] D. Li and W. L. Ng, *Optimal dynamic portfolio selection: Multi-period mean-variance formulation*, Math. Finance, 10 (2000), pp. 387–406.
[20] H. Markowitz, *Portfolio selection*, J. Fin., 7 (1952), pp. 77–91.
[21] H. Markowitz, *Portfolio Selection: Efficient Diversification of Investment*, John Wiley and Sons, New York, 1959.
[22] S. R. Pliska, *Introduction to Mathematical Finance*, Basil Blackwell, Malden, UK, 1997.
[23] P. A. Samuelson, *Lifetime portfolio selection by dynamic stochastic programming*, Rev. Econ. Statist., 51 (1969), pp. 239–246.
[24] M. Schweizer, *Approximation pricing and the variance-optimal martingale measure*, Ann. Probab., 24 (1996), pp. 206–236.
[25] H. M. Soner and N. Touzi, *Superreplication under gamma constraints*, SIAM J. Control Optim., 39 (2000), pp. 73–96.
[26] D. D. Yao, Q. Zhang, and X. Y. Zhou, *Option Pricing with Markov-Modulated Volatility*, preprint, The Chinese University of Hong Kong, Hong Kong, 2001.
[27] G. Yin, R. H. Liu, and Q. Zhang, *Recursive algorithms for stock liquidation: A stochastic optimization approach*, SIAM J. Optim., 13 (2002), pp. 240–263.
[28] G. Yin, Q. Zhang, and K. Yin, *Constrained stochastic estimation algorithms for a class of hybrid stock market models*, J. Optim. Theory Appl., 118 (2003), pp. 157–182.
[29] J. Yong and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York, 1999.
[30] Q. Zhang, *Stock trading: An optimal selling rule*, SIAM J. Control Optim., 40 (2001), pp. 64–87.
[31] X. Y. Zhou and D. Li, *Continuous-time mean-variance portfolio selection: A stochastic LQ framework*, Appl. Math. Optim., 42 (2000), pp. 19–33.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1483–1508

THE OPTIMAL TIME CONTROL OF A PHASE-FIELD SYSTEM*

LIJUAN WANG† AND GENGSHENG WANG†

Abstract. In this paper, we prove the existence of the optimal time control of a phase-field system by using the Carleman inequality, and we establish the maximum principle for the optimal time control problem governed by the phase-field system.

Key words. optimal time control, phase-field system, Carleman inequality, maximum principle

AMS subject classifications. 93B50, 93C35, 93C05

DOI. 10.1137/S0363012902405455

1. Introduction.

We shall consider the controlled phase-field system

$$
\begin{cases}
u_t(x,t) + lh_t(x,t) - k\triangle u(x,t) = m(x)w(x,t) + f_1(x) \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{in } Q_\infty = \Omega\times(0,\infty), \\
h_t(x,t) - a\triangle h(x,t) - b(h(x,t) - h^3(x,t)) - cu(x,t) \\
\qquad\qquad\qquad\qquad\qquad = m(x)v(x,t) + f_2(x) \quad \text{in } Q_\infty, \\
u(x,0) = u_0(x), \quad h(x,0) = h_0(x) \quad \text{in } \Omega, \\
u(x,t) = \overline{u}(x), \quad h(x,t) = \overline{h}(x) \qquad \text{on } \textstyle\sum_\infty = \partial\Omega\times(0,\infty),
\end{cases} \tag{1.1}
$$

where Ω is an open, bounded, and smooth set in R^n (of class C^2, for instance), $n = 1, 2, 3$, m is the characteristic function of an open subset ω of Ω, a, b, c, l, k are positive constants, u is the reduced temperature, h is the phase function defining the liquid or the solid phase, and $u_0(x), h_0(x), \overline{u}(x), \overline{v}(x), f_1$, and f_2 are given functions. Throughout this paper we shall denote $L^2(\Omega)$ by H with the usual norm denoted by $|\cdot|_2$, and we denote $H_0^1(\Omega)$ by V with the usual norm denoted by $\|\cdot\|_1$. We set $H^{2,1}(Q) = \{y \in L^2(0,T;H^2(\Omega)); y_t \in L^2(Q)\}$, where $Q = \Omega\times(0,T)$, and $H^1(0,T;H) = \{y \in L^2(0,T;H); y_t \in L^2(0,T;H)\}$. By identifying H with its own dual, we have $V \subset H \subset V^*$. $\langle\cdot,\cdot\rangle$ denotes the scalar product of H and the paring between V and V^*.

We assume that

(H_1) $u_0 \in H^1(\Omega)$, $h_0 \in H^2(\Omega)$, f_1 and $f_2 \in H$.

It is well known that (cf. [17]) under assumption (H_1), for each $T > 0$ and $w, v \in L^2(Q)$, where $Q = \Omega\times(0,T)$, system (1.1) has a unique solution $(u^{w,v}, h^{w,v}) \in H^{2,1}(Q)\times H^{2,1}(Q)$. For simplicity, we denote it by (u,h) if there is no ambiguity.

Throughout this paper, we let $(u_e, h_e) \in H^2(\Omega)\times H^2(\Omega)$ be a steady-state (equilibrium) solution to system (1.1), i.e.,

$$
\begin{cases}
-k\triangle u_e(x) = f_1(x) & \text{in } \Omega, \\
-a\triangle h_e(x) - b(h_e(x) - h_e^3(x)) - cu_e(x) = f_2(x) & \text{in } \Omega, \\
u_e(x) = \overline{u}(x), \qquad h_e(x) = \overline{h}(x) & \text{on } \partial\Omega.
\end{cases} \tag{1.2}
$$

*Received by the editors April 10, 2002; accepted for publication (in revised form) April 25, 2003; published electronically October 28, 2003. This work was supported by National Natural Science Foundation of China grant 10071028, by The Grant of Key Laboratory-Optimal Control and Discrete Mathematics of Hubei Province, and by The New Century Excellent Teacher Plan of The Ministry of National Education of China.

http://www.siam.org/journals/sicon/42-4/40545.html

†Mathematics Department and The Center for Optimal Control and Discrete Mathematics, Huazhong Normal University, Wuhan 430079, People's Republic of China (hepeijie@public.wh.hb.cn, wanggs@ ccnu.edu.cn).

Let $\rho > 0$ and $U_\rho = \{(w(x,t), v(x,t)) \in (L^\infty(0,\infty;H))^2 : |w(t)|_2, |v(t)|_2 \le \rho \text{ a.e. } t > 0\}$. U_ρ will be considered as the control set. We shall study the following optimal time control problem:

(P) $\inf\{ T : u^{w,v}(T) = u_e, h^{w,v}(T) = h_e, (w,v) \in U_\rho \}$.

We call $T^* \equiv \inf\{ T : u^{w,v}(T) = u_e, h^{w,v}(T) = h_e, (w,v) \in U_\rho \}$ the minimal time. A pair of controls $(w^*, v^*) \in U_\rho$ such that $u^*(T^*) = u_e$ and $h^*(T^*) = h_e$, where (u^*, h^*) is the solution of (1.1) corresponding to (w^*, v^*), is called a pair of optimal time controls, and (u^*, h^*) is called a pair of optimal time states. A pair of controls $(w,v) \in U_\rho$ is called admissible if there exists a $T > 0$ such that $u^{w,v}(T) = u_e$ and $h^{w,v}(T) = h_e$.

Phase-field system (1.1) models the phase transition of a large class of physical processes including the melting and solidification. The Stefan problem as well as other classical mathematical models of phase transition are limiting cases of system (1.1) (cf. [4], [6], and [17]). Physically, steady-state means that the system does not exchange energy and material with surroundings and every part of the system does not exchange energy and material with each other. When interface between solid and liquid does not move unless the system is perturbed, we say that the system (1.1) reaches its steady-state. Problem (P) is to ask a pair of controls such that the corresponding temperature and phase, which start from given temperature u_0 and phase h_0, reach the temperature u_e and phase h_e of steady-state, respectively, in the shortest time.

In this paper we study the existence of a pair of optimal time controls for problem (P) and the maximum principle of problem (P). As we know, in order to get the existence of a pair of optimal time controls, one needs to obtain the existence of a pair of admissible controls. In general, one proves the existence of an admissible control of optimal time control problem governed by parabolic systems by considering the feedback controller and the corresponding closed loop system (cf. [2], [3], and [16]). However, this method is not applicable to the phase-field system since it is quite difficult to analyze the corresponding closed loop system. The novelty of this paper is that it uses the Carleman inequality to show the existence of a pair of admissible controls. The Carleman inequality has been widely used to obtain local controllability or null controllability of parabolic differential systems, including the Navier–Stokes equation, the Boussinesq system, and the phase-field system (cf. [11], [5], and [20]). However, it seems that no one used such an inequality to obtain the existence of admissible control for an optimal time control problem governed by parabolic differential systems. We believe that the Carleman inequality will be greatly helpful to get the existence of admissible control for optimal time control problems governed by parabolic systems, since in the problem of controllability of parabolic systems, we are asked to find a control so that the corresponding trajectory of the system reaches a given state in a fixed time T, while in the problem of existence of admissible control for an optimal time control problem governed by a parabolic system, we are asked to find a control (in a bounded set in most cases) such that the corresponding trajectory of the system reaches a given state in some time T. The differences between them are that first the arriving time is fixed in the controllability problem, while it is not fixed in the optimal time control problem, and second, the controls in the controllability problem may not be in any bounded set, while the control set is bounded (in most cases) in the optimal time control problem. For other works concerning approximate and null controllability of linear and nonlinear parabolic systems, we refer the readers to [18], [13], [19], [15], [7], [14], [8], [9], [21],

[22], [12], [23], and [10]. These works give us effective ways to deal with different kinds of controllability problems governed by parabolic equations and systems. We believe that the method provided in this work can be widely used to obtain existence of admissible control for the optimal time control problem governed by parabolic systems.

It should be noted that in this paper, in order to obtain the existence of a pair of admissible controls, we put controls in both equations of the phase-field system (1.1), because the technique developed in this paper to obtain the existence of the admissible controls is based on the local controllability of the system. If we put a control only in the first equation of the system (1.1), then one needs the local null controllability to the following linearized system:

$$\begin{cases} y_t(x,t) - a\triangle y(x,t) + \xi(x,t)y(x,t) - cz(x,t) = 0 & \text{in } Q = \Omega \times (0,T), \\ z_t(x,t) - k\triangle z(x,t) + la\triangle y(x,t) - l\xi(x,t)y(x,t) + lcz(x,t) = m(x)w(x,t) & \text{in } Q, \\ y(x,0) = y^0(x), \quad z(x,0) = z^0(x) & \text{in } \Omega, \\ y(x,t) = z(x,t) = 0 & \text{on } \sum = \partial\Omega \times (0,T), \end{cases}$$

where $a, c, k, l, m(x)$ are given in (1.1), $\xi(x,t) \in L^\infty(Q), y^0(x) \in V$, and $z^0(x) \in V$. However, such local null controllability depends on the observability for the corresponding linear backward system, i.e.,

$$|p(0)|_2^2 + |q(0)|_2^2 \le C \int_{Q^\omega} q^2(x,t)\, dx\, dt$$

for all solutions (p,q) satisfying

$$\begin{cases} p_t(x,t) + (a\triangle - \xi(x,t))p(x,t) - (la\triangle - l\xi(x,t))q(x,t) = 0 & \text{in } Q, \\ q_t(x,t) + (k\triangle - lc)q(x,t) + cp(x,t) = 0 & \text{in } Q, \\ p(x,t) = q(x,t) = 0 & \text{on } \sum. \end{cases}$$

To the best of our knowledge, whether such observability holds is still an open problem. So we put two controls in both equations. However, we must mention that this is not completely natural, although it is needed for the theoretical analysis in this paper.

This paper is organized as follows. In section 2 we show the existence of a pair of optimal time controls, and in section 3 we obtain the maximum principle for problem (P).

2. Existence of optimal time control. In this section we shall prove the existence of a pair of optimal time controls if the initial datum (u_0, h_0) is close to (u_e, h_e). First, we apply the infinite dimensional Kakutani fixed point theorem to prove the existence of a pair of admissible controls. Then we show the existence of a pair of optimal time controls.

The main results obtained in this section are presented as follows.

THEOREM 2.1. *Let $T > 0$ be given and $(u_0(x), h_0(x)) \in H^1(\Omega) \times H^2(\Omega)$ be such that*

$$0 < \|u_e - u_0\|_1^2 + \|h_e - h_0\|_{H^2(\Omega)}^2 \le \min\{\rho^2, 1\} e^{-c_0(1+T+\frac{1}{T})}, \tag{2.1}$$

where c_0 is a positive constant independent of T. Then there exists a pair of controls $(w^*, v^*) \in (H^1(0,T;H))^2 \cap U_\rho$ such that the corresponding solution $(u^{w^*,v^*}, h^{w^*,v^*}) \in H^{2,1}(Q) \times (H^{2,1}(Q) \cap L^\infty(0,T;H^2(\Omega)))$ to (1.1), here $Q = \Omega \times (0,T)$, satisfies

$$u^{w^*,v^*}(T) = u_e \quad \text{and} \quad h^{w^*,v^*}(T) = h_e.$$

Theorem 2.1 amounts to saying that system (1.1) is locally exactly controllable with controls in bounded set U_ρ.

COROLLARY 2.2. *Let* $(u_0(x), h_0(x)) \in H^1(\Omega) \times H^2(\Omega)$ *be such that*

$$0 < \|u_e - u_0\|_1^2 + \|h_e - h_0\|_{H^2(\Omega)}^2 \le \min\{\rho^2, 1\} e^{-3c_0}, \tag{2.2}$$

where c_0 *is the constant in Theorem* 2.1. *Then there exists a pair of controls* $(w^*, v^*) \in (H^1(0,1;H))^2 \cap U_\rho$ *such that the corresponding solution* $(u^{w^*,v^*}, h^{w^*,v^*}) \in H^{2,1}(Q_1) \times (H^{2,1}(Q_1) \cap L^\infty(0,1;H^2(\Omega)))$ *to* (1.1), *where* $Q_1 = \Omega \times (0,1)$, *satisfies* $u^{w^*,v^*}(1) = u_e$, $h^{w^*,v^*}(1) = h_e$.

THEOREM 2.3. *Let* $(u_0(x), h_0(x)) \in H^1(\Omega) \times H^2(\Omega)$ *satisfy* (2.2). *Then there exists at least one pair of optimal time controls for problem (P).*

In order to prove the above theorems, we need the following preliminary results.

LEMMA 2.4. *Let* $\omega_0 \subset \Omega$ *be an arbitrary but fixed open subset such that* $\overline{\omega}_0 \subset \omega$. *Then there exists a function* $\widetilde{\psi} \in C^2(\overline{\Omega})$ *such that* $\widetilde{\psi} > 0$ *for all* $x \in \Omega$, $\widetilde{\psi} = 0$ *on* $\partial\Omega$, *and* $|\nabla \widetilde{\psi}(x)| > 0$ *in* $\overline{\Omega} \backslash \omega_0$.

The proof of Lemma 2.4 can be found in [11]. We omit it here.

LEMMA 2.5. *Let* $\omega_0 \subset \Omega$ *be an arbitrary but fixed open subset such that* $\overline{\omega}_0 \subset \omega$. *Then there exists a function* $\psi \in C^2(\overline{\Omega})$ *such that* $\psi > 1$ *for all* $x \in \Omega$, $\psi = 1$ *on* $\partial\Omega$, $|\nabla\psi(x)| > 0$ *in* $\overline{\Omega} \backslash \omega_0$, *and* $\|\psi\| \le \frac{10}{9}$, *where* $\|\psi\|$ *denotes* $\|\psi\|_{C(\overline{\Omega})}$.

Proof. Let $\psi = \frac{\widetilde{\psi}}{n_0} + 1$, where $\widetilde{\psi}$ is the function defined in Lemma 2.4 and n_0 is a positive integer such that $\| \widetilde{\psi} \| \le \frac{1}{9} n_0$. Then ψ is exactly what we desire. This completes the proof. □

Let ψ be given by Lemma 2.5, and set

$$\varphi(x,t) = \frac{e^{\lambda\psi}}{t(T-t)}, \qquad \alpha(x,t) = \frac{e^{\lambda\psi} - e^{\frac{9}{5}\lambda\|\psi\|}}{t(T-t)}.$$

From now on, we shall omit all x, t in the functions of x and t if there is no ambiguity.

Next we shall give the proof of the Carleman inequality for the linear backward heat equation, which is presented below for the sake of completeness and easy reference (see also [10] and [11]).

LEMMA 2.6. *There exist constants* $\lambda_0 \ge 1$ *and* $s_0 \ge 1$ *such that for* $\lambda \ge \lambda_0$ *and* $s \ge s_0(T + T^2)$, *we have*

$$\begin{aligned} &\int_Q [(s\varphi)^{-1}(p_t^2 + |\triangle p|^2) + s\varphi|\nabla p|^2 + s^3\varphi^3 p^2] e^{2s\alpha}\, dx\, dt \\ &\le C\lambda^4 \int_{Q^\omega} s^3\varphi^3 p^2 e^{2s\alpha}\, dx\, dt + C\lambda \int_Q g^2 e^{2s\alpha}\, dx\, dt \end{aligned}$$

for all solutions p *to the linear backward equation*

$$\begin{cases} p_t + b_0 \triangle p = g & \text{in } Q = \Omega \times (0,T), \\ p = 0 & \text{on } \sum = \partial\Omega \times (0,T), \end{cases} \tag{2.3}$$

where b_0 *is a positive constant,* $Q^\omega = \omega \times (0,T)$, *and* C *denotes a positive constant independent of* T, p, λ, *and* s.

Proof. Let $z = pe^{s\alpha}$; it follows that

$$\begin{cases} z_t + b_0 \triangle z + (b_0 s^2\lambda^2\varphi^2|\nabla\psi|^2 - b_0 s\lambda^2\varphi|\nabla\psi|^2)z - 2b_0 s\lambda\varphi\nabla\psi\cdot\nabla z \\ \qquad\qquad - (b_0 s\lambda\varphi\triangle\psi + s\alpha_t)z = ge^{s\alpha} & \text{in } Q, \\ z(x,0) = z(x,T) = 0 & \text{in } \Omega, \\ z = 0 & \text{on } \sum. \end{cases}$$

We set

$$\begin{cases} X(t)z = -2b_0(s\lambda\varphi\nabla\psi\cdot\nabla z + s\lambda^2\varphi|\nabla\psi|^2 z), \\ B(t)z = -b_0\triangle z - b_0 s\lambda^2\varphi|\nabla\psi|^2 z - b_0 s^2\lambda^2\varphi^2|\nabla\psi|^2 z + (b_0 s\lambda\varphi\triangle\psi + s\alpha_t)z; \end{cases}$$

it is easy to show that

$$z_t + X(t)z - B(t)z = ge^{s\alpha} \qquad \text{in } Q. \tag{2.4}$$

First we shall show that as $\lambda \geq \lambda_1$ and $s \geq s_1(T+T^2)$, where $\lambda_1 \geq 1$ and $s_1 \geq 1$ are constants independent of T,

$$\begin{aligned} &\int_Q s^3\lambda^4\varphi^3|\nabla\psi|^4 z^2\,dx\,dt + \int_Q s\lambda^2\varphi|\nabla\psi|^2|\nabla z|^2\,dx\,dt \\ &\leq C\left(\int_Q s^3\lambda^3\varphi^3 z^2\,dx\,dt + \int_Q s\lambda\varphi|\nabla z|^2\,dx\,dt + \int_Q g^2 e^{2s\alpha}\,dx\,dt\right). \end{aligned} \tag{2.5}$$

To this end, we set

$$Y = -\int_Q X(t)z\cdot B(t)z\,dx\,dt.$$

By the same arguments as those in [10] and [11], we may derive the following three inequalities:

$$\begin{aligned} &-2b_0\int_Q (s\lambda\varphi\nabla\psi\cdot\nabla z + s\lambda^2\varphi|\nabla\psi|^2 z)\cdot b_0\triangle z\,dx\,dt \\ &\geq \frac{1}{2}s\lambda^2 b_0^2\int_Q \varphi|\nabla\psi|^2|\nabla z|^2\,dx\,dt - C\int_Q s\lambda^4\varphi z^2\,dx\,dt - C\int_Q s\lambda\varphi|\nabla z|^2\,dx\,dt, \end{aligned} \tag{2.6}$$

$$\begin{aligned} &-2b_0\int_Q s\lambda\varphi(\nabla\psi\cdot\nabla z)\cdot b_0(s^2\lambda^2\varphi^2|\nabla\psi|^2 + s\lambda^2\varphi|\nabla\psi|^2)z\,dx\,dt \\ &\geq -C\int_Q s^3\lambda^3\varphi^3 z^2\,dx\,dt + 3s^3\lambda^4 b_0^2\int_Q \varphi^3|\nabla\psi|^4 z^2\,dx\,dt - C\int_Q s^2\lambda^3\varphi^2 z^2\,dx\,dt, \end{aligned} \tag{2.7}$$

and

$$\begin{aligned} &-2b_0\int_Q s\lambda\varphi\nabla\psi\cdot\nabla z(-b_0 s\lambda\varphi\triangle\psi\cdot z - s\alpha_t z)\,dx\,dt \\ &\geq -C\int_Q s^3\lambda^3\varphi^3 z^2\,dx\,dt - C\int_Q s\lambda\varphi|\nabla z|^2\,dx\,dt \\ &\quad -b_0\int_Q s^2\lambda z^2[\lambda\varphi|\nabla\psi|^2\alpha_t + \varphi\triangle\psi\alpha_t + \lambda\varphi\varphi_t|\nabla\psi|^2]\,dx\,dt, \end{aligned} \tag{2.8}$$

where $\lambda \geq 1$. Here and throughout the proof of Lemma 2.6, C denotes several positive constants independent of T.

On the other hand, we have

$$\varphi_t = (2t-T)\varphi^2 e^{-\lambda\psi}, \qquad \alpha_t = \frac{e^{\lambda\psi} - e^{\frac{9}{5}\lambda\|\psi\|}}{e^{2\lambda\psi}}(2t-T)\varphi^2.$$

The latter combined with Lemma 2.5 indicates that

$$|\varphi_t| \leq CT\varphi^2 \quad \text{and} \quad |\alpha_t| \leq CT\varphi^2,$$

which together with (2.8) yield that

$$\begin{aligned}&-2b_0\int_Q s\lambda\varphi\nabla\psi\cdot\nabla z(-b_0 s\lambda\varphi\triangle\psi\cdot z - s\alpha_t z)\,dx\,dt\\&\geq -C\int_Q s^3\lambda^3\varphi^3 z^2\,dx\,dt - C\int_Q s\lambda\varphi|\nabla z|^2\,dx\,dt - C\int_Q s^2\lambda^2 T\varphi^3 z^2\,dx\,dt,\end{aligned}$$

where $\lambda\geq 1$. It follows from (2.6), (2.7), and the latter that

$$\begin{aligned}Y\geq&\int_Q b_0^2 s^3\lambda^4\varphi^3|\nabla\psi|^4 z^2\,dx\,dt+\frac{1}{2}\int_Q b_0^2 s\lambda^2\varphi|\nabla\psi|^2|\nabla z|^2\,dx\,dt\\&-C\int_Q(s\lambda^4\varphi+s^2\lambda^4\varphi^2+s^2\lambda^2 T\varphi^3+s^3\lambda^3\varphi^3)z^2\,dx\,dt\\&-C\int_Q s\lambda\varphi|\nabla z|^2\,dx\,dt\end{aligned}\tag{2.9}$$

as $\lambda\geq 1$.

By (2.4) and the same arguments as above, we deduce that

$$\begin{aligned}2Y&\leq 2\int_Q z_t\cdot B(t)z\,dx\,dt+\int_Q g^2e^{2s\alpha}\,dx\,dt\\&\leq C\int_Q(s\lambda^2 T\varphi^2+s^2\lambda^2\varphi^3 T+s\varphi^2+sT^2\varphi^3)z^2\,dx\,dt+\int_Q g^2e^{2s\alpha}\,dx\,dt\end{aligned}$$

as $\lambda\geq 1$. The latter combined with (2.9) implies that

$$\begin{aligned}&\int_Q s^3\lambda^4\varphi^3|\nabla\psi|^4 z^2\,dx\,dt+\int_Q s\lambda^2\varphi|\nabla\psi|^2|\nabla z|^2\,dx\,dt\\&\leq\int_Q(s\lambda^4\varphi+s^2\lambda^4\varphi^2+s\varphi^2+s\lambda^2 T\varphi^2+s^2\lambda^2 T\varphi^3+s^3\lambda^3\varphi^3+sT^2\varphi^3)z^2\,dx\,dt\\&\quad+C\int_Q s\lambda\varphi|\nabla z|^2+C\int_Q g^2e^{2s\alpha}\,dx\,dt,\end{aligned}$$

from which we obtain (2.5) as desired.

Next we claim that as $\lambda\geq\lambda_2$ and $s\geq s_2(T+T^2)$, where $\lambda_2\geq\lambda_1$ and $s_2\geq s_1$ are constants independent of T,

$$\begin{aligned}&\int_Q s^3\lambda^3\varphi^3 p^2e^{2s\alpha}\,dx\,dt+\int_Q s\lambda\varphi|\nabla p|^2e^{2s\alpha}\,dx\,dt\\&\leq C\int_{Q^\omega} s^3\lambda^3\varphi^3 p^2e^{2s\alpha}\,dx\,dt+C\int_Q g^2e^{2s\alpha}\,dx\,dt.\end{aligned}\tag{2.10}$$

To this end, we observe from Lemma 2.5 and (2.5) that

$$\begin{aligned}&\int_{Q\setminus Q^{\omega_0}} s^3\lambda^4\varphi^3 z^2\,dx\,dt+\int_{Q\setminus Q^{\omega_0}} s\lambda^{\frac{7}{6}}\varphi|\nabla z|^2\,dx\,dt\\&\leq C\int_{Q^{\omega_0}}(s^3\lambda^3\varphi^3 z^2+s\lambda\varphi|\nabla z|^2)\,dx\,dt+C\int_Q g^2e^{2s\alpha}\,dx\,dt,\end{aligned}$$

where $\lambda \geq \lambda_3, s \geq s_3(T+T^2)$, $\lambda_3 \geq \lambda_1$, and $s_3 \geq s_1$ are constants independent of T. Substituting $z = pe^{s\alpha}$ into the latter inequality, after some calculation, we infer that

$$\begin{aligned} &\int_{Q\setminus Q^{\omega_0}} s^3\lambda^3\varphi^3 p^2 e^{2s\alpha}\,dx\,dt + \int_{Q\setminus Q^{\omega_0}} s\lambda\varphi|\nabla p|^2 e^{2s\alpha}\,dx\,dt \\ &\leq C\int_{Q^{\omega_0}} (s^3\lambda^3\varphi^3 p^2 + s\lambda\varphi|\nabla p|^2)e^{2s\alpha}\,dx\,dt + C\int_Q g^2 e^{2s\alpha}\,dx\,dt, \end{aligned} \tag{2.11}$$

where $\lambda \geq \lambda_4, s \geq s_4(T+T^2)$, $\lambda_4 \geq \lambda_3$, and $s_4 \geq s_3$ are constants independent of T. Taking $\chi \in C_0^\infty(\Omega)$ such that $\chi = 1$ in $\overline{\omega}_0$ and $\chi = 0$ in $\Omega\setminus\omega$, then multiplying (2.3) by $\chi\varphi p e^{2s\alpha}$ and integrating it over Q, we deduce

$$\begin{aligned} &-b_0\int_Q \chi\varphi|\nabla p|^2 e^{2s\alpha}\,dx\,dt + \frac{b_0}{2}\int_Q p^2\triangle(\chi\varphi e^{2s\alpha})\,dx\,dt - \frac{1}{2}\int_Q p^2\chi(\varphi e^{2s\alpha})_t\,dx\,dt \\ &= \int_Q g\chi\varphi p e^{2s\alpha}\,dx\,dt, \end{aligned}$$

which implies

$$\int_{Q^{\omega_0}} s\lambda\varphi|\nabla p|^2 e^{2s\alpha}\,dx\,dt \leq C\int_{Q^\omega} s^3\lambda^3\varphi^3 p^2 e^{2s\alpha}\,dx\,dt + C\int_Q g^2 e^{2s\alpha}\,dx\,dt.$$

Substituting the latter into (2.11), we deduce (2.10) as desired.

To finish the proof of Lemma 2.6, we multiply (2.3) by $(s\varphi)^{-1}p_t e^{2s\alpha}$ and then integrate it over Q to get

$$\int_Q (s\varphi)^{-1}p_t^2 e^{2s\alpha}\,dx\,dt \leq C\int_Q (s\varphi)^{-1}g^2 e^{2s\alpha}\,dx\,dt + C\int_Q s\lambda^2\varphi|\nabla p|^2 e^{2s\alpha}\,dx\,dt,$$

which together with (2.3) and (2.10) completes the proof. □

LEMMA 2.7. *The following estimate holds:*

$$\|f\|_{C([0,T])} \leq C\left(\frac{1}{T}+1\right)\|f\|_{W^{1,1}(0,T)} \qquad \forall\, f \in W^{1,1}(0,T),$$

where $C > 0$ is a constant independent of T and f.

Proof. It is well known that (cf. [1])

$$\|g\|_{C([0,1])} \leq C\|g\|_{W^{1,1}(0,1)} \qquad \forall\, g \in W^{1,1}(0,1).$$

Here and throughout the proof of Lemma 2.7, C denotes several positive constants independent of T from which we deduce that

$$\begin{aligned} &\|f\|_{C([0,T])} = \|f(Tx)\|_{C([0,1])} \leq C\|f(Tx)\|_{W^{1,1}(0,1)} \\ &\leq C\left\|f\left(\frac{T}{2}y\right)\right\|_{W^{1,1}(0,2)} \leq C\left(\frac{1}{T}+1\right)\|f\|_{W^{1,1}(0,T)}. \end{aligned}$$

This completes the proof. □

Next we prove Theorem 2.1.

Proof of Theorem 2.1. Let $h = y + h_e$ and $u = z + u_e$. We are led to prove the local null controllability of the system

$$
\left\{
\begin{array}{ll}
y_t - a\triangle y + a_1 y + b_1 y^2 + by^3 - cz = m(x)v(x,t) & \text{in } Q = \Omega \times (0,T),\\
z_t - k\triangle z + la\triangle y + a_2 y + b_2 y^2 - lby^3 + lcz & \\
\qquad = m(x)(w(x,t) - lv(x,t)) & \text{in } Q,\\
y(x,0) = h_0(x) - h_e(x) \equiv y^0(x) & \text{in } \Omega,\\
z(x,0) = u_0(x) - u_e(x) \equiv z^0(x) & \text{in } \Omega,\\
y(x,t) = z(x,t) = 0 & \text{on } \sum = \partial\Omega \times (0,T),
\end{array}
\right. \tag{2.12}
$$

where $a_1 = 3bh_e^2 - b, b_1 = 3bh_e, a_2 = lb - 3lbh_e^2$, and $b_2 = -3lbh_e$.

We shall use Kakutani's fixed point theorem to prove it. To this end, we set

$$
K = \{\xi(x,t) \in L^\infty(Q) : |\xi(x,t)| \le M \quad \text{a.e. } (x,t) \in Q\}, \tag{2.13}
$$

where $M = \|a_1\|_{L^\infty(\Omega)} + C_0\|b_1\|_{L^\infty(\Omega)} + bC_0^2$ and C_0 is the best constant such that inequality $\|y\|_{L^\infty(\Omega)} \le C_0\|y\|_{H^2(\Omega)}$ holds for all $y \in H^2(\Omega)$. We fix $\xi \in K$ and consider the solution $(y^{w,v}, z^{w,v}) \in (H^{2,1}(Q) \cap V)^2$ to the following linear system:

$$
\left\{
\begin{array}{ll}
y_t - a\triangle y + \xi y - cz = m(x)v(x,t) & \text{in } Q = \Omega \times (0,T),\\
z_t - k\triangle z + la\triangle y - l\xi y + lcz = m(x)(w(x,t) - lv(x,t)) & \text{in } Q,\\
y(x,0) = y^0(x), \qquad z(x,0) = z^0(x) & \text{in } \Omega,\\
y(x,t) = z(x,t) = 0 & \text{on } \sum = \partial\Omega \times (0,T).
\end{array}
\right. \tag{2.14}
$$

LEMMA 2.8. *For all $\xi \in K$, there exist $(w,v) \in (H^1(0,T;H))^2, (y,z) \in (H^{2,1}(Q) \cap V)^2$ satisfying* (2.14) *and such that*

$$
y(x,T) = z(x,T) = 0, \tag{2.15}
$$

$$
\|w\|^2_{H^1(0,T;H)} + \|v\|^2_{H^1(0,T;H)} \le e^{c_0^{(1)}(1+T+\frac{1}{T})}(|y^0|_2^2 + |z^0|_2^2), \tag{2.16}
$$

where $c_0^{(1)} > 0$ is a constant independent of T, ξ, y^0, and z^0.

Proof. We set $\overline{v}(x,t) = w(x,t) - lv(x,t)$. For $\varepsilon > 0$, consider the following optimal control problem:

$$
\begin{aligned}
&\text{Min}\left\{\frac{1}{2}\int_Q e^{-\frac{3}{2}s\alpha}(v^2 + \overline{v}^2)\,dx\,dt + \frac{1}{2\varepsilon}\int_\Omega (y^2(x,T) + z^2(x,T))\,dx\right\}\\
&\quad \text{subject to (2.14).}
\end{aligned} \tag{2.17}
$$

Let $((y_\varepsilon, z_\varepsilon), (v_\varepsilon, \overline{v}_\varepsilon))$ be an optimal pair for problem (2.17). (The existence follows in a standard way; see, for instance, [2].) By the maximum principle, we have that

$$
v_\varepsilon = mp_\varepsilon e^{\frac{3}{2}s\alpha}, \qquad \overline{v}_\varepsilon = mq_\varepsilon e^{\frac{3}{2}s\alpha} \quad \text{a.e. in } Q, \tag{2.18}
$$

where $(p_\varepsilon, q_\varepsilon)$ is the solution to the dual backward system

$$
\left\{
\begin{array}{ll}
(p_\varepsilon)_t + (a\triangle - \xi)p_\varepsilon - (la\triangle - l\xi)q_\varepsilon = 0 & \text{in } Q,\\
(q_\varepsilon)_t + (k\triangle - lc)q_\varepsilon + cp_\varepsilon = 0 & \text{in } Q,\\
p_\varepsilon(T) = -\frac{1}{\varepsilon}y_\varepsilon(T), \qquad q_\varepsilon(T) = -\frac{1}{\varepsilon}z_\varepsilon(T) & \text{in } \Omega,\\
p_\varepsilon(x,t) = q_\varepsilon(x,t) = 0 & \text{on } \sum.
\end{array}
\right. \tag{2.19}
$$

Taking (2.14), (2.18), and (2.19) into account, we obtain

$$\begin{aligned}(2.20)\qquad &\frac{1}{\varepsilon}\int_\Omega (y_\varepsilon^2(T)+z_\varepsilon^2(T))\,dx+\int_{Q^\omega}(p_\varepsilon^2+q_\varepsilon^2)e^{\frac{3}{2}s\alpha}\,dx\,dt\\ &\le |y^0|_2|p_\varepsilon(0)|_2+|z^0|_2|q_\varepsilon(0)|_2.\end{aligned}$$

Now we claim that as $\lambda\ge\lambda_1$ and $s\ge s_1(T+T^2)$, the following inequality holds:

$$\begin{aligned}(2.21)\qquad &\int_Q[(s\varphi)^{-1}(|(p_\varepsilon)_t|^2+|(q_\varepsilon)_t|^2+|\triangle p_\varepsilon|^2+|\triangle q_\varepsilon|^2)+s\varphi(|\nabla p_\varepsilon|^2+|\nabla q_\varepsilon|^2)\\ &\quad+s^3\varphi^3(p_\varepsilon^2+q_\varepsilon^2)]e^{2s\alpha}\,dx\,dt\le\int_{Q^\omega}(p_\varepsilon^2+q_\varepsilon^2)e^{\frac{3}{2}s\alpha}\,dx\,dt,\end{aligned}$$

where $\lambda_1\ge\lambda_0$ and $s_1\ge s_0$ are constants independent of T, and λ_0 and s_0 are the constants arising in Lemma 2.6.

To this end, we apply Lemma 2.6 to $(2.19)_1$ and $(2.19)_2$ (the first and the second equations of (2.19)), respectively, to obtain

$$\begin{aligned}(2.22)\qquad &\int_Q[(s\varphi)^{-1}(|(p_\varepsilon)_t|^2+|\triangle p_\varepsilon|^2)+s\varphi|\nabla p_\varepsilon|^2+s^3\varphi^3p_\varepsilon^2]e^{2s\alpha}\,dx\,dt\\ &\le C\lambda^4\int_{Q^\omega}s^3\varphi^3p_\varepsilon^2e^{2s\alpha}\,dx\,dt+C\lambda\int_Q q_\varepsilon^2e^{2s\alpha}\,dx\,dt+C\lambda\int_Q|\triangle q_\varepsilon|^2e^{2s\alpha}\,dx\,dt,\end{aligned}$$

$$\begin{aligned}(2.23)\qquad &\int_Q[(s\varphi)^{-1}(|(q_\varepsilon)_t|^2+|\triangle q_\varepsilon|^2)+s\varphi|\nabla q_\varepsilon|^2+s^3\varphi^3q_\varepsilon^2]e^{2s\alpha}\,dx\,dt\\ &\le C\lambda^4\int_{Q^\omega}s^3\varphi^3q_\varepsilon^2e^{2s\alpha}\,dx\,dt+C\lambda\int_Q p_\varepsilon^2e^{2s\alpha}\,dx\,dt,\end{aligned}$$

where $\lambda\ge\lambda_2$, $s\ge s_2(T+T^2)$, $\lambda_2\ge\lambda_0$, and $s_2\ge s_0$ are constants independent of T. Here and throughout the proof of Lemma 2.8, $C>0$ denotes several positive constants independent of T.

On the other hand, we have

$$\begin{cases}(t(T-t)p_\varepsilon)_t+(a\triangle-\xi)(t(T-t)p_\varepsilon)=t(T-t)(la\triangle-l\xi)q_\varepsilon\\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad+(T-2t)p_\varepsilon\quad\text{in }\ Q,\\ t(T-t)p_\varepsilon=0\qquad\text{on }\ \textstyle\sum.\end{cases}$$

By applying (2.10) to the latter system, it follows that

$$\begin{aligned}&\int_Q s^3\lambda^3\varphi e^{2\lambda\psi}p_\varepsilon^2e^{2s\alpha}\,dx\,dt+\int_Q s\lambda\varphi^{-1}e^{2\lambda\psi}|\nabla p_\varepsilon|^2e^{2s\alpha}\,dx\,dt\\ &\le C\int_{Q^\omega}s^3\lambda^3\varphi e^{2\lambda\psi}p_\varepsilon^2e^{2s\alpha}\,dx\,dt\\ &\quad+CsT^2e^{\lambda\|\psi\|}\int_Q[(s\varphi)^{-1}|\triangle q_\varepsilon|^2+s^3\varphi^3q_\varepsilon^2]e^{2s\alpha}\,dx\,dt,\end{aligned}$$

where $\lambda\ge\lambda_3$, $s\ge s_3(T+T^2)$, $\lambda_3\ge\lambda_2$, and $s_3\ge s_2$ are constants independent of T. Together with (2.23), the latter implies that

$$\begin{aligned}(2.24)\qquad &\int_Q s^3\lambda^3\varphi e^{2\lambda\psi}p_\varepsilon^2e^{2s\alpha}\,dx\,dt+\int_Q s\lambda\varphi^{-1}e^{2\lambda\psi}|\nabla p_\varepsilon|^2e^{2s\alpha}\,dx\,dt\\ &\le C\int_{Q^\omega}s^3\lambda^3\varphi e^{2\lambda\psi}p_\varepsilon^2e^{2s\alpha}\,dx\,dt+Cs^4T^2e^{\lambda\|\psi\|}\lambda^4\int_{Q^\omega}q_\varepsilon^2\varphi^3e^{2s\alpha}\,dx\,dt\end{aligned}$$

and

$$\begin{aligned}&\int_Q [(s\varphi)^{-1}(|(q_\varepsilon)_t|^2 + |\triangle q_\varepsilon|^2) + s\varphi|\nabla q_\varepsilon|^2 + s^3\varphi^3 q_\varepsilon^2]e^{2s\alpha}\,dx\,dt\\ &\le C\int_{Q^\omega} s^3\lambda^4\varphi^3 q_\varepsilon^2 e^{2s\alpha}\,dx\,dt + CT^2\int_{Q^\omega}\varphi p_\varepsilon^2 e^{2s\alpha}\,dx\,dt.\end{aligned} \tag{2.25}$$

Now (2.21) follows from (2.22) and (2.25) immediately.

Next we shall show that

$$\|v_\varepsilon\|^2_{H^1(0,T;H)} + \|\overline{v}_\varepsilon\|^2_{H^1(0,T;H)} \le e^{\frac{Cse^{\frac{9}{5}\lambda\|\psi\|}}{T^2}}(|y^0|_2^2 + |z^0|_2^2), \tag{2.26}$$

where $\lambda \ge \lambda_4$, $s \ge s_4(T + T^2 + T^3)$, $\lambda_4 \ge \lambda_1$, and $s_4 \ge s_1$ are constants independent of T.

Multiplying $(2.19)_1$ and $(2.19)_2$ by δp_ε and q_ε, respectively, where $\delta > 0$ is suitably chosen, and then integrating them on $\Omega \times (\tau, t)$, after some elementary calculation, we may obtain that

$$\begin{aligned}&|p_\varepsilon(\tau)|_2^2 + |q_\varepsilon(\tau)|_2^2 + \int_\tau^t (|\nabla p_\varepsilon|_2^2 + |\nabla q_\varepsilon|_2^2)\,ds\\ &\le C(|p_\varepsilon(t)|_2^2 + |q_\varepsilon(t)|_2^2) + C\int_\tau^t (|p_\varepsilon|_2^2 + |q_\varepsilon|_2^2)\,ds,\end{aligned}$$

where $0 \le \tau \le t \le T$. The latter combined with Gronwall's inequality indicates that

$$|p_\varepsilon(0)|_2^2 + |q_\varepsilon(0)|_2^2 \le C(|p_\varepsilon(t)|_2^2 + |q_\varepsilon(t)|_2^2)e^{CT} \qquad \forall\, 0 \le t \le T.$$

Integrating the latter on $(\frac{T}{4}, \frac{3}{4}T)$, we get that

$$\begin{aligned}&|p_\varepsilon(0)|_2^2 + |q_\varepsilon(0)|_2^2\\ &\le \frac{Ce^{CT}}{T}\int_{\frac{T}{4}}^{\frac{3}{4}T}\int_\Omega (p_\varepsilon^2 + q_\varepsilon^2)\,dx\,dt\\ &= \frac{Ce^{CT}}{T}\int_{\frac{T}{4}}^{\frac{3}{4}T}\int_\Omega s^{-3}\varphi^{-3}e^{-2s\alpha}s^3\varphi^3(p_\varepsilon^2 + q_\varepsilon^2)e^{2s\alpha}\,dx\,dt\\ &\le \frac{1}{2}e^{\frac{32s}{T^2}e^{\frac{9}{5}\lambda\|\psi\|}}\int_Q s^3\varphi^3(p_\varepsilon^2 + q_\varepsilon^2)e^{2s\alpha}\,dx\,dt,\end{aligned}$$

where $\lambda \ge \lambda_5$, $s \ge s_5(T + T^2 + T^3)$, $\lambda_5 \ge \lambda_1$, and $s_5 \ge s_1$ are constants independent of T. This together with (2.21) yields that

$$|p_\varepsilon(0)|_2^2 + |q_\varepsilon(0)|_2^2 \le \frac{1}{2}e^{\frac{32s}{T^2}e^{\frac{9}{5}\lambda\|\psi\|}}\int_{Q^\omega}(p_\varepsilon^2 + q_\varepsilon^2)e^{\frac{3}{2}s\alpha}\,dx\,dt.$$

The latter together with (2.20) shows that

$$\begin{aligned}&\frac{1}{\varepsilon}\int_\Omega (y_\varepsilon^2(T) + z_\varepsilon^2(T))\,dx + \int_{Q^\omega}(p_\varepsilon^2 + q_\varepsilon^2)e^{\frac{3}{2}s\alpha}\,dx\,dt\\ &\le e^{\frac{32s}{T^2}e^{\frac{9}{5}\lambda\|\psi\|}}(|y^0|_2^2 + |z^0|_2^2),\end{aligned} \tag{2.27}$$

which, combined with (2.18), indicates that

$$\int_Q (v_\varepsilon^2 + \overline{v}_\varepsilon^2)\,dx\,dt + \frac{1}{\varepsilon}\int_\Omega (y_\varepsilon^2(T) + z_\varepsilon^2(T))\,dx \le e^{\frac{32s}{T^2}e^{\frac{9}{5}\lambda\|\psi\|}}(|y^0|_2^2 + |z^0|_2^2). \tag{2.28}$$

Thus (2.26) follows immediately from (2.18), (2.21), (2.27), and (2.28).

Now we claim that

$$\begin{aligned}(2.29)\qquad &\|y_\varepsilon(t)\|_1^2+\|z_\varepsilon(t)\|_1^2+\|y_\varepsilon\|_{H^{2,1}(Q)}^2+\|z_\varepsilon\|_{H^{2,1}(Q)}^2\\ &\le e^{\frac{Cse^{\frac{9}{5}\lambda\|\psi\|}}{T^2}}(\|y^0\|_1^2+\|z^0\|_1^2),\end{aligned}$$

where $\lambda\ge\lambda_6$, $s\ge s_6(T+T^2+T^3)$, $\lambda_6\ge\lambda_4$, and $s_6\ge s_4$ are constants independent of T.

Multiplying $(2.14)_1$ and $(2.14)_2$ by y_ε and δz_ε in $L^2(\Omega)$, respectively, where $\delta>0$ is suitably chosen, and then integrating them over $(0,t)$, after some calculation involving Gronwall's inequality and (2.26), we obtain

$$(2.30)\qquad |y_\varepsilon(t)|_2^2+|z_\varepsilon(t)|_2^2+\int_0^T(|\nabla y_\varepsilon|_2^2+|\nabla z_\varepsilon|_2^2)\,ds\le e^{\frac{Cse^{\frac{9}{5}\lambda\|\psi\|}}{T^2}}(|y^0|_2^2+|z^0|_2^2).$$

Multiplying $(2.14)_1$ and $(2.14)_2$ by $-\triangle y_\varepsilon$ and $-\triangle z_\varepsilon$ in H, respectively, and then integrating them over $(0,t)$, after some elementary calculation, we deduce that

$$|\nabla y_\varepsilon(t)|_2^2+a\int_0^t|\triangle y_\varepsilon|_2^2\,ds\le|\nabla y^0|_2^2+C\int_0^t(|v_\varepsilon|_2^2+|y_\varepsilon|_2^2+|z_\varepsilon|_2^2)\,ds$$

and

$$|\nabla z_\varepsilon(t)|_2^2+k\int_0^t|\triangle z_\varepsilon|_2^2\,ds\le|\nabla z^0|_2^2+C\int_0^t(|\triangle y_\varepsilon|_2^2+|\overline{v}_\varepsilon|_2^2+|y_\varepsilon|_2^2+|z_\varepsilon|_2^2)\,ds,$$

which together with (2.14), (2.26), and (2.30) imply (2.29) as desired.

By (2.26) and (2.29) and by Arzela–Ascoli theorem and the Aubin compactness theorem, there exist subsequences of $\{y_\varepsilon\},\{z_\varepsilon\},\{w_\varepsilon\}$, and $\{v_\varepsilon\}$, still denoted in the same way, such that

$$\begin{aligned}&y_\varepsilon\to y\ \text{ and }\ z_\varepsilon\to z\quad \text{weakly in } H^1(0,T;H)\cap L^2(0,T;H^2(\Omega)),\\ &\qquad\qquad\qquad\quad \text{strongly in } L^2(0,T;V)\cap C([0,T];H),\\ &w_\varepsilon\to w,\qquad v_\varepsilon\to v\quad \text{weakly in } H^1(0,T;H),\end{aligned}$$

and

$$\|w_\varepsilon\|_{H^1(0,T;H)}^2+\|v_\varepsilon\|_{H^1(0,T;H)}^2\le e^{a_0s_4e^{\frac{9}{5}\lambda_4\|\psi\|}(1+T+\frac{1}{T})}(|y^0|_2^2+|z^0|_2^2),$$

where $a_0>0$ is a constant independent of T. If we set $c_0^{(1)}=a_0s_4e^{\frac{9}{5}\lambda_4\|\psi\|}$, then the latter together with (2.28) implies (2.15) and (2.16). By passing to the limit for $\varepsilon\to0$ in (2.14), we conclude that (y,z) and (w,v) satisfy (2.14). This completes the proof of Lemma 2.8. □

Proof of Theorem 2.1 (continued). We set $K_T=\{\xi(x,t)\in L^\infty(0,T;H^2(\Omega)\cap V)\cap H^1(0,T;V):\ \mu(\xi)\le1\}$, where $\mu(\xi)=\sqrt{\|\xi\|_{H^1(0,T;V)}^2+\|\xi\|_{L^\infty(0,T;H^2(\Omega))}^2}$.

Define the multivalued map $\Psi:K_T\longrightarrow L^2(Q)$ by

$$\begin{aligned}\Psi(\xi)=\{y^{w,v}\in L^2(0,T;H^2(\Omega)\cap V)\cap H^1(0,T;H):\ &y^{w,v}(T)=z^{w,v}(T)=0,\\ &\|w\|_{H^1(0,T;H)}^2+\|v\|_{H^1(0,T;H)}^2\le e^{c_0^{(1)}(1+T+\frac{1}{T})}(|y^0|_2^2+|z^0|_2^2)\},\end{aligned}$$

where $c_0^{(1)} > 0$ is the constant arising in Lemma 2.8 and $(y^{w,v}, z^{w,v})$ is the solution to the following equation:

$$
(2.31)\begin{cases} y_t - a\triangle y + (a_1 + b_1\xi + b\xi^2)y - cz = m(x)v(x,t) \ \text{ in } \ Q = \Omega \times (0,T), \\ z_t - k\triangle z + la\triangle y + (a_2 + b_2\xi - lb\xi^2)y + lcz \\ \qquad\qquad = m(x)(w(x,t) - lv(x,t)) \quad \text{in } \ Q, \\ y(x,0) = y^0(x), \qquad z(x,0) = z^0(x) \quad \text{in } \ \Omega, \\ y(x,t) = z(x,t) = 0 \quad \text{on } \ \sum = \partial\Omega \times (0,T). \end{cases}
$$

Then by Lemma 2.8, we have that

$$
\Psi(\xi) \neq \emptyset. \tag{2.32}
$$

Moreover, one can easily check that

$$
\begin{aligned} &K_T \text{ is a convex set in } \ L^2(Q) \text{ and} \\ &\qquad \Psi(\xi) \text{ is a convex set in } \ L^2(Q) \text{ for each } \ \xi \in K_T. \end{aligned} \tag{2.33}
$$

Next we shall show that

$$
K_T \text{ is a compact subset in } \ L^2(Q). \tag{2.34}
$$

Indeed, for any sequence $\{\xi_n\} \subset K_T$, by the definition of K_T, we have

$$
\|\xi_n\|^2_{H^1(0,T;V)} + \|\xi_n\|^2_{L^\infty(0,T;H^2(\Omega))} \leq 1. \tag{2.35}
$$

Thus there exists a subsequence of $\{\xi_n\}$, still denoted by itself, such that

$$
\begin{aligned} &\xi_n \to \xi \text{ weak star in } \ L^\infty(0,T;H^2(\Omega)), \\ &\qquad \text{weakly in } \ H^1(0,T;V) \text{ and strongly in } \ L^2(Q). \end{aligned} \tag{2.36}
$$

It follows from (2.35) and (2.36) that

$$
\|\xi\|^2_{H^1(0,T;V)} + \|\xi\|^2_{L^\infty(0,T;H^2(\Omega))} \leq 1,
$$

which implies $\xi \in K_T$, and (2.34) follows.

Next we prove that Ψ is upper-semicontinuous in $L^2(Q)$; i.e., if $\xi_n \in K_T$ is such that $\xi_n \to \widetilde{\xi}$ strongly in $L^2(Q)$ and $y_n \equiv y^{w_n,v_n} \in \Psi(\xi_n)$ is strongly convergent to $\widetilde{y}$ in $L^2(Q)$, then $\widetilde{y} \in \Psi(\widetilde{\xi})$.

To this end, we observe first that

$$
\|\xi_n\|^2_{H^1(0,T;V)} + \|\xi_n\|^2_{L^\infty(0,T;H^2(\Omega))} \leq 1.
$$

However, $\xi_n \to \widetilde{\xi}$ strongly in $L^2(Q)$; thus there exists a subsequence of $\{\xi_n\}$, still denoted by itself, such that

$$
\begin{aligned} &\xi_n \to \widetilde{\xi} \ \text{ weak star in } \ L^\infty(0,T;H^2(\Omega)), \\ &\qquad \text{weakly in } \ H^1(0,T;V) \text{ and strongly in } \ L^2(Q), \end{aligned} \tag{2.37}
$$

and hence

$$
\widetilde{\xi} \in K_T. \tag{2.38}
$$

Because

$$\|w_n\|^2_{H^1(0,T;H)} + \|v_n\|^2_{H^1(0,T;H)} \le e^{c_0^{(1)}(1+T+\frac{1}{T})}(|y^0|_2^2 + |z^0|_2^2), \tag{2.39}$$

by the same arguments as those in the proof of Lemma 2.8, we deduce that

$$\|y_n(t)\|_1^2 + \|z_n(t)\|_1^2 + \|y_n\|^2_{H^{2,1}(Q)} + \|z_n\|^2_{H^{2,1}(Q)} \le e^{Cc_0^{(1)}(1+T+\frac{1}{T})}(\|y^0\|_1^2 + \|z^0\|_1^2).$$

This together with (2.39) indicates that there exist subsequences of $\{y_n\}, \{z_n\}, \{w_n\}$, and $\{v_n\}$, still denoted in the same way, such that

$$\begin{array}{ll} y_n \to y, \ \ z_n \to z & \text{weakly in } \ L^2(0,T;H^2(\Omega)) \cap W^{1,2}(0,T;H), \\ & \text{strongly in } \ L^2(0,T;V) \cap C([0,T];H), \\ w_n \to \widetilde{w}, \ \ v_n \to \widetilde{v} & \text{weakly in } \ H^1(0,T;H). \end{array} \tag{2.40}$$

This implies

$$\widetilde{y} = y \quad \text{a.e. in } \ Q.$$

Passing to the limit $n \to \infty$ in (2.31) and using (2.37)–(2.40), we obtain

$$\| \widetilde{w} \|^2_{H^1(0,T;H)} + \| \widetilde{v} \|^2_{H^1(0,T;H)} \le e^{c_0^{(1)}(1+T+\frac{1}{T})}(|y^0|_2^2 + |z^0|_2^2)$$

and $\widetilde{y} \equiv y^{\widetilde{w},\widetilde{v}} \in \Psi(\widetilde{\xi})$ as desired.

To finish the proof of Theorem 2.1, we fix $\xi \in K_T$, then take $y \in \Psi(\xi)$ and multiply (scalarly in H) $(2.31)_1$ by $-\triangle y'$(here $y' = y_t$), and integrate it over $(0,t)$. After some calculation, we deduce that

$$\begin{aligned} |\triangle y(t)|_2^2 + \int_0^t |\nabla y'(s)|_2^2 \, ds \le C(\|y^0\|^2_{H^2(\Omega)} + |v(t)|_2^2 + |v(0)|_2^2 + |\nabla y(t)|_2^2) \\ + C\int_0^t (|\nabla y|_2^2 + |\nabla z|_2^2 + |\triangle y|_2^2 + |v_s|_2^2) \, ds. \end{aligned} \tag{2.41}$$

Since $y \in \Psi(\xi)$, we have that

$$\|w\|^2_{H^1(0,T;H)} + \|v\|^2_{H^1(0,T;H)} \le e^{c_0^{(1)}(1+T+\frac{1}{T})}(|y^0|_2^2 + |z^0|_2^2). \tag{2.42}$$

By the same arguments as those in Lemma 2.8, we obtain that

$$\begin{aligned} &\|y(t)\|_1^2 + \|z(t)\|_1^2 + \|y\|^2_{H^{2,1}(Q)} + \|z\|^2_{H^{2,1}(Q)} \\ &\le e^{Cc_0^{(1)}(1+T+\frac{1}{T})}(\|y^0\|_1^2 + \|z^0\|_1^2). \end{aligned} \tag{2.43}$$

Here and throughout the proof of Theorem 2.1, C denotes several positive constants independent of T. Now it follows from Lemma 2.7 and (2.41)–(2.43) that

$$\|y(t)\|^2_{L^\infty(0,T;H^2(\Omega))} + \|y\|^2_{H^1(0,T;V)} \le e^{\widetilde{a_0}c_0^{(1)}(1+T+\frac{1}{T})}(\|y^0\|^2_{H^2(\Omega)} + \|z^0\|_1^2) \tag{2.44}$$

and

$$\|w\|^2_{L^\infty(0,T;H)} + \|v\|^2_{L^\infty(0,T;H)} \le e^{\widetilde{a_0}c_0^{(1)}(1+T+\frac{1}{T})}(|y^0|_2^2 + |z^0|_2^2), \tag{2.45}$$

where $\widetilde{a_0} > 0$ is a constant independent of T.

Let $c_0 = \widetilde{a_0}\ c_0^{(1)}$. Then for $0 < \|y^0\|^2_{H^2(\Omega)} + \|z^0\|^2_1 \le \min\{\rho^2, 1\} e^{-c_0(1+T+\frac{1}{T})}$, we have

$$\|w\|^2_{L^\infty(0,T;H)} + \|v\|^2_{L^\infty(0,T;H)} \le \rho^2 \tag{2.46}$$

and

$$\mu(y) = \sqrt{\|y\|^2_{H^1(0,T;V)} + \|y\|^2_{L^\infty(0,T;H^2(\Omega))}} \le 1.$$

Hence

$$\Psi(K_T) \subset K_T.$$

By (2.32), (2.33), (2.34), and the upper-semicontinuity of Ψ, we may use the infinite dimensional Kakutani fixed point theorem to obtain that there exists $y^* \in \Psi(y^*)$ or, equivalently, that there exist $y^* \in L^\infty(0,T;H^2(\Omega)\cap V)\cap H^1(0,T;V)$ and $(w^*, v^*) \in (H^1(0,T;H))^2$, which satisfy system (2.12) and $y^*(T) = z^*(T) = 0$. Moreover, w^* and v^* satisfy estimate (2.46). This completes the proof of Theorem 2.1. □

Proof of Corollary 2.2. It is clear that $f(T) \equiv e^{-c_0(1+T+\frac{1}{T})}$ attains its supremum on $(0, +\infty)$ at $T = 1$. For $(u_0(x), h_0(x)) \in H^1(\Omega) \times H^2(\Omega)$ satisfying condition (2.2), by taking $T = 1$ in Theorem 2.1, Corollary 2.2 follows immediately. This completes the proof of Corollary 2.2. □

Now we are in the position to prove Theorem 2.3.

Proof of Theorem 2.3. We observe first that problem (P) is equivalent to the problem

$$(P') \quad \inf\{T : y^{w,v}(T) = z^{w,v}(T) = 0, \quad (w,v) \in U_\rho\} \equiv \inf Q_T,$$

where $(y^{w,v}, z^{w,v})$ is the solution of the following system:

$$\begin{cases} y_t - a\triangle y + a_1 y + b_1 y^2 + by^3 - cz \\ \qquad\qquad = m(x)v(x,t) \quad \text{in } Q_\infty = \Omega \times (0,\infty), \\ z_t - k\triangle z + la\triangle y + a_2 y + b_2 y^2 - lby^3 + lcz \\ \qquad\qquad = m(x)(w(x,t) - lv(x,t)) \quad \text{in } Q_\infty, \\ y(x,0) = h_0(x) - h_e(x) \equiv y^0(x) \quad \text{in } \Omega, \\ z(x,0) = u_0(x) - u_e(x) \equiv z^0(x) \quad \text{in } \Omega, \\ y(x,t) = z(x,t) = 0 \quad \text{on } \sum_\infty = \partial\Omega \times (0,\infty). \end{cases} \tag{2.47}$$

By Corollary 2.2, it is clear that under the assumptions of Theorem 2.3, the set Q_T is not empty. Let $T^* = \inf Q_T$; then it is clear that $0 \le T^* < \infty$ and there exist a nonincreasing sequence $\{T_m\}$ and a sequence $\{(w_m, v_m)\} \subset U_\rho$ such that $T_m \to T^*, y^{w_m,v_m}(T_m) = z^{w_m,v_m}(T_m) = 0$. Moreover, $(y^{w_m,v_m}, z^{w_m,v_m})$ is the solution to (2.47), where $v = v_m, w = w_m$.

We let $y_m \equiv y^{w_m,v_m}$ and $z_m \equiv z^{w_m,v_m}$; it is clear that

$$\begin{cases} (y_m)_t - a\triangle y_m + a_1 y_m + b_1 y_m^2 + by_m^3 \\ \qquad\qquad - cz_m = m(x)v_m \quad \text{in } Q = \Omega \times (0,T), \\ (z_m)_t - k\triangle z_m + la\triangle y_m + a_2 y_m + b_2 y_m^2 - lby_m^3 \\ \qquad\qquad + lcz_m = m(x)(w_m - lv_m) \quad \text{in } Q, \\ y_m(x,0) = y^0(x), \quad z_m(x,0) = z^0(x) \quad \text{in } \Omega, \\ y_m(x,t) = z_m(x,t) = 0 \quad \text{on } \sum = \partial\Omega \times (0,T), \end{cases} \tag{2.48}$$

where $T > T_1$ is an arbitrary but fixed constant.

Multiplying $(2.48)_1$ and $(2.48)_2$ by y_m and z_m in $L^2(\Omega)$ and integrating them over $(0,t)$, respectively, after some calculation, we obtain

$$\begin{aligned} &|y_m(t)|_2^2 + \int_0^t |\nabla y_m(s)|_2^2\,ds + \int_0^t \|y_m(s)\|_{L^4(\Omega)}^4\,ds \\ &\le C\left(|y^0|_2^2 + \int_0^t |z_m(s)|_2^2\,ds + 1\right) \end{aligned} \tag{2.49}$$

and

$$\begin{aligned} &|z_m(t)|_2^2 + \int_0^t |\nabla z_m(s)|_2^2\,ds \\ &\le C(|y^0|_2^2 + |z^0|_2^2 + 1) + C\int_0^t \int_\Omega y_m^6\,dx\,ds + C\int_0^t |z_m(s)|_2^2\,ds. \end{aligned} \tag{2.50}$$

Multiplying $(2.48)_1$ by $4y_m^3$ in $L^2(\Omega)$, integrating it over $(0,t)$, and using (2.49), we get

$$\begin{aligned} &\|y_m(t)\|_{L^4(\Omega)}^4 + 12a\int_0^t \int_\Omega y_m^2 |\nabla y_m|^2\,dx\,ds + 2b\int_0^t \int_\Omega y_m^6\,dx\,ds \\ &\le C(\|y^0\|_{L^4(\Omega)}^4 + |y^0|_2^2 + 1) + C\int_0^t |z_m(s)|_2^2\,ds, \end{aligned} \tag{2.51}$$

which, combined with (2.50), implies that

$$\begin{aligned} &|z_m(t)|_2^2 + \int_0^T |\nabla z_m(s)|_2^2\,ds \\ &\le C(|y^0|_2^2 + |z^0|_2^2 + \|y^0\|_{L^4(\Omega)}^4 + 1) \quad \forall\, t \in [0,T]. \end{aligned} \tag{2.52}$$

It follows from (2.49), (2.51), and (2.52) that

$$\begin{aligned} &\|y_m(t)\|_{L^4(\Omega)}^4 + |y_m(t)|_2^2 + \int_0^T |\nabla y_m(t)|_2^2\,dt \\ &+ \int_0^T \int_\Omega y_m^2 |\nabla y_m|^2\,dx\,dt + \int_0^T \|y_m\|_{L^6(\Omega)}^6\,dt \\ &\le C(|y^0|_2^2 + |z^0|_2^2 + \|y^0\|_{L^4(\Omega)}^4 + 1). \end{aligned} \tag{2.53}$$

Multiplying $(2.48)_1$ by $-\triangle y_m$ in $L^2(\Omega)$ and integrating it over $(0,t)$, by $(2.48)_1$, (2.52), and (2.53), after some elementary calculation, we get

$$\begin{aligned} &\|y_m(t)\|_1^2 + \|y_m\|_{H^{2,1}(Q)}^2 \\ &\le C(\|y^0\|_1^2 + |z^0|_2^2 + \|y^0\|_{L^4(\Omega)}^4 + 1) \quad \forall\, t \in [0,T]. \end{aligned} \tag{2.54}$$

Similarly, we deduce that

$$\begin{aligned} &\|z_m(t)\|_1^2 + \|z_m\|_{H^{2,1}(Q)}^2 \\ &\le C(\|y^0\|_1^2 + \|z^0\|_1^2 + \|y^0\|_{L^4(\Omega)}^4 + 1) \quad \forall\, t \in [0,T]. \end{aligned} \tag{2.55}$$

By (2.54) and (2.55), we conclude that there exist subsequences of $\{y_m\}$ and $\{z_m\}$, still denoted in the same way, such that

$$\begin{aligned} &y_m \to y \ \text{ and } \ z_m \to z \quad \text{weakly in } \ W^{1,2}([0,T];H) \cap L^2(0,T;H^2(\Omega)), \\ &\qquad\qquad\qquad\quad \text{strongly in } \ L^2(0,T;V) \cap C([0,T];H). \end{aligned} \tag{2.56}$$

Since $(w_m, v_m) \in U_\rho$, there exist subsequences of $\{w_m\}$ and $\{v_m\}$, still denoted by themselves, such that

$$w_m \to w \ \text{ and } \ v_m \to v \text{ weakly in } \ L^2(0,T;H).$$

Moreover, $(w,v) \in U_\rho$. By passing to the limit for $m \to \infty$ in (2.48), we conclude that $(y^{w,v}, z^{w,v})$ and (w,v) satisfy (2.47).

On the other hand, it follows from (2.56) that

$$\begin{aligned} |y(T^*;w,v)|_2 &\le |y(T^*;w,v) - y_m(T^*)|_2 + |y_m(T^*) - y_m(T_m)|_2 \\ &\le |y(T^*;w,v) - y_m(T^*)|_2 + C(T_m - T^*)^{\frac{1}{2}} \to 0, \end{aligned}$$

which implies that $y(T^*;w,v) = 0$. Similarly, we may obtain that $z(T^*;w,v) = 0$ from which we see that $T^* \neq 0$. Otherwise, we should have $y^0(x) = z^0(x) = 0$, which contradicts assumption (2.2). Hence problem (P') has at least one solution. This completes the proof. □

3. Maximum principle. In this section, we shall derive the maximum principle for optimal time control problem (P) (or, equivalently, (P')) in the case that $\omega \equiv \Omega$ and $al^2 - 4k \le 0$ in system (1.1). In what follows, we denote by *sgnq* the signum function of q.

We state first the maximum principle for the case that $n = 1, 2$ as follows.

THEOREM 3.1. *Let (w^*, v^*) be a pair of optimal time controls for system* (1.1) *and (u^*, h^*) be the pair of corresponding optimal states. Then there exists $(p,q) \in (C([0,T^*];H) \cap L^2(0,T^*;V) \cap W^{1,2}([0,T^*];V^*))^2$ such that*

$$\text{(3.1)} \quad \begin{cases} p_t + a\triangle p - la\triangle q - [a_1 + 2b_1(h^* - h_e) \\ \qquad\qquad +3b(h^* - h_e)^2](p - lq) = 0 \quad a.e.\ \ t \in (0,T^*), \\ q_t + k\triangle q - lcq + cp = 0 \quad a.e.\ \ t \in (0,T^*), \end{cases}$$

$$\text{(3.2)} \quad \begin{cases} w^*(t) = \rho sgnq(t) & a.e.\ \ t \in (0,T^*), \\ v^*(t) = \rho sgn(p(t) - lq(t)) & a.e.\ \ t \in (0,T^*), \end{cases}$$

$$\text{(3.3)} \quad \begin{aligned} &-\langle (A+B)(h^*(t) - h_e, u^*(t) - u_e), (p(t), q(t))\rangle \\ &\qquad +\rho(|q(t)|_2 + |p(t) - lq(t)|_2) = 1 \quad a.e.\ \ t \in (0,T^*). \end{aligned}$$

Here T^* is the minimal time, $A(y,z) = (-a\triangle y, la\triangle y - k\triangle z), B(y,z) = (a_1 y + b_1 y^2 + by^3 - cz, a_2 y + b_2 y^2 - lby^3 + lcz)$, and a_1, a_2, b_1, b_2 are given in (2.12).

For the case that $n = 3$, in order to obtain the maximum principle, we need further assumptions. To this end we shall introduce a constant η_0 and a function a_0. By Sobolev's imbedding theorem and the Poincaré inequality, we have that

$$\|y\|_{L^6(\Omega)} \le C_1 |\nabla y|_2 \quad \forall\, y \in V,$$

where $C_1 > 0$ is the best constant such that the above inequality holds. Now we define η_0 and $a_0(b,c,\rho,l)$ as follows:

$$\eta_0 = \frac{[b^2 C_1^2 (C_1^2 + 6\|h_e\| \cdot |\Omega|^{\frac{1}{6}})^2 + 4b\rho]^{\frac{1}{2}} - bC_1(C_1^2 + 6\|h_e\| \cdot |\Omega|^{\frac{1}{6}})}{4bC_1},$$

$$\begin{aligned} &a_0(b,c,\rho,l) \\ &= \exp\left\{e^{10(b\|h_e\|^2 + b + c + 1)}(l + l\|h_e\|^2 + \frac{l}{b} + 1)(\rho^2 + 4) + 4\rho^2 l^2 + 2l|\Omega|^{\frac{1}{3}} C_1^4\right\} \\ &\quad + 8b|\Omega|^{\frac{1}{3}} C_1^4, \end{aligned}$$

where here and throughout section 3, $\|h_e\|$ and $|\Omega|$ denote $\|h_e\|_{L^\infty(\Omega)}$ and the volume of Ω, respectively.

Now we present the maximum principle for the case that $n = 3$.

THEOREM 3.2. *Let (w^*, v^*) be a pair of optimal time controls for system (1.1) and (u^*, h^*) be the pair of corresponding optimal states. If h_0 satisfies $|\nabla(h_0 - h_e)|_2^2 \le \eta_0$ and (2.2), $a \ge \eta_0^{-1} a_0(b, c, \rho, l)$, then there exists $(p, q) \in (C([0, T^*]; H) \cap L^2(0, T^*; V) \cap W^{1,2}([0, T^*]; V^*))^2$ such that (3.1)–(3.3) hold.*

In order to prove Theorems 3.1 and 3.2, we approximate the optimal time control problem (P) by a free optimal time control problem defined later. To this end, we first construct the penalty functional $L_\varepsilon : R^+ \times L^\infty(0, \infty; H) \times L^\infty(0, \infty; H) \longrightarrow R^+$ as follows, which is based on the penalty functionals defined in [2] and [3] after some modification.

$$\begin{aligned} L_\varepsilon(T, w, v) = T &+ \int_0^T \left[g(w) + g(v) + \frac{\varepsilon}{2}(|w(t)|_2^2 + |v(t)|_2^2) \right] dt \\ &+ \frac{1}{2\varepsilon}(|y(T)|_2^2 + |z(T)|_2^2) \\ &+ \frac{1}{2} \int_0^T \left[\left| \int_0^t (w(s) - w^*(s))\, ds \right|_2^2 + \left| \int_0^t (v(s) - v^*(s))\, ds \right|_2^2 \right] dt, \end{aligned} \tag{3.4}$$

where (y, z) is the solution of the system

$$\begin{cases} y_t - a\triangle y + a_1 y + b_1 y^2 + by^3 - cz = v(x,t) \quad \text{in } \Omega \times (0, \infty), \\ z_t - k\triangle z + la\triangle y + a_2 y \\ \qquad\qquad + b_2 y^2 - lby^3 + lcz = w(x,t) - lv(x,t) \quad \text{in } \Omega \times (0, \infty), \\ y(x, 0) = y^0(x), \qquad z(x, 0) = z^0(x) \quad \text{in } \Omega, \\ y(x, t) = z(x, t) = 0 \quad \text{on } \partial\Omega \times (0, \infty) \end{cases} \tag{3.5}$$

and $g : H \longrightarrow \overline{R} = (-\infty, +\infty]$ is defined by

$$g(u) = \begin{cases} 0 & \text{if } |u|_2 \le \rho, \\ +\infty & \text{otherwise.} \end{cases} \tag{3.6}$$

Here a_1, a_2, b_1, and b_2 are given in (2.12).

We consider the approximating optimal control problem as follows:

(P^ε) Minimize $L_\varepsilon(T, w, v)$ over all $(T, w, v) \in R^+ \times L^\infty(0, \infty; H) \times L^\infty(0, \infty; H)$.

First, we show the existence of the optimal solutions for problem (P^ε) for each $\varepsilon > 0$.

LEMMA 3.3. *For each $\varepsilon > 0$, problem (P^ε) has at least one optimal solution.*

Proof. Let $\varepsilon > 0$ be fixed. It is clear that $\inf L_\varepsilon(T, w, v) > -\infty$.

Let $d = \inf\{L_\varepsilon(T, w, v) : (T, w, v) \in R^+ \times L^\infty(0, \infty; H) \times L^\infty(0, \infty; H)\}$ and $\{(T_n, w_n, v_n)\}$ be a minimizing sequence such that

$$d \le L_\varepsilon(T_n, w_n, v_n) \le d + \frac{1}{n}. \tag{3.7}$$

We set

$$\overline{w}_n(s) = \begin{cases} w_n(s) & \text{if } s \in [0, T_n], \\ w^*(s) & \text{otherwise} \end{cases} \tag{3.8}$$

and

$$\overline{v}_n(s) = \begin{cases} v_n(s) & \text{if } s \in [0, T_n], \\ v^*(s) & \text{otherwise.} \end{cases} \tag{3.9}$$

Let (y_n, z_n) and $(\overline{y}_n, \overline{z}_n)$ be the solutions of system (3.5) corresponding to (w_n, v_n) and $(\overline{w}_n, \overline{v}_n)$, respectively. It is clear that $(\overline{y}_n(s), \overline{z}_n(s)) = (y_n(s), z_n(s))$ for $s \in [0, T_n]$. This together with (3.4) and (3.7)–(3.9) implies that

$$d \le L_\varepsilon(T_n, \overline{w}_n, \overline{v}_n) = L_\varepsilon(T_n, w_n, v_n) \le d + \frac{1}{n}, \tag{3.10}$$

which, combined with (3.8) and (3.9), implies that

$$\limsup_{n\to\infty} T_n \le d \ \text{ and } \ (\overline{w}_n, \overline{v}_n) \in U_\rho.$$

So there exists a subsequence of $\{n\}$, still denoted by itself, such that

$$T_n \to T_0, \ \ \overline{w}_n \to \overline{w}, \ \text{ and } \ \overline{v}_n \to \overline{v} \text{ weak star in } \ L^\infty(0, \infty; H), \tag{3.11}$$

where $(\overline{w}, \overline{v}) \in U_\rho$. By the same arguments as those in the proof of Theorem 2.3 (see (2.54) and (2.55)), we get

$$\begin{aligned} &\overline{y}_n \to \overline{y} \text{ and } \ \overline{z}_n \to \overline{z} \\ &\qquad \text{weakly in } \ H^1([0, T_0+1]; H) \cap L^2(0, T_0+1; H^2(\Omega)), \\ &\qquad \text{strongly in } \ L^2(0, T_0+1; V) \cap C([0, T_0+1]; H), \end{aligned} \tag{3.12}$$

where $(\overline{y}, \overline{z})$ is the solution of (3.5) corresponding to $(\overline{w}, \overline{v})$. These together with (3.10), (3.11), and the weakly lower semicontinuity of the convex integrand indicate that

$$\begin{aligned} &T_0 + \int_0^{T_0} \left[g(\overline{w}) + g(\overline{v}) + \frac{\varepsilon}{2}(|\overline{w}(t)|_2^2 + |\overline{v}(t)|_2^2) \right] dt + \frac{1}{2\varepsilon}(|\overline{y}(T_0)|_2^2 + |\overline{z}(T_0)|_2^2) \\ &+ \frac{1}{2} \int_0^{T_0} \left[\left| \int_0^t (\overline{w}(s) - w^*(s))\, ds \right|_2^2 + \left| \int_0^t (\overline{v}(s) - v^*(s))\, ds \right|_2^2 \right] dt = d. \end{aligned} \tag{3.13}$$

Hence $L_\varepsilon(T_0, \overline{w}, \overline{v}) = d$, which completes the proof of Lemma 3.3. □

LEMMA 3.4. *Let $(T_\varepsilon, w_\varepsilon, v_\varepsilon)$ be optimal for problem (P^ε). Then for $\varepsilon \to 0$, we have*

$$T_\varepsilon \to T^*, \ \ \lim_{\varepsilon\to 0} \int_0^{T_\varepsilon} \left[\left| \int_0^t (w_\varepsilon - w^*)\, ds \right|_2^2 + \left| \int_0^t (v_\varepsilon - v^*)\, ds \right|_2^2 \right] dt = 0$$

and for all $T < T^*$,

$$w_\varepsilon \to w^* \ \text{ and } \ v_\varepsilon \to v^* \ \text{ weak star in } \ L^\infty(0, T; H),$$

$$\begin{aligned} &y_\varepsilon \to y^* \ \text{ and } \ z_\varepsilon \to z^* \\ &\qquad \text{weakly in } \ H^1([0, T]; H) \cap L^2(0, T; H^2(\Omega)), \\ &\qquad \text{strongly in } \ L^2(0, T; V) \cap C([0, T]; H), \end{aligned}$$

where $(y_\varepsilon, z_\varepsilon)$ is the solution of (3.5) *corresponding to $(w_\varepsilon, v_\varepsilon)$.*

Proof. Let

$$\overline{w}_\varepsilon(s) = \begin{cases} w_\varepsilon(s) & \text{if } s \in [0, T_\varepsilon], \\ w^*(s) & \text{otherwise,} \end{cases} \tag{3.14}$$

$$\overline{v}_\varepsilon(s) = \begin{cases} v_\varepsilon(s) & \text{if } s \in [0, T_\varepsilon], \\ v^*(s) & \text{otherwise,} \end{cases} \tag{3.15}$$

and $(\overline{y}_\varepsilon(s), \overline{z}_\varepsilon(s))$ be the solution of (3.5) corresponding to $(\overline{w}_\varepsilon(s), \overline{v}_\varepsilon(s))$. It is clear that $(\overline{y}_\varepsilon(s), \overline{z}_\varepsilon(s)) = (y_\varepsilon(s), z_\varepsilon(s))$ for $s \in [0, T_\varepsilon]$.

By optimality of $(T_\varepsilon, w_\varepsilon, v_\varepsilon)$, we have

$$L_\varepsilon(T_\varepsilon, \overline{w}_\varepsilon, \overline{v}_\varepsilon) = L_\varepsilon(T_\varepsilon, w_\varepsilon, v_\varepsilon) \le L_\varepsilon(T^*, w^*, v^*), \tag{3.16}$$

which, combined with (3.14) and (3.15), yields that

$$(\overline{w}_\varepsilon, \overline{v}_\varepsilon) \in U_\rho, \qquad \limsup_{\varepsilon \to 0} T_\varepsilon \le T^*, \tag{3.17}$$

and

$$\overline{y}_\varepsilon(T_\varepsilon) \to 0, \quad \overline{z}_\varepsilon(T_\varepsilon) \to 0 \text{ strongly in } H. \tag{3.18}$$

Thus, on a subsequence of $\{\varepsilon\}$, still denoted by itself, we have

$$T_\varepsilon \to T_0, \quad \overline{w}_\varepsilon \to \overline{w}, \text{ and } \overline{v}_\varepsilon \to \overline{v} \text{ weak star in } L^\infty(0, \infty; H), \tag{3.19}$$

where $(\overline{w}, \overline{v}) \in U_\rho$. By the same arguments as those in the proof of Theorem 2.3 (see (2.54) and (2.55)), we get

$$\begin{aligned} &\overline{y}_\varepsilon \to \overline{y} \text{ and } \overline{z}_\varepsilon \to \overline{z} \\ &\qquad \text{weakly in } H^1([0, T_0 + 1]; H) \cap L^2(0, T_0 + 1; H^2(\Omega)), \\ &\qquad \text{strongly in } L^2(0, T_0 + 1; V) \cap C([0, T_0 + 1]; H), \end{aligned} \tag{3.20}$$

where $(\overline{y}, \overline{z})$ is the solution of (3.5) corresponding to $(\overline{w}, \overline{v})$. These together with (3.14), (3.15), (3.18), and (3.19) indicate that for all $T < T_0$,

$$w_\varepsilon \to \overline{w} \text{ and } v_\varepsilon \to \overline{v} \text{ weak star in } L^\infty(0, T; H), \tag{3.21}$$

$$\begin{aligned} &y_\varepsilon \to \overline{y} \text{ and } z_\varepsilon \to \overline{z} \\ &\qquad \text{weakly in } H^1([0, T]; H) \cap L^2(0, T; H^2(\Omega) \cap V), \\ &\qquad \text{strongly in } L^2(0, T; V) \cap C([0, T]; H), \end{aligned} \tag{3.22}$$

$$\overline{y}(T_0) = 0, \qquad \overline{z}(T_0) = 0.$$

Hence $(\overline{w}, \overline{v})$ is admissible.

By (3.16), (3.17), (3.19), (3.21), (3.22), and the weakly lower semicontinuity of the convex integrand, we get that $T_0 = T^*, (\overline{w}(s), \overline{v}(s)) = (w^*(s), v^*(s))$, and $(\overline{y}(s), \overline{z}(s)) = (y^*(s), z^*(s))$ for a.e. $s < T^*$.

This completes the proof of Lemma 3.4. □

LEMMA 3.5. *Let $(T_\varepsilon, w_\varepsilon, v_\varepsilon)$ be optimal for problem (P^ε). Then there exists $(p_\varepsilon, q_\varepsilon) \in (H^1([0,T_\varepsilon];H) \cap C([0,T_\varepsilon];V))^2$ such that*

$$
\begin{cases}
p_\varepsilon' + a\triangle p_\varepsilon - la\triangle q_\varepsilon \\
\qquad -(a_1 + 2b_1 y_\varepsilon + 3by_\varepsilon^2)(p_\varepsilon - lq_\varepsilon) = 0 & \text{a.e. } t \in (0,T_\varepsilon), \\
q_\varepsilon' + k\triangle q_\varepsilon - lcq_\varepsilon + cp_\varepsilon = 0 & \text{a.e. } t \in (0,T_\varepsilon), \\
p_\varepsilon(T_\varepsilon) = -\frac{1}{\varepsilon} y_\varepsilon(T_\varepsilon) & \text{in } \Omega, \\
q_\varepsilon(T_\varepsilon) = -\frac{1}{\varepsilon} z_\varepsilon(T_\varepsilon) & \text{in } \Omega,
\end{cases} \tag{3.23}
$$

$$
q_\varepsilon(t) \in \partial g(w_\varepsilon(t)) + \varepsilon w_\varepsilon(t) + \int_t^{T_\varepsilon} ds \int_0^s (w_\varepsilon - w^*)\, d\tau \quad \forall\, t \in (0,T_\varepsilon), \tag{3.24}
$$

$$
\begin{aligned}
&p_\varepsilon(t) - lq_\varepsilon(t) \\
&\in \partial g(v_\varepsilon(t)) + \varepsilon v_\varepsilon(t) + \int_t^{T_\varepsilon} ds \int_0^s (v_\varepsilon - v^*)\, d\tau \quad \forall\, t \in (0,T_\varepsilon),
\end{aligned} \tag{3.25}
$$

$$
\begin{aligned}
1 \le{}& \frac{\varepsilon}{2}(|w_\varepsilon(T_\varepsilon)|_2^2 + |v_\varepsilon(T_\varepsilon)|_2^2) + \rho|p_\varepsilon(T_\varepsilon) - lq_\varepsilon(T_\varepsilon) - \varepsilon v_\varepsilon(T_\varepsilon)|_2 \\
&+\rho|q_\varepsilon(T_\varepsilon) - \varepsilon w_\varepsilon(T_\varepsilon)|_2 - \langle (A+B)(y_\varepsilon(T_\varepsilon), z_\varepsilon(T_\varepsilon)), (p_\varepsilon(T_\varepsilon), q_\varepsilon(T_\varepsilon))\rangle \\
&\le 1 + \frac{1}{2}\left[\left| \int_0^{T_\varepsilon} (w_\varepsilon - w^*)\, ds \right|_2^2 + \left| \int_0^{T_\varepsilon} (v_\varepsilon - v^*)\, ds \right|_2^2 \right],
\end{aligned} \tag{3.26}
$$

where $\partial g(u) = \{w \in H, \langle w, u - v\rangle \ge 0$ for all $v \in H,\ |v|_2 \le \rho\}$.

Proof. Since $(T_\varepsilon, w_\varepsilon, v_\varepsilon)$ is optimal for problem (P^ε), we have

$$
L_\varepsilon(T_\varepsilon, w_\varepsilon, v_\varepsilon) \le L_\varepsilon(T_\varepsilon, w_\varepsilon + \lambda w, v_\varepsilon + \lambda v)
$$

for all $(w,v) \in (L^\infty(0,\infty;H))^2$ and $\lambda > 0$, which yields

$$
\begin{aligned}
&\int_0^{T_\varepsilon} [g'(w_\varepsilon; w) + g'(v_\varepsilon; v) + \varepsilon(\langle w_\varepsilon, w\rangle + \langle v_\varepsilon, v\rangle)]\, dt \\
&+\frac{1}{\varepsilon}(\langle \varphi_\varepsilon(T_\varepsilon), y_\varepsilon(T_\varepsilon)\rangle + \langle \psi_\varepsilon(T_\varepsilon), z_\varepsilon(T_\varepsilon)\rangle) \\
&+ \int_0^{T_\varepsilon} \left[\left\langle \int_0^t (w_\varepsilon - w^*)\, ds, \int_0^t w\, ds \right\rangle + \left\langle \int_0^t (v_\varepsilon - v^*)\, ds, \int_0^t v\, ds \right\rangle \right] dt \ge 0,
\end{aligned} \tag{3.27}
$$

where g' is the directional derivative of g, and $(\varphi_\varepsilon, \psi_\varepsilon) \in (H^1([0,T_\varepsilon];H) \cap L^2(0,T_\varepsilon; H^2(\Omega) \cap V))^2$ is the solution to the variational system of (3.5) at $(y_\varepsilon, z_\varepsilon)$, i.e.,

$$
\begin{cases}
\varphi_\varepsilon' - a\triangle\varphi_\varepsilon + (a_1 + 2b_1 y_\varepsilon + 3by_\varepsilon^2)\varphi_\varepsilon - c\psi_\varepsilon = v & \text{in } \Omega \times (0,T_\varepsilon), \\
\psi_\varepsilon' - k\triangle\psi_\varepsilon + la\triangle\varphi_\varepsilon \\
\qquad +(a_2 + 2b_2 y_\varepsilon - 3lby_\varepsilon^2)\varphi_\varepsilon + lc\psi_\varepsilon = w - lv & \text{in } \Omega \times (0,T_\varepsilon), \\
\varphi_\varepsilon = \psi_\varepsilon = 0 & \text{on } \partial\Omega \times (0,T_\varepsilon), \\
\varphi_\varepsilon(x,0) = \psi_\varepsilon(x,0) = 0 & \text{in } \Omega.
\end{cases} \tag{3.28}
$$

Let $(p_\varepsilon, q_\varepsilon) \in (H^1([0,T_\varepsilon];H) \cap L^2(0,T_\varepsilon; H^2(\Omega) \cap V))^2$ be the solution to (3.23). (The existence of the solution to (3.23) follows the same arguments as in [17].)

Multiplying $(3.28)_1$ (the first equation of (3.28)) by p_ε scalarly in H, then integrating it over $(0, T_\varepsilon)$, and using $(3.23)_1$ (the first equation of (3.23)), we obtain

$$\begin{aligned}&\langle \varphi_\varepsilon(T_\varepsilon), p_\varepsilon(T_\varepsilon)\rangle\\ &= \int_0^{T_\varepsilon} \langle \varphi_\varepsilon, la\triangle q_\varepsilon + (a_2 + 2b_2 y_\varepsilon - 3lby_\varepsilon^2) q_\varepsilon\rangle\, dt\\ &\quad + \int_0^{T_\varepsilon} \langle v, p_\varepsilon\rangle\, dt + \int_0^{T_\varepsilon} \langle c\psi_\varepsilon, p_\varepsilon\rangle\, dt.\end{aligned} \tag{3.29}$$

Similarly, we have

$$\begin{aligned}&\langle \psi_\varepsilon(T_\varepsilon), q_\varepsilon(T_\varepsilon)\rangle\\ &= -\int_0^{T_\varepsilon} \langle \varphi_\varepsilon, la\triangle q_\varepsilon + (a_2 + 2b_2 y_\varepsilon - 3lby_\varepsilon^2) q_\varepsilon\rangle\, dt\\ &\quad + \int_0^{T_\varepsilon} \langle w - lv, q_\varepsilon\rangle\, dt - \int_0^{T_\varepsilon} \langle c\psi_\varepsilon, p_\varepsilon\rangle\, dt,\end{aligned}$$

which, combined with (3.27) and (3.29), indicates that

$$\begin{aligned}&\int_0^{T_\varepsilon} [g'(w_\varepsilon; w) + g'(v_\varepsilon; v) + \varepsilon(\langle w_\varepsilon, w\rangle + \langle v_\varepsilon, v\rangle)]\, dt\\ &- \int_0^{T_\varepsilon} \langle v, p_\varepsilon - lq_\varepsilon\rangle\, dt - \int_0^{T_\varepsilon} \langle w, q_\varepsilon\rangle\, dt\\ &+ \int_0^{T_\varepsilon} \left[\left\langle \int_0^t (w_\varepsilon - w^*)\, ds, \int_0^t w\, ds\right\rangle + \left\langle \int_0^t (v_\varepsilon - v^*)\, ds, \int_0^t v\, ds\right\rangle\right] dt \ge 0.\end{aligned} \tag{3.30}$$

Letting $v = 0$ in (3.30), we get

$$\begin{aligned}&\int_0^{T_\varepsilon} [g'(w_\varepsilon; w) + \varepsilon\langle w_\varepsilon, w\rangle]\, dt - \int_0^{T_\varepsilon} \langle w, q_\varepsilon\rangle\, dt\\ &+ \int_0^{T_\varepsilon} \left\langle w(t), \int_t^{T_\varepsilon} ds \int_0^s (w_\varepsilon - w^*)\, d\tau\right\rangle dt \ge 0 \quad \forall\, w \in L^\infty(0, \infty; H).\end{aligned} \tag{3.31}$$

Letting $w = 0$ in (3.30), we deduce that

$$\begin{aligned}&\int_0^{T_\varepsilon} [g'(v_\varepsilon; v) + \varepsilon\langle v_\varepsilon, v\rangle]\, dt + \int_0^{T_\varepsilon} \langle lq_\varepsilon - p_\varepsilon, v\rangle\, dt\\ &+ \int_0^{T_\varepsilon} \left\langle v(t), \int_t^{T_\varepsilon} ds \int_0^s (v_\varepsilon - v^*)\, d\tau\right\rangle dt \ge 0 \quad \forall\, v \in L^\infty(0, \infty; H).\end{aligned}$$

This together with (3.31) shows (3.24) and (3.25).

It remains to prove (3.26). We note first that

$$L_\varepsilon(T_\varepsilon, w_\varepsilon, v_\varepsilon) \le L_\varepsilon(T_\varepsilon - \lambda, w_\varepsilon, v_\varepsilon) \quad \forall\, \lambda \in (0, T_\varepsilon),$$

which indicates that

$$\begin{aligned}&\frac{\varepsilon}{2}\int_0^{T_\varepsilon} (|w_\varepsilon(t)|_2^2 + |v_\varepsilon(t)|_2^2)\, dt + \frac{1}{2\varepsilon}(|y_\varepsilon(T_\varepsilon)|_2^2 + |z_\varepsilon(T_\varepsilon)|_2^2)\\ &\le -\lambda + \frac{\varepsilon}{2}\int_0^{T_\varepsilon - \lambda} (|w_\varepsilon(t)|_2^2 + |v_\varepsilon(t)|_2^2)\, dt\\ &\quad + \frac{1}{2\varepsilon}(|y_\varepsilon(T_\varepsilon - \lambda)|_2^2 + |z_\varepsilon(T_\varepsilon - \lambda)|_2^2).\end{aligned} \tag{3.32}$$

On the other hand, since $(\varepsilon I + \partial g)^{-1}$ is Lipschitz continuous on H, it follows from (3.24) and (3.25) that $w_\varepsilon, v_\varepsilon$ is Hölder continuous on $[0, T_\varepsilon]$. Then it follows from (3.5) that $(y'_\varepsilon, z'_\varepsilon) \in (C([0, T_\varepsilon]; H))^2$ (cf. [17]). Thus by (3.32), we get that

$$\frac{\varepsilon}{2}(|w_\varepsilon(T_\varepsilon)|_2^2 + |v_\varepsilon(T_\varepsilon)|_2^2) - \langle y'_\varepsilon(T_\varepsilon), p_\varepsilon(T_\varepsilon)\rangle - \langle z'_\varepsilon(T_\varepsilon), q_\varepsilon(T_\varepsilon)\rangle \le -1. \tag{3.33}$$

Let

$$w_\varepsilon^\lambda(s) = \begin{cases} w_\varepsilon(T_\varepsilon) & \text{if } \ s \in (T_\varepsilon, T_\varepsilon + \lambda), \\ w_\varepsilon(s) & \text{otherwise} \end{cases}$$

and

$$v_\varepsilon^\lambda(s) = \begin{cases} v_\varepsilon(T_\varepsilon) & \text{if } \ s \in (T_\varepsilon, T_\varepsilon + \lambda), \\ v_\varepsilon(s) & \text{otherwise.} \end{cases}$$

By the optimality of $(T_\varepsilon, w_\varepsilon, v_\varepsilon)$, we get

$$L_\varepsilon(T_\varepsilon, w_\varepsilon, v_\varepsilon) \le L_\varepsilon(T_\varepsilon + \lambda, w_\varepsilon^\lambda, v_\varepsilon^\lambda) \quad \forall\, \lambda > 0,$$

which indicates that

$$\begin{aligned} &\frac{\varepsilon}{2}(|w_\varepsilon(T_\varepsilon)|_2^2 + |v_\varepsilon(T_\varepsilon)|_2^2) - \langle y'_\varepsilon(T_\varepsilon), p_\varepsilon(T_\varepsilon)\rangle - \langle z'_\varepsilon(T_\varepsilon), q_\varepsilon(T_\varepsilon)\rangle \\ &\ge -1 - \frac{1}{2}\left[\left|\int_0^{T_\varepsilon}(w_\varepsilon - w^*)\,ds\right|_2^2 + \left|\int_0^{T_\varepsilon}(v_\varepsilon - v^*)\,ds\right|_2^2\right]. \end{aligned} \tag{3.34}$$

It follows from (3.24) and (3.25) that

$$\rho|q_\varepsilon(T_\varepsilon) - \varepsilon w_\varepsilon(T_\varepsilon)|_2 = \langle w_\varepsilon(T_\varepsilon), q_\varepsilon(T_\varepsilon) - \varepsilon w_\varepsilon(T_\varepsilon)\rangle$$

and

$$\rho|p_\varepsilon(T_\varepsilon) - lq_\varepsilon(T_\varepsilon) - \varepsilon v_\varepsilon(T_\varepsilon)|_2 = \langle v_\varepsilon(T_\varepsilon), p_\varepsilon(T_\varepsilon) - lq_\varepsilon(T_\varepsilon) - \varepsilon v_\varepsilon(T_\varepsilon)\rangle,$$

which, combined with (3.5), (3.33), and (3.34), imply (3.26) as desired. This completes the proof of Lemma 3.5. □

Now we turn to the proof of Theorem 3.1.

Proof of Theorem 3.1. By (3.26) and the condition $al^2 - 4k \le 0$, we deduce that

$$\begin{aligned}
&\tfrac{\varepsilon}{2}(|w_\varepsilon(T_\varepsilon)|_2^2 + |v_\varepsilon(T_\varepsilon)|_2^2) \\
&\quad +\rho|q_\varepsilon(T_\varepsilon) - \varepsilon w_\varepsilon(T_\varepsilon)|_2 + \rho|p_\varepsilon(T_\varepsilon) - lq_\varepsilon(T_\varepsilon) - \varepsilon v_\varepsilon(T_\varepsilon)|_2 \\
&\le 1 + \tfrac{1}{2}[|\textstyle\int_0^{T_\varepsilon}(w_\varepsilon - w^*)\,ds|_2^2 + |\int_0^{T_\varepsilon}(v_\varepsilon - v^*)\,ds|_2^2] + \langle a_1 y_\varepsilon(T_\varepsilon) - cz_\varepsilon(T_\varepsilon), \\
&\quad p_\varepsilon(T_\varepsilon) - lq_\varepsilon(T_\varepsilon)\rangle + \langle by_\varepsilon^3(T_\varepsilon) + 3bh_e y_\varepsilon^2(T_\varepsilon), p_\varepsilon(T_\varepsilon) - lq_\varepsilon(T_\varepsilon)\rangle \\
&\quad - \tfrac{a|\nabla y_\varepsilon(T_\varepsilon)|_2^2 - la\int_\Omega \nabla y_\varepsilon(T_\varepsilon)\cdot\nabla z_\varepsilon(T_\varepsilon)\,dx + k|\nabla z_\varepsilon(T_\varepsilon)|_2^2}{\varepsilon} \\
&\le 1 + \tfrac{1}{2}[|\textstyle\int_0^{T_\varepsilon}(w_\varepsilon - w^*)\,ds|_2^2 + |\int_0^{T_\varepsilon}(v_\varepsilon - v^*)\,ds|_2^2] + \langle a_1 y_\varepsilon(T_\varepsilon) - cz_\varepsilon(T_\varepsilon), \\
&\quad p_\varepsilon(T_\varepsilon) - lq_\varepsilon(T_\varepsilon)\rangle + \langle by_\varepsilon^3(T_\varepsilon) + 3bh_e y_\varepsilon^2(T_\varepsilon), p_\varepsilon(T_\varepsilon) - lq_\varepsilon(T_\varepsilon)\rangle \\
&\quad - \tfrac{a|\nabla y_\varepsilon(T_\varepsilon)|_2^2 - la|\nabla y_\varepsilon(T_\varepsilon)|_2|\nabla z_\varepsilon(T_\varepsilon)|_2 + k|\nabla z_\varepsilon(T_\varepsilon)|_2^2}{\varepsilon} \\
&= 1 + \tfrac{1}{2}[|\textstyle\int_0^{T_\varepsilon}(w_\varepsilon - w^*)\,ds|_2^2 + |\int_0^{T_\varepsilon}(v_\varepsilon - v^*)\,ds|_2^2] + \langle a_1 y_\varepsilon(T_\varepsilon) - cz_\varepsilon(T_\varepsilon), \\
&\quad p_\varepsilon(T_\varepsilon) - lq_\varepsilon(T_\varepsilon)\rangle + \langle by_\varepsilon^3(T_\varepsilon) + 3bh_e y_\varepsilon^2(T_\varepsilon), p_\varepsilon(T_\varepsilon) - lq_\varepsilon(T_\varepsilon)\rangle \\
&\quad - \tfrac{a(|\nabla y_\varepsilon(T_\varepsilon)|_2 - 2^{-1}l|\nabla z_\varepsilon(T_\varepsilon)|_2)^2 + 4^{-1}(4k - al^2)|\nabla z_\varepsilon(T_\varepsilon)|_2^2}{\varepsilon} \\
&\le 1 + \tfrac{1}{2}[|\textstyle\int_0^{T_\varepsilon}(w_\varepsilon - w^*)\,ds|_2^2 + |\int_0^{T_\varepsilon}(v_\varepsilon - v^*)\,ds|_2^2] + \langle a_1 y_\varepsilon(T_\varepsilon) - cz_\varepsilon(T_\varepsilon), \\
&\quad p_\varepsilon(T_\varepsilon) - lq_\varepsilon(T_\varepsilon)\rangle + \langle by_\varepsilon^3(T_\varepsilon) + 3bh_e y_\varepsilon^2(T_\varepsilon), p_\varepsilon(T_\varepsilon) - lq_\varepsilon(T_\varepsilon)\rangle \\
&\quad + \tfrac{al^2 - 4k}{4\varepsilon}|\nabla z_\varepsilon(T_\varepsilon)|_2^2 \\
&\le 1 + \tfrac{1}{2}[|\textstyle\int_0^{T_\varepsilon}(w_\varepsilon - w^*)\,ds|_2^2 + |\int_0^{T_\varepsilon}(v_\varepsilon - v^*)\,ds|_2^2] + \langle a_1 y_\varepsilon(T_\varepsilon) - cz_\varepsilon(T_\varepsilon), \\
&\quad p_\varepsilon(T_\varepsilon) - lq_\varepsilon(T_\varepsilon)\rangle + \langle by_\varepsilon^3(T_\varepsilon) + 3bh_e y_\varepsilon^2(T_\varepsilon), p_\varepsilon(T_\varepsilon) - lq_\varepsilon(T_\varepsilon)\rangle.
\end{aligned} \tag{3.35}$$

Now we claim that

$$\frac{\varepsilon}{2}(|w_\varepsilon(T_\varepsilon)|_2^2 + |v_\varepsilon(T_\varepsilon)|_2^2) + |p_\varepsilon(T_\varepsilon)|_2 + |q_\varepsilon(T_\varepsilon)|_2 \le C. \tag{3.36}$$

Indeed, by the same way as in the proof of Lemma 3.4, we derive that

$$\left|\int_0^{T_\varepsilon} (w_\varepsilon - w^*)\,ds\right|_2^2 + \left|\int_0^{T_\varepsilon} (v_\varepsilon - v^*)\,ds\right|_2^2 \to 0 \ \text{ as } \ \varepsilon \to 0, \tag{3.37}$$

and as $n = 1, 2$,

$$y_\varepsilon(T_\varepsilon) \to 0, \ \ z_\varepsilon(T_\varepsilon) \to 0 \ \text{ weakly in } \ V \ \text{ and strongly in } \ L^6(\Omega).$$

These together with (3.35) yield (3.36).

Multiplying $(3.23)_2$ by q_ε scalarly in H and integrating it over (t, T_ε), after some elementary calculation, we obtain that

$$|q_\varepsilon(t)|_2^2 + \int_t^{T_\varepsilon} |\nabla q_\varepsilon(s)|_2^2\,ds \le C + C\int_t^{T_\varepsilon} |p_\varepsilon(s)|_2^2\,ds \quad \forall\, t \in [0, T_\varepsilon]. \tag{3.38}$$

Similarly, we deduce that

$$|p_\varepsilon(t)|_2^2 + \int_t^{T_\varepsilon} |\nabla p_\varepsilon(s)|_2^2\,ds \le C + C\int_t^{T_\varepsilon} (|p_\varepsilon(s)|_2^2 + |q_\varepsilon(s)|_2^2)\,ds \quad \forall\, t \in [0, T_\varepsilon],$$

which, combined with (3.38), implies that

$$|p_\varepsilon(t)|_2^2 + |q_\varepsilon(t)|_2^2 + \int_0^{T_\varepsilon} |\nabla p_\varepsilon(s)|_2^2\,ds + \int_0^{T_\varepsilon} |\nabla q_\varepsilon(s)|_2^2\,ds \le C \quad \forall\, t \in [0, T_\varepsilon].$$

In particular, the latter implies that $\{p'_\varepsilon\}$ and $\{q'_\varepsilon\}$ are bounded in $L^2(0, T_\varepsilon; V^*)$. Therefore, there exists $(p, q) \in (L^2(0, T^*; V) \cap W^{1,2}([0, T^*]; V^*))^2 \subset (C([0, T^*]; H))^2$ such that on a subsequence of $\{\varepsilon\}$, still denoted by itself, we have that for all $T < T^*$,

$$\begin{aligned} &p_\varepsilon \to p, \ \ q_\varepsilon \to q \ \text{ strongly in } \ L^2(0, T; H), \ \text{ weakly in } \ L^2(0, T; V), \\ &\qquad p'_\varepsilon \to p' \ \text{ and } \ q'_\varepsilon \to q' \ \text{ weakly in } \ L^2(0, T; V^*). \end{aligned} \tag{3.39}$$

Hence, by passing to the limit for $\varepsilon \to 0$ in (3.23)–(3.25), we obtain (3.1) and (3.2) as desired.

It remains to prove (3.3). By (3.5) and (3.23), it follows that

$$\begin{aligned} &\frac{d}{dt}\langle (A + B)(y_\varepsilon(t), z_\varepsilon(t)), (p_\varepsilon(t), q_\varepsilon(t))\rangle \\ &= \langle (v_\varepsilon(t), w_\varepsilon(t) - l v_\varepsilon(t)), (p'_\varepsilon(t), q'_\varepsilon(t))\rangle \quad \text{a.e. } \ t \in (0, T_\varepsilon). \end{aligned} \tag{3.40}$$

Set

$$U_\varepsilon(t) = \int_t^{T_\varepsilon} ds \int_0^s (w_\varepsilon(\tau) - w^*(\tau))\,d\tau \quad \forall\, t \in [0, T_\varepsilon],$$

$$\overline{U}_\varepsilon(t) = \int_t^{T_\varepsilon} ds \int_0^s (v_\varepsilon(\tau) - v^*(\tau))\,d\tau \quad \forall\, t \in [0, T_\varepsilon],$$

and then (3.24) and (3.25) can be rewritten as

$$\begin{cases} w_\varepsilon(t) = (\varepsilon + \partial g)^{-1}(q_\varepsilon(t) - U_\varepsilon(t)), & \text{a.e. } t \in (0, T_\varepsilon), \\ v_\varepsilon(t) = (\varepsilon + \partial g)^{-1}(p_\varepsilon(t) - lq_\varepsilon(t) - \overline{U}_\varepsilon(t)), & \text{a.e. } t \in (0, T_\varepsilon), \end{cases}$$

and therefore it follows from (3.40) that

$$\begin{aligned} &\frac{d}{dt}(\langle (A+B)(y_\varepsilon(t), z_\varepsilon(t)), (p_\varepsilon(t), q_\varepsilon(t))\rangle \\ &\quad -g_\varepsilon^*(q_\varepsilon(t) - U_\varepsilon(t)) - g_\varepsilon^*(p_\varepsilon(t) - lq_\varepsilon(t) - \overline{U}_\varepsilon(t))) \\ &= \langle (w_\varepsilon(t), v_\varepsilon(t)), (U_\varepsilon'(t), \overline{U}_\varepsilon'(t))\rangle \quad \text{a.e. } t \in (0, T_\varepsilon), \end{aligned}$$

where $g_\varepsilon^*(p) = \sup\{(p,u) - \frac{\varepsilon}{2}|u|_2^2 : \ |u|_2 \le \rho\}$, which implies

$$\begin{aligned} &\langle (A+B)(y_\varepsilon(T_\varepsilon), z_\varepsilon(T_\varepsilon)), (p_\varepsilon(T_\varepsilon), q_\varepsilon(T_\varepsilon))\rangle - g_\varepsilon^*(q_\varepsilon(T_\varepsilon)) - g_\varepsilon^*(p_\varepsilon(T_\varepsilon) - lq_\varepsilon(T_\varepsilon)) \\ &= \langle (A+B)(y_\varepsilon(t), z_\varepsilon(t)), (p_\varepsilon(t), q_\varepsilon(t))\rangle \\ &\quad -g_\varepsilon^*(q_\varepsilon(t) - U_\varepsilon(t)) - g_\varepsilon^*(p_\varepsilon(t) - lq_\varepsilon(t) - \overline{U}_\varepsilon(t)) \\ &\quad + \int_t^{T_\varepsilon} \langle (w_\varepsilon(s), v_\varepsilon(s)), (U_\varepsilon'(s), \overline{U}_\varepsilon'(s))\rangle \, ds \quad \forall\, t \in [0, T_\varepsilon). \end{aligned}$$

This together with (3.26), Lemma 3.4, and (3.37) shows that

$$\begin{aligned} &\langle (A+B)(y_\varepsilon(t), z_\varepsilon(t)), (p_\varepsilon(t), q_\varepsilon(t))\rangle \\ &\to \rho|q(t)|_2 + \rho|p(t) - lq(t)|_2 - 1 \quad \text{as } \ \varepsilon \to 0 \ \forall\, t \in (0, T^*). \end{aligned} \tag{3.41}$$

On the other hand, by Lemma 3.4 and (3.39), we see that

$$\begin{aligned} &\langle (A+B)(y_\varepsilon(t), z_\varepsilon(t)), (p_\varepsilon(t), q_\varepsilon(t))\rangle \\ &\to \langle (A+B)(y^*(t), z^*(t)), (p(t), q(t))\rangle \quad \text{a.e. in } \ (0, T^*). \end{aligned}$$

Then by (3.41), we get (3.3) as desired.

This completes the proof of Theorem 3.1. □

Finally, we prove Theorem 3.2.

Proof of Theorem 3.2. We only need to prove that (3.36) holds. The rest of the proof is the same as the proof of Theorem 3.1.

Indeed, by the same arguments as those in the proof of Theorem 2.3, we obtain

$$|\nabla y_\varepsilon(t)|_2^2 \le |\nabla y^0|_2^2 + \frac{a_0(b,c,\rho,l)}{a},$$

which yields that

$$\begin{aligned} &|by_\varepsilon^3(T_\varepsilon) + 3bh_e y_\varepsilon^2(T_\varepsilon)|_2 \\ &\le b|y_\varepsilon^3(T_\varepsilon)|_2 + 3b\|h_e\| \cdot |y_\varepsilon^2(T_\varepsilon)|_2 \\ &\le b\|y_\varepsilon(T_\varepsilon)\|_{L^6(\Omega)}^3 + 3b\|h_e\| \cdot |\Omega|^{\frac{1}{6}} \cdot \|y_\varepsilon(T_\varepsilon)\|_{L^6(\Omega)}^2 \\ &\le bC_1^2 |\nabla y_\varepsilon(T_\varepsilon)|_2^2 (C_1 |\nabla y_\varepsilon(T_\varepsilon)|_2 + 3\|h_e\| \cdot |\Omega|^{\frac{1}{6}}) \\ &\le bC_1^2 \left[C_1 \left(|\nabla y^0|_2^2 + \frac{a_0(b,c,\rho,l)}{a} \right)^{\frac{1}{2}} + 3\|h_e\| \cdot |\Omega|^{\frac{1}{6}} \right] \left(|\nabla y^0|_2^2 + \frac{a_0(b,c,\rho,l)}{a} \right). \end{aligned}$$

So for $|\nabla y_0|_2^2 \le \eta_0$ and $a \ge \eta_0^{-1} a_0(b, c, \rho, l)$, one may derive that

$$\begin{aligned} &|by_\varepsilon^3(T_\varepsilon) + 3bh_e y_\varepsilon^2(T_\varepsilon)|_2 \\ &\le bC_1^2[(2\eta_0)^{\frac{1}{2}} C_1 + 3\|h_e\| \cdot |\Omega|^{\frac{1}{6}}]2\eta_0 \\ &\le bC_1^2[2\eta_0 + C_1^2 + 6\|h_e\| \cdot |\Omega|^{\frac{1}{6}}]\eta_0 \\ &= 2bC_1^2\eta_0^2 + bC_1^2(C_1^2 + 6\|h_e\| \cdot |\Omega|^{\frac{1}{6}})\eta_0 \\ &\le \frac{\rho}{2}. \end{aligned} \tag{3.42}$$

On the other hand, it follows from Lemma 3.4 that

$$y_\varepsilon(T_\varepsilon) \to 0, \quad z_\varepsilon(T_\varepsilon) \to 0 \ \text{ weakly in } \ V \ \text{ and strongly in } \ H,$$

which, combined with (3.35), (3.37), and (3.42), imply (3.36). This completes the proof of Theorem 3.2. □

Remark. It should be noted that in order to obtain the maximum principle for problem (P), we put controls on the whole domain Ω and in both equations of the system (1.1) in section 3. This is because, in general, one approximates an optimal time control problem by a free optimal time control problem in order to derive the maximum principle (see, for instance, [2] and [16]). Thus one needs to estimate the values of the adjoint states at T_ε. However, such an estimate seems to depend on the values of the controls on the whole domain Ω. More precisely, for the phase-field system, in order to get the boundness of $\{(p_\varepsilon(T_\varepsilon), q_\varepsilon(T_\varepsilon))\}$ in $L^2(\Omega) \times L^2(\Omega)$, one needs to have both controls in two equations, which are defined on Ω.

Acknowledgment. The authors thank the referee for his (or her) careful reading of this paper and valuable suggestions.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] V. Barbu, *Analysis and Control of Nonlinear Infinite Dimensional Systems*, Academic Press, New York, 1993.
[3] V. Barbu, *The time optimal control of Navier-Stokes equations*, Systems Control Lett., 30 (1997), pp. 93–100.
[4] V. Barbu, *Local controllability of the phase field system*, Nonlinear Anal., 50 (2002), pp. 363–372.
[5] V. Barbu, *Controllability of parabolic and Navier-Stokes equations*, Sci. Math. Jpn., 6 (2002), pp. 143–211.
[6] G. Caginalp, *An analysis of a phase field model of a free boundary*, Arch. Ration. Mech. Anal., 92 (1986), pp. 205–245.
[7] C. Fabre, J. P. Puel, and E. Zuazua, *Approximate controllability of the semilinear heat equation*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 31–61.
[8] E. Fernandez-Cara, *Null controllability of the semilinear heat equations*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 87–107.
[9] E. Fernandez-Cara, *Null controllability for semilinear parabolic equations with critical growth of the nonlinearity*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 1371–1376.
[10] E. Fernandez and E. Zuazua, *The cost of approximate controllability for heat equations: The linear case*, Adv. Differential Equations, 5 (2000), pp. 465–514.
[11] A. V. Fursikov and O. Yu. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Ser. 34, Seoul National University, Korea, 1996.
[12] A. V. Fursikov and O. Yu. Imanuvilov, *Local exact boundary controllability of the Boussinesq equation*, SIAM J. Control Optim., 36 (1998), pp. 391–421.
[13] J. Henry, *Etude de la contróllabilité de certaines équations paraboliques*, Thèse d'Etat, Université Paris-VI, Paris, France, 1978.
[14] G. Lebeau and L. Robbiano, *Contrôle exact de l'eq́uation de la chaleur*, Comm. Partial Differential Equations, 20 (1995), pp. 335–356.

[15] J. L. Lions, *Remarks on approximate controllability*, J. Anal. Math., 59 (1992), pp. 103–116.
[16] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser Boston, Cambridge, MA, 1995.
[17] C. Morosanu and D. Motreanu, *A generalized phase-field system*, J. Math. Anal. Appl., 237 (1999), pp. 515–540.
[18] D. L. Russell, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., 52 (1973), pp. 189–211.
[19] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.
[20] L. J. Wang and G. S. Wang, *Local internal controllability of the Boussinesq system*, Nonlinear Anal., 53 (2003), pp. 637–652.
[21] E. Zuazua, *Finite dimensional null controllability of the semilinear heat equation*, J. Math. Pures Appl. (9), 76 (1997), pp. 237–264.
[22] E. Zuazua, *Approximate controllability of the semilinear heat equation: Boundary control*, in Proceedings of the International Symposium of Computational Science for the 21st Century, Dedicated to Prof. Roland Glowinski on the Occasion of His 60th Birthday, John Wiley and Sons, Chichester, UK, 1997, pp. 738–747.
[23] E. Zuazua, *Approximate controllability for semilinear heat equations with globally Lipschitz nonlinearities*, Control Cybernet., 28 (1999), pp. 665–683.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 4, pp. 1509–1522

ON THE GENERICITY OF THE OBSERVABILITY OF UNCONTROLLED DISCRETE NONLINEAR SYSTEMS*

J.-C. VIVALDA†

Abstract. In this paper, it is proven that, generically, a nonlinear discrete system given on a compact manifold X is observable.

Key words. observability, discrete-time systems, transversality

AMS subject classifications. 57R40, 93B07, 93B27, 93C55

DOI. 10.1137/S0363012902403997

1. Introduction. This paper deals with the question of genericity of observability for discrete-time nonlinear system such that

$$\left\{\begin{array}{rcl} x_{k+1} & = & f(x_k), \\ y_k & = & h(x_k), \\ x_k \in X, & & y_k \in \mathbb{R}, \end{array}\right. \tag{1}$$

where X is a C^∞ connected n-dimensional manifold, f is a diffeomorphism, and $h : X \to \mathbb{R}$ is a C^∞ mapping.

There are many notions of observability; among them the weaker is the following. We will say that (1) is observable if, given two initial conditions x_0 and $\bar{x}_0$, there exists $k \in \mathbb{N}^*$ such that $h(x_k) \neq h(\bar{x}_k)$. Our main result is about the stronger notion of strong observability defined below.

DEFINITION 1. *Denoting by f^s, $f \circ \cdots \circ f$ s times, we say that system* (1) *is strongly observable if the map*

$$\begin{array}{rccl} \Theta_{f,h} : & X & \longrightarrow & \mathbb{R}^{2n+1}, \\ & x & \longmapsto & (h(x), h \circ f(x), \ldots, h \circ f^{2n}(x)) \end{array}$$

is an embedding.

We will prove that the set of pairs $(f, h) \in \mathrm{Diff}(X) \times C^\infty(X, \mathbb{R})$ such that system (1) is strongly observable is residual (for the C^∞ topology).

On this subject, we have to mention an important work from Gauthier and Kupka; in [4] the authors (together with Hammouri) proved the genericity of the observability for uncontrolled continuous-time systems; in [5] the same result is established for controlled continuous-time systems.

We want to compare this work with works on the same subject in the scientific literature: we begin with a paper written by Aeyels [2]. In this article, the author considers an uncontrolled system given on a compact manifold X such as

$$\left\{\begin{array}{l} \dot{x} = f(x), \\ y = h(x) \end{array}\right. \tag{2}$$

*Received by the editors March 13, 2002; accepted for publication (in revised form) April 20, 2003; published electronically October 28, 2003. Research for this work was supported by Inria-Lorraine (CONGE project) and UMR CNRS 7122.

http://www.siam.org/journals/sicon/42-4/40399.html

†Inria-Lorraine, ISGMP Bât. A, Université de Metz, Ile du Saulcy, 57045 Metz Cedex 01, France (vivalda@loria.fr).

and works with the discretized vector field f. He introduces the notion of P-observability.

DEFINITION 2 (Aeyels). *A time $T > 0$ being given, P is a finite subset of $[0, T]$. System (2) is said to be P-observable iff for each $(x, y) \in X^2$, $x \neq y$, there exists a $t_i \in P$ such that $h \circ \Phi_{t_i}(x) \neq h \circ \Phi_{t_i}(y)$, where Φ denotes the flow of f.*

Two results are proved in this paper.

THEOREM 1 (Aeyels). *There exists an open and dense set of vector fields such that, a vector field f in this set being fixed, the subset of functions h belonging to $C^r(X, \mathbb{R})$ such that the system (f, h) is P-observable is open and dense in $C^r(X, \mathbb{R})$. This is true for almost any finite subset P of $(2 \dim X + 1)$ points in $[0, T]$.*

The second theorem is the dual of this: the same result is stated with h fixed. The tools used to prove these results are also the tools of transversality theory. Our result is a little bit more general because we work with a diffeomorphism f which does not necessarily derive from the flow of a vector field. Moreover, the result we prove is about a stronger notion of observability. Notice, however, that in a concluding remark of the above-mentioned paper, the author mentions that it is possible to extend the result to an embedding context. The proof is not explicitly written but can be made along the same lines.

In [8], the author proves exactly the same result as the one of this paper; nevertheless, its proof is, in our opinion, oversimplified and does not seem to use the same tools as the ones used in this article. We have to mention also the paper [7], where a proof of this result is given; the proof is based on the tools of transversality theory and is similar to the one supplied here.

All the works mentioned in this introduction make extensive use of the tools of the transversality theory; as explained hereafter, we will use the same tools to prove our result.

2. Preliminaries. In this section we recall some theorems from differential topology which will be intensively used in the proof of the main result of this paper. For details on the C^∞ Whitney topology, the reader is referred to the book "Stable Mappings and their Singularities" (see [6]).

If X and Y are two smooth manifolds, $J^k(X, Y)$ will denote, as usual, the set of k-jets from X to Y, $\alpha : J^k(X, Y) \to X$ is the source map, and $\beta : J^k(X, Y) \to Y$ is the target map. If f is in $C^\infty(X, Y)$—the space of smooth maps from X to Y—$j^k f$ denotes the k-jet of f. Recall that the set $C^\infty(X, Y)$ endowed with the Whitney topology is a Baire space, and so every residual set of $C^\infty(X, Y)$ (i.e., every countable intersection of open dense subsets) is dense.

The notion of transversality is of paramount importance for our purpose, and we recall below its definition.

DEFINITION 3. *Let f be a smooth mapping between two smooth manifolds X and Y, W a submanifold of Y, and x a point in X. We will say that f intersects W* transversely *at x if either*

(i) $f(x) \notin W$ *or*

(ii) $f(x) \in W$ *and* $T_{f(x)}Y = T_{f(x)}W + df_x(T_xX)$,

T_xX denoting the tangent space to X at x and df_x the Jacobian of f at x. We will say that f intersects W transversely if it intersects W transversely at x for all x in W. We will use the symbol $\pitchfork$ to denote the transversality.

The following theorem states a result of genericity (see [6]).

THEOREM 2 (Thom transversality theorem). *Let X and Y be smooth manifolds*

and W a submanifold of $J^k(X,Y)$. Let

$$T_W = \{f \in C^\infty(X,Y) \mid j^k f \pitchfork W\}.$$

Then T_W is a residual subset of $C^\infty(X,Y)$ in the C^∞ topology. Moreover, if W is closed, then T_W is open.

The following result generalizes the above theorem to multijet spaces. We first define the set $X^{(s)} = \{\,(x_1,\dots,x_s) \in X^s \;\; x_i \neq x_j \text{ for } 1 \leq i < j \leq s\,\}$ and the mapping

$$\begin{aligned} \alpha^s : \big(J^k(X,Y)\big)^s &\longrightarrow X^s, \\ (\sigma_1,\dots,\sigma_s) &\longmapsto \big(\alpha(\sigma_1),\dots,\alpha(\sigma_s)\big), \end{aligned}$$

and we let $J^k_s(X,Y) = (\alpha^s)^{-1}(X^{(s)})$, $J^k_s(X,Y)$ being a submanifold of $\Big(J^k(X,Y)\Big)^s$.

For $f \in C^\infty(X,Y)$, we can define

$$\begin{aligned} j^k_s f : \qquad X^{(s)} \qquad &\longrightarrow J^k_s(X,Y), \\ (x_1,\dots,x_s) \quad &\longmapsto \big(j^k f(x_1),\dots,j^k f(x_s)\big). \end{aligned}$$

THEOREM 3 (multijet transversality theorem). *Let W be a submanifold of $J^k_s(X,Y)$, and let*

$$T_W = \{f \in C^\infty(X,Y) \mid j^k_s f \pitchfork W\}.$$

Then T_W is a residual subset of $C^\infty(X,Y)$ in the C^∞ topology. Moreover, if W is compact, then T_W is open.

We will use also a transversality theorem due to Abraham and Robbin (see [1]). Let $\mathcal{A}, X$, and Y be C^r manifolds, $C^r(X,Y)$ the set of C^r maps from X to Y, and $p : \mathcal{A} \to C^r(X,Y)$ a map.

For $a \in \mathcal{A}$, we write ρ_a, the C^r map

$$\begin{aligned} \rho_a : X &\longrightarrow Y, \\ x &\longmapsto \rho_a(x) = \rho(a)(x), \end{aligned}$$

and we say that ρ is a C^r representation if the evaluation map

$$\begin{aligned} \mathrm{ev}_\rho : \mathcal{A} \times X &\longrightarrow Y, \\ (a,x) &\longmapsto \rho_a(x) = \rho(a)(x) \end{aligned}$$

is a C^r map from $\mathcal{A} \times X$ to Y.

THEOREM 4 (Abraham transversal density theorem). *Let $\mathcal{A}, X, Y$ be C^r manifolds, $\rho : \mathcal{A} \to C^r(X,Y)$ a C^r representation, $W \subset Y$ a submanifold (not necessarily closed), and $\mathrm{ev}_\rho : \mathcal{A} \times X \to Y$ the evaluation map. Define $\mathcal{A}_W \subset \mathcal{A}$ by*

$$\mathcal{A}_W = \{a \in \mathcal{A} \mid \rho_a \pitchfork W\}.$$

Assume that

(i) *X has a finite dimension n and W has a finite codimension q in Y;*
(ii) *$\mathcal{A}$ and X are second countable;*
(iii) *$r > \max(0, n-q)$;*
(iv) *$\mathrm{ev}_\rho \pitchfork W$.*

Then $\mathcal{A}_W$ is residual in $\mathcal{A}$.

Notice that manifold $\mathcal{A}$ is not necessarily finite dimensional; it may be a Banach space or an open subset of a Banach space.

Finally, we will need the following theorem that can also be found in [1].

THEOREM 5 (openness of transversal intersection). *Let $\mathcal{A}$, X, and Y be C^r manifolds with X finite dimensional, $W \subset Y$ a closed C^r submanifold, K a compact subset of X, and $\rho : \mathcal{A} \to C^r(X, Y)$ a C^r representation. Then the subset $\mathcal{A}_{KW} \subset \mathcal{A}$ defined by*

$$\mathcal{A}_{KW} = \{a \in \mathcal{A} \mid \rho_a \pitchfork_x \ W \text{ for } x \in K\}$$

is open.

3. Main result. In what follows in this paper, X will denote a compact manifold and $\mathrm{Diff}(X)$ will denote the set of diffeomorphisms on X. There are many notions of observability; among them the weaker is the following. We will say that (1) is observable if, given two initial conditions x_0 and $\bar{x}_0$, there exists $k \in \mathbb{N}^*$ such that $h(x_k) \neq h(\bar{x}_k)$. Our main result is about the stronger notion of strong observability defined below.

DEFINITION 4. *Denoting by f^s, $f \circ \cdots \circ f$ s times, we say that system* (1) *is strongly observable if the map*

$$\begin{aligned} \Theta_{f,h} : \ X &\longrightarrow \mathbb{R}^{2n+1}, \\ x &\longmapsto (h(x), h \circ f(x), \ldots, h \circ f^{2n}(x)) \end{aligned}$$

is an embedding.

We state the following theorem about the discrete system of type (1).

THEOREM 6. *The set of pair $(f, h) \in \mathrm{Diff}(X) \times \mathscr{C}^\infty(X, \mathbb{R})$ such that system* (1) *is strongly observable is residual.*

The proof of this theorem will be given through Lemmas 1–6. In Lemmas 1–3, we prove that the set $\mathscr{C}_{2n}(X)$ of diffeomorphisms whose periodic points with period $\leq 2n$ and which are cocyclic at each periodic point is open dense. In Lemmas 4–7, we prove that a diffeomorphism $f \in \mathscr{C}_{2n}(X)$ being given, the set of maps $h \in C^\infty(X, \mathbb{R})$ such that $\Theta_{f,h}$ is an immersion at each s-periodic point of f and is injective when restricted to the set of periodic points of f is dense. In Lemmas 5–6, we conclude that the set of pairs (f, h) such that $\Theta_{f,h}$ is an embedding is dense in $\mathrm{Diff}(X) \times C^\infty(X, \mathbb{R})$.

3.1. Periodic points. Let f be a diffeomorphism of X; the point $x \in X$ is said to be a periodic point of f with period s (we will also say s-periodic point of f) if $f^s(x) = x$ and $f^p(x) \neq x$ whenever $0 < p < s$. We will say that the s-tuple $(x_1, \ldots, x_s)$ is a cycle of length s (or an s-cycle) for f if $x_{i+1} = f(x_i)$ for $i = 1, \ldots, s-1$ and $f(x_s) = x_1$; this is equivalent to saying that x_i is an s'-periodic point of f (with s' a divisor of s). Let $\Gamma_s(f)$ denote the set of all periodic points of f with period $\leq s$.

LEMMA 1. *For every $r \in \mathbb{N}^*$, there exists an open dense set $\mathscr{P}_r(X)$ of diffeomorphisms of X such that if $f \in \mathscr{P}_r(X)$, $\Gamma_r(f)$ is finite.*

Proof. Let us denote by Ω_r the set of roots of unity with order no greater than r ($\omega \in \Omega_r \Leftrightarrow \omega^e = 1$ with $1 \leq e \leq r$) and by $\mathscr{P}_r(X)$ the set of diffeomorphisms f such that if x is a s-periodic point of f with $s \leq r$, none of the eigenvalues of df_x^s belongs to Ω_r.

First we claim that if f belongs to $\mathscr{P}_r$, $\Gamma_r(f)$ is finite. Indeed, we notice first that the limit of a sequence of periodic points is also a periodic point, so $\Gamma_r(f)$ is closed.

Then we will see that every periodic point in $\Gamma_r(f)$ is isolated (among the points of this set). Suppose on the contrary that $x_0 \in \Gamma_r(f)$ is not isolated; then there exists a sequence $(x_n)_{n\geq 1}$ of points in $\Gamma_r(f)$ whose limit is x_0; we can suppose that all the x_n are periodic with the same period $s \leq r$, and so we have $f^s(x_0) = x_0$, which proves that x_0 is an s'-periodic point of f (with s' a divisor of s). In other words, x_0 is a nonisolated fixed point for f^s, but it is well known that if x_0 is a nonisolated fixed point of a mapping g, dg_x admits the number 1 as an eigenvalue. We deduce that 1 is an eigenvalue of $df^s_{x_0}$, which implies that $df^{s'}_{x_0}$ has an eigenvalue in Ω_r, but this is in contradiction with the definition of $\mathscr{P}_r$.

Density of $\mathscr{P}_r$. We will prove now that $\mathscr{P}_r$ is open dense. For the density part of the lemma, it suffices to prove that the set of $f \in \mathrm{Diff}(X)$ such that df^s_x has no eigenvalue in Ω_r if x is an s-periodic point of f is residual. Manifold X being second countable and metrizable, there exists a countable family $\mathscr{F}$ of open subsets whose union is equal to X and which satisfies

$$\forall (x,y) \in X^2, x \neq y \Rightarrow \exists (p,q) \in \mathbb{N}^2 \mid x \in U_p, y \in U_q, U_p \cap U_q = \emptyset.$$

We will assume that a coordinate function Φ_p is associated with each U_p.

Let $s \leq r$ and $\mathscr{A}_s$ be the set of s-tuples $(M_1, \dots, M_s)$ of matrices of $\mathcal{M}_n(\mathbb{R})$ such that the product $M_1 \cdots M_s$ has at least one eigenvalue in Ω_r; clearly, $\mathscr{A}_s$ is an algebraic set and thus, from Whitney's theorem (see [9]), a finite union of smooth submanifolds of $(\mathcal{M}_n(\mathbb{R}))^s : \mathscr{A}_s = W_1 \cup \cdots \cup W_N$. Let $(U_1, \Phi_1), \dots, (U_s, \Phi_s)$ be charts of X with $U_i \in \mathscr{F}$ and $U_i \cap U_j = \emptyset$ if $i \neq j$. We consider the chart $(\mathscr{J}, \gamma)$ of $J^1_s(X,X)$, where $\mathscr{J} = J^1(U_1,U_2) \times \cdots \times J^1(U_{s-1},U_s) \times J^1(U_s,U_1)$ and γ is the mapping from $\mathscr{J}$ to

$$\mathscr{U} = \Phi_1(U_1) \times \Phi_2(U_2) \times \cdots \times \Phi_{s-1}(U_{s-1}) \times \Phi_s(U_s) \times \Phi_s(U_s) \times \Phi_1(U_1) \times (\mathcal{M}_n(\mathbb{R}))^s$$

defined by

$$\begin{aligned}\gamma(\sigma_1, \dots, \sigma_s) = \Big(&\phi_1\big(\alpha(\sigma_1)\big), \phi_2\big(\beta(\sigma_1)\big), \dots, \phi_s\big(\alpha(\sigma_s)\big), \phi_1\big(\beta(\sigma_s)\big),\\ &d\big(\phi_2 \circ f_1 \circ \phi_1^{-1}\big)_{\phi_1(\alpha(\sigma_1))}, \dots, d\big(\phi_1 \circ f_s \circ \phi_s^{-1}\big)_{\phi_s(\alpha(\sigma_s))}\Big),\end{aligned}$$

the f_i's representing the σ_i's. Let W_i be one of the submanifolds appearing in the decomposition of the algebraic manifold $\mathcal{A}$ described above. We denote by $(x_1, x_2', x_2, x_3', \dots, x_{s-1}, x_s', x_s, x_1', M_1^s)$ an element of $\gamma(\mathcal{J})$, and we consider $V_i \subset \gamma(\mathcal{J})$ defined by the relations

$$\begin{cases} x_i = x_i' \text{ for } i = 1, \dots, s, \\ M_1^s \in W_i. \end{cases}$$

V_i is a submanifold of $\gamma(\mathcal{J})$ of codimension $\operatorname{codim} V_i = sn + d_i$ ($d_i = \operatorname{codim} W_i$). Thus, $\tilde{V}_i = \gamma^{-1}(V_i)$ is a submanifold of $J^1_s(X,X)$ with the same codimension as V_i.

For $f \in \mathrm{Diff}(X)$, to say that the mapping $j^1_s f : X^{(s)} \to J^1_s(X,X)$ is transverse to $\tilde{V}_i$ means that $j^1_s f(x) \notin \tilde{V}_i$ for all x because the codimension of $\tilde{V}_i$ is equal to $sn + d_i$ and is greater than sn, the dimension of $X^{(s)}$.

Let us consider the countable set $\{\mathcal{J}_1, \mathcal{J}_2, \dots\}$ of charts of $J^1_s(X,X)$ defined from the s-tuples $(U_1, \dots, U_s)$ as above. To each chart $\mathcal{J}_k$ are associated N manifolds $\tilde{V}_i^k$ $(i = 1, \dots, N)$; thanks to the multijet transversality theorem, we can assert that the set of $\mathscr{R}^k$ of diffeomorphisms f such that the mapping $j^1_s f : X^{(s)} \to J^1_s(X,X)$ is

transverse to $\tilde{V}_i^k$ is residual, so the set $\mathscr{R}^k$ of f such that the mapping $j_s^1 f : X^s \to J_s^1(X,X)$ is transverse to each submanifold $\tilde{V}_i^k$ $(i = 1, \dots, N)$ is residual as a finite intersection of residual sets, and $\mathscr{R} = \bigcap_{k\geq 1} \mathscr{R}^k$ is also residual.

Now, let f be an element of $\mathscr{R}$ and suppose that x is a periodic point of order $s \leq r$ for f. If we put

$$\begin{cases} x_1 = x, \\ x_2 = f(x), \\ \vdots \\ x_s = f^{s-1}(x). \end{cases}$$

Then $x_i \neq x_j$ for $i \neq j$ and $f(x_s) = x_1$.

From the definition of $\mathscr{F}$, it is possible to find charts $(U_1, \phi_1), \dots, (U_s, \phi_s)$ such that $U_i \in \mathscr{F}$, $U_i \cap U_j = \emptyset$ if $i \neq j$ and $x_i \in U_i$ for $i = 1, \dots, s$.

To say that the map $j_s^1 f$ is transverse to $\tilde{V}_i$ at the point $(x_1, \dots, x_s)$ means that $j_s^1 f(x_1, \dots, x_s)$ does not belong to $\tilde{V}_i$. From the definition of $\tilde{V}_i$, this means that $\gamma\big(j_s^1 f(x_1, \dots, x_s)\big) \notin V_i$; therefore $\big(d\big(\phi_2 \circ f \circ \phi_1^{-1}\big)_{\phi_1(x_1)}, \dots d\big(\phi_1 \circ f \circ \phi_s^{-1}\big)_{\phi_s(x_s)}\big) \notin W_i$ (for $i = 1, \dots; N$), which implies that the s-tuple of matrices

$$\Big(d\big(\phi_2 \circ f \circ \phi_1^{-1}\big)_{\phi_1(x_1)}, \dots d\big(\phi_1 \circ f \circ \phi_s^{-1}\big)_{\phi_s(x_s)}\Big)$$

does not belong to $\mathscr{A}_s$, and so none of the eigenvalues of df_x^s is in Ω_r if x is an s-periodic point of f.

Openness of $\mathscr{P}_r$. Let f_0 be in $\mathscr{P}_r$; the s-periodic points of f_0 are isolated $(s \leq r)$ and X is compact, so $\Gamma_r(f_0)$ is finite. By definition of the topology of $\mathrm{C}^\infty(X,X)$ and since X is compact, we have

$$\forall \varepsilon > 0, \exists\, U_\varepsilon \mid g \in U_\varepsilon \Rightarrow d(f(x), g(x)) < \varepsilon \ \forall x \in X,$$

where U_ε is an open neighborhood of f in $C^\infty(X,X)$ and d is a distance compatible with the topology of X. For $s \leq r$, the number of s-cycles of f is finite; we will denote them by

$$\begin{array}{cccc} x_{11}, & x_{12}, & \dots, & x_{1s}, \\ x_{21}, & x_{22}, & \dots, & x_{2s}, \\ & & \vdots & \\ x_{N1}, & x_{N2}, & \dots, & x_{Ns}. \end{array}$$

We choose $\rho > 0$ such that $B(x_{ij}, \rho) \cap B(x_{i'j'}, \rho) = \emptyset$ if $x_{ij} \neq x_{i'j'}$ ($B(x_{ij}, \rho)$ denotes the open ball of center x_{ij} and radius ρ); to each $B(x_{ij}, \rho)$ we associate a chart ϕ_{ij}.

We set

$$\begin{aligned} B_i =& \Big(B(x_{i1}, \rho) \times \dots \times B(x_{is}, \rho)\Big) \cup \Big(B(x_{i2}, \rho) \times \dots \times B(x_{is}, \rho) \times B(x_{i1}, \rho)\Big), \\ & \cup \dots \cup \Big(B(x_{is}, \rho) \times B(x_{i1}, \rho) \times \dots \times B(x_{is-1}, \rho)\Big), \\ K =& X^s \setminus \bigcup_{i=1}^{N} B_i, \\ \beta =& \inf_{(x_1, \dots, x_s) \in K} \{d(f(x_1), x_2) + \dots + d(f(x_{s-1}), x_s) + d(f(x_s), x_1\}. \end{aligned}$$

We have $\beta \neq 0$. Indeed, if $\beta = 0$, then by continuity of f and since K is compact, there exists $(x_1, \dots, x_s) \in K$ satisfying $f(x_1) = x_2, \dots, f(x_s) = x_1$, which means that $(x_1, \dots, x_s)$ is an s-cycle for f which is impossible because all the s-cycles belong to $\bigcup_{i=1}^{N} B_i$.

Let $\varepsilon > 0$ be such that $s\varepsilon < \beta$, let $g \in U_\varepsilon$ and x_1 be a periodic point of g of order s, and put $x_2 = g(x_1), \dots, x_s = g(x_{s-1})$; we have $g(x_s) = x_1$.

If $(x_1, \dots, x_s) \notin \bigcup_{i=1}^{N} B_i$, then $d(f(x_1), x_2) + \cdots + d(f(x_{s-1}), x_s) + d(f(x_s), x_1) \geq \beta$. Since $x_i = g(x_{i-1})$, this implies $\beta \leq d(f(x_1), g(x_1)) + \cdots + d(f(x_{s-1}), g(x_{s-1})) + d(f(x_s), g(x_s)) < s\epsilon < \beta$, which is a contradiction.

We conclude that all the periodic points of g are in the balls $B(x_{ij}, \rho)$. Now if ρ and ε are chosen small enough, it is clear that none of the eigenvalues of $dg^s x_1$ is contained in Ω_r.

Remark 1. If f is in $\mathscr{P}_r$, X being compact, the set of periodic points of f of order $\leq r$ is finite.

3.2. Cocyclic diffeomorphisms.

DEFINITION 5. *Let $f \in \mathrm{Diff}(X)$, and let x be a periodic point of order $s \leq n-1$ of f; we will say that f is* cocyclic *at x if it is possible to find s covectors $h_1, \dots, h_s$ ($h_i \in \left(T_{f^{i-1}(x)}X\right)^*$) such that the covectors*

$$h_1,\ h_2 \cdot df_x, \dots, h_s \cdot df_x^{s-1}, \dots,\ h_1 \cdot df_x^{sq}, \dots, h_{r+1} \cdot df_x^{sq+r}$$

are linearly independent ($n - 1 = sq + r$ with $0 \leq r < s$).

We will denote by $\mathscr{C}_r(X)$ the subset of $\mathscr{P}_r(X)$ constituted by the diffeomorphisms which are cocyclic at all their s-periodic points for all $s \leq r$.

LEMMA 2. *The set $\mathscr{C}_r(X)$ is residual.*

Proof. Let $s \leq r$, in $(M_n(\mathbb{R}))^s$, we consider the set $\mathscr{M}$ of s-tuples $(A_1, A_2, \dots, A_s)$ for which there exist s covectors $C_1, C_2, \dots, C_s$ such that the n covectors

$$\begin{array}{ccccccc}
C_1, & C_2A_1, & C_3A_2A_1, & \dots, & \dots, & C_sA_{s-1}\dots A_1, \\
C_1\tilde{A}, & C_2A_1\tilde{A}, & C_3A_2A_1\tilde{A}, & \dots, & \dots, & C_sA_{s-1}\dots A_1\tilde{A}, \\
 & & \vdots & & & \\
C_1\tilde{A}^{q-1}, & C_2A_1\tilde{A}^{q-1}, & C_3A_2A_1\tilde{A}^{q-1}, & \dots, & \dots, & C_sA_{s-1}\dots A_1\tilde{A}^{q-1}, \\
C_1\tilde{A}^{q}, & C_2A_1\tilde{A}^{q}, & C_3A_2A_1\tilde{A}^{q}, & \dots, & C_rA_{r-1}\dots A_1\tilde{A}^{q} &
\end{array}$$

are linearly independent ($n = sq + r$ with $0 \leq r < s$, and we put $\tilde{A} = A_s \cdots A_1$).

We notice that the set $\mathcal{A} = (M_n(\mathbb{R}))^s \setminus \mathscr{M}$ is algebraic, and so it is a finite union of smooth submanifolds of $(M_n(\mathbb{R}))^s$. Now the proof uses exactly the same arguments as in the proof of the previous lemma. The codimension of each submanifold W_i is at least equal to 1 since the complement of $\mathcal{A}$ is an open nonempty subset of $(M_n(\mathbb{R}))^s$. □

LEMMA 3. *The set $\mathscr{C}_r(X)$ is open.*

Proof. The proof is analogous to the one of Lemma 1. From this last lemma, we know that all the periodic points of order $\leq r$ of f are isolated, so the same is true for all the s-cycles of f; these s-cycles will be denoted by

$$\begin{array}{cccc}
x_{11}, & x_{12}, & \dots, & x_{1s}, \\
x_{21}, & x_{22}, & \dots, & x_{2s}, \\
 & & \vdots & \\
x_{N1}, & x_{N2}, & \dots, & x_{Ns}.
\end{array}$$

If diffeomorphism g is closed enough to f, all the periodic points of g (of order $\leq s$) will be closed to the periodic points of f, and the differentials of g at a periodic point will be closed to the differential of f at some periodic point x_{ij}, from which we can deduce that $g \in \mathscr{C}_r$ as soon as g is closed enough to f. □

3.3. Genericity of the injectivity.

LEMMA 4. *For a given finite set $E = \{x_1^0, \dots, x_p^0\}$ of points of X, the set $\mathscr{H}(E)$ of mappings $h \in C^\infty(X, \mathbb{R})$ satisfying $h(x_i^0) \neq h(x_j^0)$ for all $i \neq j$ is open and dense in $C^\infty(X, \mathbb{R})$.*

Proof. Once again we will use the multijet transversality theorem with

$$J_p^0(X, \mathbb{R}) = \{(x_1, r_1, \dots, x_p, r_p) \in (X \times \mathbb{R})^p \ : \ x_i \neq x_j \text{ if } i \neq j\}.$$

For a pair $(i, j) \in \{1, \dots, p\}$ with $i \neq j$ let

$$W_{ij} = \{(x_1, r_1, \dots, x_p, r_p) \in J_p^0(X, \mathbb{R}) \mid x_1 = x_1^0, \dots, x_p = x_p^0, r_i = r_j\}.$$

W_{ij} is a closed submanifold of $J_p^0(X, \mathbb{R})$ with codimension $np + 1$. Since this codimension is greater than $\dim X^{(p)}$, to say that the mapping $j_p^0 h$

$$\begin{array}{rccc} j_p^0 h : & X^{(p)} & \longrightarrow & J_p^0(X, \mathbb{R}), \\ & (x_1, \dots, x_p) & \longmapsto & \big(x_1, h(x_1), \dots, x_p, h(x_p)\big) \end{array}$$

is transverse to W_{ij} means that $j_p^0 h(x_1, \dots, x_p) \notin W_{ij}$ for all $(x_1, \dots, x_p)$, which is equivalent to saying that $h(x_i^0) \neq h(x_j^0)$. So the set $\mathscr{H}(E)$ of mappings $h \in C^\infty(X, \mathbb{R})$ such that $h(x_i^0) \neq h(x_j^0)$ for every pair (i, j) such that $1 \leq i < j \leq p$ is residual. Now, it is easy to see that if h is in $\mathscr{H}(E)$, the same is true for every mapping $\bar{h}$ closed enough to h, which proves that $\mathscr{H}(E)$ is open. □

LEMMA 5. *Let f be an element of $\mathscr{C}_{2n}(X)$, and recall that $\Gamma_{2n}(f)$ denotes the set of periodic points of f of order $\leq 2n$. The set $\mathscr{B}(f)$ defined by*

$$\mathscr{B}(f) = \{\, h \in C^\infty(X, \mathbb{R}) \mid \big(h(x), \dots, h \circ f^{2n}(x)\big) \neq \big(h(x_0), \dots, h \circ f^{2n}(x_0)\big) \\ \forall (x, x_0) \in \big(X \setminus \Gamma_{2n}(f)\big) \times \Gamma_{2n}(f) \,\}$$

is open dense.

Proof. We begin by proving the density part of this lemma; to this end, we need the Abraham's theorem (Theorem 4) that we recalled in the introduction.

In this proof, we consider that $C^\infty(X, \mathbb{R})$ is included in $C^r(X, \mathbb{R})$ endowed with the C^r topology (which is a Banach space); it is sufficient to prove the lemma for every (large enough) finite r to get the C^∞ result.

For each period $s = 1, 2, \dots, 2n$, we consider the C^r representation

$$\rho^s : C^\infty(X, \mathbb{R}) \to C^r(X, \mathbb{R}^{n+1})$$

defined by

$$\rho^s(h)(x) = \big(h(x) - h \circ f^s(x), h \circ f(x) - h \circ f^{s+1}(x), \dots, h \circ f^n(x) - h \circ f^{n+s}(x)\big).$$

The set $\tilde{X} = X \setminus \Gamma_{2n}(f)$ is a manifold as an open subset of X. We take $W = \{0\} \subset \mathbb{R}^{n+1}$, and we will show that ev_{ρ^s} is transverse to W.

$$\begin{array}{rcl} \mathrm{ev}_{\rho^s} : C^\infty(X,\mathbb{R}) \times \tilde{X} & \longrightarrow & \mathbb{R}^{n+1}, \\ (h,x) & \longmapsto & (h(x) - h\circ f^s(x), \dots, h\circ f^n(x) - h\circ f^{n+s}(x)). \end{array}$$

The differential of ev_{ρ^s} at (h_0, x_0) is given by

$$d\ \mathrm{ev}_{\rho^s}\,|_{(h_0,x_0)}.(h,\xi) = (\psi_0, \psi_1, \dots, \psi_n),$$

where $\psi_i = h\circ f^i(x_0) + d(h_0\circ f^i)_{x_0}\cdot\xi - h\circ f^{i+s}(x_0) - d(h\circ f^{i+s})_{x_0}\cdot\xi$.

This differential is surjective; indeed, let $u = (u_0, u_1, \dots, u_n) \in \mathbb{R}^{n+1}$; we will show that there exists a map $h : X \to \mathbb{R}$ such that $d\,\mathrm{ev}_{\rho^s}\,|_{(h_0,x_0)}.(h,0) = u$.

To this end, it is enough to show that there exists h satisfying the following system:

$$\left\{\begin{array}{rcl} h(x_0) - h\circ f^s(x_0) & = & u_0, \\ h\circ f(x_0) - h\circ f^{s+1}(x_0) & = & u_1, \\ & \vdots & \\ h\circ f^n(x_0) - h\circ f^{n+s}(x_0) & = & u_n. \end{array}\right. \tag{3}$$

Put $x_i = f^i(x_0)$; if $s \le n$, the points $x_0, x_1, \dots, x_{n+s}$ are all distinct because $x_0 \in \tilde{X} = X \setminus \Gamma_{2n}(f)$ so that either x_0 is not a periodic point or it is a periodic point with order $> 2n$. Obviously, the system

$$\left\{\begin{array}{rcl} y_0 - y_s & = & u_0, \\ y_1 - y_{s+1} & = & u_1, \\ & \vdots & \\ y_n - y_{n+s} & = & u_n \end{array}\right. \tag{4}$$

of $n+1$ equations and $n+s+1$ unknown $y_0, \dots, y_{n+s}$ has always at least one solution. If $\bar{y}_0, \dots, \bar{y}_{n+s}$ is a solution of (4), then it is always possible to find a map h such that $h(x_i) = \bar{y}_i$, and then h is a solution of (3).

If $s > n$, we consider the two lists

$$l_1 = \{x_0, \dots, x_n, x_{n+1}, \dots, x_{2n}\} \text{ and } l_2 = \{x_{2n+1}, \dots, x_{n+s}\}.$$

The elements of l_2 are all distinct because the equality $x_{2n+i} = x_{2n+j}$ with $1 \le i < j \le s-n$ implies $x_{j-i} = x_0$, which in turn implies that x_0 is a periodic point of period $\le j - i \le n-1$.

If $l_1 \cap l_2 = \emptyset$, we can conclude as above, so assume that $l_1 \cap l_2 \neq \emptyset$; then there exists $0 \le i \le 2n$ and $1 \le j \le s-n$ such that $x_i = x_{2n+j}$, which implies $x_{2n+j-i} = x_0$, and so x_0 is periodic with a period dividing $2n+j-i$. Notice also that $j-i$ must be positive since $x_0 \notin \Gamma_{2n}(f)$, and let $r_0 = \min\{r \mid x_{2n+r} = x_0\}$ $(0 < r_0 \le s-n)$. Now the points $x_0, \dots, x_{2n}, \dots, x_{2n+r_0-1}$ are all distinct, while the other ones satisfy

$$\left\{\begin{array}{rcl} x_{2n+r_0} & = & x_0, \\ & \vdots & \\ x_{n+s} & = & x_{s-n-r_0}. \end{array}\right.$$

Consider now the following system:

$$\left\{\begin{array}{rcl} y_0 - y_s & = & u_0, \\ & \vdots & \\ y_{2n+r_0-s-1} - y_{2n+r_0-1} & = & u_{2n+r_0-s-1}, \\ y_{2n+r_0-s} - y_0 & = & u_{2n+r_0-s-1}, \\ & \vdots & \\ y_n - y_{s-n-r_0} & = & u_n. \end{array}\right.$$

The rank of the matrix of this linear system ($n+1$ equations, $3n+r_0-s+1$ unknowns) is equal to $n+1$, so this system has at least one solution. If $\bar{y}_0, \ldots, \bar{y}_n, \bar{y}_s, \ldots, \bar{y}_{2n+r_0-1}$ is a solution, then in order to satisfy (3), it is sufficient to choose h such that $h(x_i) = \bar{y}_i$ for $0 \leq i \leq n$ and $s \leq i \leq 2n + r_0 - 1$.

So, in every case, there exists a map h such that (3) holds which proves that the evaluation map ev_{ρ_s} is transverse to $W = \{0\}$. So the set $\mathscr{B}^s(f) \subset C^\infty(X, \mathbb{R})$ such that $(\rho^s)_h$ is transverse to W is residual. To say that the map $(\rho^s)_h$ is transverse to $W = \{0\}$ means that $\rho^s(h)(x) \notin W$ because $\operatorname{codim} W = n + 1 > \dim X$. Let $\mathscr{B}(f) = \bigcap_{s=1}^{2n} \mathscr{B}^s(f)$, $\mathscr{B}(f)$ is residual, and, by construction, if $h \in \mathscr{B}(f)$, then for all $(x, x_0) \in (X \setminus \Gamma_{2n}(f) \times \Gamma_{2n}(f)$ one has

$$(h(x), h \circ f(x), \ldots, h \circ f^{2n}(x)) \neq (h(x_0), \ldots, h \circ f^{2n}(x_0))$$

because there exist $s \leq 2n$ such that $h \circ f^i(x_0) - h \circ f^{i+s}(x_0) = 0$ for all $0 \leq i \leq n$.

The remaining part of the proof is devoted to demonstrating the openness of $\mathscr{B}(f)$; to this end, we need Theorem 5 of openness of transversal intersection.

Let $h_0 \in \mathscr{B}(f)$, for every $\bar{x} \in \Gamma_{2n}(f)$, the covectors $(dh_{0\bar{x}}, d(h_0 \circ f)_{\bar{x}}, \ldots, d(h_0 \circ f^{n-1})_{\bar{x}})$ being linearly independent. This ensures the existence of a positive number $r(\bar{x}, h_0)$ such that the map

$$\theta_{h_0} : x \longmapsto (h_0(x), h_0 \circ f(x), \ldots, h_0 \circ f^{n-1}(x))$$

is a diffeomorphism from $B(\bar{x}, r(\bar{x}, h_0))$ to its image by θ_{h_0}. Then there exists an open set $\theta(h_0, \bar{x}) \subset C^\infty(X, \mathbb{R})$ such that, for all h in this set, the map

$$\theta_h : x \longmapsto (h(x), h \circ f(x), \ldots, h \circ f^{n-1}(x))$$

is a diffeomorphism from $B(\bar{x}, \frac{r(\bar{x}, h_0)}{2})$ to its image.

Furthermore, $s(\bar{x})$ denoting the period of $\bar{x}$, by continuity, there exists $r'(\bar{x}, h_0)$ $(0 < r' < r/2)$ such that $f^s(x) \in B(\bar{x}, r/2)$ for all $x \in B(\bar{x}, r')$; also we can choose $r(\bar{x}, h_0)$ so small that $B(\bar{x}, r(\bar{x}, h_0)) \cap B(\bar{y}, r(\bar{y}, h_0)) = \emptyset$ whenever $\bar{x}, \bar{y} \in \Gamma_{2n}(f)$ and $\bar{x} \neq \bar{y}$.

Now, in order to apply the above theorem, we consider the compact set

$$K = X \setminus \bigcup_{\bar{x} \in \Gamma_{2n}(f)} B(\bar{x}, r'(\bar{x}, h_0))$$

and the representation ρ^s defined in the first part of this proof. We set $C_K^s = \{\, h \in C^\infty(X, \mathbb{R}) \mid \rho_h^s \pitchfork_x \{0\} \text{for all } x \in K \,\}$; thanks to the above-mentioned theorem, we can assert that C_K^s is open and the set

$$\mathscr{U}_{h_0} = \left(\bigcap_{s=1}^{2n} C_K^s \right) \cap \left(\bigcap_{\bar{x} \in \Gamma_{2n}(f)} \theta(h_0, \bar{x}) \right)$$

is open and contains h_0.

We will now show that $\mathscr{U}_{h_0}$ is included in $\mathscr{B}(f)$; to this end, it suffices to show that the equality

$$(h(x) - h \circ f^s(x), \dots, f \circ f^n(x) - h \circ f^{n+s}(x)) = (0, \dots, 0) \tag{5}$$

implies, if $h_0 \in \mathscr{U}_{h_0}$, that $x \in \Gamma_{2n}(f)$. So let $x \in X$ and $h \in \mathscr{U}_{h_0}$:

(i) if $x \in K$, then by construction of $\mathscr{U}_{h_0}$, x cannot satisfy the above equality (notice that $\rho_h^s \pitchfork_x \{0\}$ means $\rho_h^s(x) \neq 0$);

(ii) if $x \notin K$, then $x \in B(\bar{x}, r'(\bar{x}, h_0))$ for some $\bar{x} \in \Gamma_{2n}(f)$, which implies that $f^s(x) \in B(\bar{x}, r(\bar{x}, h_0)/2)$. Now the map θ_h is a diffeomorphism from $B(\bar{x}, r/2)$ to its image, so equality (5) is equivalent to $\theta_h(x) = \theta_h(f^s(x))$, which leads to $x = f^s(x)$. This implies that x is a periodic point for f and thus $x = \bar{x} \in \Gamma_{2n}(f)$ because x is the unique periodic point of f which belongs to $B(\bar{x}, r_{\bar{x}})$. □

LEMMA 6. *Let $f \in \mathscr{C}_{2n}(X)$ be a given diffeomorphism. Then the set of mappings $h \in C^\infty(X, \mathbb{R})$ for which the map $\Theta_{f,h}$ is injective is residual.*

Proof. We define the set $\tilde{X}^{(2)} = \{(x_1, x_2) \in (X \setminus \Gamma_{2n}(f))^2 \mid x_1 \neq x_2\}$, and we consider the map $\rho : C^\infty(X, \mathbb{R}) \to C^r\big(\tilde{X}^{(2)}, \mathbb{R}^{2n+1}\big)$ defined by

$$\rho(h)(x_1, x_2) = \big(h(x_1) - h(x_2), h \circ f(x_1) - h \circ f(x_2), \dots, h \circ f^{2n}(x_1) - h \circ f^{2n}(x_2)\big)$$

and the evaluation map

$$\begin{array}{rccl} \mathrm{ev}_\rho & : & C^\infty(X, \mathbb{R}) \times \tilde{X}^{(2)} & \longrightarrow & \mathbb{R}^{2n+1}, \\ & & (h, x) & \longmapsto & \Big(h(x_1) - h(x_2), h \circ f(x_1) - h \circ f(x_2), \dots, \\ & & & & \dots, h \circ f^{2n}(x_1) - h \circ f^{2n}(x_2)\Big). \end{array}$$

Using the same reasoning as in the proof of density of $\mathscr{B}(f)$ (Lemma 5), we can prove that the linear tangent map of ev_ρ, computed at any point of $C^\infty(X, \mathbb{R}) \times \tilde{X}^{(2)}$, is surjective. This implies that ev_ρ is a submersion on $C^\infty(X, \mathbb{R}) \times \tilde{X}^{(2)}$. Hence ev_ρ is transverse to $\{(0, \dots, 0)\} \subset \mathbb{R}^{2n+1}$. From Theorem 4, we deduce that the set of h for which $\rho(h)(x_1, x_2) \neq 0$ for all $(x_1, x_2) \in \tilde{X}^{(2)} \times \tilde{X}^{(2)}$ is residual. On the other hand, if $(x_1, x_2) \in (X \setminus \Gamma_{2n}(f)) \times \Gamma_{2n}(f)$, then $\Theta_{f,h}(x_1) \neq \Theta_{f,h}(x_2)$ for $h \in \mathscr{B}(f)$, which is open and dense (Lemma 5). Finally, if $(x_1, x_2) \in \Gamma_{2n}(f) \times \Gamma_{2n}(f)$, then $\Theta_{f,h}(x_1) \neq \Theta_{f,h}(x_2)$ for $h \in \mathscr{H}(\Gamma_{2n}(f))$, which is open and dense. This allows us to conclude that the set of h for which $\Theta_{f,h}$ is injective is residual. □

3.4. Mapping $\Theta_{f,h}$ is generically an immersion.

LEMMA 7. *Let $f \in \mathscr{C}(X)$ be a given mapping; then the set of mappings $h \in C^\infty(X, \mathbb{R})$, for which the covectors*

$$dh_x, d(h \circ f)_x, \dots, d(h \circ f^{n-1})_x$$

are linearly independent for every periodic point of f, is open and dense in $C^\infty(X, \mathbb{R})$.

Proof. Let $f \in C^\infty(X, X)$ be a mapping with a finite number of periodic points of order $\leq r$, and let x be a periodic point of order $s \leq r$. It is sufficient to show the density of mapping $h \in \mathrm{C}^\infty(X, \mathbb{R})$ that satisfies

$$dh_x, d(h \circ f)_x, \dots, d(h \circ f^{n-1})_x \text{ are linearly independent.}$$

To show this, we can adapt the proof of Lemmas 2 and 3 by noticing that for given matrices $A_1, \dots, A_s$, the set of covectors $C_1, C_2, \dots, C_s$ such that the n covectors

$$\begin{array}{ccccccc}
C_1, & C_2A_1, & C_3A_2A_1, & \dots, & \dots, & C_sA_{s-1}\dots A_1, \\
C_1\tilde{A}, & C_2A_1\tilde{A}, & C_3A_2A_1\tilde{A}, & \dots, & \dots, & C_sA_{s-1}\dots A_1\tilde{A}, \\
 & & \vdots & & & \\
C_1\tilde{A}^{q-1}, & C_2A_1\tilde{A}^{q-1}, & C_3A_2A_1\tilde{A}^{q-1}, & \dots, & \dots, & C_sA_{s-1}\dots A_1\tilde{A}^{q-1}, \\
C_1\tilde{A}^{q}, & C_2A_1\tilde{A}^{q}, & C_3A_2A_1\tilde{A}^{q}, & \dots, & C_rA_{r-1}\dots A_1\tilde{A}^{q} &
\end{array}$$

are linearly dependent is an algebraic subset of $((\mathbb{R}^n)^*)^s$.

The proof of the openness is analogous to the proof of Lemma 4. □

For a given $f \in \mathcal{C}_r(X)$, we denote by $\mathscr{A}_r(f)$ the subset of $\mathrm{C}^\infty(X,\mathbb{R})$ verifying the conclusion of the above lemma.

LEMMA 8. *Let E be an n-dimensional vector space, and, for $k > n$, let $V(E,k)$ denote the set of k-tuples $(v_1,\dots,v_k) \in E^k$ such that* $\operatorname{rank}(v_1,\dots,v_k) < n$. *We claim that $V(E,k)$ is a finite union of smooth submanifolds of E^k, say, $W_1,\dots,W_l$; moreover, if* $\dim W_i = \max(\dim W_1,\dots,\dim W_l)$, *then* $\operatorname{codim} W_i = k - n + 1$.

Proof. An element $(v_1,\dots,v_k)$ of E^k can be identified with the matrix of $M_{n\times k}(\mathbb{R})$ whose columns are the v_i. Therefore, to prove the lemma it is sufficient to prove that the set of matrices with rank $< n$ is a finite union of smooth submanifolds of $M_{n\times k}(\mathbb{R})$. The proof of this result (as well as the computation of the least codimension) can be found in [6, pp. 60–61]. □

LEMMA 9. *For a given $f \in \mathscr{C}_r(X)$, the set of mappings $h \in C^\infty(X,\mathbb{R})$ such that the map*

$$\begin{array}{rcl}
\Theta_{f,h} : X & \longrightarrow & \mathbb{R}^{2n+1}, \\
x & \longmapsto & \big(h(x), h\circ f(x),\dots,h\circ f^{2n}(x)\big)
\end{array}$$

is an immersion is residual.

Proof. The map $\pi : T^*X \to X$ denoting the canonical projection, we consider the following set:

$$(T^*X)^{\otimes k} = \{\, (p_1,\dots,p_k) \in (T^*X)^k \;\mid \pi(p_1) = \pi(p_2) = \dots = \pi(p_k) \,\}.$$

Using the same notations as in Lemma 8, we set $\mathcal{V}(k,T^*X) = \bigcup_{x\in X} V(k,T^*_xX)$. Thanks to this lemma we know that $\mathcal{V}(k,T^*X)$ is a union of submanifolds of $(T^*X)^{\otimes k}$, the codimension of the highest dimensional submanifold being $k-n+1$.

Let $f \in \mathscr{C}_{2n}(X)$ and $\tilde{X} = X \setminus \Gamma_{2n}(f)$, and consider the map

$$\rho \;:\; \mathscr{A}_{2n}(f) \to \mathrm{C}^r\left(\tilde{X}, (T^*X)^{\otimes 2n+1}\right)$$

defined by $\rho(h)(x) = \big(dh_x, d(h\circ f)_x,\dots,d(h\circ f^{2n})_x\big)$. We will show that the evaluation map ev_ρ defined by

$$\begin{array}{rcl}
ev_\rho \;:\; \mathscr{A}(f)\times\tilde{X} & \longrightarrow & (T^*X)^{\otimes 2n+1}, \\
(h,x) & \longmapsto & \big(dh_x, d(h\circ f)_x,\dots,d(h\circ f^{2n})_x\big)
\end{array}$$

is transverse to $\mathcal{V}(2n+1,T^*X)$ (i.e., transverse to every submanifold of the union $\mathcal{V}(k,T^*X)$). To prove this, it is sufficient to show that ev_ρ is a submersion at every point $(h,x) \in \mathscr{A}_{2n}(f)\times\tilde{X}$. For a given $\bar{x} \in \tilde{X}$ the map

$$\begin{array}{rcl}
\mathrm{ev}_{\rho\bar{x}} \;:\; \mathscr{A}_{2n}(f) & \longrightarrow & (T^*X)^{\otimes 2n+1}, \\
h & \longmapsto & \big(dh_{\bar{x}}, d(h\circ f)_{\bar{x}},\dots,d(h\circ f^{2n})_{\bar{x}}\big)
\end{array} \tag{6}$$

is linear, so its expression at $\bar{h}$ coincides with the one of its linear tangent map at $\bar{h}$. This tangent map is surjective; indeed, $\bar{x} \notin \Gamma_{2n}(f)$, so $f^i(\bar{x}) \neq f^j(\bar{x})$ for all $i, j \in \{0, 1, \dots, 2n\}$ $(i \neq j)$. Therefore, for a given $(p_0, \dots, p_{2n}) \in (T^*_{\bar{x}}X)^{\otimes 2n+1}$, it is possible to find $h \in \mathscr{A}_{2n}(f)$ in such a way that $d(h \circ f^i)_{\bar{x}} = dh_{f^i(\bar{x})} \circ df^i(\bar{x}) = p_i$. We showed that the map $ev_{\rho\bar{x}}$ defined by (6) is a submersion at every $h \in \mathscr{A}_{2n}(f)$; this implies that ev_ρ is a submersion at every point $(h, x) \in \mathscr{A}_{2n}(f) \times \tilde{X}$ and thus ev_ρ is transverse to $\mathcal{V}(2n+1, T^*X)$. Thanks to Theorem 4, we can conclude that the set of h such that ρ_h is transverse to $\mathcal{V}(2n+1, T^*X)$ is residual. Here, transversality means nonintersection for the codimension of the highest dimensional submanifold contained in $\mathcal{V}(2n+1, T^*X)$ is $2n+1-n+1 = n+2 > \dim \tilde{X}$. Therefore, $\rho_h \pitchfork \mathcal{V}(2n+1, T^*X)$ implies $\operatorname{rank}\left(dh_x, d(h \circ f)_x, \dots, d(h \circ f^{2n})_x\right) = n$ for all $x \in \tilde{X}$ and thus the set of h such that $\Theta_{f,h}$ is an immersion on $\tilde{X}$ is residual. On the other hand, the set of h such that $\Theta_{f,h}$ is an immersion at each point of $\Gamma_{2n}(f)$ is $\mathscr{A}_{2n}(f)$, which is open and dense (Lemma 7). Hence the set of $h \in \mathrm{C}^r(X, \mathbb{R})$ such that $\Theta_{f,h}$ is an immersion on $X = \tilde{X} \cup \Gamma_{2n}(f)$ is residual in $C^r(X, \mathbb{R})$. The result being true for all $r \in \mathbb{R}$ it is still true for $r = \infty$. □

3.5. Proof of Theorem 6. A diffeomorphism $f \in \mathscr{C}_{2n}(f)$ being given, we saw that the set of maps h such that $\Theta_{f,h}$ is an injective immersion (and so an embedding) is dense in $C^\infty(X, \mathbb{R})$. Now the set $\mathscr{C}_{2n}(f)$ is dense in $\mathrm{Diff}(X)$, so we can conclude that the set of pairs $(f, h) \in \mathrm{Diff}(X) \times C^\infty(X, \mathbb{R})$ such that $\Theta_{f,h}$ is an embedding is dense.

4. Counterexample. In this section, we provide a counterexample to speculations that the observability could be generic even if map f is not a diffeomorphism. We will show that, even if the number of observations is equal to the dimension of manifold X, the property of observability is not generic. This counterexample is very similar to the one given in [3] for a slightly different purpose; in fact, it is based on the same idea: we will choose f and h such that (f, h) is an immersion with normal crossings; for such an immersion, a slight perturbation does not eliminate the crossings.

As compact manifold X, we take the n-dimensional torus $\mathbb{T}^n = (S^1)^n$, and we will produce a pair $(h, f) \in C^\infty(X, \mathbb{R}^n) \times C^\infty(X, X)$ such that there exists an open neighborhood $\mathscr{O}$ of (h, f) such that all pairs $(\bar{h}, \bar{f}) \in \mathscr{O}$ are not observable.

We take

$$\begin{aligned} h : \quad \mathbb{T}^n &\longrightarrow \mathbb{R}^n, \\ (e^{i\theta_1}, \dots, e^{i\theta_n}) &\longrightarrow (\sin\theta_1, \dots, \sin\theta_n) \end{aligned}$$

and

$$\begin{aligned} f : \quad \mathbb{T}^n &\longrightarrow \mathbb{T}^n, \\ (e^{i\theta_1}, \dots, e^{i\theta_n}) &\longrightarrow (e^{2i\theta_1}, e^{i(\theta_1+\theta_2)}, \dots, e^{i(\theta_{n-1}+\theta_n)}). \end{aligned}$$

In order to prove our claim, it suffices to prove that the mapping (h, f) is not one-to-one but is stable. As a matter of fact, (f, h) stable implies that the mapping $(\bar{f}, \bar{h})$ is equivalent to (f, h) if it belongs to some neighborhood of (f, h) (see [6]); so there exist $\varphi \in \mathrm{Diff}(X)$ and $\psi \in \mathrm{Diff}(X \times \mathbb{R}^n)$ such that $(\bar{f}, \bar{h}) = \psi \circ (f, h) \circ \varphi$, which implies that $(\bar{f}, \bar{h})$ is not one-to-one as (f, h). Obviously the fact that $(\bar{f}, \bar{h})$ is not one-to-one implies nonobservability: there exist $x_0 \neq x_1$ in X such that $\bar{h}(x_0) = \bar{h}(x_1)$ and $\bar{f}(x_0) = \bar{f}(x_1)$ and points x_0 and x_1 are indistinguishable.

We investigate now the injectivity of f, taking two points $x_0 = (e^{i\theta_1}, \dots, e^{i\theta_n})$ and $x_0' = (e^{i\theta_1'}, \dots, e^{i\theta_n'})$ in $\mathbb{T}^n$; they will have the same image under mapping (f, h) iff

$$\begin{cases} \theta_1' &= \theta_1 + k_1\pi, \\ \theta_2' &= \theta_2 + \theta_1 - \theta_1' + 2k_2\pi, \\ &\vdots \\ \theta_n' &= \theta_n + \theta_{n-1} - \theta_{n-1}' + 2k_n\pi, \\ \sin\theta_1' &= \sin\theta_1, \\ &\vdots \\ \sin\theta_n' &= \sin\theta_n, \end{cases}$$

which is equivalent to $\theta_i' = \theta_i$ or $\theta_i = \varepsilon_i\pi$ and $\theta_i' = (1-\varepsilon_i)\pi$ $(i = 1, \dots, n,\ \varepsilon_i \in \{0, 1\})$. This proves that (f, h) is not one-to-one. In order to prove that (f, h) is stable, it suffices to show that (f, h) is an immersion with normal crossings (see [6]). We recall that a mapping g between two manifolds X and Y is called *immersion with normal crossings* iff

(i) it is an immersion;

(ii) letting $g^{(s)} : X^{(s)} \to Y^s$, the restriction of g^s to $X^{(s)}$ and denoting by ΔY^s the subset of Y^s: $\Delta Y^s = \{(y, \dots, y) \mid y \in Y\}$, we ask that $g^{(s)} \pitchfork \Delta Y^s$ for every $s > 1$.

Notice that for our mapping, we have to check the second point of this definition only for $s = 2$ since it is impossible to have $(f, h)(x_1) = \cdots = (f, h)(x_p)$ for more than two points. Now, the computations to prove that (f, h) verifies the two preceding points are tedious but do not present any kind of difficulty, and so they are not reproduced here.

REFERENCES

[1] R. Abraham and J.W. Robbin, *Transversal Mappings and Flows*, W. A. Benjamin, New York, 1967.

[2] D. Aeyels, *Generic observability of differentiable systems*, SIAM J. Control Optim., 19 (1981), pp. 595–603.

[3] D. Aeyels, *On the number of samples necessary to achieve observability*, Systems Control Lett., 1 (1981), pp. 92–94.

[4] J.-P. Gauthier, H. Hammouri, and I. Kupka, *Observers for nonlinear systems*, in Proceedings of the IEEE 30th CDC, Brighton, UK, 1991, pp. 1483–1489.

[5] J.-P. Gauthier and I. Kupka, *Observability for systems with more outputs than inputs and asymptotic observers*, Math. Z., 223 (1996), pp. 47–78.

[6] M. Golubitsky and V. Guillemin, *Stable Mappings and Their Singularities*, Springer-Verlag, New York, 1986.

[7] J. Stark, *Delay embedding for forced systems* – I. *Deterministic forcing*, J. Nonlinear Sci., 9 (1999), pp. 255–332.

[8] F. Takens, *Detecting strange attractors in turbulence*, in Dynamical Systems and Turbulence, D.A. Rand and L.S. Young, eds., Lecture Notes in Math. 898, Springer-Verlag, Berlin, New York, 1981, pp. 366–381.

[9] H. Whitney, *Elementary structure of real algebraic varieties*, Ann. of Math. (2), 66 (1957), pp. 545–556.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1523–1544

THE TOPOLOGICAL ASYMPTOTIC FOR THE HELMHOLTZ EQUATION*

BESSEM SAMET†, SAMUEL AMSTUTZ†, AND MOHAMED MASMOUDI†

Abstract. The aim of the topological sensitivity analysis is to obtain an asymptotic expansion of a functional with respect to the creation of a small hole in the domain. In this paper such an expansion is obtained for the Helmholtz equation with a Dirichlet condition on the boundary of a circular hole. Some applications of this work to waveguide optimization are presented.

Key words. topological optimization, shape optimization, topological gradient, topological asymptotic, Helmholtz equation, waveguides, adjoint equation

AMS subject classifications. 49Q10, 49Q12, 78A25, 78A40, 78A45, 78A50, 35J05

DOI. S0363012902406801

1. Introduction. Classical shape optimization methods are based on the perturbation of the boundary of the initial shape. The initial and the final shapes have the same topology. The aim of topological optimization is to find an optimal shape without any a priori assumption about the topology of the structure. Many important contributions in this field are concerned with structural mechanics and, in particular, the minimization of the compliance (external work) subject to a volume constraint. In view of the fact that the optimal structure generally has a large number of small holes, most authors [3, 5, 15] have considered composite material optimization. Using the homogenization theory, Allaire and Kohn [3] exhibit a class of laminated materials with an explicit expression for the optimal material at any point of the structure. The range of application of this approach is quite restricted. For this reason, global optimization techniques like genetic algorithms and simulated annealing are used in order to solve more general problems [26]. Unfortunately, these methods are very slow.

The topological gradient has been introduced by Schumacher [27] to minimize a cost function $j(\Omega) = J(\Omega, u_\Omega)$, where u_Ω is the solution to a PDE defined in the domain Ω. The idea is to create a spherical hole $B(x, \varepsilon)$ of radius ε around a point x in Ω. Generally, an asymptotic expansion of the function j can be obtained in the following form:

$$j(\Omega \setminus \overline{B(x,\varepsilon)}) - j(\Omega) = f(\varepsilon)g(x) + o(f(\varepsilon)). \tag{1.1}$$

The function $f(\varepsilon)$ is positive and tends to zero with ε. We call this expansion the topological asymptotic. To minimize the criterion, we have to create holes where g is negative. The optimality condition $g \geq 0$ in Ω is exactly what Buttazzo and Dal Maso [6] have obtained for the Laplace equation, using a relaxed formulation. The topological gradient $g(x)$ has been computed by Schumacher [27] in the case of compliance minimization with Neumann condition on the boundary of the hole. In the same context, Sokolowski [25] gave some mathematical justifications in the

*Received by the editors April 30, 2002; accepted for publication (in revised form) April 2, 2003; published electronically November 6, 2003.
http://www.siam.org/journals/sicon/42-5/40680.html
†UFR MIG, Université Paul Sabatier and CNRS UMR 5640 MIP, 118 route de Narbonne 31062 Toulouse Cedex 4, France (samet@mip.ups-tlse.fr, amstutz@mip.ups-tlse.fr, masmoudi@mip.ups-tlse.fr).

plane stress case and generalized it to various cost functions. A topological sensitivity framework using an adaptation of the adjoint method and a truncation technique has been introduced in [16] in the case of an homogeneous Dirichlet condition imposed on the boundary of a circular hole. The fundamental property of the adjoint technique is to provide the variation of a function with respect to a parameter by using a solution u_Ω and an adjoint state p_Ω which do not depend on the chosen parameter. From the numerical viewpoint, only two systems have to be solved for obtaining $g(x)$ for all $x \in \Omega$. This observation leads to very efficient numerical algorithms. In [10, 11, 12], the topological sensitivity has been obtained in the contexts of linear elasticity, the Poisson equation, and the Stokes problem with general shape functions and arbitrary shaped holes. These publications are concerned with PDE operators whose symbols are homogeneous polynomials.

In this paper, we are interested in the differential operator

$$P = \sum_{i=1}^{2} \frac{\partial^2}{\partial {x_i}^2} + k^2,$$

whose symbol is not homogenous. First, an adaptation of the adjoint method to the topological context is proposed in section 2 for the operator P. Next, a waveguide problem, the truncation method, and the explicit expression of the topological asymptotic are presented in section 3. Finally, an optimization algorithm and some applications of the topological gradient to waveguide optimization are given in section 4. This work was done in collaboration with Alcatel Space Industries.

2. A generalized adjoint method. In this section, the adjoint method is adapted to topological optimization. Let $\mathcal{V}$ be a fixed complex Hilbert space. For $\varepsilon \geq 0$, let $a_\varepsilon(.,.)$ be a sesquilinear and continuous form on $\mathcal{V}$ and l_ε be a semilinear and continuous form on $\mathcal{V}$. We consider the following assumptions.

Hypothesis 1. There exists a sesquilinear and continuous form δ_a, a semilinear and continuous form δ_l, and a real function $f(\varepsilon) > 0$ defined on ${\mathbb{R}^*}_+$ such that

$$\lim_{\varepsilon \to 0} f(\varepsilon) = 0, \tag{2.1}$$

$$\|a_\varepsilon - a_0 - f(\varepsilon)\delta_a\|_{\mathcal{L}_2(\mathcal{V})} = o(f(\varepsilon)), \tag{2.2}$$

$$\|l_\varepsilon - l_0 - f(\varepsilon)\delta_l\|_{\mathcal{L}(\mathcal{V})} = o(f(\varepsilon)), \tag{2.3}$$

where $\mathcal{L}(\mathcal{V})$ (respectively, $\mathcal{L}_2(\mathcal{V})$) denotes the space of continuous and semilinear (respectively, sesquilinear) forms on $\mathcal{V}$.

Hypothesis 2. There exists a constant $\alpha > 0$ such that

$$\inf_{u \neq 0} \sup_{v \neq 0} \frac{|a_0(u,v)|}{\|u\|_{\mathcal{V}} \|v\|_{\mathcal{V}}} \geq \alpha.$$

We say that a_0 satisfies the inf-sup condition.

According to (2.2), there exists a constant $\beta > 0$ (independent of ε) such that

$$\inf_{u \neq 0} \sup_{v \neq 0} \frac{|a_\varepsilon(u,v)|}{\|u\|_{\mathcal{V}} \|v\|_{\mathcal{V}}} \geq \beta \quad \forall \varepsilon \geq 0.$$

For $\varepsilon \geq 0$, we suppose that the following problem has one solution: find $u_\varepsilon \in \mathcal{V}$ such that

$$a_\varepsilon(u_\varepsilon, v) = l_\varepsilon(v) \qquad \forall v \in \mathcal{V}. \tag{2.4}$$

According to Hypothesis 2, this solution is unique. We have the following lemma.

LEMMA 2.1. *If Hypotheses* 1 *and* 2 *are satisfied, then*

$$\|u_\varepsilon - u_0\|_{\mathcal{V}} = O(f(\varepsilon)).$$

Proof. It follows from Hypothesis 2 that there exists $v_\varepsilon \in \mathcal{V}, v_\varepsilon \neq 0$, such that

$$\beta\|u_\varepsilon - u_0\|_{\mathcal{V}}\|v_\varepsilon\|_{\mathcal{V}} \leq |a_\varepsilon(u_\varepsilon - u_0, v_\varepsilon)|,$$

which implies

$$\begin{aligned}
&\beta\|u_\varepsilon - u_0\|_{\mathcal{V}}\|v_\varepsilon\|_{\mathcal{V}}\\
&\leq |a_\varepsilon(u_0, v_\varepsilon) - l_\varepsilon(v_\varepsilon)|\\
&= |a_\varepsilon(u_0, v_\varepsilon) - (l_\varepsilon - l_0 - f(\varepsilon)\delta_l)(v_\varepsilon) - l_0(v_\varepsilon) - f(\varepsilon)\delta_l(v_\varepsilon)|\\
&= |(a_\varepsilon(u_0, v_\varepsilon) - a_0(u_0, v_\varepsilon)) - (l_\varepsilon - l_0 - f(\varepsilon)\delta_l)(v_\varepsilon) - f(\varepsilon)\delta_l(v_\varepsilon)|\\
&\leq |a_\varepsilon(u_0, v_\varepsilon) - a_0(u_0, v_\varepsilon) - f(\varepsilon)\delta_a(u_0, v_\varepsilon)| + |l_\varepsilon(v_\varepsilon) - l_0(v_\varepsilon) - f(\varepsilon)\delta_l(v_\varepsilon)|\\
&\quad + f(\varepsilon)(|\delta_a(u_0, v_\varepsilon)| + |\delta_l(v_\varepsilon)|).
\end{aligned}$$

Using Hypothesis 1, we obtain

$$\beta\|u_\varepsilon - u_0\|_{\mathcal{V}}\|v_\varepsilon\|_{\mathcal{V}} \leq \left(o(f(\varepsilon)) + f(\varepsilon)(\|\delta_a\|_{\mathcal{L}_2(\mathcal{V})}\|u_0\|_{\mathcal{V}} + \|\delta_l\|_{\mathcal{L}(\mathcal{V})})\right)\|v_\varepsilon\|_{\mathcal{V}}. \qquad \square$$

Consider now a cost function $j(\varepsilon) = J(u_\varepsilon)$, where the functional J satisfies

$$J(u + h) = J(u) + \Re(L_u(h)) + o(\|h\|_{\mathcal{V}}) \qquad \forall u, h \in \mathcal{V}. \tag{2.5}$$

Here, L_u is a linear and continuous form on $\mathcal{V}$. We suppose that the following problem has a unique solution p_0, called the adjoint state: find $p_0 \in \mathcal{V}$ such that

$$a_0(v, p_0) = -L_{u_0}(v) \qquad \forall v \in \mathcal{V}. \tag{2.6}$$

For $\varepsilon \geq 0$, we define the Lagrangian operator $\mathcal{L}_\varepsilon$ by

$$\mathcal{L}_\varepsilon(u, v) = J(u) + a_\varepsilon(u, v) - l_\varepsilon(v) \qquad \forall u, v \in \mathcal{V}.$$

The next theorem gives the asymptotic expansion of $j(\varepsilon)$.

THEOREM 2.2. *If Hypotheses* 1 *and* 2 *are satisfied, then*

$$j(\varepsilon) - j(0) = f(\varepsilon)\Re(\delta_{\mathcal{L}}(u_0, p_0)) + o(f(\varepsilon)), \tag{2.7}$$

where u_0 is the solution to (2.4) *with $\varepsilon = 0$, p_0 is the adjoint state solution to problem* (2.6)*, and*

$$\delta_{\mathcal{L}}(u, v) = \delta_a(u, v) - \delta_l(v) \quad \forall u, v \in \mathcal{V}.$$

Proof. We have that

$$j(\varepsilon) = \mathcal{L}_\varepsilon(u_\varepsilon, v) \qquad \forall \varepsilon \geq 0, \quad \forall v \in \mathcal{V}.$$

Next, choosing $v = p_0$, we obtain

$$\begin{aligned}
j(\varepsilon) - j(0) &= \mathcal{L}_\varepsilon(u_\varepsilon, p_0) - \mathcal{L}_0(u_0, p_0)\\
&= J(u_\varepsilon) - J(u_0) + a_\varepsilon(u_\varepsilon, p_0) - a_0(u_0, p_0) + l_0(p_0) - l_\varepsilon(p_0)\\
&= J(u_\varepsilon) - J(u_0) + \Re(a_\varepsilon(u_\varepsilon, p_0) - a_0(u_0, p_0)) - \Re(l_\varepsilon(p_0) - l_0(p_0))\\
&= J(u_\varepsilon) - J(u_0) + \Re(a_\varepsilon(u_\varepsilon, p_0) - a_0(u_\varepsilon, p_0) + a_0(u_\varepsilon - u_0, p_0))\\
&\quad -\Re(l_\varepsilon(p_0) - l_0(p_0) - f(\varepsilon)\delta_l(p_0)) - f(\varepsilon)\Re(\delta_l(p_0)).
\end{aligned}$$

Using (2.5), we have that

$$J(u_\varepsilon) - J(u_0) = \Re(L_{u_0}(u_\varepsilon - u_0)) + o(\|u_\varepsilon - u_0\|_{\mathcal{V}}).$$

Hence,

$$\begin{aligned} &j(\varepsilon) - j(0) \\ &= \Re(a_\varepsilon(u_\varepsilon, p_0) - a_0(u_\varepsilon, p_0)) + \Re(a_0(u_\varepsilon - u_0, p_0) + L_{u_0}(u_\varepsilon - u_0)) + o(\|u_\varepsilon - u_0\|_{\mathcal{V}}) \\ &\quad -\Re(l_\varepsilon(p_0) - l_0(p_0) - f(\varepsilon)\delta_l(p_0)) - f(\varepsilon)\Re(\delta_l(p_0)). \end{aligned}$$

Using that p_0 is the adjoint solution, we obtain

$$\begin{aligned} j(\varepsilon) - j(0) &= \Re(a_\varepsilon(u_\varepsilon, p_0) - a_0(u_\varepsilon, p_0)) + o(\|u_\varepsilon - u_0\|_{\mathcal{V}}) \\ &\quad -\Re(l_\varepsilon(p_0) - l_0(p_0) - f(\varepsilon)\delta_l(p_0)) - f(\varepsilon)\Re(\delta_l(p_0)) \\ &= \Re((a_\varepsilon - a_0)(u_0, p_0)) + \Re((a_\varepsilon - a_0)(u_\varepsilon - u_0, p_0)) + o(\|u_\varepsilon - u_0\|_{\mathcal{V}}) \\ &\quad -\Re(l_\varepsilon(p_0) - l_0(p_0) - f(\varepsilon)\delta_l(p_0)) - f(\varepsilon)\Re(\delta_l(p_0)). \end{aligned}$$

It follows from Hypothesis 1 that

$$\begin{aligned} j(\varepsilon) - j(0) = f(\varepsilon)\Re(\delta_a(u_0, p_0)) + o(f(\varepsilon)) + f(\varepsilon)\Re(\delta_a(u_\varepsilon - u_0, p_0)) + o(f(\varepsilon))\|u_\varepsilon - u_0\|_{\mathcal{V}} \\ + o(\|u_\varepsilon - u_0\|_{\mathcal{V}}) - f(\varepsilon)\Re(\delta_l(p_0)). \end{aligned}$$

Finally, from Lemma 2.1 and the hypothesis $\lim_{\varepsilon\to 0} f(\varepsilon) = 0$, we have

$$j(\varepsilon) = j(0) + f(\varepsilon)\Re(\delta_a(u_0, p_0) - \delta_l(p_0)) + o(f(\varepsilon)),$$

since δ_a is continuous by assumption. ∎

3. A waveguide problem. In this section, we study a problem of a waveguide as a component of a spatial antenna feeding system. Because the waveguide $\mathcal{O}$ has a uniform thickness, $\mathcal{O} = \Omega\times]a, b[$, $\Omega \subset \mathbb{R}^2$, and the electric field has a vertical polarization (normal to Ω), the three-dimensional problem can be reduced to a two-dimensional problem in Ω, called the H-plane model. We assume that Ω is a domain of $\mathbb{R}^2$ with a regular boundary $\Gamma = \Gamma_0 \cup \Gamma_1 \cup \cdots \cup \Gamma_N$, $N \in \mathbb{N}^*$. We denote by u_Ω the normal component to Ω of the electric field. It is a solution to the Helmholtz problem:

$$\left\{ \begin{array}{lll} \Delta u_\Omega + k^2 u_\Omega & = 0 & \text{in } \Omega, \\ u_\Omega & = 0 & \text{on } \Gamma_0, \\ \partial_n u_\Omega - iku_\Omega & = h_j & \text{on } \Gamma_j, j = 1, 2, \ldots, N, \end{array} \right. \tag{3.1}$$

where $\partial_n u_\Omega$ is the normal derivative of u_Ω, $k \in \{k \in \mathbb{C}^*/\Im(k) \geq 0\}$, and $h_j \in H_{00}^{\frac{1}{2}}(\Gamma_j)'$ for all $j \in \{1, 2, \ldots, N\}$. The first boundary condition means that Γ_0 is a perfect metallic surface. When $h_j = 0$, the last equation is an approximate absorbing boundary condition (the normal incident plane waves are completely absorbed). When $h_j \neq 0$, it is a transmission condition. We prove in section 5.1 that problem (3.1) has one and only one solution in the Hilbert space

$$\mathcal{V}_\Omega = \{u \in H^1(\Omega), u = 0 \text{ on } \Gamma_0\}. \tag{3.2}$$

Here and in the following, all the Sobolev spaces involve complex-valued functions.

For a given $x \in \Omega$, let us consider the perforated open set $\Omega_\varepsilon = \Omega\backslash\overline{B(x, \varepsilon)}$, where x is a point of Ω and $B(x, \varepsilon)$ is the ball of center x and of radius ε (see Figure 1). We

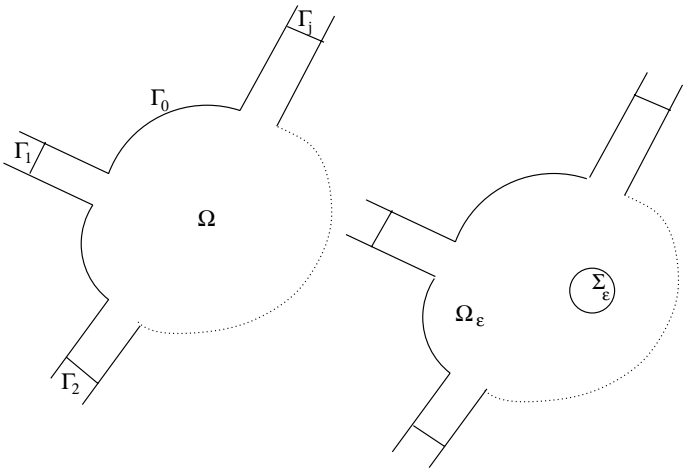


FIG. 1. *The initial domain and the same domain after the perforation.*

assume that $\varepsilon > 0$ is small enough, and we denote $\Sigma_\varepsilon = \partial B(x,\varepsilon)$. Our aim is to get the sensitivity analysis of u_{Ω_ε}, being the unique solution (see section 5.1) to

$$\left\{\begin{array}{lll} \Delta u_{\Omega_\varepsilon} + k^2 u_{\Omega_\varepsilon} & = 0 & \text{in } \Omega_\varepsilon, \\ u_{\Omega_\varepsilon} & = 0 & \text{on } \Gamma_0, \\ u_{\Omega_\varepsilon} & = 0 & \text{on } \Sigma_\varepsilon, \\ \partial_n u_{\Omega_\varepsilon} - iku_{\Omega_\varepsilon} & = h_j & \text{on } \Gamma_j, j = 1,2,\ldots,N, \end{array}\right. \tag{3.3}$$

with respect to ε at $\varepsilon = 0$. The solution of problem (3.3) is defined on the variable open set Ω_ε; thus it belongs to a functional space which depends on ε. Hence, if we want to derive the asymptotic expansion of a function of the form

$$j(\varepsilon) = J(u_{\Omega_\varepsilon}), \tag{3.4}$$

we cannot apply directly the tools of section 2, which require a fixed functional space. In classical shape optimization, this requirement can be satisfied with the help of a domain parameterization technique [13, 20, 17]. This technique involves a fixed domain and a bi-Lipshitz map between this domain and the modified one. In the topology optimization context, such a map does not exist between Ω and Ω_ε. However, a functional space independent of ε can be constructed by using a domain truncation technique.

3.1. The domain truncation. Let $R > \varepsilon$ be such that the ball $B(x,R)$ is included in Ω. The boundary of $B(x,R)$ is denoted by Σ_R. The truncated domain $\Omega\backslash\overline{B(x,R)}$ is denoted by Ω_R, and D_ε denotes the corona $B(x,R)\backslash\overline{B(x,\varepsilon)}$ (see Figure 2).

For a $\Psi \in H^{\frac{1}{2}}(\Sigma_R)$, we consider u_Ψ^ε the solution to the problem

$$\left\{\begin{array}{lll} \Delta u_\Psi^\varepsilon + k^2 u_\Psi^\varepsilon & = 0 & \text{in } D_\varepsilon, \\ u_\Psi^\varepsilon & = \Psi & \text{on}\Sigma_R, \\ u_\Psi^\varepsilon & = 0 & \text{on } \Sigma_\varepsilon \end{array}\right. \tag{3.5}$$

and the *Dirichlet-to-Neumann* operator

$$\begin{array}{rccc} T^\varepsilon : & H^{1/2}(\Sigma_R) & \longrightarrow & H^{-1/2}(\Sigma_R), \\ & \Psi & \longmapsto & T^\varepsilon\Psi = \nabla u_\Psi^\varepsilon . n_{|\Sigma_R}, \end{array}$$

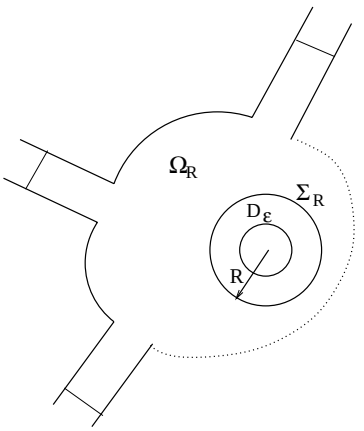


FIG. 2. *The truncated domain.*

where $n_{|\Sigma_R}$ denotes the outward normal to the boundary Σ_R. Using the Poincaré inequality, we obtain that, for $\varepsilon < R < (\sqrt{2}|k|)^{-1}$, problem (3.5) is coercive. Hence it has one and only one solution.

We consider the truncated problem: find u_ε such that

$$\left\{\begin{array}{lll} \Delta u_\varepsilon + k^2 u_\varepsilon & = 0 & \text{in } \Omega_R, \\ u_\varepsilon & = 0 & \text{on } \Gamma_0, \\ \partial_n u_\varepsilon + T^\varepsilon u_\varepsilon & = 0 & \text{on } \Sigma_R, \\ \partial_n u_\varepsilon - iku_\varepsilon & = h_j & \text{on } \Gamma_j, j = 1, 2, \ldots, N. \end{array}\right. \tag{3.6}$$

The variational formulation associated to problem (3.6) is the following: find $u_\varepsilon \in \mathcal{V}_R$ such that

$$a_\varepsilon(u_\varepsilon, v) = l(v) \qquad \forall v \in \mathcal{V}_R, \tag{3.7}$$

where the functional space $\mathcal{V}_R$, the sesquilinear form a_ε, and the semilinear form l are defined by

$$\mathcal{V}_R = \{u \in H^1(\Omega_R), u = 0 \text{ on } \Gamma_0\}, \tag{3.8}$$

$$a_\varepsilon(u, v) = \int_{\Omega_R} \nabla u.\overline{\nabla v}\, dx - k^2 \int_{\Omega_R} u\overline{v}\, dx + \int_{\Sigma_R} (T^\varepsilon u)\overline{v}\, d\gamma(x) \tag{3.9}$$

$$-ik \sum_{j=1}^{N} \int_{\Gamma_j} u\overline{v}\, d\gamma(x), \tag{3.10}$$

$$l(v) = \sum_{j=1}^{N} \int_{\Gamma_j} h_j \overline{v}\, d\gamma(x). \tag{3.11}$$

Here, $\nabla u.\overline{\nabla v} = \sum_{i=1}^{2} \frac{\partial u}{\partial x_i} \frac{\partial \overline{v}}{\partial x_i}$ and $d\gamma(x)$ is the Lebesgue measure on the boundary. The following result is standard in PDE theory.

PROPOSITION 3.1. *Problem* (3.6) *has one and only one solution in* $\mathcal{V}_R$ *which is the restriction to* Ω_R *of the solution to* (3.3).

Proof. Existence: Applying the definition of T^ε, we prove that the restriction to Ω_R of the solution to (3.3) is a solution to (3.6).

Uniqueness: Any solution u to problem (3.6) can be extended in Ω_ε to the solution to problem (3.3): we use the solution u^ε_Ψ to (3.5) with $\Psi = u_{|\Sigma_R}$. ∎

We have now at our disposal the fixed Hilbert space $\mathcal{V}_R$ required by section 2. We assume that the function J is defined in a neighbor part of Γ. Then we have

$$j(\varepsilon) = J(u_{\Omega_\varepsilon}) = J(u_\varepsilon) \qquad \forall \varepsilon \geq 0. \tag{3.12}$$

3.2. Variation of the sesquilinear form. The variation of the sesquilinear form $a_\varepsilon - a_0$ reads

$$a_\varepsilon(u,v) - a_0(u,v) = \int_{\Sigma_R} \left((T^\varepsilon - T^0)u\right) \overline{v}\, d\gamma(x). \tag{3.13}$$

Hence, the problem reduces to the computation of $(T^\varepsilon - T^0)\Psi$ for $\Psi = u_{|\Sigma_R}$.
We have the following proposition.

PROPOSITION 3.2. *The solution u^ε_Ψ to problem* (3.5) *and the operator T^ε are given by the explicit expressions:*

$$u^\varepsilon_\psi(r,\theta) = \sum_{n\in\mathbb{Z}} \frac{J_n(kr)Y_n(k\varepsilon) - J_n(k\varepsilon)Y_n(kr)}{J_n(kR)Y_n(k\varepsilon) - Y_n(kR)J_n(k\varepsilon)} \psi_n e^{in\theta}$$

and

$$T^\varepsilon\psi = k \sum_{n\in\mathbb{Z}} \frac{J_n'(kR)Y_n(k\varepsilon) - J_n(k\varepsilon)Y_n'(kR)}{J_n(kR)Y_n(k\varepsilon) - Y_n(kR)J_n(k\varepsilon)} \psi_n e^{in\theta}, \tag{3.14}$$

where (r,θ) are the polar coordinates in $\mathbb{R}^2$, (Ψ_n) are the Fourier coefficients of Ψ, and (J_n) and (Y_n) are, respectively, the Bessel functions of the first and the second kind.

Proof. We have in polar coordinates

$$u^\varepsilon_\psi(r,\theta) = \sum_{n\in\mathbb{Z}} c_n(r) e^{in\theta},$$

where $c_n(r)$ satisfies the differential equation:

$$\frac{d^2c_n}{dr^2} + \frac{1}{r}\frac{dc_n}{dr} + \left(k^2 - \frac{n^2}{r^2}\right) c_n(r) = 0 \qquad \forall n \in \mathbb{Z},$$

and thus c_n is a linear combination of J_n and Y_n Bessel functions:

$$c_n(r) = a_n J_n(kr) + b_n Y_n(kr) \quad \forall n \in \mathbb{Z}.$$

Using the boundary conditions, we obtain

$$a_n = \frac{Y_n(k\varepsilon)}{J_n(kR)Y_n(k\varepsilon) - Y_n(kR)J_n(k\varepsilon)}\psi_n, \quad b_n = \frac{-J_n(k\varepsilon)}{J_n(kR)Y_n(k\varepsilon) - Y_n(kR)J_n(k\varepsilon)}\psi_n. \qquad ∎$$

In particular, for $\varepsilon = 0$ we have the following proposition.

PROPOSITION 3.3. *The solution u^0_Ψ and the operator T^0 are given by the explicit expressions*

$$u^0_\psi(r,\theta) = \sum_{n\in\mathbb{Z}} \frac{J_n(kr)}{J_n(kR)} \psi_n e^{in\theta}$$

and

$$T^0\psi = k\sum_{n\in\mathbb{Z}} \frac{J_n'(kR)}{J_n(kR)}\psi_n e^{in\theta}, \tag{3.15}$$

where u_ψ^0 *is the solution to* (3.5) *for* $\varepsilon = 0$.

For $\Psi \in H^s(\Sigma_R)$, let

$$\|\psi\|_{s,\Sigma_R}^2 = \sum_{n\in\mathbb{Z}} |\psi_n|^2(1+|n|)^{2s} \tag{3.16}$$

be the norm of Ψ in this space. The so defined norm is equivalent to the usual norm of $H^s(\Sigma_R)$. We introduce the operator:

$$\begin{array}{rccc} \delta_T: & H^{1/2}(\Sigma_R) & \longrightarrow & H^{-1/2}(\Sigma_R), \\ & \Psi & \longmapsto & \delta_T\Psi = \frac{1}{RJ_0^2(kR)}\Psi_0. \end{array}$$

We have the following lemma.

LEMMA 3.4. *We have that*

$$\left\| T^\varepsilon - T^0 - \frac{-1}{\log(\varepsilon)}\delta_T \right\|_{\mathcal{L}(H^{1/2}(\Sigma_R);H^{-1/2}(\Sigma_R))} = o\left(\frac{-1}{\log(\varepsilon)}\right).$$

Proof. Let $\Psi \in H^{\frac{1}{2}}(\Sigma_R)$. Using the series (3.14) and (3.15), we obtain

$$\begin{aligned}(T^\varepsilon - T^0)\psi &= k\sum_{n\in\mathbb{Z}} \frac{J_n'(kR)Y_n(k\varepsilon) - J_n(k\varepsilon)Y_n'(kR)}{J_n(kR)Y_n(k\varepsilon) - Y_n(kR)J_n(k\varepsilon)}\psi_n e^{in\theta} - k\sum_{n\in\mathbb{Z}} \frac{J_n'(kR)}{J_n(kR)}\psi_n e^{in\theta} \\ &= k\sum_{n\in\mathbb{Z}^*} \frac{J_n'(kR)Y_n(k\varepsilon) - J_n(k\varepsilon)Y_n'(kR)}{J_n(kR)Y_n(k\varepsilon) - Y_n(kR)J_n(k\varepsilon)}\psi_n e^{in\theta} - k\sum_{n\in\mathbb{Z}^*} \frac{J_n'(kR)}{J_n(kR)}\psi_n e^{in\theta} \\ &\quad -k\frac{Y_0'(kR)J_0(kR) - Y_0(kR)J_0'(kR)}{J_0^2(kR)} \frac{J_0(k\varepsilon)J_0(kR)}{J_0(kR)Y_0(k\varepsilon) - Y_0(kR)J_0(k\varepsilon)}\psi_0.\end{aligned}$$

We have that [1]

$$\begin{aligned}\frac{Y_0'(kR)J_0(kR) - Y_0(kR)J_0'(kR)}{J_0^2(kR)} &= \frac{W\{J_0(kR), Y_0(kR)\}}{J_0^2(kR)} \\ &= \frac{2}{\pi kR}\frac{1}{J_0^2(kR)},\end{aligned}$$

where W is the Wronskian. Then

$$\begin{aligned}(T^\varepsilon - T^0)\psi &= k\sum_{n\in\mathbb{Z}^*} \frac{J_n(k\varepsilon)Y_n(kR)}{Y_n(k\varepsilon)J_n(kR) - Y_n(kR)J_n(k\varepsilon)}\left(\frac{J_n'(kR)}{J_n(kR)} - \frac{Y_n'(kR)}{Y_n(kR)}\right)\psi_n e^{in\theta} \\ &\quad - \frac{2}{\pi}\frac{J_0(k\varepsilon)J_0(kR)}{J_0(kR)Y_0(k\varepsilon) - Y_0(kR)J_0(k\varepsilon)}\frac{1}{RJ_0^2(kR)}\psi_0. \end{aligned} \tag{3.17}$$

We have the following formula [1]:

$$Y_0(k\varepsilon) = \frac{2}{\pi}\left(\log\left(\frac{k\varepsilon}{2}\right) + \gamma\right)J_0(k\varepsilon) + \varepsilon\alpha(\varepsilon), \tag{3.18}$$

where γ denotes Euler's constant and $\alpha(\varepsilon) \to 0$ when $\varepsilon \to 0$. We insert (3.18) into (3.17):

$$(T^\varepsilon - T^0)\psi = \varepsilon R_\varepsilon \Psi + \frac{-1}{\log(\varepsilon)} \left(1 + \frac{M}{\log(\varepsilon)} + \varepsilon\theta(\varepsilon)\right)^{-1} \delta_T \Psi,$$

where M is a constant independent of ε, $\theta(\varepsilon) \to 0$ when $\varepsilon \to 0$ and

$$R_\varepsilon \psi = \sum_{n\in\mathbb{Z}^*} \frac{k}{\varepsilon} \frac{J_n(k\varepsilon)Y_n(kR)}{Y_n(k\varepsilon)J_n(kR) - Y_n(kR)J_n(k\varepsilon)} \left(\frac{J_n'(kR)}{J_n(kR)} - \frac{Y_n'(kR)}{Y_n(kR)}\right) \psi_n e^{in\theta}.$$

Then

$$\left(T^\varepsilon - T^0 - \frac{-1}{\log(\varepsilon)}\delta_T\right)\psi = \varepsilon R_\varepsilon \psi + O(1)\left(\frac{-1}{\log(\varepsilon)}\right)^2 \frac{1}{RJ_0^2(kR)}\psi_0.$$

Using (3.16), we have

$$\| R_\varepsilon \psi \|^2_{-\frac{1}{2};\Sigma_R} = \sum_{n\in\mathbb{Z}^*} \frac{|k|^2}{\varepsilon^2} \left|\frac{J_n(k\varepsilon)Y_n(kR)}{Y_n(k\varepsilon)J_n(kR) - Y_n(kR)J_n(k\varepsilon)}\right|^2 \cdot \left|\frac{J_n'(kR)}{J_n(kR)(1+|n|)} - \frac{Y_n'(kR)}{Y_n(kR)(1+|n|)}\right|^2 (1+|n|)|\psi_n|^2.$$

Let us prove that there exists a constant $c > 0$ (independent of Ψ and ε) such that for all $0 < \varepsilon < \varepsilon_0 < R$,

$$\|R_\varepsilon \psi\|_{-\frac{1}{2};\Sigma_R} \leq c \| \psi\|_{\frac{1}{2};\Sigma_R}.$$

We have [1]

$$\frac{1}{1+|n|}\frac{J_n'(kR)}{J_n(kR)} = -\frac{1}{1+|n|}\frac{J_{n+1}(kR)}{J_n(kR)} + \frac{n}{1+|n|}\frac{1}{kR}$$

and for $n \to \infty$

$$J_n(z) \sim (2\pi n)^{-\frac{1}{2}} \left(\frac{ez}{2n}\right)^n.$$

Then

$$\lim_{n\to\infty} \frac{1}{1+|n|}\frac{J_{n+1}(kR)}{J_n(kR)} = 0$$

and

$$\left|\frac{1}{1+|n|}\frac{J_n'(kR)}{J_n(kR)}\right| \leq c \quad \forall n \in \mathbb{Z}^*.$$

Here and in what follows, c is a positive constant independent of the data (e.g., of ε and n). Similarly, we have

$$\left|\frac{1}{1+|n|}\frac{Y_n'(kR)}{Y_n(kR)}\right| \leq c \quad \forall n \in \mathbb{Z}^*.$$

Hence,

$$\left|\frac{J_n'(kR)}{J_n(kR)(1+|n|)} - \frac{Y_n'(kR)}{Y_n(kR)(1+|n|)}\right| \leq c \quad \forall n \in \mathbb{Z}^*.$$

We denote

$$f_n(\varepsilon) = \frac{1}{\varepsilon}\left|\frac{J_n(k\varepsilon)Y_n(kR)}{Y_n(k\varepsilon)J_n(kR) - Y_n(kR)J_n(k\varepsilon)}\right|.$$

We have also

$$f_n(\varepsilon) = \left|\frac{\varepsilon J_n(kR)Y_n(k\varepsilon)}{J_n(k\varepsilon)Y_n(kR)} - \varepsilon\right|^{-1}.$$

We show in section 5.3 that there exist n_0 and ε_0 such that

$$\left|\varepsilon\frac{J_n(kR)}{J_n(k\varepsilon)}\right| \geq c\left(\frac{R}{\varepsilon}\right)^{n-1} \qquad \forall n \geq n_0, \quad \forall \varepsilon < \varepsilon_0 \tag{3.19}$$

and

$$\left|\frac{Y_n(k\varepsilon)}{Y_n(kR)}\right| \geq c\left(\frac{R}{\varepsilon}\right)^{n} \qquad \forall n \geq n_0, \quad \forall \varepsilon < \varepsilon_0. \tag{3.20}$$

Using (3.19) and (3.20), we obtain

$$\left|\frac{\varepsilon Y_n(k\varepsilon)J_n(kR)}{J_n(k\varepsilon)Y_n(kR)}\right| \geq c \qquad \forall n \geq n_0, \quad \forall \varepsilon < \varepsilon_0$$

and

$$f_n(\varepsilon) \leq c \qquad \forall n \geq n_0, \quad \forall \varepsilon < \varepsilon_0.$$

For $p \in \{1, 2, \ldots, n_0 - 1\}$, we have $f_p(\varepsilon) \to 0$ when $\varepsilon \to 0$. Then

$$f_n(\varepsilon) \leq c \quad \forall n \in \mathbb{Z}^*, \quad \forall \varepsilon < \varepsilon_0.$$

Hence

$$\|R_\varepsilon \psi\|_{-\frac{1}{2},\Sigma_R} \leq c \parallel \psi \parallel_{\frac{1}{2};\Sigma_R} \quad \forall \psi \in H^{\frac{1}{2}}(\Sigma_R).$$

This completes the proof. □

From this lemma we obtain the following proposition.

PROPOSITION 3.5. *Let δ_a be the sesquilinear and continuous form defined on $\mathcal{V}_R$ by*

$$\delta_a(u,v) = \frac{u^{mean}}{J_0(kR)}\frac{\overline{v^{mean}}}{J_0(kR)},$$

where u^{mean} and v^{mean} denote, respectively, the mean values of u and v on Σ_R. We have

$$\left|a_\varepsilon(u,p) - a_0(u,p) - \frac{-2\pi}{\log(\varepsilon)}\delta_a(u,p)\right| = o\left(\frac{-1}{\log(\varepsilon)}\right)\|u\|_{\mathcal{V}_R}\|p\|_{\mathcal{V}_R} \qquad \forall u,p \in \mathcal{V}_R.$$

3.3. The asymptotic expansion. We prove in section 5.2 that the sesquilinear form a_0 satisfies Hypothesis 2 (inf-sup condition).
The adjoint problem is the following: find $p_\Omega \in \mathcal{V}_\Omega$ such that

$$\int_\Omega (\nabla v.\overline{\nabla p_\Omega} - k^2 v\overline{p_\Omega})\,dx - ik\sum_{j=1}^{N}\int_{\Gamma_j} v\overline{p_\Omega}\,d\gamma(x) = -L_{u_\Omega}(v) \quad \forall v \in \mathcal{V}_\Omega. \tag{3.21}$$

This problem has one and only one solution (see section 5.1). If $L_{u_\Omega} \in H_{00}^{\frac{1}{2}}(\Gamma_m)'$, $m \in \{1, 2, \ldots, N\}$, the strong formulation of problem (3.21) is

$$\begin{cases} \Delta p_\Omega + \overline{k}^2 p_\Omega &= 0 & \text{in } \Omega, \\ p_\Omega &= 0 & \text{on } \Gamma_0, \\ \partial_n p_\Omega + i\overline{k} p_\Omega &= -L_{u_\Omega} & \text{on } \Gamma_m, \\ \partial_n p_\Omega + i\overline{k} p_\Omega &= 0 & \text{on } \Gamma_j, j \in \{1, 2, \ldots, N\}\backslash\{m\}. \end{cases} \tag{3.22}$$

Hence, all the assumptions of section 2 are satisfied and we can apply the adjoint method. Then we have the following theorem.

THEOREM 3.6. *The function j has the following asymptotic expansion:*

$$j(\varepsilon) - j(0) = \frac{-2\pi}{\log(\varepsilon)}\Re(u_\Omega(x)\overline{p_\Omega}(x)) + o\left(\frac{-1}{\log(\varepsilon)}\right).$$

Proof. Using Theorem 2.2, we obtain

$$j(\varepsilon) - j(0) = \frac{-2\pi}{\log(\varepsilon)}\Re(\delta_a(u_0, p_0)) + o\left(\frac{-1}{\log(\varepsilon)}\right),$$

where u_0 is the solution to (3.7) for $\varepsilon = 0$ and p_0 is the solution to the adjoint problem

$$a_0(v, p_0) = -L_{u_0}(v) \qquad \forall v \in \mathcal{V}_R. \tag{3.23}$$

As observed in Proposition 3.1, u_0 is the restriction to Ω_R of u_Ω. Let us prove that the same property holds for p_0 and p_Ω. For $v \in \mathcal{V}_\Omega$, we denote by p_R and v_R the restriction of p_Ω and v to Ω_R. On the one hand, we have

(3.24)

$$\begin{aligned} &\int_\Omega (\nabla v.\overline{\nabla p_\Omega} - k^2 v\overline{p_\Omega})\,dx - ik\sum_{j=1}^{N}\int_{\Gamma_j} v\overline{p_\Omega}\,d\gamma(x) \\ &= \int_{\Omega_R} (\nabla v_R.\overline{\nabla p_R} - k^2 v_R\overline{p_R})\,dx - ik\sum_{j=1}^{N}\int_{\Gamma_j} v_R\overline{p_R}\,d\gamma(x) + \int_{D_0} (\nabla v.\overline{\nabla p_\Omega} - k^2 v\overline{p_\Omega})\,dx \\ &= \int_{\Omega_R} (\nabla v_R.\overline{\nabla p_R} - k^2 v_R\overline{p_R})\,dx - ik\sum_{j=1}^{N}\int_{\Gamma_j} v_R\overline{p_R}\,d\gamma(x) + \int_{\Sigma_R} (T^0 v_R)\overline{p_R}\,d\gamma(x) \\ &= a_0(v_R, p_R). \end{aligned}$$

On the other hand, due to the fact that J is defined in a neighbor part of Γ, we have that $J(u) = J(u_R)$ for all $u \in \mathcal{V}_\Omega$. Hence

$$L_{u_\Omega}(v) = L_{u_0}(v_R). \tag{3.25}$$

Then, gathering (3.24), (3.21), and (3.25), we obtain

$$a_0(v_R, p_R) = -L_{u_0}(v_R) \qquad \forall v_R \in \mathcal{V}_R,$$

which proves that p_R is the solution to (3.23). Then p_0 is the restriction to Ω_R of p_Ω. It remains to prove that $\delta_a(u_{\Omega|\Omega_R}, p_{\Omega|\Omega_R}) = u_\Omega(x).p_\Omega(x)$. Using that u_Ω is the solution to the Helmholtz equation in the ball $B(x, R)$, we obtain

$$u_\Omega(x) = \frac{u_{\Omega|\Sigma_R}^{mean}}{J_0(kR)}.$$

Similarly, we have

$$\overline{p}_\Omega(x) = \frac{\overline{p_{\Omega|\Sigma_R}^{mean}}}{J_0(kR)}.$$

Hence

$$\begin{aligned}\delta_a(u_0, p_0) &= \delta_a(u_{\Omega|\Omega_R}, p_{\Omega|\Omega_R}) \\ &= u_\Omega(x)\overline{p_\Omega(x)}.\end{aligned}$$

This completes the proof. ∎

Then the topological gradient is

$$g = \Re(u_\Omega \overline{p_\Omega}).$$

4. Numerical results.

4.1. T-shaped waveguide. We use the topological gradient to design an H-plane T-shaped waveguide. The geometric constraints are shown in Figure 3(a). The input Γ_1 is excited by the TE10 mode (see the second boundary condition of (4.1)): the excitation is given by

$$u_e(y) = \cos\left(\frac{\pi y}{d}\right) \quad \forall y \in \Gamma_1.$$

We follow the two ideas [22]:

- the initial guess is the free space;
- instead of minimizing the reflected energy, we maximize the transmitted energy on Γ_2 and Γ_3.

At the beginning, only the input and output channels have metallic boundaries. In order to use the finite element method, the design domain is delimited by a fictitious boundary Γ_4 on which an absorbing condition is imposed (see Figure 3(b)). The problem is modelized as follows:

$$\left\{\begin{array}{lll} \Delta u + k^2 u = 0 & & \text{in } \Omega, \\ u = 0 & & \text{on } \Gamma_0, \\ \partial_n u - ik'u = \partial_n u_e - ik'u_e & & \text{on } \Gamma_1, \\ \partial_n u - ik'u = 0 & & \text{on } \Gamma_2, \Gamma_3, \\ \partial_n u - iku = 0 & & \text{on } \Gamma_4, \end{array}\right. \tag{4.1}$$

where $k^2 = k'^2 + \frac{\pi^2}{d^2}$, d being the length of Γ_1. The perfect conduction on the metallic boundary leads to the first boundary condition $u = 0$ on Γ_0. The third

boundary condition prevents reflections on Γ_2, Γ_3. The last equation is an approximate absorbing boundary condition. Here and in the following, we take $k = 10$.

The cost function to maximize is

$$J(u) = |S_{12}(u)|^2 + |S_{13}(u)|^2,$$

where $S_{1j}(u)$ is given by

$$S_{1j}(u) = \int_{\Gamma_j} u_{|\Gamma_j} \cos\left(\frac{\pi x}{d}\right)\, dx, \quad j \in \{2, 3\}.$$

The adjoint state is the solution to

$$\left\{\begin{array}{lll} \Delta\bar{p} + k^2\bar{p} &= 0 & \text{in } \Omega, \\ \bar{p} &= 0 & \text{on } \Gamma_0, \\ \partial_n\bar{p} - ik'\bar{p} &= 0 & \text{on } \Gamma_1, \\ \partial_n\bar{p} - ik'\bar{p} &= -2\overline{S_{12}(u)}\cos\left(\frac{\pi x}{d}\right) & \text{on } \Gamma_2, \\ \partial_n\bar{p} - ik'\bar{p} &= -2\overline{S_{13}(u)}\cos\left(\frac{\pi x}{d}\right) & \text{on } \Gamma_3, \\ \partial_n\bar{p} - ik\bar{p} &= 0 & \text{on } \Gamma_4. \end{array}\right. \tag{4.2}$$

Then the topological gradient is $g = \Re(u\bar{p})$ (see Figure 4(b)). We are interested in the relative loss of energy

$$P(u) = \frac{E_e - (E_2 + E_3)(u)}{E_e},$$

where E_e is the entering energy and $E_j(u)$ is the outgoing energy through Γ_j, $j \in \{2, 3\}$.

We present here the topological optimization procedure. The underlying idea is the following: in the ℓth step of the process, if $\overline{x}$ is such that the topological gradient is higher than a certain value t_ℓ, we insert at this point a Dirichlet node (metal). The constant t_ℓ is chosen by the user, which allows him to take into account other constraints, for example the feasibility. The process is stopped when the topological gradient is everywhere negative in the design domain or when the shape suits the designer. The algorithm is as follows.

- Initialization: choose the initial domain Ω_0, and set $\ell = 0$. The domain Ω_0 is meshed and it is identified with the set of the nodes: $\Omega_0 = \{x_k, k \in \{1, 2, \ldots, n\}\}$. The grid is fixed during the process.
- Repeat:
  1. compute u_ℓ, p_ℓ the direct and adjoint solutions in the domain Ω_ℓ,
  2. compute the topological gradient $g_\ell = \Re(u_\ell\overline{p_\ell})$,
  3. set $\Omega_{\ell+1} = \Omega_\ell \backslash \{x_k, g_\ell(x_k) \geq t_{\ell+1}\}$,
  4. $\ell \leftarrow \ell + 1$.

Figure 4 shows the isovalues of $|u|$ and the topological gradient for the initial geometry. In this case, 94.4% of the energy is lost. After two iterations, the loss is reduced to 2.02% (see Figure 5) and the topological gradient is everywhere negative. The last step consists of smoothing the boundary of the domain by inserting some metal where $|u|$ is close to zero. The loss of energy of this waveguide is equal to 1.5% (see Figure 6). The convergence history is given by Figure 7.

4.2. L-shaped waveguide. Here, we use the topological gradient like a decision help system to build a junction between two rectangular waveguides. The initial

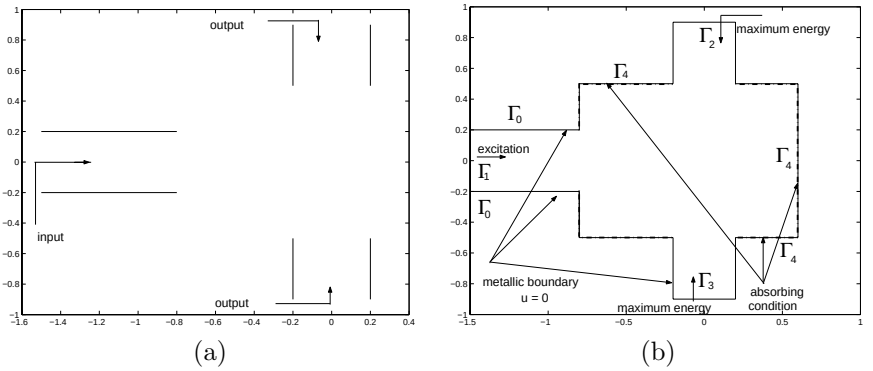


FIG. 3. *The initial geometry* (a) *and the design domain* (b).

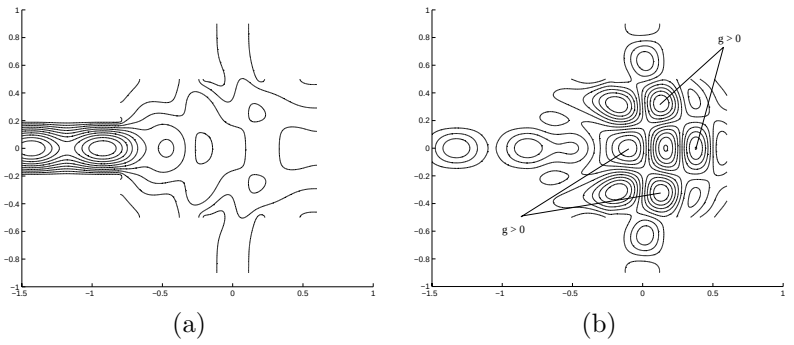


FIG. 4. *Modulus of the electric field* (a) *and topological gradient* (b).

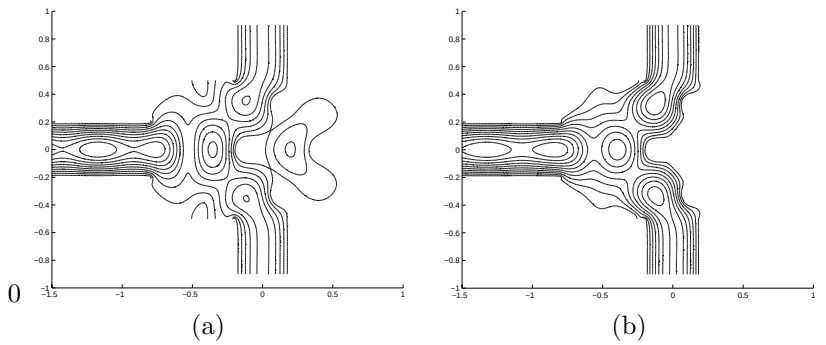


FIG. 5. *Modulus of the electric fields obtained after a first iteration* (a) *and after two iterations* (b).

geometry and the design domain are given by Figure 8. The cost function to maximize is

$$J(u) = |S_{12}(u)|^2.$$

Figure 9(a) shows the isovalues of $|u|$ for the initial geometry. In this case, 95.43% of the energy is lost. We observe that the topological gradient is high on a quarter

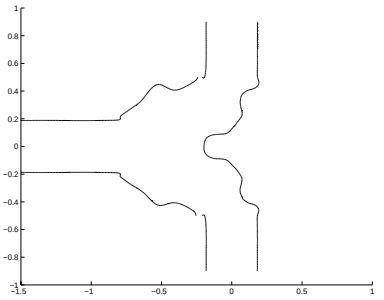


FIG. 6. *Final geometry.*

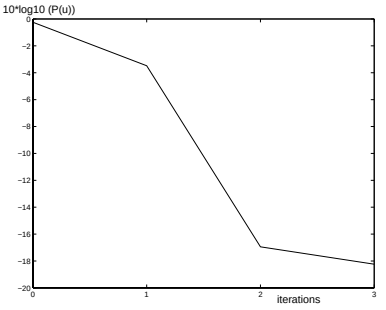
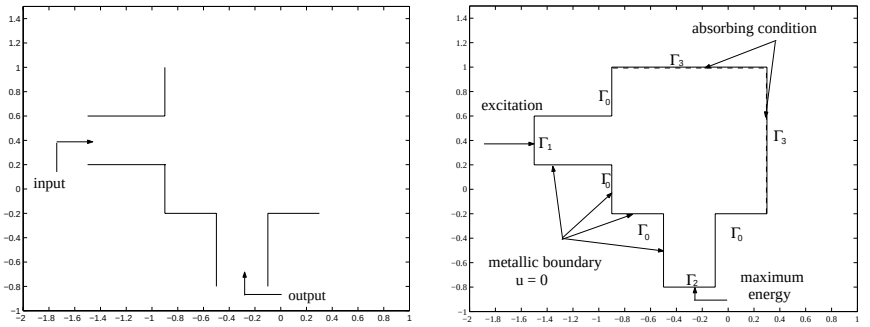


FIG. 7. *Convergence history.*



(a) (b)

FIG. 8. *The initial geometry* (a) *and the design domain* (b).

of circle where we decide to put metal (see Figure 9(b)). The loss of energy of the obtained waveguide is now equal to 0.34% (see Figure 10).

4.3. U-shaped waveguide. Here, the initial guess is a metallic cavity. The geometry of the waveguide is shown in Figure 11. The cost function to maximize is

$$J(u) = |S_{12}(u)|^2.$$

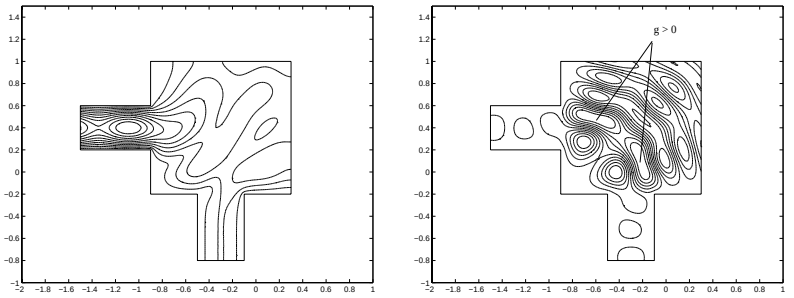


FIG. 9. *Modulus of the electric field* (a) *and topological gradient* (b).

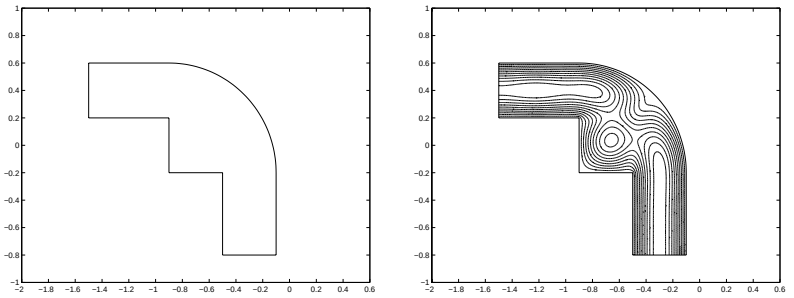


FIG. 10. *Final geometry* (a) *and modulus of the electric field* (b).

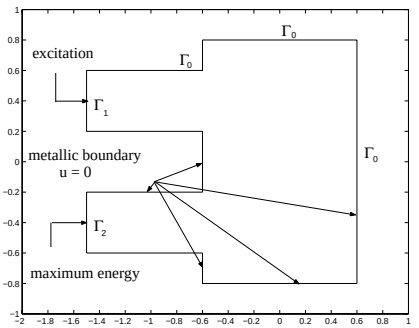
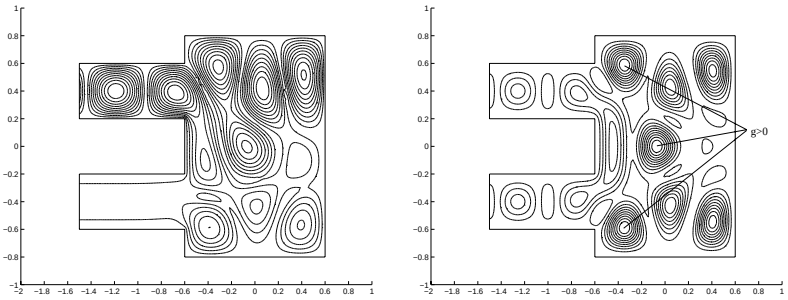


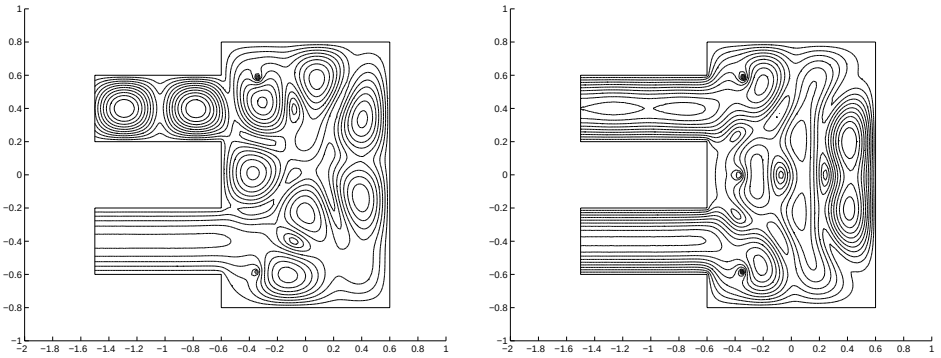
FIG. 11. *Geometry of the initial guide.*

Figure 12(a) shows the isovalues of $|u|$ for the initial geometry. In this case, 88.45% of the energy is reflected. There are three local maximas of the topological gradient (see Figure 12(b)). At each local maxima, we introduce a pointwise Dirichlet condition (a metallic plot). The new energy distribution is shown in Figure 13(a). The loss of energy is now equal to 39.19%. A new analysis is performed: after the introduction of another metallic plot, we obtain the design of Figure 13(b). The objective is fulfilled; the loss of energy is equal to 0.7%. For feasibility reasons, we decide not to insert additional plots.



(a) (b)

FIG. 12. *Modulus of the electric field* (a) *and topological gradient* (b).



(a) (b)

FIG. 13. *Modulus of the electric fields obtained after a first iteration* (a) *and after two iterations* (b).

5. Appendix.

5.1. Existence and uniqueness of the solution. Here we establish the existence and uniqueness of the solution to problem (3.1). Replacing Ω with Ω_ε, the argumentation would be the same for problem (3.3). Without any loss of generality, we suppose here that $N = 1$. The variational form of problem (3.1) is the following: find $u \in \mathcal{V}_\Omega$ satisfying

$$a(u, v) = l(v) \quad \forall v \in \mathcal{V}_\Omega, \tag{5.1}$$

where the functional space $\mathcal{V}_\Omega$, the sesquilinear form a, and the semilinear form l are defined by

$$\begin{aligned} \mathcal{V}_\Omega &= \{v \in H^1(\Omega), v = 0 \text{ on } \Gamma_0\}, \\ a(u, v) &= \int_\Omega (\nabla u . \overline{\nabla v} - k^2 u \overline{v})\, dx - ik \int_{\Gamma_1} u \overline{v}\, d\gamma(x), \\ l(v) &= \int_{\Gamma_1} g \overline{v}\, d\gamma(x). \end{aligned}$$

We split a in the following form:

$$a(u, v) = b(u, v) + c(u, v), \tag{5.2}$$

where b and c are defined by

$$b(u,v) = \int_\Omega (\nabla u.\overline{\nabla v} + u\overline{v})\, dx, \tag{5.3}$$

$$c(u,v) = -(1+k^2)\int_\Omega u\overline{v}\, dx - ik\int_{\Gamma_1} u\overline{v}\, d\gamma(x). \tag{5.4}$$

We recall the following result which is a consequence of the Lax–Milgram theorem.

LEMMA 5.1. *For all $f \in \mathcal{V}'_\Omega$, there exists a unique $u_f \in \mathcal{V}_\Omega$ such that*

$$b(u_f, v) = \langle f, v\rangle_{\mathcal{V}'_\Omega, \mathcal{V}_\Omega}.$$

The operator $f \mapsto u_f$ is continuous from $\mathcal{V}'_\Omega$ to $\mathcal{V}_\Omega$.

We define

$$\begin{array}{rccc} \mathcal{C}: & \mathcal{V}_\Omega & \longrightarrow & \mathcal{V}_\Omega, \\ & u & \longmapsto & \mathcal{C}u \end{array}$$

such that

$$b(\mathcal{C}u, v) + c(u,v) = 0 \quad \forall v \in \mathcal{V}_\Omega. \tag{5.5}$$

We have the following lemma.

LEMMA 5.2. *The operator $\mathcal{C}$ is compact.*

Proof. By Lemma 5.1, it suffices to prove that the operator

$$u \longmapsto c(u,.)$$

from $\mathcal{V}_\Omega$ to $\mathcal{V}'_\Omega$ is compact. Let (u_i) be a sequence bounded in $\mathcal{V}_\Omega$. The imbeddings $\mathcal{V}_\Omega \to L^2(\Omega)$ and $H^{\frac{1}{2}}_{00}(\Gamma_1) \to L^2(\Gamma_1)$ are compact; then there exists a subsequence always denoted by (u_i) such that

$$u_i \to w_1 \text{ in } L^2(\Omega)$$

and

$$\gamma_0 u_i \to w_2 \text{ in } L^2(\Gamma_1).$$

Then

$$c(u_i,.) \to l^{w_2}_{w_1} \text{ in } \mathcal{V}'_\Omega,$$

where $l^{w_2}_{w_1}$ is defined by

$$\left\langle l^{w_2}_{w_1}, v\right\rangle_{\mathcal{V}'_\Omega, \mathcal{V}_\Omega} = -(1+k^2)\int_\Omega w_1\overline{v}\, dx - ik\int_{\Gamma_1} w_2\overline{v}\, d\gamma(x) \quad \forall v \in \mathcal{V}_\Omega.$$

Hence the operator $\mathcal{C}$ is compact. ∎

Using (5.5), problem (5.1) can be written as follows: find $u \in \mathcal{V}_\Omega$ such that

$$b((I-\mathcal{C})u, v) = l(v) \quad \forall v \in \mathcal{V}_\Omega. \tag{5.6}$$

We have the following lemma.

LEMMA 5.3. *For $k \in \{k \in \mathbb{C}^*/\Im(k) \geq 0\}$, the following problem has no nontrivial solution: find $u \in \mathcal{V}_\Omega$ such that*

$$a(u,v) = 0 \quad \forall v \in \mathcal{V}_\Omega. \tag{5.7}$$

Proof. Let u be a solution to problem (5.7). For $v = u$, we have

$$a(u,u) = 0.$$

Then

$$\int_\Omega |\nabla u|^2\, dx - k^2 \int_\Omega |u|^2\, dx - ik \int_{\Gamma_1} |u|^2\, d\gamma(x) = 0. \tag{5.8}$$

By writing $k = k_1 + ik_2$, where $(k_1, k_2) \in \mathbb{R}^2$ and using (5.8), we obtain

$$\int_\Omega |\nabla u|^2\, dx - (k_1^2 - k_2^2) \int_\Omega |u|^2\, dx + k_2 \int_{\Gamma_1} |u|^2\, d\gamma(x) = 0 \tag{5.9}$$

and

$$k_1 \int_{\Gamma_1} |u|^2\, d\gamma(x) + 2k_1 k_2 \int_\Omega |u|^2\, dx = 0. \tag{5.10}$$

Two cases can arise:

- First case: $k_2 > 0$. If $k_1 = 0$, using (5.9) we obtain

$$\int_\Omega |\nabla u|^2\, dx + k_2^2 \int_\Omega |u|^2\, dx + k_2 \int_{\Gamma_1} |u|^2\, d\gamma(x) = 0.$$

  Then $u = 0$ in Ω. If $k_1 \neq 0$, using (5.10) we obtain

$$\int_{\Gamma_1} |u|^2\, d\gamma(x) + 2k_2 \int_\Omega |u|^2\, dx = 0.$$

  Then $u = 0$ in Ω.
- Second case: $k_2 = 0$ and $k_1 \neq 0$. Using (5.10), we obtain

$$u = 0 \quad \text{on } \Gamma_1.$$

  Let $\tilde{\Omega}$ be a regular domain containing Ω and so that $\Gamma_0 \subset \partial\tilde{\Omega}$. Extending u by zero in $\tilde{\Omega} \setminus \Omega$, we obtain a function $\tilde{u}$ that satisfies

$$\Delta\tilde{u} + k^2\tilde{u} = 0 \quad \text{in } \mathcal{D}'(\tilde{\Omega}).$$

  This extension is analytic; it is equal to zero in an open subset of a connected domain; thus $\tilde{u} = 0$ in $\tilde{\Omega}$.

This completes the proof. ∎

By Lemmas 5.2 and 5.3, and by using the Fredholm alternative, we obtain the following result.

THEOREM 5.4. *For $k \in \{k \in \mathbb{C}^*/\Im(k) \geq 0\}$, problem* (5.1) *has one and only one solution.*

5.2. The inf-sup condition. Our aim is to prove that the sesquilinear form a_0 defined by (3.9) for $\varepsilon = 0$ satisfies the inf-sup condition (see Hypothesis 2). We have the following lemma.

LEMMA 5.5. *The sesquilinear form a defined in* (5.1) *satisfies the* inf-sup *condition.*

Proof. Let u $\in \mathcal{V}_\Omega$. We set $v = (I - \mathcal{C})u$, where $\mathcal{C}$ is the operator defined by (5.5). According to (5.5), we have

$$\begin{aligned} a(u,v) &= b(v,v) \\ &= \|(I-\mathcal{C})u\|_{\mathcal{V}_\Omega}\|v\|_{\mathcal{V}_\Omega} \\ &\geq \alpha\|u\|_{\mathcal{V}_\Omega}\|v\|_{\mathcal{V}_\Omega}, \end{aligned}$$

where $\alpha = \|(I-\mathcal{C})^{-1}\|^{-1}_{\mathcal{L}(\mathcal{V}_\Omega,\mathcal{V}_\Omega)}$. Thus the sesquilinear form a satisfies the inf-sup condition. ∎

We have the following result.

PROPOSITION 5.6. *The sesquilinear form a_0 satisfies the* inf-sup *condition.*

Proof. We have

$$a_0(u,v) = \int_{\Omega_R} (\nabla u.\overline{\nabla v} - k^2 u\overline{v})\, dx + \int_{\Sigma_R} (T^0 u)\overline{v}\, d\gamma(x) - ik\int_{\Gamma_1} u\overline{v}\, d\gamma(x) \quad \forall u, v \in \mathcal{V}_R.$$

For all $u \in \mathcal{V}_R$ we set

$$\tilde{u} = \begin{cases} u \text{ in } \Omega_R, \\ u^0_\psi \text{ in } B(x,R), \end{cases}$$

where $\psi = u_{|\Sigma_R}$ and u^0_ψ is the solution to

$$\begin{cases} \Delta u^0_\psi + k^2 u^0_\psi = 0 & \text{in } B(x,R), \\ u^0_\psi = \psi & \text{on } \Sigma_R. \end{cases}$$

It can easily be proved that

$$a_0(u, v_{|\Omega_R}) = a(\tilde{u}, v) \quad \forall u \in \mathcal{V}_R, \quad \forall v \in \mathcal{V}_\Omega.$$

According to Lemma 5.5, there exists $v \in \mathcal{V}_\Omega, v \neq 0$, such that

$$\begin{aligned} a_0(u, v_{|\Omega_R}) = a(\tilde{u}, v) &\geq \alpha\|\tilde{u}\|_{\mathcal{V}_\Omega}\|v\|_{\mathcal{V}_\Omega} \\ &\geq \alpha\|u\|_{\mathcal{V}_R}\|v_{|\Omega_R}\|_{\mathcal{V}_R}. \end{aligned}$$

This completes the proof. ∎

5.3. Some useful inequalities. We have the following proposition.

PROPOSITION 5.7. *There exists $c > 0$ such that*

$$\left|\varepsilon\frac{J_n(kR)}{J_n(k\varepsilon)}\right| \geq c\left(\frac{R}{\varepsilon}\right)^{n-1} \quad \forall n \geq n_0, \quad \forall \varepsilon < \varepsilon_0.$$

Proof. The Bessel function $J_n(z)$ is defined by

$$J_n(z) = \left(\frac{1}{2}z\right)^n \sum_{p=0}^{+\infty} \frac{(-\frac{1}{4}z^2)^p}{p!\Gamma(n+p+1)}.$$

Then we have

$$\begin{aligned}
\varepsilon \frac{J_n(kR)}{J_n(k\varepsilon)} &= \varepsilon \left(\frac{R}{\varepsilon}\right)^n \frac{\displaystyle\sum_{p=0}^{+\infty} \frac{(-\frac{1}{4}k^2R^2)^p}{p!\Gamma(n+p+1)}}{\displaystyle\sum_{p=0}^{+\infty} \frac{(-\frac{1}{4}k^2\varepsilon^2)^p}{p!\Gamma(n+p+1)}} \\
&= \varepsilon \left(\frac{R}{\varepsilon}\right)^n \frac{(\Gamma(n+1))^{-1} + \displaystyle\sum_{p=1}^{+\infty} \frac{(-\frac{1}{4}k^2R^2)^p}{p!\Gamma(n+p+1)}}{(\Gamma(n+1))^{-1} + \displaystyle\sum_{p=1}^{+\infty} \frac{(-\frac{1}{4}k^2\varepsilon^2)^p}{p!\Gamma(n+p+1)}} \\
&= \varepsilon \left(\frac{R}{\varepsilon}\right)^n \frac{1 + \displaystyle\sum_{p=1}^{+\infty} \frac{n!}{p!(n+p)!} \left(-\frac{1}{4}k^2R^2\right)^p}{1 + \displaystyle\sum_{p=1}^{+\infty} \frac{n!}{p!(n+p)!} \left(-\frac{1}{4}k^2\varepsilon^2\right)^p} \\
&= \left(\frac{R}{\varepsilon}\right)^{n-1} u_n(\varepsilon),
\end{aligned}$$

where $u_n(\varepsilon)$ is defined by

$$u_n(\varepsilon) = \frac{R + \displaystyle\sum_{p=1}^{+\infty} \frac{Rn!}{p!(n+p)!} \left(-\frac{1}{4}k^2R^2\right)^p}{1 + \displaystyle\sum_{p=1}^{+\infty} \frac{n!}{p!(n+p)!} \left(-\frac{1}{4}k^2\varepsilon^2\right)^p}.$$

It is easy to see that the series which intervene in the expression of $u_n(\varepsilon)$ converge normally with respect to (n, ε). Hence, we have

$$\lim_{(n,\varepsilon)\to(\infty,0)} u_n(\varepsilon) = R.$$

Using the limit definition, there exists $c > 0$ such that

$$|u_n(\varepsilon)| \geq c \quad \forall n \geq n_0, \quad \forall \varepsilon < \varepsilon_0.$$

This completes the proof. □

By the same techniques we obtain the following result.

PROPOSITION 5.8. *There exists $c > 0$ such that*

$$\left|\frac{Y_n(k\varepsilon)}{Y_n(kR)}\right| \geq c \left(\frac{R}{\varepsilon}\right)^n \qquad \forall n \geq n_0, \quad \forall \varepsilon < \varepsilon_0.$$

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1964.

[2] G. Allaire and R. Kohn, *Optimal bounds on the effective behavior of a mixture of two well-ordered elastic materials*, Quart. Appl. Math., 51 (1993), pp. 643–674.

[3] G. Allaire and R. Kohn, *Optimal design for minimum weight and compliance in plane stress using extremal microstructures*, Eur. J. Mech. A Solids, 12 (1993), pp. 839–878.

[4] M. Becker, *Optimisation topologique de structure en variables discrètes*, Technical report, Université de Lièges, Lièges, Belgium, 1996.

[5] M. Bendsoe, *Optimal Topology Design of Continuum Structure: An Introduction*, Technical report, Department of Mathematics, Technical University of Denmark, Lyngby, Denmark, 1996.

[6] G. Buttazzo and G. Dal Maso, *Shape optimization for Dirichlet problems: Relaxed formulation and optimality conditions*, Appl. Math. Optim., 23 (1991), pp. 17–49.

[7] J. Céa, *Conception optimale ou identification de forme, calcul rapide de la dérivée directionnelle de la fonction coût*, RAIRO Modél. Math. Anal. Numér., 20 (1986), pp. 371–402.

[8] J. Céa, A. Gioan, and J. Michel, *Quelques résultats sur l'identification de domaines*, Calcolo, 10 (1973), pp. 207–232.

[9] R. Dautray and J.-L. Lions, *Analyse mathématique et calcul numérique pour les sciences et les techniques*, Collection CEA, Masson, Paris, 1987.

[10] S. Garreau, Ph. Guillaume, and M. Masmoudi, *The topological asymptotic for PDE systems: The elasticity case*, SIAM J. Control Optim., 39 (2001), pp. 1756–1778.

[11] Ph. Guillaume and K. Sid Idris, *The topological asymptotic expansion for the Dirichlet problem*, SIAM J. Control Optim., 41 (2002), pp. 1042–1072.

[12] Ph. Guillaume and K. Sid Idris, *Topological sensitivity and shape optimization for the Stokes equations*, Numer. Math., submitted.

[13] P. Guillaume and M. Masmoudi, *Computation of high order derivatives in optimal shape design*, Numer. Math., 67 (1994), pp. 231–250.

[14] P. Guillaume, *Dérivées d'ordre supérieur en conception optimale de forme*, Thèse, Université Paul Sabatier, Toulouse, France, 1994.

[15] J. Jacobsen, N. Olhoff, and E. Ronholt, *Generalized Shape Optimization of Three-Dimensional Structures Using Materials with Optimum Microstructures*, Technical report, Institute of Mechanical Engineering, Aalborg University, Aalborg, Denmark, 1996.

[16] M. Masmoudi, *The topological asymptotic*, in Computational Methods for Control Applications, R. Glowinski, H. Kawarada, and J. Periaux, eds., GAKUTO Internat. Ser. Math. Sci. Appl. 16, Tokyo, Japan, 2001, pp. 53–72.

[17] M. Masmoudi, *Outils pour la conception optimale de formes*, Thèse d'état, Université de Nice, Sophia-Antipolis, France, 1987.

[18] M. Masmoudi, *Numerical solution for exterior problems*, Numer. Math., 51 (1987), pp. 87–101.

[19] M. Masmoudi, *Résolution numérique de problèmes exterieurs*, Thèse présentée à l'université de Nice, Sophia-Antipolis, France, 1979.

[20] F. Murat and J. Simon, *Sur le contrôle par un domaine géométrique*, Thèse d'état, Université Pierre et Marie Curie, Paris, 1976.

[21] F. Murat and L. Tartar, *Calcul des variations et homogénéisation*, in Les méthodes de l'homogénéisation: Théorie et Applications en Physique, Eyrolles, Paris, 1985, pp. 319–369.

[22] P. Mader, *Optimisation topologique pour la conception de composants guidés*, Thèse, Université Paul Sabatier, Toulouse, France, 2002.

[23] J.-C. Nédélec, *Acoustic and Electromagnetic Equations*, Springer, New York, 2000.

[24] P. A. Raviart and J. M. Thomas, *Introduction à l'analyse numérique des équations aux dérivées partielles*, Masson, Paris, 1988.

[25] J. Sokolowski and A. Żochowski, *On the topological derivative in shape optimization*, SIAM J. Control Optim., 37 (1999), pp. 1251–1272.

[26] M. Shoenauer, L. Kallel, and F. Jouve, *Mechanics inclusions identification by evolutionary computation*, Revue européenne des éléments finis, 5 (1996), pp. 619–648.

[27] A. Schumacher, *Topologieoptimierung von Bauteilstrukturen unter Verwendung von Lopch-positionierungkrieterien*, Ph.D. thesis, Universität-Gesamthochschule Siegen, Siegen, Germany, 1995.

[28] Ch. Vassallo, *Théorie des guides d'ondes électromagnétiques*, Collection Technique et scientifique des Telécommunications, Eyrolles, Paris, 1985.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1545–1558

PARTIAL HEDGING UNDER TRANSACTION COSTS*

KENJI KAMIZONO†

Abstract. In this paper, we consider the problem of minimizing expected shortfall for a contingent claim in a continuous-time, multiasset financial market in the presence of proportional transaction costs. By generalizing the convex duality technique of Cvitanić [*SIAM J. Control. Optim.*, 38 (2000), pp. 1050–1066] to the case of multivariate contingent claim, we establish the existence of an optimal trading strategy and describe it in terms of an appropriate dual optimization problem.

Key words. hedging, expected shortfall, transaction costs

AMS subject classifications. 91B28, 91B30, 91B70, 93E20

DOI. S0363012902405145

1. Introduction. In a standard, complete financial market, every contingent claim can be replicated through some admissible portfolio, and the minimum initial capital required for replication defines the hedging price of the contingent claim. Although there is no such unique hedging price in the presence of proportional transaction costs, an investor who is to cover a contingent claim at the terminal time can still eliminate the entire risk of shortfall by putting up a sufficient amount of initial capital, the least amount of which is called the *upper hedging price*. The upper hedging price is thus the price required by the most conservative investor. It has been pointed out, however, that the superreplication of a contingent claim often requires a large amount of initial capital. For example, in the case of simple European call-options, except for the transaction cost at initial time, the upper hedging price is just the initial stock price, which is a trivial upper bound on the value of the option; and the optimal strategy for hedging a call-option is the so-called *buy-and-hold* strategy; see, for example, [5], [13], and [2].

With such a situation in mind, we consider the problem of *minimization of expected shortfall*, the idea of which goes back to [4], [6], and [3]. We generalize the results of these works to the general case of financial markets with several risky assets under proportional transaction costs. The prototype and some results in the superreplicating problem of such markets can be found in [9]. The paper [7] recently considered similar problems by directly tackling the primal problem. We take an alternative approach via convex duality.[1] An advantage of the recourse to an appropriate dual problem is that we can solve the problem with minimum use of the martingale theory. In fact, we do not need the semimartingality of the asset price process. All we need, in fact, is the $K(S(T))$-Fatou closedness of the set $\mathscr{C}_{x,y}$ in Definition 2.4 below. A sufficient condition for $\mathscr{C}_{x,y}$ to be $K(S(T))$-Fatou closed is that the asset price process S is a continuous semimartingale for which there exists an equivalent martingale measure. However, our result still holds true even for nonsemimartingale price processes if the set $\mathscr{C}_{x,y}$ is $K(S(T))$-Fatou closed.

This paper is organized as follows. In section 2, we set up our financial market

*Received by the editors April 9, 2002; accepted for publication (in revised form) May 5, 2003; published electronically November 6, 2003.
http://www.siam.org/journals/sicon/42-5/40514.html

†Faculty of Economics, Nagasaki University, 4-2-1 Katafuchi, Nagasaki, Nagasaki 850-8506, Japan (k-kamiz@net.nagasaki-u.ac.jp).

[1] In the review process of this paper, the results of [7] were generalized for nonsemimartingale price processes; see [8].

model with proportional transaction costs. The model is essentially the same as [9], except that we exclude bartering between the risky assets. The dissertation [11] dealt with the general case with bartering allowed but with an additional restriction on the contingent claim. In section 3, we present our main result and proofs. There, we generalize the convex duality techniques of [3] to the case of vector-valued contingent claims.

2. The model. Throughout this paper, we fix a filtered probability space $(\Omega, \mathscr{F}, \{\mathscr{F}(t)\}_{t\in[0,T]}, \mathbb{P})$ satisfying the usual conditions. Our financial market consists of one risk-free asset with price normalized to be one and d risky assets with price given by a $(0,\infty)^d$-valued, measurable, adapted process $S \triangleq \{(S_1(t),\dots,S_d(t))\}_{t\in[0,T]}$. By changing the physical unit of the assets, if necessary, we assume $S(0) = \mathbf{1}$, where $\mathbf{1}$ denotes the vector in $\mathbb{R}^d$ with each component equal to 1. We do not make any additional assumptions on the process S such as semimartingality or the existence of an equivalent martingale measure.

We regard the risk-free asset as money and assume that an investor must purchase risky assets through money; that is, bartering between two risky assets is impossible. Furthermore, purchase (respectively, sale) of the ith risky asset is subject to a proportional transaction cost of rate λ_i (respectively, μ_i); an investor is allowed to buy (respectively, sell) the ith asset at time t for price $(1+\lambda_i)S_i(t)$ (respectively, $(1-\mu_i)S_i(t)$). The numbers $\lambda_i > 0$ and $0 < \mu_i < 1$ are assumed to be constant throughout the investment period.

A *trading strategy* is simply an $\mathbb{R}^d$-valued adapted process $A \triangleq \{A_i(t)\}_{t\in[0,T]}$ such that each component A_i has right-continuous paths of bounded variation with the property

$$\int_{[0,T]} S_i(t)d|A_i|(t) < \infty \quad \text{a.s.},$$

where $|A_i| \triangleq \{|A_i|(t)\}_{t\in[0,T]}$ denotes the total variation process of A_i. We set $A(0-) \triangleq 0$. As is well known, we have a decomposition $A_i = A_i^{\oplus} - A_i^{\ominus}$ into the positive variation $A^{\oplus}$ and the negative variation $A^{\ominus}$. These processes are right-continuous, nondecreasing adapted processes. The random variable $A_i^{\oplus}(t)$ (respectively, $A_i^{\ominus}(t)$) signifies the cumulative number-of-shares of the ith asset, purchased (respectively, sold) up to time t. Given a trading strategy A, we define the *portfolio-holdings process* to be the process $(X^A, A) \triangleq \{(X^A(t), A(t))\}_{t\in[0,T]}$ with $X^A(t)$ given by

$$X^A(t) \triangleq \sum_{i=1}^{d} \int_{[0,t]} S_i(u)\{(1-\mu_i)dA_i^{\ominus}(u) - (1+\lambda_i)dA_i^{\oplus}(u)\}. \tag{2.1}$$

The random variable $X^A(t)$ signifies the amount of money held at time t by an investor who has chosen a trading strategy A with no initial-holdings.

Our formulation on transaction costs is essentially the same as [9] with the additional assumption that

$$(1+\lambda^{i1})(1+\lambda^{1j}) < (1+\lambda^{ij}) \quad \forall i \neq j \tag{2.2}$$

in their notation. Under this assumption, it is more efficient to use money as the medium of exchange than to barter, and inefficient strategies involving bartering will be excluded by optimality. Notice also that since the price process S need not be a semimartingale, there is no stochastic differential equation for the portfolio-holdings

process $(X^A, S_1A_1, \ldots, S_dA_d)$ in terms of amount. In this paper, therefore, we always express portfolio-holdings in terms of the number-of-shares.

DEFINITION 2.1 (the solvency region). *For each $s \in (0,\infty)^d$, we define the* solvency region $K(s)$ *to be the set of vectors $(x,y) \in \mathbb{R} \times \mathbb{R}^d$ such that*

$$\ell(s,x,y) \triangleq x + \sum_{i=1}^{d} s_i\{(1-\mu_i)y_i^+ - (1+\lambda^i)y_i^-\} \geq 0. \tag{2.3}$$

We also define the positive polar $K^*(s)$ *of the set $K(s)$ by*

$$K^*(s) \triangleq \{(\xi,\eta) \in \mathbb{R} \times \mathbb{R}^d \mid \xi x + \eta \cdot y \geq 0 \quad \forall (x,y) \in K(s)\}. \tag{2.4}$$

It is easy to see that the sets $K(s)$ and $K^*(s)$ are convex polyhedral cones in $\mathbb{R} \times \mathbb{R}^d$ such that $K^*(s) \subseteq \mathbb{R}_+ \times \mathbb{R}^d_+ \subseteq K(s)$. The economic significance of the solvency region $K(s)$ is the following. If an investor with a portfolio-holdings vector $(x,y) \in \mathbb{R} \times \mathbb{R}^d$ liquidates his portfolio-holdings when the price vector of the risky assets is s, then after this liquidation has taken place the amount of money which he will be holding is equal to $\ell(s,x,y)$. Therefore, the portfolio-holdings vector (x,y) belongs to the solvency region $K(s)$ if and only if the value of the portfolio-holdings vector (x,y) is "nonnegative." We call the function $\ell : (0,\infty)^d \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ the *liquidation function.*

At this point, we remark that as portfolio-holdings vectors are expressed in terms of number-of-shares, our notations such as $K(S(T))$ and $K^*(S(T))$ correspond to $\hat{K}$ and $\hat{K^*}$ of [9] via what they called the "hat" operator; by applying the hat operator, one can divide primal (respectively, multiply dual) random vectors by the diagonal matrix with diagonal elements $1, S_1(T), \ldots, S_d(T)$. Thus, except for easy line-by-line modifications, we may use their results, for example Theorem 2.6.

Given a set-valued random variable $\omega \to E(\omega)$, where $E(\omega)$ is a Borel set of some Euclidean space for each ω, we denote by $\mathbb{L}^0(E)$ the set of all $\mathscr{F}(T)$-measurable random variables ξ such that $\xi(\omega) \in E(\omega)$ almost surely. The following notion of the "boundedness from below" of a random vector was given by [9].

DEFINITION 2.2. *Given a random vector $\theta \in \mathbb{L}^0((0,\infty)^d)$, a random vector $(G,H) \in \mathbb{L}^0(\mathbb{R}) \times \mathbb{L}^0(\mathbb{R}^d)$ is said to be* $K(\theta)$-bounded from below *if there exists a constant $\kappa \geq 0$ such that*

$$(G,H) + \kappa(1,\mathbf{1}) \in \mathbb{L}^0(K(\theta)). \tag{2.5}$$

We denote by $\mathbb{L}^0_{b,\theta}$ the set of all random vectors $(G,H) \in \mathbb{L}^0(\mathbb{R}) \times \mathbb{L}^0(\mathbb{R}^d)$ that are $K(\theta)$-bounded from below.

In this paper, we do not make any restriction on the trading strategy except for a mild boundedness-from-below condition. In particular, the portfolio-holdings vector $(X^A(t), A(t))$ may go out of the solvency region $K(S(t))$. The dissertation [11] discussed the case where the *nonbankruptcy* condition "$(X^A(t), A(t)) \in K(S(t))$ $\forall t \in [0,T]$" is imposed.

DEFINITION 2.3. *A trading strategy A is said to be* admissible *if the random vector $(X^A(t), A(t))$ is $K(S(t))$-bounded from below uniformly in t, in the sense that there exists a constant $\kappa \geq 0$ with the property*

$$(X^A(t), A(t)) + \kappa(1,\mathbf{1}) \in K(S(t)) \quad a.s. \tag{2.6}$$

for every $t \in [0,T]$. We denote by $\mathscr{A}$ the set of all admissible trading strategies.

A *contingent claim* is simply a random vector $(G,H)\in\mathbb{L}^0_{b,S(T)}$. Here the random variable G (respectively, H_i) denotes the amount of money (respectively, the number-of-shares of the ith risky asset) that the claim holder receives at time T. Similarly, an *initial-holdings vector* is simply a vector $(x,y)\in\mathbb{R}\times\mathbb{R}^d$, where x denotes the initial amount of money held and y_i the initial number-of-shares of the ith risky asset held.

DEFINITION 2.4. *Let (x,y) be an initial-holdings vector. A contingent claim (G,H) is said to be* hedgeable *with (x,y) if there exists an admissible trading strategy A such that*

$$(x+X^A(T)-G,\,y+A(T)-H)\in\mathbb{L}^0(K(S(T))). \tag{2.7}$$

We denote by $\mathscr{C}_{x,y}$ the set of all contingent claims that are hedgeable with (x,y).

It is easy to see that the set $\mathscr{C}_{x,y}$ is convex and $K(S(T))$-solid in $\mathbb{L}^0_{b,S(T)}$; that is, we have $\big(\mathscr{C}_{x,y}-\mathbb{L}^0(K(S(T)))\big)\cap\mathbb{L}^0_{b,S(T)}\subseteq\mathscr{C}_{x,y}$. We make the following assumption about the set $\mathscr{C}_{x,y}$.

Assumption 2.5. The set $\mathscr{C}_{x,y}$ is $K(S(T))$-Fatou closed; that is, if $\{(G^n,H^n)\}_{n\in\mathbb{N}}$ is a sequence in $\mathscr{C}_{x,y}$ such that $(G^n,H^n)+\kappa(1,\mathbf{1})\in\mathbb{L}^0(K(S(T)))\ \forall n\in\mathbb{N}$ and if $(G^n,H^n)\to(G,H)\in\mathbb{L}^0_{b,S(T)}$ $(n\to\infty)$ almost surely, then we have $(G,H)\in\mathscr{C}_{x,y}$.

Note that since $\mathscr{C}_{x,y}=(x,y)+\mathscr{C}_{0,0}$, $K(S(T))$-Fatou closedness of $\mathscr{C}_{\bar{x},\bar{y}}$ for some $(\bar{x},\bar{y})$ implies that of $\mathscr{C}_{x,y}$ $\forall(x,y)$. The following theorem of [9] gives a sufficient condition for $\mathscr{C}_{x,y}$ to be $K(S(T))$-Fatou closed; see also Lemma 3.5 of [10] for a more general result.

THEOREM 2.6 (Kabanov–Last [9]). *If the asset price process S is a continuous semimartingale and if there exists an equivalent martingale measure for S, then the set $\mathscr{C}_{x,y}$ is $K(S(T))$-Fatou closed.*

Let $(x,y)\in(0,\infty)\times(0,\infty)^d$ be a "positive" initial-holdings vector. For such (x,y), we consider the set $\mathscr{D}_{x,y}$, given by

$$\mathscr{D}_{x,y}\triangleq\Big\{(\Gamma,\Phi)\in\mathbb{L}^0(K^*(S(T)))\;\Big|\;\mathbb{E}[\Gamma G+\Phi\cdot H]\le 1\quad\forall(G,H)\in\mathscr{C}_{x,y}\cap\mathbb{L}^0(K(S(T)))\Big\}. \tag{2.8}$$

It is easy to see that the set $\mathscr{D}_{x,y}$ is convex, bounded in $\mathbb{L}^1(\mathbb{R})\times\mathbb{L}^1(\mathbb{R}^d)$, and closed in almost-sure convergence. Furthermore, if the set $\mathscr{C}_{x,y}$ is $K(S(T))$-Fatou closed, then by using the results of [9], especially Theorem 4.1, we can prove the next lemma, for which we leave the proof to the readers. Again, we should keep in mind that our model is related to that of [9] via (2.2) and the so-called hat operator. Note also that the key Lemma 3.1 of [9] does not involve the stochastic integral with respect to S and thus is still valid without semimartingality of S.

LEMMA 2.7. *Let Assumption* 2.5 *hold, and let $(x,y)\in(0,\infty)\times(0,\infty)^d$. If $(G,H)\in\mathbb{L}^0_{b,S(T)}$ satisfies*

$$\mathbb{E}[\Gamma G+\Phi\cdot H]\le 1\quad\forall(\Gamma,\Phi)\in\mathscr{D}_{x,y}, \tag{2.9}$$

then $(G,H)\in\mathscr{C}_{x,y}$.

3. The partial hedging problem. We fix a contingent claim $(G,H)\in\mathbb{L}_{b,S(T)}$ and consider a financial institute which will issue this contingent claim (G,H). The financial institute, which is thus going to face a random obligation (G,H) at time T, may allocate some initial-holdings vector $(x,y)\in\mathbb{R}\times\mathbb{R}^d$ and wants to minimize

the expected shortfall at time T. Here, the shortfall is defined to be the amount of additional money that is necessary for the financial institute to pay the full amount of the obligation. In other words, if the financial institute chooses a trading strategy A, the shortfall will be

$$\ell\big(S(T), x + X^A(T) - G, y + A(T) - H\big)^-.$$

The optimal value of the expected-shortfall minimization problem is then given by

$$V_{x,y} \triangleq \inf_{A\in\mathscr{A}} \mathbb{E}\Big[\ell\big(S(T), x + X^A(T) - G, y + A(T) - H\big)^-\Big]. \tag{3.1}$$

As a matter of fact, the following lemma implies that minimizing the expected shortfall is equivalent to minimizing the expected value of a weighted sum of the shortfall in each asset.

LEMMA 3.1. *The optimal value $V_{x,y}$ given by* (3.1) *can also be written as*

$$V_{x,y} = \inf_{A\in\mathscr{A}} \mathbb{E}\Big[\big(G - x - X^A(T)\big)^+ + \sum_{i=1}^d (1+\lambda_i) S_i(T)\big(H_i - y_i - A_i(T)\big)^+\Big]. \tag{3.2}$$

Proof. Denote the right-hand side of (3.1) (respectively, (3.2)) by $V^1_{x,y}$ (respectively, $V^2_{x,y}$). Then, by comparing the random variables inside the expectations, we may easily see that $V^1_{x,y} \le V^2_{x,y}$. To see the reverse, for given $A \in \mathscr{A}$, let

$$\rho \triangleq \min\Big\{k \ge 1 \;\Big|\; x - X^A(T-) - G + \sum_{i=1}^d S_i(T)(1-\mu_i)\big(y_i + A_i(T-) - H_i\big)^+ < \sum_{i=1}^k S_i(T)(1+\lambda_i)\big(y_i + A_i(T-) - H_i\big)^-\Big\}$$

on the set $B \triangleq \{\omega \in \Omega \mid \ell\big(S(T), x + X^A(T-) - G, y + A(T-) - H\big) < 0\}$, and define the process $\tilde{A}(\cdot)$ by setting $\tilde{A}(t) \triangleq A(t)$ for $t \in [0, T)$,

$$\tilde{A}_i(T) \triangleq \begin{cases} H_i - y_i & \text{if } 1 \le i \le \rho - 1, \\ A_\rho(T-) + \frac{1}{(1+\lambda_\rho)S_\rho(T)}\Big\{x + X^A(T-) - G \\ \quad + \sum_{j=1}^d S_j(T)(1-\mu_j)\big(y_j + A_j(T-) - H_j\big)^+ \\ \quad - \sum_{j=1}^{\rho-1} S_j(T)(1+\lambda_j)\big(y_j + A_j(T-) - H_j\big)^-\Big\} & \text{if } i = \rho, \\ A_i(T-) - \big(y_i + A_i(T-) - H_i\big)^+ & \text{otherwise,} \end{cases}$$

on B and $\tilde{A}_i(T) \triangleq H_i - y_i$ for $i = 1, \dots, d$ on B^c. Then it is easy to check that $\tilde{A} \in \mathscr{A}$ and

$$\big(G - x - X^{\tilde{A}}(T)\big)^+ + \sum_{i=1}^d S_i(T)(1+\lambda_i)\big(H_i - y_i - \tilde{A}_i(T)\big)^+ = \ell\big(S(T), x + X^A(T-) - G, y + A(T-) - H\big)^-.$$

Since we easily see that the right-hand side is dominated by $\ell\big(S(T), x + X^A(T) - G, y + A(T) - H\big)^-$, we obtain $V^2_{x,y} \le V^1_{x,y}$. This completes the proof. ∎

In what follows, we shall work on the optimization problem (3.2), which is equivalent to (3.1) by Lemma 3.1 above. For the contingent claim $(G,H)\in\mathbb{L}_{b,S(T)}$ in consideration, we make the following assumptions.

Assumption 3.2. The contingent claim $(G,H)\in\mathbb{L}_{b,S(T)}$ satisfies

$$\mathbb{E}\Big[|G|+\sum_{i=1}^{d}S_i(T)|H_i|\Big]<\infty \quad \text{and} \tag{3.3a}$$

$$\sup_{(\Gamma,\Phi)\in\mathscr{D}_{1,\mathbf{1}}}\mathbb{E}[\Gamma G+\Phi\cdot H]<\infty. \tag{3.3b}$$

It is easy to see that

$$V_{x,y}=\inf_{(\xi,\eta)\in\mathscr{C}_{x,y}}\mathbb{E}\Big[(G-\xi)^+ +\sum_{i=1}^{d}(1+\lambda_i)S_i(T)(H_i-\eta_i)^+\Big], \tag{3.4}$$

which is finite by (3.3a) since $(x,y)\in\mathscr{C}_{x,y}$. We can also see that if there exists $(\hat{\xi},\hat{\eta})\in\mathscr{C}_{x,y}$ that attains the infimum in (3.4), then any trading strategy $A\in\mathscr{A}$ satisfying $(X^A(T)-\hat{\xi},A(T)-\hat{\eta})\in\mathbb{L}^0(K(S(T)))$, which will exist, will be optimal for the original optimization problem (3.2).

In what follows, we fix some "positive" initial-holdings vector $(x,y)\in(0,\infty)\times(0,\infty)^d$. Notice, however, that we do not lose generality by this positivity restriction since we may always translate the contingent claim (G,H) by $(x-1,y-\mathbf{1})$. In particular, the conditions in Assumption 3.2 are not affected by this translation. Notice also that from (3.3b) and the $\mathbb{L}^1$-boundedness of the set $\mathscr{D}_{x,y}$, we have

$$\alpha_{x,y}(G,H)\triangleq\sup_{(\Gamma,\Phi)\in\mathscr{D}_{x,y}}\mathbb{E}[\Gamma G+\Phi\cdot H]<\infty. \tag{3.5}$$

We now define the *optimal value function* $V_{x,y}(\cdot)$ by

$$V_{x,y}(\alpha)\triangleq V_{\alpha x,\alpha y},\quad 0<\alpha<\alpha_{x,y}(G,H). \tag{3.6}$$

Noting that $\mathscr{C}_{\alpha x,\alpha y}=\alpha\mathscr{C}_{x,y}$, we see

(3.7)

$$V_{x,y}(\alpha)=\inf_{(\xi,\eta)\in\mathscr{C}_{x,y}}\mathbb{E}\Big[(G-\alpha\xi)^+ +\sum_{i=1}^{d}(1+\lambda_i)S_i(T)(H_i-\alpha\eta_i)^+\Big]\quad\forall 0<\alpha<\alpha_{x,y}(G,H).$$

We adopt the convex-duality approach to solve the optimization problem (3.7). As a conjugate of the objective function, we consider the function $\tilde{R}(\cdot,\cdot;\omega)\colon\mathbb{R}\times\mathbb{R}^d\to$

$[-\infty,\infty)$ given by

(3.8)
$$\begin{aligned}
&\tilde{R}(z,w;\omega)\\
&\triangleq \inf_{(u,v)\in\mathbb{R}\times\mathbb{R}^d}\Big[\big(G(\omega)-u\big)^+ + \sum_{i=1}^d (1+\lambda_i)S_i(T,\omega)\big(H_i(\omega)-v_i\big)^+ + zu + w\cdot v\Big]\\
&= \inf_{u\in\mathbb{R}}\Big[\big(G(\omega)-u\big)^+ + zu\Big]\\
&\qquad + \sum_{i=1}^d (1+\lambda_i)S_i(T,\omega)\inf_{v_i\in\mathbb{R}}\left[\big(H_i(\omega)-v_i\big)^+ + \frac{w_i}{(1+\lambda_i)S_i(T,\omega)}v_i\right]\\
&= \left\{\begin{array}{ll} zG(\omega)+w\cdot H(\omega) & \text{if } 0\le z\le 1 \text{ and } 0\le w_i\le (1+\lambda_i)S_i(T,\omega)\\ & \forall i=1,\dots d,\\ -\infty & \text{otherwise}\end{array}\right\}.
\end{aligned}$$

Here, the infimum in the second line are attained at

$$\begin{array}{ll} u\in[G(\omega),\infty) & \text{if}\quad z=0,\\ u=G(\omega) & \text{if}\quad 0<z<1,\\ u\in(-\infty,G(\omega)] & \text{if}\quad z=1\end{array}\tag{3.9}$$

and

$$\begin{array}{ll} v_i\in[H_i(\omega),\infty) & \text{if}\quad w_i=0,\\ v_i=H_i(\omega) & \text{if}\quad 0<w_i<(1+\lambda_i)S_i(T),\\ v_i\in(-\infty,H_i(\omega)] & \text{if}\quad w_i=(1+\lambda_i)S_i(T),\end{array}\tag{3.10}$$

respectively.

Now, alongside the space $\mathscr{C}_{x,y}(\alpha)\triangleq\alpha\mathscr{C}_{x,y}$, where $\alpha>0$, we shall define the space $\mathscr{D}_{x,y}(\beta)\triangleq\beta\mathscr{D}_{x,y}$ for each $\beta>0$. Then, for given $(\xi,\eta)\in\mathbb{L}^0_{b,S(T)}$, we have $(\xi,\eta)\in\mathscr{C}_{x,y}(\alpha)$ and

(3.11)
$$\begin{aligned}
\mathbb{E}\Big[\big(G-\xi\big)^+ + \sum_{i=1}^d (1+\lambda_i)S_i(T)\big(H_i-\eta_i\big)^+\Big] &\ge \mathbb{E}[\tilde{R}(\beta\Gamma,\beta\Phi)] - \beta\mathbb{E}[\Gamma\xi+\Phi\cdot\eta]\\
&\ge \mathbb{E}[\tilde{R}(\beta\Gamma,\beta\Phi)] - \beta\alpha,
\end{aligned}$$

provided that

$$\mathbb{E}[\xi\Gamma+\eta\cdot\Phi]\le\alpha\quad\forall(\Gamma,\Phi)\in\mathscr{D}_{x,y}.\tag{3.12}$$

Moreover, the inequalities in (3.11) hold as an equality if and only if we have

$$\begin{aligned}
&0\le\beta\Gamma\le 1,\\
&0\le\beta\Phi_i\le(1+\lambda_i)S_i(T)\quad\forall i=1,\dots,d,
\end{aligned}\tag{3.13a}$$

$$\begin{aligned}
&\xi = G1_{\{0<\beta\Gamma<1\}} + U1_{\{\beta\Gamma=0 \text{ or } \beta\Gamma=1\}},\\
&\eta_i = H_i1_{\{0<\beta\Phi_i<(1+\lambda_i)S_i(T)\}} + V_i1_{\{\beta\Phi_i=0 \text{ or } \beta\Phi_i=(1+\lambda_i)S_i(T)\}}\quad\forall i=1,\dots,d
\end{aligned}\tag{3.13b}$$

almost surely, and

$$\mathbb{E}[\Gamma\xi + \Phi\cdot\eta] = \alpha, \tag{3.13c}$$

where U and V_i are any random variables satisfying

$$\begin{array}{c} \left\{ \begin{array}{lll} U \geq G & \text{on} & \{\beta\Gamma = 0\}, \\ U \leq G & \text{on} & \{\beta\Gamma = 1\}, \end{array} \right\} \quad \text{and} \\ \left\{ \begin{array}{lll} V_i \geq H_i & \text{on} & \{\beta\Phi_i = 0\}, \\ V_i \leq H_i & \text{on} & \{\beta\Phi_i = (1+\lambda_i)S_i(T)\} \end{array} \right\} \quad \forall i = 1,\dots,d \end{array} \tag{3.13d}$$

almost surely. For each $\beta > 0$, the dual optimization problem is given by

$$\tilde{V}_{x,y}(\beta) \triangleq \sup_{(\Gamma,\Phi)\in\mathscr{D}_{x,y}} \mathbb{E}[\tilde{R}(\beta\Gamma, \beta\Phi)]. \tag{3.14}$$

Our main theorem is the following.

MAIN THEOREM. *Let $(G,H) \in \mathbb{L}^0_{b,S(T)}$ be a contingent claim satisfying the conditions in Assumption* 3.2*, and let $(x,y) \in (0,\infty)\times(0,\infty)^d$ be a positive initial-holdings vector. Then we have the following:*

(i) *For each $\beta > 0$, there exists a random vector $(\hat{\Gamma},\hat{\Phi}) \equiv (\hat{\Gamma}(\beta),\hat{\Phi}(\beta)) \in \mathscr{D}_{x,y}$ that attains the supremum in* (3.14)*. Such $(\hat{\Gamma},\hat{\Phi})$ necessarily satisfies* (3.13a) *as well.*

(ii) *For each $0 < \alpha < \alpha_{x,y}(G,H)$, there exists a number $\hat{\beta} > 0$ that attains the supremum*

$$\sup_{\beta>0}\Big[\tilde{V}_{x,y}(\beta) - \beta\alpha\Big]. \tag{3.15}$$

(iii) *The functions $V_{x,y}(\cdot)$ and $\tilde{V}_{x,y}(\cdot)$ are conjugate of each other; that is, we have*

$$\begin{aligned} V_{x,y}(\alpha) &= \sup_{\beta>0}\Big[\tilde{V}_{x,y}(\beta) - \beta\alpha\Big] \quad \forall 0 < \alpha < \alpha_{x,y}(G,H), \\ \tilde{V}_{x,y}(\beta) &= \inf_{0<\alpha<\alpha_{x,y}(G,H)}\Big[V_{x,y}(\alpha) + \alpha\beta\Big] \quad \forall \beta > 0. \end{aligned} \tag{3.16}$$

(iv) *For each $0 < \alpha < \alpha_{x,y}(G,H)$, there exists a random vector $(U,V) \in \mathbb{L}^0(\mathbb{R})\times\mathbb{L}^0(\mathbb{R}^d)$ satisfying* (3.13d) *almost surely such that the random vector $(\hat{\xi},\hat{\eta})$ given by the right-hand side of* (3.13b) *with $\beta \equiv \hat{\beta}$ and $(\Gamma,\Phi) \equiv (\hat{\Gamma},\hat{\Phi})$ satisfies* (3.13c) *as well and thereby attains the infimum in* (3.7)*.*

4. Proof of main theorem. We shall give our proof of the main theorem in the form of a series of lemmas. The results in this section are generalization of those of [3] to the case of a vector-valued contingent claim.

LEMMA 4.1. *For each $\beta > 0$ there exists a random vector $(\hat{\Gamma},\hat{\Phi}) \equiv (\hat{\Gamma}(\beta),\hat{\Phi}(\beta)) \in \mathscr{D}_{x,y}$ that attains the supremum in* (3.14)*. Moreover, such $(\hat{\Gamma},\hat{\Phi})$ necessarily satisfies* (3.13a) *as well.*

Proof. First of all, note that since $(0,0) \in \mathscr{D}_{x,y}$, we have $\tilde{V}_{x,y}(\beta) \geq \mathbb{E}[\tilde{R}(0,0)] = 0 > -\infty$. This, in particular, gives the second assertion of the lemma. Also, from (3.5) and the last equality of (3.8), we see $\tilde{V}_{x,y}(\beta) \leq \beta\alpha_{x,y}(G,H) < \infty$.

To see the existence of a maximizer $(\hat{\Gamma},\hat{\Phi})$, take a sequence $\{(\Gamma^n,\Phi^n)\}_{n=1}^{\infty} \subseteq \mathscr{D}_{x,y}$ such that $\lim_{n\to\infty} \uparrow \mathbb{E}[\tilde{R}(\beta\Gamma^n,\beta\Phi^n)] = \tilde{V}_{x,y}(\beta)$. Then, from the L^1-boundedness of

the set $\mathscr{D}_{x,y}$ and Komlós's theorem, we may take a subsequence $\{(\Gamma^{n_k}, \Phi^{n_k})\}_{k=1}^{\infty}$ of $\{(\Gamma^n, \Phi^n)\}_{n=1}^{\infty}$ such that the sequence $\{(\Theta^k, \Psi^k)\}_{k=1}^{\infty}$ defined by

$$(\Theta^k, \Psi^k) \triangleq \frac{1}{k}\sum_{i=1}^{k}(\Gamma^{n_i}, \Phi^{n_i}), \quad k \geq 1,$$

converges, almost surely, to some random vector $(\hat{\Gamma}, \hat{\Phi}) \in \mathbb{L}^0(\mathbb{R}_+) \times \mathbb{L}^0(\mathbb{R}_+^d)$; see, for example, [12]. By the convexity and closedness in almost-sure convergence of $\mathscr{D}_{x,y}$, we have $(\Theta^k, \Psi^k) \in \mathscr{D}_{x,y}$, $\forall k \in \mathbb{N}$, and $(\hat{\Gamma}, \hat{\Phi}) \in \mathscr{D}_{x,y}$. Then the concavity of the functional $(\Gamma, \Phi) \mapsto \mathbb{E}[\tilde{R}(\beta\Gamma, \beta\Phi)]$ implies

$$\tilde{V}_{x,y}(\beta) \geq \mathbb{E}[\tilde{R}(\beta\Theta^k, \beta\Psi^k)] \geq \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[\tilde{R}(\beta\Gamma^{n_i}, \beta\Phi^{n_i})] \quad \forall k \in \mathbb{N}. \tag{4.1}$$

Let $k \to \infty$. Then, being the Cesàro average of the set of numbers $\{\mathbb{E}[\tilde{R}(\beta\Gamma^{n_i}, \beta\Phi^{n_i})]\}_{i=1}^{k}$, the right-hand side of (4.1) converges to $\tilde{V}_{x,y}(\beta)$, which implies that $\lim_{k\to\infty}\mathbb{E}[\tilde{R}(\beta\Theta^k, \beta\Psi^k)] = \tilde{V}_{x,y}(\beta)$.

Finally, from the inequalities (3.3a) and (3.13a) and from the last equality of (3.8), we see $|\tilde{R}(\beta\Gamma^k, \beta\Phi^k)| \leq C|G + S(T)\cdot H|$ for large $k \in \mathbb{N}$ with some suitable constant $C > 0$, and then, from the dominated convergence theorem, $\mathbb{E}[\tilde{R}(\beta\hat{\Gamma}, \beta\hat{\Phi})] = \lim_{k\to\infty}\mathbb{E}[\tilde{R}(\beta\Theta^k, \beta\Psi^k)]$. Therefore, $\tilde{V}_{x,y}(\beta) = \mathbb{E}[\tilde{R}(\beta\hat{\Gamma}, \beta\hat{\Phi})]$. □

For each $\alpha \in (0, \alpha_{x,y}(G,H))$, put

$$\psi_{x,y}(\beta;\alpha) \triangleq \tilde{V}_{x,y}(\beta) - \beta\alpha, \quad \beta > 0. \tag{4.2}$$

LEMMA 4.2. *Let $\alpha \in (0, \alpha_{x,y}(G,H))$. Then the function $\psi_{x,y}(\cdot;\alpha)$, defined by (4.2), is continuous and concave on $(0,\infty)$ and satisfies*

$$\lim_{\beta\to 0}\frac{\psi_{x,y}(\beta;\alpha)}{\beta} = \alpha_{x,y}(G,H) - \alpha > 0, \tag{4.3}$$

as well as

$$\lim_{\beta\to\infty}\frac{\psi_{x,y}(\beta;\alpha)}{\beta} = -\alpha. \tag{4.4}$$

Proof. The concavity of $\psi_{x,y}(\cdot\,;\alpha)$ follows from that of the function $\tilde{R}(\cdot,\cdot)$ and from the fact $k\beta_1\mathscr{D}_{x,y} + (1-k)\beta_2\mathscr{D}_{x,y} \subseteq (k\beta_1 + (1-k)\beta_2)\mathscr{D}_{x,y}$ $\forall 0 \leq k \leq 1$ and $\forall \beta_1, \beta_2 > 0$.

To prove the continuity of $\psi_{x,y}(\cdot\,;\alpha)$, it suffices to prove that $\lim_{\beta\downarrow 0}\tilde{V}_{x,y}(\beta) = 0$. To prove this equation, notice that since $(\hat{\Gamma}, \hat{\Phi})$ satisfies (3.13a), we have $\tilde{V}_{x,y}(\beta) \geq 0$ $\forall \beta > 0$. This, together with the fact that $\tilde{V}_{x,y}(\beta) \leq \beta\alpha_{x,y}(G,H)$ and $\alpha_{x,y}(G,H) < \infty$, gives $\lim_{\beta\downarrow 0}\tilde{V}_{x,y}(\beta) = 0$.

Next, we prove (4.3). Since

$$\frac{\tilde{V}_{x,y}(\beta)}{\beta} = \sup_{(\Gamma,\Phi)\in\mathscr{D}_{x,y}}\mathbb{E}\Big[\frac{\tilde{R}(\beta\Gamma,\beta\Phi)}{\beta}\Big] \leq \sup_{(\Gamma,\Phi)\in\mathscr{D}_{x,y}}\mathbb{E}[\Gamma G + \Phi\cdot H] = \alpha_{x,y}(G,H),$$

we have

$$\overline{\lim_{\beta\downarrow 0}}\frac{\psi_{x,y}(\beta;\alpha)}{\beta} = \overline{\lim_{\beta\downarrow 0}}\frac{\tilde{V}_{x,y}(\beta)}{\beta} - \alpha \leq \alpha_{x,y}(G,H) - \alpha. \tag{4.5}$$

To show the reverse inequality for the limit inferior, we take, for each $\varepsilon > 0$, a random vector $(\Gamma^\varepsilon, \Phi^\varepsilon) \in \mathscr{D}_{x,y}$ such that

$$\mathbb{E}[\Gamma^\varepsilon G + \Phi^\varepsilon \cdot H] > \sup_{(\Gamma,\Phi)\in\mathscr{D}_{x,y}} \mathbb{E}[\Gamma G + \Phi \cdot H] - \varepsilon = \alpha_{x,y}(G,H) - \varepsilon.$$

Then, for each $\beta > 0$, we have

$$\begin{aligned}\frac{\psi_{x,y}(\beta;\alpha)}{\beta} &= \sup_{(\Gamma,\Phi)\in\mathscr{D}_{x,y}} \mathbb{E}\Big[\Gamma G 1_{\{\beta\Gamma\le 1\}} + \sum_{i=1}^d \Phi_i H_i 1_{\{\beta\Phi_i \le (1+\lambda_i)S_i(T)\}}\Big] - \alpha \\ &\ge \mathbb{E}\Big[\Gamma^\varepsilon G 1_{\{\beta\Gamma^\varepsilon\le 1\}} + \sum_{i=1}^d \Phi_i^\varepsilon H_i 1_{\{\beta\Phi_i^\varepsilon \le (1+\lambda_i)S_i(T)\}}\Big] - \alpha.\end{aligned}$$

Letting $\beta \downarrow 0$, we obtain from Fatou's lemma

$$\varliminf_{\beta\downarrow 0} \frac{\psi_{x,y}(\beta;\alpha)}{\beta} \ge \mathbb{E}[\Gamma^\varepsilon G + \Phi^\varepsilon \cdot H] - \alpha > \alpha_{x,y}(G,H) - \varepsilon - \alpha.$$

Since $\varepsilon > 0$ is arbitrary, we conclude that

$$\varliminf_{\beta\downarrow 0} \frac{\psi_{x,y}(\beta;\alpha)}{\beta} \ge \alpha_{x,y}(G,H) - \alpha. \tag{4.6}$$

In conjunction with (4.5), the inequality (4.6) now gives (4.3).

It remains to prove (4.4). Since $V_{x,y}(\alpha) \ge \psi_{x,y}(\beta;\alpha)$ and $\tilde{V}_{x,y}(\beta) \ge 0$, we have

$$\frac{V_{x,y}(\alpha)}{\beta} \ge \frac{\psi_{x,y}(\beta;\alpha)}{\beta} = \frac{\tilde{V}_{x,y}(\beta)}{\beta} - \alpha \ge -\alpha.$$

Then, since $0 \le V_{x,y}(\alpha) \le C\mathbb{E}|G + S(T)\cdot H| < \infty$ with some suitable constant $C > 0$, we may let $\beta \to \infty$ to obtain

$$0 \ge \varlimsup_{\beta\to\infty} \frac{\tilde{V}_{x,y}(\beta)}{\beta} - \alpha \ge -\alpha.$$

By letting $\alpha \downarrow 0$, we obtain

$$\lim_{\beta\to\infty} \frac{\tilde{V}_{x,y}(\beta)}{\beta} = 0.$$

It follows that

$$\lim_{\beta\to\infty} \frac{\psi_{x,y}(\beta;\alpha)}{\beta} = \lim_{\beta\to\infty} \frac{\tilde{V}_{x,y}(\beta)}{\beta} - \alpha = -\alpha.$$

This completes the proof of the lemma. □

LEMMA 4.3. *The function $\psi_{x,y}(\cdot\,;\alpha)$ attains its supremum at some $0 < \hat{\beta} < \infty$.*

Proof. Suppose that $\sup_{0<\beta<\infty} \psi_{x,y}(\beta;\alpha) = \lim_{\beta\uparrow\infty} \psi_{x,y}(\beta;\alpha) > \psi_{x,y}(\beta;\alpha)\ \forall 0 < \beta < \infty$. Then there exists a sequence $\{\beta_n\}_{n=1}^\infty \subseteq (0,\infty)$ such that $\beta_n \uparrow \infty$ and the sequence $\{\psi_{x,y}(\beta_n;\alpha)\}_{n=1}^\infty$ is increasing. However, in view of the previous lemma, in particular (4.4), this is impossible. Therefore, either the function $\psi_{x,y}(\cdot\,;\alpha)$ attains the supremum at some $\hat{\beta} > 0$, or else we have $\psi_{x,y}(\beta;\alpha) \le \psi_{x,y}(0;\alpha) = 0\ \forall \beta > 0$.

Suppose the latter is true. Then $\psi_{x,y}(\beta;\alpha)/\beta \le 0\ \forall \beta > 0$. But this is again impossible because of (4.3). Therefore the function $\psi_{x,y}(\cdot\,;\alpha)$ must attain its supremum at some $0 < \hat{\beta} < \infty$. □

We now proceed to prove the main theorem. First, (i) and (ii) have already been proved in Lemmas 4.1 and 4.3, respectively. From (3.11), we have

$$V_{x,y}(\alpha) \ge \sup_{\beta>0}\Big[\tilde{V}_{x,y}(\beta) - \beta\alpha\Big] \quad \forall 0 < \alpha < \alpha_{x,y}(G,H). \tag{4.7}$$

Once we have established (iv), the reverse inequality of (4.7) follows as well, which yields by duality the second equality of (3.16). It therefore remains to prove (iv).

We begin with defining the space $\mathbb{L}$ by

$$\mathbb{L} \triangleq \mathbb{R} \times \mathbb{L}^1(\mathbb{R}) \times \mathbb{L}^1(\mathbb{R}^d) \tag{4.8}$$

equipped with the norm

$$\|(\beta,K,L)\| \triangleq |\beta| + \mathbb{E}[|K| + |L|], \quad (\beta,K,L) \in \mathbb{L}. \tag{4.9}$$

Define also the subset $\mathscr{G}$ by

$$\mathscr{G} \triangleq \{(\beta,K,L) \in \mathbb{L} \mid \beta \ge 0,\ (K,L) \in \beta\mathscr{D}_{x,y}\}. \tag{4.10}$$

We see easily that $\mathscr{G}$ is a convex cone of $\mathbb{L}$. The set $\mathscr{G}$ is also closed in the norm topology of $\mathbb{L}$. To see this, let $\{(\beta^n,\beta^n\Gamma^n,\beta^n\Phi^n)\}_{n=1}^{\infty}$ be any sequence in $\mathscr{G}$ that converges to some $(\beta,K,L) \in \mathbb{L}$. Then we have $\lim_{n\to\infty}\beta^n = \beta$, $\lim_{n\to\infty}\mathbb{E}|\beta^n\Gamma^n - K| = 0$, and $\lim_{n\to\infty}\mathbb{E}|\beta^n\Phi^n - L| = 0$. If $\beta = 0$, then $\lim_{n\to\infty}\mathbb{E}|\beta^n\Gamma^n| = \lim_{n\to\infty}\beta^n\mathbb{E}|\Gamma^n| = 0$ because of the $\mathbb{L}^1$-boundedness of $\mathscr{D}_{x,y}$, and we are done. If $\beta \neq 0$, then we have

$$\begin{aligned}
\mathbb{E}\left|\Gamma^n - \frac{K}{\beta}\right| &\le \frac{1}{\beta}\Big(\mathbb{E}|\beta\Gamma^n - \beta^n\Gamma^n| + \mathbb{E}|\beta^n\Gamma^n - K|\Big) \\
&= \frac{1}{\beta}\Big(|\beta - \beta^n|\mathbb{E}|\Gamma^n| + \mathbb{E}|\beta^n\Gamma^n - K|\Big) \\
&\to 0 \quad \text{as} \quad z \to \infty,
\end{aligned}$$

again by the $\mathbb{L}^1$-boundedness of $\mathscr{D}_{x,y}$. Similarly, we can show that $\lim_{n\to\infty}\mathbb{E}|\Phi^n - L/\beta| = 0$. Take a subsequence $\{(\Gamma^{n_k},\Phi^{n_k})\}_{k=1}^{\infty}$ of $\{(\Gamma^n,\Phi^n)\}_{n=1}^{\infty}$ such that $(\Gamma^{n_k},\Phi^{n_k}) \to (K/\beta, L/\beta)$ as $k \to \infty$ almost surely. Then Fatou's lemma gives

$$\mathbb{E}\Big[\frac{K}{\beta}\xi + \frac{L}{\beta}\cdot\eta\Big] \le \varliminf_{k\to\infty}\mathbb{E}[\Gamma^{n_k}\xi + \Phi^{n_k}\cdot\eta] \le 1 \quad \forall(\xi,\eta) \in \mathscr{C}_{x,y} \cap (\mathbb{L}^0(\mathbb{R}_+) \times \mathbb{L}^0(\mathbb{R}^d_+)),$$

which implies that $(K/\beta, L/\beta) \in \mathscr{D}_{x,y}$ and hence that $(\beta,K,L) \in \mathscr{G}$.

On the space $\mathbb{L}$, we consider the functional $\tilde{J}$ given by

$$\tilde{J}(\beta,K,L) \triangleq -\mathbb{E}[\tilde{R}(K,L)] + \beta\alpha, \quad (\beta,K,L) \in \mathbb{L}, \tag{4.11}$$

where $0 < \alpha < \alpha_{x,y}(G,H)$ is a fixed constant. Then it is easy to see that the functional $\tilde{J}$ is convex and proper. From (3.3a) and the dominated convergence theorem, we can easily see that $\tilde{J}$ is lower semicontinuous under the norm topology of $\mathbb{L}$. Furthermore, from Lemmas 4.1 and 4.3, we know that $\tilde{J}$ attains the infimum over $\mathscr{G}$ at $(\hat{\beta},\hat{\beta}\hat{\Gamma},\hat{\beta}\hat{\Phi})$, which belongs to $\mathscr{G} \setminus \{(0,0,0)\}$. Therefore, from standard results on

convex optimization (for example, Corollary 4.6.3 of [1]), it follows that there exists a pair $(\hat{\gamma}, \hat{Y}, \hat{Z})$ in the dual space $\mathbb{L}^* = \mathbb{R} \times \mathbb{L}^\infty(\mathbb{R}) \times \mathbb{L}^\infty(\mathbb{R}^d)$ that satisfies

$$-(\hat{\gamma}, \hat{Y}, \hat{Z}) \in \partial \tilde{J}(\hat{\beta}, \hat{\beta}\hat{\Gamma}, \hat{\beta}\hat{\Phi}) \tag{4.12}$$

and

$$(\hat{\gamma}, \hat{Y}, \hat{Z}) \in N(\hat{\beta}, \hat{\beta}\hat{\Gamma}, \hat{\beta}\hat{\Phi}), \tag{4.13}$$

where $\partial \tilde{J}(\hat{\beta}, \hat{\beta}\hat{\Gamma}, \hat{\beta}\hat{\Phi})$ and $N(\hat{\beta}, \hat{\beta}\hat{\Gamma}, \hat{\beta}\hat{\Phi})$ are the subdifferential of $\tilde{J}$ and the normal cone of $\mathscr{G}$ at $(\hat{\beta}, \hat{\beta}\hat{\Gamma}, \hat{\beta}\hat{\Phi})$, respectively. These sets are given by

$$\begin{aligned} \partial \tilde{J}(\hat{\beta}, \hat{\beta}\hat{\Gamma}, \hat{\beta}\hat{\Phi}) \triangleq \Big\{ (\gamma, Y, Z) \in \mathbb{L}^* \;\Big|\; & \tilde{J}(\beta, K, L) \geq \tilde{J}(\hat{\beta}, \hat{\beta}\hat{\Gamma}, \hat{\beta}\hat{\Phi}) + (\beta - \hat{\beta})\gamma \\ & + \mathbb{E}[(K - \hat{\beta}\hat{\Gamma})Y] + \mathbb{E}[(L - \hat{\beta}\hat{\Phi}) \cdot Z] \quad \forall (\beta, K, L) \in \mathbb{L} \Big\} \end{aligned} \tag{4.14}$$

and

$$\begin{aligned} N(\hat{\beta}, \hat{\beta}\hat{\Gamma}, \hat{\beta}\hat{\Phi}) \triangleq \Big\{ (\gamma, Y, Z) \in \mathbb{L}^* \;\Big|\; & (\beta - \hat{\beta})\gamma + \mathbb{E}[(\beta\Gamma - \hat{\beta}\hat{\Gamma})Y] \\ & + \mathbb{E}[(\beta\Phi - \hat{\beta}\hat{\Phi}) \cdot Z] \leq 0 \quad \forall (\beta, \beta\Gamma, \beta\Phi) \in \mathscr{G} \Big\}; \end{aligned} \tag{4.15}$$

see, for example, Propositions 4.4.4 and 4.3.3 of [1]. By definition, (4.12) and (4.13) are equivalent to

$$\begin{aligned} -\mathbb{E}[\tilde{R}(K, L)] + \beta\alpha \geq & -\mathbb{E}[\tilde{R}(\hat{\beta}\hat{\Gamma}, \hat{\beta}\hat{\Phi})] + \hat{\beta}\alpha - (\beta - \hat{\beta})\hat{\gamma} \\ & - \mathbb{E}[(K - \hat{\beta}\hat{\Gamma})\hat{Y}] - \mathbb{E}[(L - \hat{\beta}\hat{\Phi}) \cdot \hat{Z}] \quad \forall (\beta, K, L) \in \mathbb{L} \end{aligned} \tag{4.16}$$

and

$$(\beta - \hat{\beta})\hat{\gamma} + \mathbb{E}[(\beta\Gamma - \hat{\beta}\hat{\Gamma})\hat{Y}] + \mathbb{E}[(\beta\Phi - \hat{\beta}\hat{\Phi}) \cdot \hat{Z}] \leq 0 \quad \forall (\beta, \beta\Gamma, \beta\Phi) \in \mathscr{G}, \tag{4.17}$$

respectively. We claim that this $(\hat{Y}, \hat{Z})$ satisfies

$$\mathbb{E}[\hat{\Gamma}\hat{Y} + \hat{\Phi}\hat{Z}] = \alpha, \tag{4.18a}$$

$$\mathbb{E}[\Gamma\hat{Y} + \Phi\hat{Z}] \leq \alpha \quad \forall (\Gamma, \Phi) \in \mathscr{D}_{x,y} \tag{4.18b}$$

and can be written as the right-hand side (3.13b) with $\beta \equiv \hat{\beta}$, $(\Gamma, \Phi) \equiv (\hat{\Gamma}, \hat{\Phi})$, and with some $(U, V) \in \mathbb{L}^0(\mathbb{R}) \times \mathbb{L}^0(\mathbb{R}^d)$ satisfying (3.13d); once these claims are established, we may conclude that $(\hat{Y}, \hat{Z})$ serves as an optimal solution to (3.7); the $K(S(T))$-boundedness from below of $(\hat{Y}, \hat{Z})$ is trivial because $(\hat{Y}, \hat{Z}) \in \mathbb{L}^\infty(\mathbb{R}) \times \mathbb{L}^\infty(\mathbb{R}^d)$.

First, by taking $(\beta, K, L) = (\beta, 0, 0)$ in (4.16), we see

$$\infty > \mathbb{E}[\tilde{R}(\hat{\beta}\hat{\Gamma}, \hat{\beta}\hat{\Phi})] - \hat{\beta}\mathbb{E}[\hat{\Gamma}\hat{Y} + \hat{\Phi} \cdot \hat{Z}] \geq -(\beta - \hat{\beta})(\alpha + \hat{\gamma}) \quad \forall \beta \in \mathbb{R},$$

which implies $\hat{\gamma} = -\alpha$; otherwise, we could make the middle-hand side arbitrarily large by letting $\beta \to \pm\infty$.

To prove (4.18a) and (4.18b), note first that since $\hat{\gamma} = -\alpha$, letting $\beta \downarrow 0$ in the inequality (4.17) and then dividing by $\hat{\beta} > 0$ yield

$$\mathbb{E}[\hat{\Gamma}\hat{Y} + \hat{\Phi}\hat{Z}] \geq \alpha.$$

The reverse inequality also follows from the inequality (4.17) with $\hat{\gamma} = -\alpha$ and $(\beta, \beta\Gamma, \beta\Phi) = (\hat{\beta} + \varepsilon, (\hat{\beta} + \varepsilon)\hat{\Gamma}, (\hat{\beta} + \varepsilon)\hat{\Phi})$ for some $\varepsilon > 0$. This establishes (4.18a). Finally, letting $\beta = \hat{\beta}$ in (4.17) and then using (4.18a), we obtain (4.18b).

It remains to show that $(\hat{Y}, \hat{Z})$ can be written as the right-hand side (3.13b) with $\beta \equiv \hat{\beta}$, $(\Gamma, \Phi) \equiv (\hat{\Gamma}, \hat{\Phi})$, and with some $(U, V) \in \mathbb{L}^0(\mathbb{R}) \times \mathbb{L}^0(\mathbb{R}^d)$ satisfying (3.13d). For this end, we define a random vector $(A, B) \in \mathbb{L}^0(\mathbb{R}) \times \mathbb{L}^0(\mathbb{R}^d)$ by

$$\hat{Y} = G1_{\{0<\hat{\beta}\hat{\Gamma}<1\}} + A, \tag{4.19a}$$

$$\hat{Z}_i = H_i 1_{\{0<\hat{\beta}\hat{\Phi}_i<(1+\lambda_i)S_i(T)\}} + B_i, \quad i = 1, \dots, d. \tag{4.19b}$$

Then, for every $(K, L) \in \mathbb{L}^1(\mathbb{R}) \times \mathbb{L}^1(\mathbb{R}^d)$ satisfying

$$0 \le K \le 1 \quad \text{and} \quad 0 \le L_i \le (1 + \lambda_i)S_i(T) \quad \forall i = 1, \dots, d \quad \text{a.s.}, \tag{4.20}$$

the inequality (4.16) with $\hat{\gamma} = -\alpha$ gives

$$\begin{aligned} \mathbb{E}\Big[(K - \hat{\beta}\hat{\Gamma})(A - G1_{\{\hat{\beta}\hat{\Gamma}=1, \text{ or } \hat{\beta}\hat{\Gamma}=0\}}) \\ + \sum_{i=1}^{d} (B_i - H_i 1_{\{\hat{\beta}\hat{\Phi}_i=(1+\lambda_i)S_i(T) \text{ or } \hat{\beta}\hat{\Phi}_i=0\}})(L_i - \hat{\beta}\hat{\Phi}_i)\Big] \ge 0. \end{aligned} \tag{4.21}$$

Taking $L_i = \hat{\beta}\hat{\Phi}_i$ for $i = 1, \dots, d$ in (4.21), we obtain

$$\mathbb{E}\Big[(K - \hat{\beta}\hat{\Gamma})(A - G1_{\{\hat{\beta}\hat{\Gamma}=1 \text{ or } \hat{\beta}\hat{\Gamma}=0\}})\Big] \ge 0. \tag{4.22}$$

Suppose $\mathbb{P}\big(0 < \hat{\beta}\hat{\Gamma} < 1 \text{ and } A > 0\big) > 0$. Then, with $K = 1_{\{\hat{\beta}\hat{\Gamma}=1\}} + \hat{\beta}\hat{\Gamma}1_{\{\hat{\beta}\hat{\Gamma}<1 \text{ and } A\le 0\}}$, we would obtain from (4.22)

$$\mathbb{E}[-\hat{\beta}\hat{\Gamma}A1_{\{0<\hat{\beta}\hat{\Gamma}<1 \text{ and } A>0\}}] \ge 0,$$

which would be impossible because the integrand would be nonpositive almost surely and strictly negative with positive probability. Therefore, it must be that

$$A \le 0 \quad \text{on} \quad \{0 < \hat{\beta}\hat{\Gamma} < 1\}.$$

Similarly, supposing $\mathbb{P}\big(0 < \hat{\beta}\hat{\Gamma} < 1 \text{ and } A < 0\big) > 0$, we can derive a contradiction by taking $K = \hat{\beta}\hat{\Gamma} - (1 - \hat{\beta}\hat{\Gamma})1_{\{0<\hat{\beta}\hat{\Gamma}<1 \text{ and } A<0\}}$. Therefore,

$$A = 0 \quad \text{on} \quad \{0 < \hat{\beta}\hat{\Gamma} < 1\}. \tag{4.23}$$

In conjunction with (4.23), the inequality (4.22) implies

$$\mathbb{E}\Big[(K - 1)(A - G)1_{\{\hat{\beta}\hat{\Gamma}=1\}}\Big] + \mathbb{E}\Big[K(A - G)1_{\{\hat{\beta}\hat{\Gamma}=0\}}\Big] \ge 0. \tag{4.24}$$

Now, suppose $\mathbb{P}(A > G \text{ and } \hat{\beta}\hat{\Gamma} = 1) > 0$. Then there would exist $\delta > 0$ such that

$$\mathbb{E}\Big[(A - G)1_{\{A>G \text{ and } \hat{\beta}\hat{\Gamma}=1\}}\Big] > \delta,$$

and with $K = 1_{\{A\le G \text{ and } \hat{\beta}\hat{\Gamma}=1\}}$ it would follow from (4.24) that

$$\mathbb{E}\Big[-(A - G)1_{\{A>G \text{ and } \hat{\beta}\hat{\Gamma}=1\}}\Big] \ge 0,$$

a contradiction. Therefore,

$$A \leq G \quad \text{on} \quad \{\hat{\beta}\hat{\Gamma} = 1\}. \tag{4.25}$$

Finally, suppose $\mathbb{P}\big(A < G \text{ and } \hat{\beta}\hat{\Gamma} = 0\big) > 0$. Then there would exist $\delta > 0$ such that

$$\mathbb{E}\Big[(A - G)1_{\{A<G \text{ and } \hat{\beta}\hat{\Gamma}=0\}}\Big] < -\delta,$$

and with $K = 1_{\{A<G \text{ and } \hat{\beta}\hat{\Gamma}=0\}} + 1_{\{\hat{\beta}\hat{\Gamma}=1\}}$ it would follow from (4.24) that

$$\mathbb{E}\Big[(A - G)1_{\{A<G \text{ and } \hat{\beta}\hat{\Gamma}=0\}}\Big] \geq 0,$$

a contradiction. Therefore,

$$A \geq G \quad \text{on} \quad \{\hat{\beta}\hat{\Gamma} = 0\}. \tag{4.26}$$

The inequalities (4.23), (4.25), and (4.26) now imply that $\hat{Y}$ can be written as the right-hand side of the first equation of (3.13b).

Similarly, for each fixed $i = 1, \ldots, d$, we can show, by taking $K = 0$, and $L_j = 0$ for $j \neq i$ in (4.21), that $\hat{Z}_i$ can be written as the right-hand side of the second equation of (3.13b). This completes the proof.

Acknowledgment. This work is developed from part of my Ph.D. dissertation at Columbia University. I would like to express my deep gratitude to my adviser, Professor Ioannis Karatzas, for his guidance throughout my years at Columbia University.

REFERENCES

[1] J. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, Wiley, New York, 1984.
[2] B. Bouchard and N. Touzi, *Explicit solution of the multivariate super-replication problem under transaction costs*, Ann. Appl. Probab., 10 (2000), pp. 685–708.
[3] J. Cvitanić, *Minimizing expected loss of hedging in incomplete and constrained markets*, SIAM J. Control Optim., 38 (2000), pp. 1050–1066.
[4] J. Cvitanić and I. Karatzas, *On dynamic measures of risk*, Finance Stoch., 3 (1999), pp. 451–482.
[5] M. H. A. Davis and J. M. C. Clark, *A note on super-replicating strategies*, Philos. Trans. Roy. Soc. London Ser. A, 347 (1994), pp. 485–494.
[6] H. Föllmer and P. Leukert, *Efficient hedging: Cost versus shortfall risk*, Finance Stoch., 4 (2000), pp. 117–146.
[7] P. Guasoni, *Risk minimization under transaction costs*, Finance Stoch., 6 (2002), pp. 91–113.
[8] P. Guasoni, *Optimal investment with transaction costs and without semimartingales*, Ann. Appl. Probab., 12 (2002), pp. 1227–1246.
[9] Y. Kabanov and G. Last, *Hedging under transaction costs in currency markets: A continuous-time model*, Math. Finance, 12 (2002), pp. 63–70.
[10] Y. Kabanov and C. Stricker, *Hedging of contingent claims under transaction costs*, in Advances in Finance and Stochastics, Essays in Honour of Dieter Sondermann, K. Sandmann and P. Schönbucher, eds., Springer, Berlin, 2002, pp. 125–136.
[11] K. Kamizono, *Hedging and Optimization under Transaction Costs*, Ph.D. thesis, Department of Mathematics, Columbia University, New York, NY, 2001.
[12] M. Schwartz, *New proofs of a theorem of Komlós*, Acta Math. Hungar., 47 (1986), pp. 181–185.
[13] H. M. Soner, S. E. Shreve, and J. Cvitanić, *There is no nontrivial hedging portfolio for option pricing with transaction costs*, Ann. Appl. Probab., 5 (1995), pp. 327–355.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1559–1577

STUDY OF THE OPTIMAL HARVESTING CONTROL AND THE OPTIMALITY SYSTEM FOR AN ELLIPTIC PROBLEM*

M. DELGADO†, J. A. MONTERO‡, AND A. SUÁREZ†

Abstract. An optimal harvesting problem with a concave nonquadratic cost functional and a diffusive degenerate elliptic logistic state equation type is investigated. Under certain assumptions, we prove the existence and uniqueness of an optimal control. A characterization of the optimal control via the optimality system is also derived, which leads to approximating the optimal control.

Key words. degenerate logistic equation, singular eigenvalue problems, optimal control, optimality system

AMS subject classifications. 49J20, 49K20, 35J65, 92D25

DOI. S0363012902410903

1. Introduction. In this work we consider the optimal harvesting control of a species whose state is governed by the degenerate elliptic logistic equation; i.e.,

$$\begin{cases} -\Delta u = (a-f)u^{\alpha} - bu^{\beta} & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where Ω is a bounded and regular domain of $\mathbb{R}^N$, $N \geq 1$. Here a, f, and b are bounded functions. In particular, a is strictly positive, b is nonnegative and nontrivial, $a - f$ can change sign, and α and β satisfy

$$0 < \alpha < 1, \qquad \alpha < \beta. \tag{1.2}$$

The solutions of (1.1) can be regarded as the steady states solutions of the corresponding time-dependent model. In such a case, $u(x)$ stands for the population density and Ω for the inhabiting area. Since the population is subject to homogeneous Dirichlet boundary conditions, we are assuming that the environment surrounding Ω is lethal. In such a model, the positive function $b(x)$ describes the introspecific pressure of the species and $a(x)$ represents the growth rate of the species. The function $f(x)$ will be considered nonnegative and denotes the distribution of control harvesting of the species by reducing the growth rate. Equation (1.1), under the change of variables $w^m = u$, is a particular case of

$$\begin{cases} -\Delta w^m = (a-f)w - bw^2 & \text{in } \Omega, \\ w = 0 & \text{on } \partial\Omega. \end{cases} \tag{1.3}$$

This model was introduced in population dynamics by Gurtin and MacCamy in [11] for describing the dynamics of biological populations whose mobility depends upon their

*Received by the editors July 5, 2002; accepted for publication (in revised form) April 10, 2003; published electronically November 6, 2003.

http://www.siam.org/journals/sicon/42-5/41090.html

†Dpto. Ecuaciones Diferenciales y Análisis Numérico, Fac. Matemáticas, C/ Tarfia s/n C. P. 41012, Univ. Sevilla, Spain (madelgado@us.es, suarez@us.es). The work of these authors was supported by the Spanish Ministry of Science and Technology under grant BFM2000-0797.

‡Dpto. Análisis Matemático, C. P. 18071, Univ. Granada, Spain (jmontero@goliat.ugr.es). The work of this author was partially supported by "Junta de Andalucía" (FQM116) and DGESIC (PB98-1343).

density. In this context, $m > 1$ (nonlinear slow diffusion) means that the diffusion is slower than in the linear case $m = 1$, giving rise to more realistic biological results; see [11].

One of the main differences between the degenerate case ($m > 1$) and the nondegenerate one ($m = 1$) is that in the first case the strong maximum principle does not hold in general. So, unlike the nondegenerate case, three kinds of solutions appear: the trivial solution, the strictly positive solutions (the species can survive in the whole domain), and the nonnegative and nontrivial solutions, which are zero in a region of Ω. This region is called *dead core*.

Equation (1.1) has been studied previously for $b = 0$ in [1] and [2] and for b strictly positive in [8] and [19] and references therein. However, very little is known in the case that b can vanish in some region. In our knowledge, this problem has been analyzed only in [9] in the particular case $a - f$ equals a constant. We generalize these results and prove that there exists a maximal nonnegative solution of (1.1), which will be denoted by u_f. Moreover, when f is such that the function $a - f$ is positive, we show that (1.1) possesses a unique positive solution which is linearly asymptotically stable.

After studying in detail the state equation, our main goal is to analyze the optimal control criteria, that is, maximize the payoff functional

$$J(f) := \int_\Omega (\lambda u_f h(f) - k(f)), \tag{1.4}$$

where h and k are regular functions, and $\lambda > 0$ will be considered as a parameter. Here, J represents the difference between economic revenue measured by $\int_\Omega \lambda u_f h(f)$ and the control cost measured by $\int_\Omega k(f)$. The parameter λ describes the quotient between the price of the species and the cost of the control. This functional includes the special case (quadratic functional)

$$h(t) = t \qquad \text{and} \qquad k(t) = t^2,$$

which seems to have been introduced in population dynamics in [17] (see also [6], [15], and references therein).

We say that $f \in L^\infty_+(\Omega)$ is an optimal control if

$$J(f) = \sup_{g \in L^\infty_+(\Omega)} J(g).$$

This control problem is a generalization of the one studied in detail in [6], [17], and [18], where $\alpha = 1$, $\beta = 2$, $h(t) = t$, and $k(t) = t^2$.

In [7], the authors analyzed the case $0 < \alpha < 1 \le \beta$, b strictly positive, and the cost functional (1.4) under more restrictive monotony assumptions on functions h, k. There, the controls are restricted to the set

$$\mathcal{D} := \{f \in L^\infty_+(\Omega) : f \le a \quad \text{a.e. in } \Omega\}.$$

If $f \in \mathcal{D}$, then the maximal solution of (1.1) is strictly positive. In such a case, the existence and uniqueness of optimal control in $\mathcal{D}$ for λ sufficiently small are proved.

In this work, we assume only (1.2), b nonnegative and nontrivial, and our control space is $L^\infty_+(\Omega)$. So, u_f can have dead cores depending on the control $f \in L^\infty_+(\Omega)$ chosen. In this framework, we show that there exists an optimal control in $L^\infty_+(\Omega)$ for any $\lambda > 0$. When λ is smaller than a determined bound, we can express the optimal

control in terms of u_f and, if λ is small enough, then the optimal control is unique. In such a case, our assumptions imply that if f is an optimal control, then the dead core for u_f is empty. See [20], where a related problem is studied and where the dead core is allowed to exist.

In order to obtain the uniqueness result, we will use two different ways. First, we follow an argument described in [6] proving that the map $f \mapsto J(f)$ is Fréchet differentiable and strictly concave. The Fréchet derivability of the map $f \mapsto J(f)$ is rather more difficult than in the case $m = 1$, because it involves both linear elliptic and eigenvalue problems with potentials which blow up in a neighborhood of $\partial\Omega$. These difficulties have been solved by using results of singular eigenvalue problems from [4]; see also [12]. Second, we express the unique optimal control in terms of the solution of the optimality system, and we give an alternative proof of the uniqueness for the optimal control via the optimality system. This is an interesting point in the optimal control problems, because it allows us to approximate the optimal control by a constructive scheme which provides us with a sequence of functions converging to some special solutions of the optimality system. The uniqueness of solution of the optimality system was not considered in [17], but it was studied in [6] in the particular case $m = 1$ and the quadratic functional. Here, we present an alternative and shorter proof of the uniqueness, which can be applied to the case studied in [6]. Again, the second alternative presents another technical difficulty that must be overcome: the optimality system is a reaction-diffusion system with a singular reaction term. We present the subsupersolution method for these kinds of systems which provides us with an iterative method to approach the solution of the nonlinear system; see [5], [12] for the case of one equation.

An outline of this work is as follows: in section 2 we introduce some notations and collect some results concerning the existence and uniqueness of the principal eigenvalue and the corresponding solution for linear elliptic problems with unbounded potentials. In section 3 we study (1.1). We show the existence of a maximal nonnegative solution and, under stronger restrictions on the coefficients, the existence and uniqueness of a positive solution of (1.1). In section 4 we prove the existence of optimal control for functional J and show that for λ sufficiently small the functional J is Fréchet differentiable and strictly concave. Then, we deduce easily the uniqueness of optimal control. In the last section we characterize the optimal control. This characterization provides us with the optimality system. Finally, we prove the uniqueness of the positive solution of the optimality system and an iterative scheme based on alternating monotone sequences to approach its solution. As is remarked in recent related works (see [16], [17, Remark 4.1]), it is interesting to give conditions to guarantee the convergence of the method to the solution of the optimal control problem.

2. Preliminaries and notations. Let Ω be a bounded domain in $\mathbb{R}^N$ with a smooth boundary $\partial\Omega$. For any $f \in L^\infty(\Omega)$ we denote

$$f_M := \operatorname{ess\,sup} f, \qquad f_L := \operatorname{ess\,inf} f,$$

and define the sets

$$L^\infty_+(\Omega) := \{f \in L^\infty(\Omega) : f_L \geq 0\} \quad L^\infty_-(\Omega) := \{f \in L^\infty(\Omega) : f_M \leq 0\}.$$

Moreover, we denote $C^1_0(\overline{\Omega}) = \{u \in C^1(\overline{\Omega}) : u = 0 \quad \text{on } \partial\Omega\}$ and by P_+ its nonnegative cone, whose interior is

$$\operatorname{int}(P_+) := \{u \in C^1_0(\overline{\Omega}) : \ u > 0 \quad \text{in } \Omega,\ \partial u/\partial n < 0 \quad \text{on } \partial\Omega\},$$

where n is the outward unit normal at $\partial\Omega$.

In this section we primarily consider the singular eigenvalue problem

$$\begin{cases} -\Delta u + M(x)u = \sigma u & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{2.1}$$

where

$$M \in L^\infty_{loc}(\Omega) \quad \text{verifying} \quad M(x)d_\Omega(x) \in L^\infty(\Omega), \tag{HM}$$

and $d_\Omega(x) := dist(x, \partial\Omega)$.

The following result, whose proof can be found in [13], shows that (2.1) is well defined in $H^1_0(\Omega)$.

LEMMA 2.1. *Let $\varphi \in W^{1,q}_0(\Omega)$ for some $1 < q < \infty$. Then there exists a constant $C > 0$ such that*

$$\left\| \frac{\varphi}{d_\Omega} \right\|_q \leq C \|\varphi\|_{W^{1,q}(\Omega)}.$$

Although (2.1) is not included in the singular eigenvalue problem studied in [4], we can do some minor changes to the proofs of Theorem 3.4 and Lemma 3.5 in [4] to conclude the existence and uniqueness of the principal eigenvalue of (2.1) and its associated eigenfunction. In the following result, we collect these results and some properties of the principal eigenvalue; see [7].

THEOREM 2.2. *Assume that M satisfies* (HM). *Then there exists a unique principal eigenvalue (i.e., a real eigenvalue with an associated positive eigenfunction $\varphi_1(-\Delta + M)$). We denote it by $\sigma_1(-\Delta + M)$. Moreover, $\varphi_1(-\Delta + M) \in W^{2,p}(\Omega)$ for all $p > 1$, and so $\varphi_1(-\Delta + M) \in \text{int}(P_+)$. Furthermore, we have the following:*

1. *Assume that M_i, $i = 1, 2$, satisfy* (HM) *and $M_1 \leq M_2$. Then*

$$\sigma_1(-\Delta + M_1) \leq \sigma_1(-\Delta + M_2).$$

2. *Assume that M_n, M, $n \in \mathbb{N}$, satisfy* (HM) *with*

$$\int_\Omega M_n \varphi^2 \to \int_\Omega M\varphi^2 \quad \text{as } n \to \infty \text{ and for all } \varphi \in H^1_0(\Omega). \tag{2.2}$$

Then,

$$\sigma_1(-\Delta + M_n) \to \sigma_1(-\Delta + M) \quad \text{as } n \to \infty.$$

In the particular case $M \equiv 0$, we denote $\sigma_1 := \sigma_1(-\Delta)$ and $\varphi_1 = \varphi_1(-\Delta)$ normalized such that $\|\varphi_1\|_\infty = 1$.

When M verifies (HM), the following strong maximum principle is satisfied.

LEMMA 2.3. *Let $u \in W^{2,p}(\Omega) \cap C^1(\overline{\Omega})$, $p > 1$, be such that $u \geq 0$ in Ω, $u \neq 0$, and*

$$(-\Delta + M)u \geq 0 \quad \text{a.e. in } \Omega, \qquad u \geq 0 \quad \text{on } \partial\Omega.$$

Then $u(x) > 0$ for all $x \in \Omega$ and $(\partial u/\partial n)(x_0) < 0$ for all $x_0 \in \partial\Omega$, where $u(x_0) = 0$.

Proof. Assume there exists $x_0 \in \Omega$ such that $u(x_0) = 0$. By hypothesis, we can take $x_1 \in \Omega$, where $u(x_1) > 0$ and a subdomain regular $\Omega_1 \subset \Omega$ such that $x_0, x_1 \in \Omega_1$. But $M \in L^\infty(\Omega_1)$, and so the strong maximum principle leads us to a contradiction.

On the other hand, applying Lemma 3.6 in [4] with $\rho(s) = s^{-1}$, we get that $(\partial u/\partial n)(x_0) < 0$ for all $x_0 \in \partial\Omega$ such that $u(x_0) = 0$. □

The following technical result will help us to prove the positivity of the principal eigenvalue.

PROPOSITION 2.4. *Assume that M satisfies* (HM) *and that there exists $\varphi \in W^{2,p}_{loc}(\Omega) \cap C^0_0(\overline{\Omega})$, $p > N$ such that $\varphi > 0$ in Ω and for all subdomain $\Omega' \subset \overline{\Omega}' \subset \Omega$ it holds $(-\Delta + M)\varphi := F$ with $F_L > 0$ in Ω'. Then, $\sigma_1(-\Delta + M) > 0$.*

Proof. From the Krein–Rutman theorem, it is well known that if $-\Delta + M$ satisfies the strong maximum principle in Ω, then $\sigma_1(-\Delta + M) > 0$. Let $v \in W^{2,p}(\Omega) \cap C^1(\overline{\Omega})$ be such that $v \neq 0$, and

$$(-\Delta + M)v \geq 0 \quad \text{in a.e. } \Omega, \qquad v \geq 0 \quad \text{on } \partial\Omega.$$

We have to prove that $v > 0$ in Ω and $\partial v/\partial n(x) < 0$ for all $x \in \partial\Omega$ such that $v(x) = 0$. For each $\epsilon > 0$ and $K > 0$, we define

$$w := v + \epsilon + \epsilon K\varphi \in C^0(\overline{\Omega}).$$

And so, for any $\epsilon > 0$, there exists $\gamma(\epsilon) > 0$ such that $w > 0$ in $\Omega_\epsilon := \{x \in \Omega : d_\Omega(x) < \gamma(\epsilon)\}$. Moreover,

$$(-\Delta + M)w \geq \epsilon(M + KF) > 0 \quad \text{a.e. in } \Omega\backslash\overline{\Omega}_\epsilon \tag{2.3}$$

for K sufficiently large. Moreover, since φ is a strict supersolution in $\Omega\backslash\overline{\Omega}_\epsilon$, we can apply Corollary 2.4 in [3] and obtain that $w > 0$ in $\Omega\backslash\overline{\Omega}_\epsilon$. Thus, we get that $w > 0$ in $\Omega\backslash\overline{\Omega}_\epsilon$. Hence, $w > 0$ in Ω for all $\epsilon > 0$, and we obtain that $v \geq 0$ in Ω. Now, it suffices to apply Lemma 2.3. □

Given M verifying (HM) and $f \in L^\infty(\Omega)$ we consider the problem

$$\begin{cases} -\Delta u + M(x)u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{2.4}$$

Observe that by Lemma 2.1, (2.4) is well defined in $H^1_0(\Omega)$. The following result (whose proof can be found in [7]) shows that (2.4) possesses a unique solution in $C^1_0(\overline{\Omega})$; it provides a useful estimate and properties of the solution.

THEOREM 2.5. *Assume that M satisfies* (HM) *and $\sigma_1(-\Delta + M) > 0$. Then, there exists a unique solution $u \in C^{1,\kappa}(\overline{\Omega})$ for some $\kappa \in (0,1)$ of* (2.4). *Moreover, there exists a constant $K > 0$ (independent of f) such that*

$$\|u\|_{C^{1,\kappa}(\overline{\Omega})} \leq K\|f\|_\infty. \tag{2.5}$$

Furthermore, the following properties hold:

1. *Consider $f_i \in L^\infty(\Omega)$, $i = 1, 2$, with $f_1 \leq f_2$, and let u_i, $i = 1, 2$, be the respective solutions of* (2.4). *Then, $u_1 \leq u_2$.*
2. *Assume that M_i, $i = 1, 2$, satisfy* (HM), *$\sigma_1(-\Delta + M_1) > 0$, and $M_1 \leq M_2$. Let u_i, $i = 1, 2$, be the respective solutions of* (2.4) *with $f \in L^\infty_+(\Omega)$. Then, $u_2 \leq u_1$.*

Note: Similar results to the previous ones have been obtained in [12] when $M \in C^1(\Omega)$, $Md^\gamma_\Omega \in L^\infty(\Omega)$ for $\gamma \in (0, 2)$, and the operator is not necessarily self-adjoint.

3. The state equation. Consider the equation

$$\begin{cases} -\Delta u = (a-f)u^\alpha - bu^\beta & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{3.1}$$

and assume that

$$\begin{gathered} 0<\alpha<1, \quad \alpha<\beta, \quad a,b\in L^\infty_+(\Omega)\backslash\{0\}, \quad f\in L^\infty(\Omega), \\ a_L>0,\ (a-f)_M>0. \end{gathered} \tag{H1}$$

Observe that if $(a-f)_M \le 0$, then, by the maximum principle, (3.1) does not possess a nonnegative and nontrivial solution. This justifies the hypothesis $(a-f)_M > 0$.

In order to study (3.1), we consider the porous medium equation

$$\begin{cases} -\Delta w = \mu w^\alpha & \text{in } \Omega, \\ w = 0 & \text{on } \partial\Omega, \end{cases} \tag{3.2}$$

where $\mu \in \mathbb{R}$. The following lemma holds.

LEMMA 3.1. *Assume $0<\alpha<1$. The porous medium equation* (3.2) *has a nontrivial and nonnegative solution if and only if $\mu>0$. If $\mu>0$, there exists a unique solution, denoted w_μ, which is strictly positive and $w_\mu \in C^{2,\alpha}(\overline{\Omega})$. Moreover, it verifies*

$$\epsilon_0\varphi_1 \le w_\mu \le K_0 e \quad \text{in } \Omega, \tag{3.3}$$

where e is the unique positive solution of

$$-\Delta e = 1 \quad \text{in } \Omega, \qquad e=0 \quad \text{on } \partial\Omega,$$

and $\epsilon_0^{1-\alpha} = \mu/\sigma_1$, $K_0^{1-\alpha} = \mu\|e\|_\infty^\alpha$.

The results of the existence and uniqueness of a positive solution of (3.2) are well known; see [1], for instance. Estimate (3.3) can be obtained easily by the subsupersolution method.

The following result shows that (3.1) has a maximal nonnegative solution.

THEOREM 3.2. *Assume* (H1). *There exists a unique maximal nonnegative solution u_f of* (3.1). *Moreover, by elliptic regularity $u_f \in W^{2,p}(\Omega)$, for all $p>1$, and so $u_f \in C^{1,\kappa}(\overline{\Omega})$, with $0<\kappa\le 1-N/p$. Furthermore, we have the following a priori bound:*

$$\|u_f\|_\infty \le ((a-f)_M\|e\|_\infty)^{1/(1-\alpha)}. \tag{3.4}$$

Finally, the map $f \mapsto u_f$ is nonincreasing.

Proof. Let u be a weak solution of (3.1); then by (H1) and elliptic regularity it follows that $u \in C_0^1(\overline{\Omega})$. So, there exists $K>0$ sufficiently large such that

$$u \le Ke \qquad \text{in } \Omega,$$

and the pair (u, Ke) is a subsupersolution of (3.2) with $\mu = (a-f)_M$. By the uniqueness of positive solution of (3.2) it follows that

$$u \le w_{(a-f)_M}.$$

The existence of positive a priori bounds and that $u \equiv 0$ is a solution of (3.1) imply the existence of a nonnegative maximal solution of (3.1). By (3.3) we get the bound (3.4).

Let $f_1, f_2 \in L^\infty(\Omega)$ be such that $f_1 \leq f_2$. It is clear that the pair (u_{f_2}, Ke) is a subsupersolution of (3.1) for $f = f_1$ for $K > 0$ sufficiently large. So, there exists a solution u such that $u_{f_2} \leq u \leq Ke$. The maximality of u_{f_1} completes the proof. □

Note: From (3.4) we get, for each maximal nonnegative solution of (3.1), a uniform upper bound, i.e.,

$$u_f \leq ((a-f)_M \|e\|_\infty)^{1/(1-\alpha)} \leq (a_M \|e\|_\infty)^{1/(1-\alpha)} := \mathcal{K}, \tag{3.5}$$

for any $f \in L^\infty_+(\Omega)$.

Observe that u_f would be eventually the trivial solution. The following result shows that this cannot occur in a subset of $L^\infty(\Omega)$. We define

$$\mathcal{C} := \{f \in L^\infty(\Omega) : (a-f)_L > 0\}.$$

In the following result we prove the existence and uniqueness of positive solution of (3.1) when $f \in \mathcal{C}$.

PROPOSITION 3.3. *Assume* (H1), *and let* $f \in \mathcal{C}$. *Then, there exists a unique nontrivial and nonnegative solution,* u_f, *of* (3.1). *Moreover,* u_f *is strictly positive; in fact,*

$$\epsilon_f \varphi_1 \leq u_f \quad \text{in } \Omega, \tag{3.6}$$

where ϵ_f *satisfies*

$$\epsilon_f^{1-\alpha} \sigma_1 + \epsilon_f^{\beta-\alpha} b_M = (a-f)_L. \tag{3.7}$$

Moreover, u_f *is linearly asymptotically stable, i.e.,*

$$\sigma_1(-\Delta + M_f) > 0, \tag{3.8}$$

where

$$M_f := -\alpha(a-f)u_f^{\alpha-1} + \beta b u_f^{\beta-1}. \tag{3.9}$$

Furthermore, the map $f \in \mathcal{C} \mapsto u_f$ *is continuous.*

Note: Observe that by (H1), (3.7) possesses a unique positive solution.

Proof. For the existence of solution, it is not hard to show that $(\epsilon_f \varphi_1, w_{(a-f)_M})$ is a subsupersolution of (3.1) for $\epsilon_f > 0$ defined in (3.7).

Observe that by the strong maximum principle for $f \in \mathcal{C}$, any nontrivial and nonnegative solution u of (3.1) is strictly positive; this means that $u \in int(P_+)$.

The uniqueness of positive solution follows as in Theorem 1 of [9] and the continuity of the map $f \mapsto u_f$ as in Theorem 3.3 of [7].

It remains to prove (3.8). First observe that M_f satisfies (HM). Indeed, by (3.6), there exists a positive constant C (independent of f) such that

$$C\epsilon_f d_\Omega \leq u_f \quad \text{in } \Omega. \tag{3.10}$$

Thus, since $\alpha < 1$, we have that

$$\begin{aligned} |M_f| d_\Omega &= |u_f^{\alpha-1} d_\Omega| - \alpha(a-f) + \beta b u_f^{\beta-\alpha}| \\ &\leq |C^{\alpha-1}\epsilon_f^{\alpha-1} d_\Omega^{\alpha-1} d_\Omega| - \alpha(a-f) + \beta b u_f^{\beta-\alpha}| \leq K \end{aligned}$$

for some $K > 0$. Therefore, M_f satisfies (HM), and $\sigma_1(-\Delta + M_f)$ is well defined. Observe that $u_f^\alpha \in W^{2,p}_{loc}(\Omega) \cap C^0_0(\overline{\Omega})$ for all $p > 1$, and it satisfies

$$(-\Delta + M_f)(u_f^\alpha) = \alpha(1-\alpha)u_f^{\alpha-2}|\nabla u_f|^2 + (\beta-\alpha) b u_f^{\alpha+\beta-1} > 0 \quad \text{in } \Omega,$$

and thus we can apply Proposition 2.4 and conclude that $\sigma_1(-\Delta + M_f) > 0$. □

4. Existence and uniqueness of optimal control. For $\lambda > 0$ we consider the functional $J : L^\infty_+(\Omega) \mapsto \mathbb{R}$,

$$J(g) := \int_\Omega (\lambda h(g) u_g - k(g)),$$

where $h \in C^1(\mathbb{R}^+;\mathbb{R}^+)$, $k \in C^2(\mathbb{R}^+;\mathbb{R}^+)$; $h(s) = 0$ if and only if $s = 0$, and $k(s) = 0$ if and only if $s = 0$. Function h is concave, and k is a strictly convex function satisfying $k''(s) \geq k_0 > 0$ for some k_0. Note that h', k' are Lipschitz continuous functions on a bounded set. We assume

$$\text{(H2)} \qquad \lim_{t\to 0} \frac{k(t)}{h(t)} = 0, \qquad \lim_{t\to+\infty} \frac{k(t)}{h(t)} = +\infty.$$

Observe that the particular case $h(t) = t$ and $k(t) = t^2$, studied in [6], [15], [17], and [18], is in the setting of our functional. Also, we remove some hypotheses of monotonous type involving functions k and h considered in [7]. The idea will be to show that the integrand of functional $J(f)$ must be positive if f is an optimal control.

In the first part of this section we want to prove the existence of the optimal control under hypothesis (H2). First, we prove that the optimal controls are bounded.

LEMMA 4.1. *Assume* (H2). *If $f \in L^\infty_+(\Omega)$ is an optimal control, then*

$$\lambda u_f(x) h(f(x)) \geq k(f(x)) \text{ a.e. in } \Omega. \tag{4.1}$$

Moreover, if $f \in L^\infty_+(\Omega)$ is an optimal control, then

$$0 \leq f \leq T_\lambda,$$

where

$$T_\lambda := \sup\left\{ t \in \mathbb{R}^+ : \frac{k(t)}{h(t)} = \lambda \mathcal{K} \right\},$$

and $\mathcal{K}$ is the uniform bound defined in (3.5).

Note: By the hypotheses imposed on h and k and (H2), it follows that $T_\lambda > 0$ and that $T_\lambda \to 0$ as $\lambda \downarrow 0$.

Proof. Suppose that $f \in L^\infty_+(\Omega)$ is an optimal control and (4.1) is not true. Then, there exists $\Omega_1 \subset \Omega$ with $|\Omega_1| > 0$ (positive measure) such that

$$\lambda u_f(x) h(f(x)) < k(f(x)) \quad \forall x \in \Omega_1. \tag{4.2}$$

Now, by defining a new control $\overline{f}$ as

$$\overline{f}(x) = \begin{cases} f(x) & \text{if } x \in \Omega \setminus \Omega_1, \\ 0 & \text{if } x \in \Omega_1, \end{cases}$$

and taking into account that $u_{\overline{f}} \geq u_f$ in Ω, we obtain

$$\begin{aligned} J(f) = & \int_{\Omega_1} \lambda u_f(x) h(f(x)) - k(f(x)) + \int_{\Omega\setminus\Omega_1} \lambda u_f(x) h(f(x)) - k(f(x)) \\ < & \int_{\Omega\setminus\Omega_1} \lambda u_f(x) h(f(x)) - k(f(x)) \leq \int_{\Omega\setminus\Omega_1} \lambda u_{\overline{f}}(x) h(f(x)) - k(f(x)) \\ = & \int_{\Omega\setminus\Omega_1} \lambda u_{\overline{f}}(x) h(\overline{f}(x)) - k(\overline{f}(x)) = J(\overline{f}). \end{aligned}$$

But f is an optimal control. So, previous inequality shows that (4.2) is absurd. It also shows that $f \leq T_\lambda$ follows from the definition of T_λ, Theorem 3.2, and (4.1). □

THEOREM 4.2. *Assume* (H2). *There exists an optimal control; i.e.,* $f \in L^\infty_+(\Omega)$ *such that*

$$J(f) = \sup_{g \in L^\infty_+(\Omega)} J(g).$$

Moreover, the benefit is positive, i.e., $\sup_{g \in L^\infty_+(\Omega)} J(g) > 0$.

Proof. By (3.5) and Lemma 4.1, it follows that

$$s := \sup_{g \in L^\infty_+(\Omega)} J(g) < +\infty,$$

and so there exists a maximizing sequence $f_n \in L^\infty_+(\Omega)$. By similar reasoning to that used in the previous lemma, we can suppose that $0 \leq f_n \leq T_\lambda$. Then, there exists a subsequence, relabelled by f_n, such that

$$f_n \rightharpoonup f \in [0, T_\lambda] \quad \text{in } L^2(\Omega).$$

By (3.5), we can prove that

$$u_{f_n} \to u_* \quad \text{in } H^1_0(\Omega), \tag{4.3}$$

where u_* is a positive solution of (3.1) (possibly not the maximal positive solution). In any case, we have $u_f \geq u_*$.

Now, taking into account the concavity of the functions h and $-k$, it follows that

$$J(f) \geq \limsup \int_\Omega \lambda h(f_n) u_{f_n} - k(f_n) = s,$$

and so we have the existence of an optimal control.

The optimal benefit is positive by following an argument like that used in [7]. In fact, it is clear, from the asymptotic properties of the functions h and k, that $J(\epsilon) > 0$ by taking $\epsilon \in \mathbb{R}^+$ small enough. □

Now, we are going to prove that, for λ sufficiently small, there exists a unique optimal control. For that we will use the argument described in section 6 in [6]. In summary, by Lemma 4.1 we know that the optimal controls belong to a convex, $[0, T_\lambda]$. Moreover, we will show that J is Fréchet continuously differentiable and strictly concave in $[0, T_\lambda]$. Hence, the uniqueness of optimal control is a direct consequence. The first step is the following result, which provides us with the Gâteaux derivative of the map $f \in \mathcal{C} \mapsto u_f \in \text{int}(P_+)$. Its proof is similar to Lemma 3.5 in [7], and so we omit it.

LEMMA 4.3. *Let* $f \in \mathcal{C}$, $g \in L^\infty(\Omega)$, *and* $\epsilon \simeq 0$ *be such that* $f + \epsilon g \in \mathcal{C}$. *Then,*

$$\frac{u_{f+\epsilon g} - u_f}{\epsilon} \rightharpoonup \xi_{f,g} \quad \text{in } H^1_0(\Omega) \text{ as } \epsilon \to 0,$$

where $\xi_{f,g}$ *is the unique solution of*

$$\begin{cases} -\Delta \xi + M_f(x)\xi = -g u_f^\alpha & \text{in } \Omega, \\ \xi = 0 & \text{on } \partial\Omega. \end{cases} \tag{4.4}$$

Observe that (4.4) has a unique solution because $\sigma_1(-\Delta + M_f) > 0$ (see (3.8)) and Theorem 2.5.

Now, we can prove the following proposition (see Proposition 4.4 in [7]).

PROPOSITION 4.4. *Let $J : \mathcal{C} \subset L^\infty(\Omega) \mapsto \mathbb{R}$ be. Then J is Fréchet continuously differentiable and*

$$J'(f)(g) = \int_\Omega (\lambda h'(f)u_f - \lambda u_f^\alpha P_f - k'(f))g \quad \forall f \in \mathcal{C}\ \forall g \in L^\infty(\Omega), \tag{4.5}$$

where for any $f \in \mathcal{C}$, $P_f \in C_0^1(\overline{\Omega})$ is the unique solution of

$$\begin{cases} -\Delta P_f + M_f(x)P_f = h(f) & \text{in } \Omega, \\ P_f = 0 & \text{on } \partial\Omega, \end{cases} \tag{4.6}$$

and M_f is defined in (3.9).

Note: Since M_f satisfies (HM) and by (3.8), it follows from Theorem 2.5 that the existence and uniqueness of $P_f \in C_0^1(\overline{\Omega})$.

Observe that by the note following Lemma 4.1, there exists $\lambda_0 > 0$ such that

$$a_L > T_\lambda \quad \text{for } \lambda < \lambda_0. \tag{4.7}$$

Following the argument of Theorem 3.1 in [17] (using now (4.7) and Proposition 4.4) we obtain the following corollary.

COROLLARY 4.5. *Assume* (H2). *Let $f \in L^\infty_+(\Omega)$ be an optimal control. Then for $\lambda < \lambda_0$,*

$$k'(f) = \lambda(h'(f)u_f - u_f^\alpha P_f)^+.$$

In order to prove that J is strictly concave in $[0, T_\lambda]$, we will show that maps involved in J' are Lipschitz continuous. This result was proven in [7] when $\beta \geq 1$. Since the Lipschitz character of the maps involved is crucial in this work (see, for example, the proof of Lemma 5.4), we present a complete proof of this result for the reader's convenience.

THEOREM 4.6. *Assume* (H2). *There exists $\Lambda > 0$ such that for $0 < \lambda < \Lambda$ the maps*

$$f \in [0, T_\lambda] \mapsto u_f,\ P_f,\ u_f^\alpha P_f \in L^\infty(\Omega),$$

are Lipschitz continuous, with the Lipschitz constants independent of λ.

Proof. Let $f, g \in [0, T_\lambda]$ be; by the monotony of the map $f \mapsto u_f$, it follows that $u_{T_\lambda} \leq u_f, u_g \leq u_0$. Moreover, for $\lambda < \lambda_0$ (defined in (4.7)), $u_{T_\lambda} > 0$, and so

$$0 < u_{T_\lambda} \leq u_f, u_g \leq u_0$$

for $\lambda < \lambda_0$. Hereafter, we take $\lambda < \lambda_0$. By the mean value theorem,

$$\begin{aligned} &u_f^\alpha - u_g^\alpha = \alpha\theta^{\alpha-1}(f,g)(u_f - u_g), \quad u_f^\beta - u_g^\beta = \beta\eta^{\beta-1}(f,g)(u_f - u_g) \quad \text{with} \\ &0 < u_{T_\lambda} \leq \min\{u_f, u_g\} \leq \theta(f,g), \eta(f,g) \leq \max\{u_f, u_g\} \leq u_0. \end{aligned} \tag{4.8}$$

Let $w := u_f - u_g$ be. Then, w satisfies

$$\begin{cases} (-\Delta + N(f,g))w = (g - f)u_g^\alpha & \text{in } \Omega, \\ w = 0 & \text{on } \partial\Omega, \end{cases}$$

where

$$N(f,g) := -\alpha(a-f)\theta^{\alpha-1}(f,g) + \beta b\eta^{\beta-1}(f,g).$$

Using $f \geq 0$ and (4.8), it follows that

$$N(f,g) \geq -\alpha a\theta^{\alpha-1}(f,g) + \beta b\eta^{\beta-1}(f,g) \geq m_\lambda,$$

where

$$m_\lambda := \begin{cases} -\alpha a u_{T_\lambda}^{\alpha-1} + b\beta u_{T_\lambda}^{\beta-1} & \text{if } \beta \geq 1, \\ -\alpha a u_{T_\lambda}^{\alpha-1} + b\beta u_0^{\beta-1} & \text{if } \beta < 1. \end{cases} \tag{4.9}$$

It is not hard to show that m_λ satisfies (HM). Moreover, we claim that as $\lambda \downarrow 0$,

$$\int_\Omega m_\lambda \varphi^2 \to \int_\Omega (-\alpha a u_0^{\alpha-1} + b\beta u_0^{\beta-1})\varphi^2 \quad \forall \varphi \in H_0^1(\Omega). \tag{4.10}$$

Indeed, for $\varphi \in H_0^1(\Omega)$ and using (3.10) we have

$$\int_\Omega (u_{T_\lambda}^{\alpha-1} - u_0^{\alpha-1})\varphi^2 = \int_\Omega (u_{T_\lambda}^{\alpha} - u_{T_\lambda} u_0^{\alpha-1})\frac{\varphi}{u_{T_\lambda}}\varphi \leq C\epsilon_{T_\lambda}^{-1} \|u_{T_\lambda}^{\alpha} - u_{T_\lambda} u_0^{\alpha-1}\|_\infty \int_\Omega \frac{\varphi}{d_\Omega}\varphi,$$

where ϵ_{T_λ} is defined in (3.7). By the continuity of the map $f \mapsto u_f$, Lemma 2.1, and the fact that ϵ_{T_λ} does not tend to 0 as $\lambda \downarrow 0$, we obtain that

$$\int_\Omega (u_{T_\lambda}^{\alpha-1} - u_0^{\alpha-1})\varphi^2 \to 0 \qquad \text{as } \lambda \downarrow 0.$$

Reasoning similarly with the other terms, (4.10) is proved. So, by Theorem 2.2 we obtain that

$$\sigma_1(-\Delta + N(f,g)) \geq \sigma_1(-\Delta + m_\lambda) \to \sigma_1(-\Delta - \alpha a u_0^{\alpha-1} + b\beta u_0^{\beta-1}) > 0 \tag{4.11}$$

as $\lambda \downarrow 0$. This last inequality follows by (3.8) because $f \equiv 0 \in \mathcal{C}$. Hence, using the monotony of the map $\lambda \mapsto T_\lambda$, there exists $\lambda_1 > 0$ such that

$$N(f,g) \geq m_\lambda \geq m_{\lambda_1} \tag{4.12}$$

and

$$\sigma_1(-\Delta + N(f,g)) \geq \sigma_1(-\Delta + m_{\lambda_1} > 0 \quad \text{for } \lambda < \lambda_1. \tag{4.13}$$

So, by (4.12) we get

$$(-\Delta + m_{\lambda_1})w \leq (g-f)u_g^\alpha,$$

and hence, using (4.13), Theorem 2.5, and (3.5), it follows that

$$\|u_f - u_g\|_\infty = \|w\|_\infty \leq \|w\|_{C^1(\overline{\Omega})} \leq C\|f-g\|_\infty. \tag{4.14}$$

This shows that the map $f \mapsto u_f$ is Lipschitz.

Now, take $f \in [0, T_\lambda]$. Using the monotony of the map $f \mapsto u_f$, we have that

$$M_f \geq m_{\lambda_1}. \tag{4.15}$$

Thus, by Theorem 2.5 we obtain that

$$P_f \le \mathcal{P} \quad \text{in } \Omega, \tag{4.16}$$

where $\mathcal{P} \in C_0^1(\overline{\Omega})$ is the unique solution of

$$\begin{cases} -\Delta u + m_{\lambda_1} u &= T \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{cases}$$

and $T := \max_{r\in[0,T_{\lambda_1}]} h(r)$.

We will prove now that the map $f \mapsto P_f$ is Lipschitz. Let $f, g \in [0, T_\lambda]$ and $z := P_f - P_g$ be. Then z satisfies

$$-\Delta z + M_f z = T(f,g) \quad \text{in } \Omega, \qquad z = 0 \quad \text{on } \partial\Omega,$$

where

$$T(f,g) = h(f) - h(g) + P_g[\alpha(a-f)(u_f^{\alpha-1} - u_g^{\alpha-1}) - \beta b(u_f^{\beta-1} - u_g^{\beta-1})] + \alpha(g-f)P_g u_g^{\alpha-1}.$$

Applying again the mean value theorem, we get

$$\begin{gathered} u_f^{\alpha-1} - u_g^{\alpha-1} = (\alpha-1)\xi^{\alpha-2}(f,g)(u_f - u_g), \\ u_f^{\beta-1} - u_g^{\beta-1} = (\beta-1)\eta^{\beta-2}(f,g)(u_f - u_g), \\ 0 < u_{T_\lambda} \le \min\{u_f, u_g\} \le \xi(f,g), \eta(f,g) \le \max\{u_f, u_g\} \le u_0. \end{gathered} \tag{4.17}$$

Hence,

$$T(f,g) = h(f) - h(g) + P_g[\alpha(\alpha-1)(a-f)\xi^{\alpha-2} - \beta(\beta-1)b\eta^{\beta-2}](u_f - u_g) + \alpha(g-f)P_g u_g^{\alpha-1}.$$

By a similar argument to the one used in the proof of (4.14), we obtain

$$\|P_f - P_g\|_\infty = \|z\|_\infty \le C\|T(f,g)\|_\infty. \tag{4.18}$$

Since $\mathcal{P} \in C_0^1(\overline{\Omega})$, it follows that

$$|\mathcal{P}(x)| \le d_\Omega(x)\|\mathcal{P}\|_{C^1(\overline{\Omega})}. \tag{4.19}$$

So, using (3.10), (4.16), and (4.19), we obtain

$$\begin{aligned} \|(f-g)P_g u_g^{\alpha-1}\|_\infty &\le C\|f-g\|_\infty \|P_g u_{T_\lambda}^{\alpha-1}\|_\infty \le C\|f-g\|_\infty \|\mathcal{P} d_\Omega^{\alpha-1}\|_\infty \\ &\le C\|f-g\|_\infty \|d_\Omega^\alpha\|_\infty \|\mathcal{P}\|_{C^1(\overline{\Omega})} \le C\|f-g\|_\infty, \end{aligned}$$

with C independent of f and g.
On the other hand, since $u_f - u_g \in C_0^1(\overline{\Omega})$ and using (4.14), (4.16), (4.17), and (4.19),

$$\begin{aligned} \|(a-f)P_g\xi^{\alpha-2}(u_f - u_g)\|_\infty &\le C\|\mathcal{P}\xi^{\alpha-2}(u_f - u_g)\|_\infty \\ &\le C\|\mathcal{P}\|_{C^1(\overline{\Omega})}\|d_\Omega^\alpha\|_\infty \|u_f - u_g\|_{C^1(\overline{\Omega})} \le C\|f-g\|_\infty \end{aligned}$$

with C independent of f and g.

Analogously it can be treated the term $P_g\eta^{\beta-2}(u_f - u_g)$. Then, since h is Lipschitz in $[0, T_\lambda]$ and by (4.18), it follows that the map $f \mapsto P_f$ is Lipschitz.

Let $f, g \in [0, T_\lambda]$ be. By (4.8), we have

$$\|(u_f^\alpha - u_g^\alpha)P_f\|_\infty = \|\alpha\xi^{\alpha-1}P_f(u_f - u_g)\|_\infty \le C\|\mathcal{P}\|_{C^1(\overline{\Omega})}\|f-g\|_\infty \le C\|f-g\|_\infty,$$

and so

$$\|u_f^\alpha P_f - u_g^\alpha P_g\|_\infty \le \|(u_f^\alpha - u_g^\alpha)P_f\|_\infty + \|u_g^\alpha(P_f - P_g)\|_\infty \le C\|f-g\|_\infty.$$

This completes the proof. □

We can conclude the main result about uniqueness of optimal control of this section with the following theorem.

THEOREM 4.7. *Assume* (H2). *Then, there exists* $\Lambda_0 > 0$ *such that if* $\lambda < \Lambda_0$, *there exists a unique optimal control.*

5. The optimality system and the approximation to the optimal control. In this section, we deduce the optimality system in the special cases $h(t) = t$ and $k(t) = t^2$, which satisfy clearly (H2). The optimality system will be used to demonstrate the uniqueness of the optimal control in a different way and provides an iterative method to approach it. In this case, we know that

$$T_\lambda = \lambda\mathcal{K} \quad \text{and} \quad \lambda_0 = \frac{a_L}{\mathcal{K}},$$

where $\mathcal{K}$ is defined in (3.5). Moreover, by Corollary 4.5, for $\lambda < \lambda_0$, if f is an optimal control, then

$$f = \frac{\lambda}{2} u_f (1 - u_f^{\alpha-1} P_f)^+. \tag{5.1}$$

Let ψ be the unique positive solution of

$$\begin{cases} -\Delta\psi + m_{\lambda_1}\psi = \mathcal{K} & \text{in } \Omega, \\ \psi = 0 & \text{on } \partial\Omega, \end{cases} \tag{5.2}$$

where m_{λ_1} is defined in (4.9) and satisfies (4.12) and (4.13). So, if f is an optimal control it follows by Lemma 4.1 that $f \in [0, \lambda\mathcal{K}]$. On the other hand, by (4.15) and Theorem 2.5, we get that

$$P_f \le \lambda\psi \qquad \text{for } \lambda \le \lambda_1. \tag{5.3}$$

As a consequence of (5.3) we obtain the following proposition (see Proposition 5.2 and Corollary 5.3 in [7]).

PROPOSITION 5.1. *Assume* (H1). *There exists a constant* $\Lambda_1 > 0$ *such that if* $\lambda \le \Lambda_1$, *then*

$$P_f \le u_f^{1-\alpha}. \tag{5.4}$$

So, if f *is an optimal control, we have that*

$$f = \frac{\lambda}{2} u_f (1 - u_f^{\alpha-1} P_f). \tag{5.5}$$

As a consequence, any optimal control f *may be expressed as in* (5.5), *where the pair* $(u_f, P_f) := (u, P)$ *satisfies*

$$\begin{cases} -\Delta u = u^\alpha(a - \frac{\lambda}{2}u + \frac{\lambda}{2}u^\alpha P - bu^{\beta-\alpha}) \text{ in } \Omega, \\ -\Delta P + (-\alpha a u^{\alpha-1} + \beta b u^{\beta-1})P = \frac{\lambda}{2}(u - u^\alpha P(1+\alpha) + \alpha u^{2\alpha-1}P^2) \text{ in } \Omega, \\ u = P = 0 \text{ on } \partial\Omega, \end{cases} \tag{5.6}$$

and $u > 0$.

The former result says that, when λ is small enough, if f is an optimal control, then (u_f, P_f) is a solution of (5.6). We are going to prove now that, for a range of λ, there exists a unique positive solution of (5.6) verifying $u^{1-\alpha} \geq P$, and so the unique optimal control will be

$$f = \frac{\lambda}{2}(u - u^\alpha P).$$

THEOREM 5.2 (uniqueness of optimal control). *Assume* (H1). *There exists* $\Lambda_2 > 0$ *such that for* $\lambda \leq \Lambda_2$, (5.6) *possesses a unique positive solution* (u, P) *satisfying* $u^{1-\alpha} \geq P$.

Proof. We define the following map:

$$\mathcal{T} : I := [0, \lambda\mathcal{K}] \subset L^\infty_+(\Omega) \mapsto L^\infty_+(\Omega), \qquad f \mapsto \mathcal{T}(f) = \frac{\lambda}{2}(u_f - u_f^\alpha P_f).$$

By Theorem 4.6, for $\lambda < \Lambda$, $\mathcal{T}$ is a Lipschitz continuous function with Lipschitz constant of type $\lambda\mathcal{L}/2$, where $\mathcal{L}$ is the corresponding one for the function $f \mapsto u_f - u_f^\alpha P_f$. So, we can choose $\Lambda_2 := \min\{\Lambda, \frac{2}{\mathcal{L}}\}$ such that for $\lambda \leq \Lambda_2$, $\mathcal{T}$ is a contractive function.

Assume that there exist two positive solutions (u_i, P_i), $i = 1, 2$, of (5.6) with $u_i^{1-\alpha} \geq P_i$. We define

$$f_i = \frac{\lambda}{2}(u_i - u_i^\alpha P_i) \in I, \quad i = 1, 2.$$

Hence, by (5.6) and Proposition 3.3 we have that

$$u_i = u_{f_i}, \qquad P_i = P_{f_i}, \quad \Rightarrow \mathcal{T}(f_i) = f_i, \quad i = 1, 2.$$

Since $\mathcal{T}$ is contractive, it follows that $f_1 = f_2$, and again by Proposition 3.3 we have that $u_{f_1} = u_{f_2}$; hence $u_1 = u_2$, and so $P_1 = P_2$. This completes the proof. □

Now, we use the optimal control characterization obtained by formula (5.5) to give an iterative procedure to approach it. The idea is to be near the solution of the optimality system by sub- and supersolutions (see other papers related with similar problems [6], [14], [15], [17]). The interest here, besides the degeneration of the second equation of the optimality system, is that we prove the convergence of the method by a different argument than that used in the mentioned references. We start this part with some notation. We define, for simplicity, the following functions:

$$B(x, u, p) = \left[a(x) - \frac{\lambda}{2}(u - u^\alpha p)\right] u^\alpha - b(x)u^\beta \text{ for } x \in \Omega,$$

and, taking into account the monotony properties of the second equation of optimality system (5.6), we define

$$D(x, u, p) = \begin{cases} \frac{\lambda}{2}u - \beta p b(x) u^{\beta-1} & \text{if } \beta < 1,\ 2\alpha - 1 < 0, \\ \frac{\lambda}{2}u - \beta p b(x) u^{\beta-1} + \frac{\lambda}{2}\alpha u^{2\alpha-1}p^2 & \text{if } \beta < 1,\ 2\alpha - 1 \geq 0, \\ \frac{\lambda}{2}u & \text{if } \beta \geq 1,\ 2\alpha - 1 < 0, \\ \frac{\lambda}{2}u + \frac{\lambda}{2}\alpha u^{2\alpha-1}p^2 & \text{if } \beta \geq 1,\ 2\alpha - 1 \geq 0 \end{cases}$$

and

$$C(x,u,p)=\begin{cases} p(\alpha a(x)u^{\alpha-1}-\frac{\lambda}{2}u^{\alpha}(\alpha+1))+\frac{\lambda}{2}\alpha u^{2\alpha-1}p^2 & \text{if } \beta<1,\ 2\alpha-1<0,\\ p(\alpha a(x)u^{\alpha-1}-\frac{\lambda}{2}u^{\alpha}(\alpha+1)) & \text{if } \beta<1,\ 2\alpha-1\geq 0,\\ p(\alpha a(x)u^{\alpha-1}-\beta b(x)u^{\beta-1} & \\ -\frac{\lambda}{2}u^{\alpha}(\alpha+1))+\frac{\lambda}{2}\alpha u^{2\alpha-1}p^2 & \text{if } \beta\geq 1,\ 2\alpha-1<0,\\ p(\alpha a(x)u^{\alpha-1}-\beta b(x)u^{\beta-1}-\frac{\lambda}{2}u^{\alpha}(\alpha+1)) & \text{if } \beta\geq 1,\ 2\alpha-1\geq 0. \end{cases}$$

We are interested in the solutions, (u,p), of optimality system (5.6) that satisfy $u_{\lambda\mathcal{K}}\leq u\leq u_0$ and $0\leq p\leq\lambda\psi$ (recall (5.3)). Consequently, we can reduce the study for solutions that satisfy $(u,p)\in[u_{\lambda\mathcal{K}},u_0]\times[0,\lambda\psi]$. Therefore, there exist a constant $K>0$ and a function $M_1(x), x\in\Omega$, satisfying hypothesis (HM) such that

$$\begin{array}{ll} B(x,u,p)+Ku^{\alpha} & (\nearrow u,\nearrow p)\,,\\ C(x,u,p)+M_1(x)p & (\searrow u,\nearrow p)\,,\\ D(x,u,p)+M_1(x)p & (\nearrow u,\nearrow p)\,; \end{array}$$

i.e., $B(x,u,p)+Ku^{\alpha}$ is increasing in u for fixed $x\in\Omega$ and $0\leq p\leq\lambda\psi$ and increasing in p for fixed $x\in\Omega$ and $u_{\lambda\mathcal{K}}\leq u\leq u_0$. The other terms are interpreted analogously.

Definition 5.3 (subsupersolutions). *The functions $u_1,p_1,u^1,p^1\in L^{\infty}(\Omega)\cap H^1(\Omega)$ are said to be a system of subsupersolutions for optimality system* (5.6) *if they verify*

$$\begin{cases} u_1(x)\leq u^1(x),\quad p_1(x)\leq p^1(x) & \text{a.e. in } \Omega,\\ p_1\leq 0\leq p^1 & \text{on } \partial\Omega, \end{cases} \tag{5.7}$$

and there exists a positive constant k such that

$$0<kd_{\Omega}(x)\leq u_1(x)\leq u^1(x)\quad \text{a.e. in } \Omega \tag{5.8}$$

and, for any $\phi\in H_0^1(\Omega)$, $\phi\geq 0$,

$$\int_{\Omega}\nabla u^1\cdot\nabla\phi\geq\int_{\Omega}B(x,u^1,p^1)\phi,\qquad \int_{\Omega}\nabla u_1\cdot\nabla\phi\leq\int_{\Omega}B(x,u_1,p_1)\phi,$$

$$\int_{\Omega}\nabla p^1\cdot\nabla\phi\geq\int_{\Omega}C(x,u_1,p^1)\phi+\int_{\Omega}D(x,u^1,p^1)\phi,$$

$$\int_{\Omega}\nabla p_1\cdot\nabla\phi\leq\int_{\Omega}C(x,u^1,p_1)\phi+\int_{\Omega}D(x,u_1,p_1)\phi.$$

Recall that a function $v\in H^1(\Omega)$ is said to be less than or equal to $w\in H^1(\Omega)$ on $\partial\Omega$ when $(v-w)^+=\max\{0,v-w\}\in H_0^1(\Omega)$.

It is not difficult to prove that, under the hypothesis of Theorem 5.2, there exists a $\Lambda_3>0$ such that if $\lambda\leq\Lambda_3$, then the functions

$$u_1=u_{\lambda\mathcal{K}},\quad p_1\equiv 0,\quad u^1=u_0,\quad p^1=\lambda\psi, \tag{5.9}$$

are a system of subsupersolutions for the optimality system (5.6) in the sense of Definition 5.3. We show only the case $p^1=\lambda\psi$ when $\beta\geq 1$, $2\alpha-1\geq 0$. The other cases are similar. It is not hard to show that $p^1=\lambda\psi$ is a supersolution if

$$\lambda(\mathcal{K}-m_{\lambda_1}\psi)\geq\lambda\psi\left[\alpha a u_{\lambda\mathcal{K}}^{\alpha-1}-\beta b u_{\lambda\mathcal{K}}^{\beta-1}-\frac{\lambda}{2}u_{\lambda\mathcal{K}}^{\alpha}(\alpha+1)\right]+\frac{\lambda}{2}u_0+\frac{\lambda}{2}\alpha u_0^{2\alpha-1}(\lambda\psi)^2$$

or, equivalently,

$$\mathcal{K} \geq \psi \left[m_{\lambda_1} + \alpha a u_{\lambda\mathcal{K}}^{\alpha-1} - \beta b u_{\lambda\mathcal{K}}^{\beta-1} \right] - \frac{\lambda}{2} u_{\lambda\mathcal{K}}^{\alpha}(\alpha+1)\psi + \frac{1}{2}u_0 + \frac{\alpha}{2}\lambda^2 u_0^{2\alpha-1}\psi^2.$$

Recalling the definition of m_{λ_1}, for $\lambda \leq \lambda_1$ we have that $m_{\lambda_1} + \alpha a u_{\lambda\mathcal{K}}^{\alpha-1} - \beta b u_{\lambda\mathcal{K}}^{\beta-1} \leq 0$, and by (3.5) $u_0 \leq \mathcal{K}$; it is enough to take λ small to obtain that p^1 is a supersolution.

Now, we define by induction, for $n \geq 2$, the sequences $\{u_n\}$, $\{u^n\}$, $\{p_n\}$, $\{p^n\} \in H_0^1(\Omega)$ as

$$-\Delta u_n + K(u_n)^\alpha = B(x, u_{n-1}, p_{n-1}) + K(u_{n-1})^\alpha \quad \text{in } \Omega, \tag{5.10}$$

$$-\Delta u^n + K(u^n)^\alpha = B(x, u^{n-1}, p^{n-1}) + K(u^{n-1})^\alpha \quad \text{in } \Omega, \tag{5.11}$$

$$-\Delta p_n + M_1(x)p_n = C(x, u^n, p_{n-1}) + D(x, u_n, p_{n-1}) + M_1(x)p_{n-1} \quad \text{in } \Omega, \tag{5.12}$$

$$-\Delta p^n + M_1(x)p^n = C(x, u_n, p^{n-1}) + D(x, u^n, p^{n-1}) + M_1(x)p^{n-1} \quad \text{in } \Omega. \tag{5.13}$$

Observe that sequences $\{u_n\}$, $\{u^n\}$, are well defined because the map $u \mapsto Ku^\alpha$ is continuous and strictly increasing and such that $B(x,u,p) + Ku^\alpha$ is also increasing in u when the other variables are fixed. (See more details in [5], [10].)

On the other hand, fixed u_1, u^1 and, thanks to (5.8), the problems (5.12) and (5.13) are in the setting of (2.4), and so by Theorem 2.5 it follows the existence and uniqueness of p_2 and p^2 and such that $p_2 \leq p^2$ and so on. We note that for (5.10)–(5.11) and (5.12)–(5.13), the subsupersolutions method works (cf. [12]). The standard method gives us the following order relation:

$$u_1 \leq u_2 \leq \cdots \leq u_n \leq u^n \leq u^{n-1} \leq \cdots \leq u^1,$$
$$p_1 \leq p_2 \leq \cdots \leq p_n \leq p^n \leq p^{n-1} \leq \cdots \leq p^1$$

and

$$u_n \nearrow u,\ u^n \searrow v,\ p_n \nearrow p,\ p^n \searrow q$$

(pointwise), where u, v, p, q belong to $C^{1,\delta}(\Omega)$, for any $\delta \in (0,1)$, and satisfy the system

$$\begin{cases} -\Delta u = B(x,u,p) & \text{in } \Omega, \\ -\Delta v = B(x,v,q) & \text{in } \Omega, \\ -\Delta p = C(x,v,p) + D(x,u,p) & \text{in } \Omega, \\ -\Delta q = C(x,u,q) + D(x,v,q) & \text{in } \Omega, \\ u = v = p = q = 0 & \text{on } \partial\Omega \end{cases} \tag{5.14}$$

and

$$\begin{aligned} u_1 = u_{\lambda\mathcal{K}} \leq u,\ v \leq u^1 = u_0 \quad &\text{in } \Omega \\ p_1 = 0 \leq p,\ q \leq p^1 = \lambda\psi \leq u_{\lambda\mathcal{K}}^{1-\alpha} \quad &\text{in } \Omega. \end{aligned} \tag{5.15}$$

Clearly, if (u,p) is the solution of optimality system (5.6), then (u,u,p,p) is a solution of (5.14). So, to complete the iterative approximation and the convergence of the sequences $\{u_n\}$, $\{u^n\}$, $\{p_n\}$, $\{p^n\}$ to the unique solution, (u,p), of the optimality system, it is sufficient to prove the uniqueness of the solution for system (5.14), under conditions (5.15). To do it, we need the following technical lemma.

LEMMA 5.4. *Assume* (H1). *Then*

$$\forall f,\ g \in [0, \lambda\mathcal{K}] \subset L_+^\infty(\Omega),$$

it is possible to define the function $P_{f,g}$ as the unique positive solution of the problem

$$\begin{cases} -\Delta P = C(x,u_f,P) + D(x,u_g,P) & \text{in } \Omega, \\ P = 0 & \text{on } \partial\Omega, \end{cases} \tag{5.16}$$

satisfying

$$0 \le P_{f,g} \le \lambda\psi, \tag{5.17}$$

provided that the parameter λ is small enough and the function ψ is defined in (5.2). *Moreover, the map defined before, $(f,g) \in [0,\lambda\mathcal{K}] \times [0,\lambda\mathcal{K}] \mapsto P_{f,g} \in L^\infty(\Omega)$, is Lipschitz continuous.*

An analogous result is obtained interchanging u_f and u_g in (5.16).

Proof. We consider the case $\beta \ge 1$, $2\alpha - 1 \ge 0$. The other cases have similar proofs. Observe that, in this case, (5.16) is

$$\begin{cases} -\Delta P + \left[-\alpha a u_f^{\alpha-1} + \beta b u_f^{\beta-1} + \frac{\lambda}{2} u_f^\alpha (1+\alpha) \right] P = \frac{\lambda}{2} u_g + \frac{\lambda}{2} \alpha u_g^{2\alpha-1} P^2 \text{ in } \Omega, \\ P = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{5.18}$$

Taking into account Theorem 2.5, condition (5.17), and the definition of the function ψ, we can use the subsupersolution method with $p_* \equiv 0$ as subsolution and $p^* \equiv \lambda\psi$ as supersolution, provided $\lambda > 0$ small. Thus, the existence of positive solution of (5.18) is proved. The uniqueness of a contradiction argument follows. Suppose that P, Q are two solutions under above requirements; then $P - Q$ satisfies

$$(-\Delta + M_1(x))(P - Q) = 0,$$

where

$$M_1 = -\alpha a u_f^{\alpha-1} + \beta b u_f^{\beta-1} + \frac{\lambda}{2} u_f^\alpha (1+\alpha) - \frac{\lambda}{2} \alpha u_g^{2\alpha-1} (P+Q).$$

Observe that M_1 satisfies (HM). Now, using that $P, Q \le \lambda\psi$, we obtain that

$$M_1 \ge -\alpha a u_{T_\lambda}^{\alpha-1} + \beta b u_{T_\lambda}^{\beta-1} + \frac{\lambda}{2} u_{T_\lambda}^\alpha (1+\alpha) - \lambda^2 \alpha u_0^{2\alpha-1} \psi,$$

and so we can prove the existence of a function N satisfying (HM) and $\lambda_0 > 0$ such that for $\lambda \le \lambda_0$

$$M_1 \ge N \qquad \text{in } \Omega \text{ and} \qquad \sigma_1(-\Delta + N) > 0. \tag{5.19}$$

Hence, the previous equation has the unique solution $(P - Q) \equiv 0$.

To show the Lipschitzian character of the map $(f,g) \mapsto P_{f,g}$, let $P_{f,g}$ be the solution of problem (5.18) satisfying (5.17). Denote $\overline{q} = P_{\overline{f},\overline{g}}$ and $q = P_{f,g}$. Then, some calculus gives

$$\begin{aligned} &(-\Delta + M(x))(q - \overline{q}) = R_{f,\overline{f},g,\overline{g}} := \alpha a q (u_f^{\alpha-1} - u_{\overline{f}}^{\alpha-1}) \\ &-\beta b q (u_f^{\beta-1} - u_{\overline{f}}^{\beta-1}) - \tfrac{\lambda}{2}(\alpha+1) q (u_f^\alpha - u_{\overline{f}}^\alpha) + \tfrac{\lambda}{2}(u_g - u_{\overline{g}}) + \tfrac{\lambda}{2} \alpha q^2 (u_g^{2\alpha-1} - u_{\overline{g}}^{2\alpha-1}), \end{aligned}$$

where

$$M = -\alpha a u_{\overline{f}}^{\alpha-1} + \beta b u_{\overline{f}}^{\beta-1} + \frac{\lambda}{2}(\alpha+1)u_{\overline{f}}^{\alpha} - \frac{\lambda}{2}\alpha(q+\overline{q})u_{\overline{g}}^{2\alpha-1}.$$

As in the proof of (5.19), we can prove the existence of a function N satisfying (HM) such that for $M \geq N$ in Ω and $\sigma_1(-\Delta + N) > 0$ for small λ. Thus, by Theorem 2.5 we get that

$$\|q - \overline{q}\|_\infty = ||P_{f,g} - P_{\overline{f},\overline{g}}||_\infty \leq \|P_{f,g} - P_{\overline{f},\overline{g}}\|_{C^1(\overline{\Omega})} \leq C\|R_{f,\overline{f},g,\overline{g}}\|_\infty.$$

Now, using a similar argument to the one used in the proof of Theorem 4.6 to obtain a bound of $\|T(f,g)\|_\infty$, we have

$$\left\|\alpha a q(u_f^{\alpha-1} - u_{\overline{f}}^{\alpha-1}) + \beta b q(u_f^{\beta-1} - u_{\overline{f}}^{\beta-1}) + \tfrac{\lambda}{2}(\alpha+1)q\left(u_f^{\alpha} - u_{\overline{f}}^{\alpha}\right)\right\|_\infty \leq C\|f - \overline{f}\|_\infty,$$
$$\left\|\tfrac{\lambda}{2}(u_g - u_{\overline{g}}) + \tfrac{\lambda}{2}\alpha q^2\left(u_g^{2\alpha-1} - u_{\overline{g}}^{2\alpha-1}\right)\right\|_\infty \leq C\|g - \overline{g}\|_\infty,$$

and so $\|R_{f,\overline{f},g,\overline{g}}\|_\infty \leq C\left\{\|f - \overline{f}\|_\infty + \|g - \overline{g}\|_\infty\right\}$. Finally, we have

$$\|P_{f,g} - P_{\overline{f},\overline{g}}\|_\infty \leq C\left\{\|f - \overline{f}\|_\infty + \|g - \overline{g}\|_\infty\right\} \tag{5.20}$$

for a convenient positive constant C. □

THEOREM 5.5. *Assume* (H1). *There exists a positive constant* Λ_4 *such that if* $\lambda \leq \Lambda_4$, *then the system* (5.14)–(5.15) *has a unique solution.*

Proof. The main idea is simple. We will use the Lipschitzian character of the solutions of system (5.14)–(5.15) with respect to the controls and an argument similar to the one used in Theorem 5.2.

Suppose (u_i, v_i, p_i, q_i), for $i = 1, 2$, are two solutions of system (5.14)–(5.15). We define, for $i = 1, 2$,

$$f_i = \frac{\lambda}{2}[u_i - u_i^{\alpha} p_i], \qquad g_i = \frac{\lambda}{2}[v_i - v_i^{\alpha} q_i]. \tag{5.21}$$

Now, taking into account the previous notation, we have for $i = 1, 2$,

$$u_i = u_{f_i}, \quad v_i = u_{g_i}, \quad p_i = P_{g_i,f_i}, \quad q_i = P_{f_i,g_i}$$

and

$$f_i = \frac{\lambda}{2}[u_{f_i} - u_{f_i}^{\alpha} P_{g_i,f_i}], \qquad g_i = \frac{\lambda}{2}[u_{g_i} - u_{g_i}^{\alpha} P_{f_i,g_i}].$$

We know (recall Theorem 4.6 and Lemma 5.4) that the operator $T : [0, \lambda\mathcal{K}] \times [0, \lambda\mathcal{K}] \to L^\infty(\Omega) \times L^\infty(\Omega)$, defined as

$$T(f,g) = \left(\frac{\lambda}{2}[u_f - u_f^{\alpha} P_{g,f}], \frac{\lambda}{2}[u_g - u_g^{\alpha} P_{f,g}]\right),$$

is Lipschitz continuous, with constant $\lambda C/2$, where $C > 0$ is the Lipschitz constant of map $(f,g) \mapsto \left(u_f - u_f^{\alpha} P_{f,g}, u_g - u_g^{\alpha} P_{g,f}\right)$ and verifies $T(f_i, g_i) = (f_i, g_i)$. Therefore, by taking $\lambda < \min\{\Lambda_1, \frac{2}{C}\}$, T is a contraction and consequently has an unique fixed point. So, $(f_1, g_1) = (f_2, g_2)$. Then, we have $u_1 = u_2$, $v_1 = v_2$, and finally $p_1 = p_2$ and $q_1 = q_2$. □

REFERENCES

[1] D.G. Aronson and L.A. Peletier, *Large time behaviour of solutions of the porous medium equation in bounded domains,* J. Differential Equations, 39 (1981), pp. 378–412.

[2] C. Bandle, M.A. Pozio, and A. Tesei, *The asymptotic behavior of the solutions of degenerate parabolic equations*, Trans. Amer. Math. Soc., 303 (1987), pp. 487–501.

[3] H. Berestycki, L. Nirenberg, and S.R.S. Varadhan, *The principal eigenvalue and maximum principle for second order elliptic operators in general domains,* Comm. Pure Appl. Math., 47 (1994), pp. 47–92.

[4] M. Bertsch and R. Rostamian, *The principle of linearized stability for a class of degenerate diffusion equations,* J. Differential Equations, 57 (1985), pp. 373–405.

[5] A. Cañada and J.L. Gámez, *Existence of solutions for some semilinear degenerate elliptic systems with applications to populations dynamics*, Differential Equations Dynam. Systems, 3 (1995), pp. 189–204.

[6] A. Cañada, J.L. Gámez, and J.A. Montero, *Study of an optimal control problem for diffusive nonlinear elliptic equations of logistic type,* SIAM J. Control Optim., 36 (1998), pp. 1171–1189.

[7] M. Delgado, J.A. Montero, and A. Suárez, *Optimal control for the degenerate elliptic logistic equation,* Appl. Math. Optim., 45 (2002), pp. 325–345.

[8] M. Delgado and A. Suárez, *On the existence of dead cores for degenerate Lotka-Volterra models,* Proc. Roy. Soc. Edinburgh Sect. A, 130 (2000), pp. 743–766.

[9] M. Delgado and A. Suárez, *On the structure of the positive solutions of the logistic equation with nonlinear diffusion*, J. Math. Anal. Appl., 268 (2002), pp. 200–216.

[10] J.I. Díaz , *Nonlinear Partial Differential Equations and Free Boundaries. Vol.* I. *Elliptic Equations*, Pitman, Boston, 1985.

[11] M.E. Gurtin and R.C. MacCamy, *On the diffusion of biological populations,* Math. Biosci., 33 (1977), pp. 35–49.

[12] J. Hernández, F. Mancebo, and J.M. Vega de Prada, *On the linearization of some singular nonlinear elliptic problems and applications,* Ann. Inst. H. Poincaré Anal. Non Linéaire, 19 (2002), pp. 777–813.

[13] A. Kufner, *Weighted Sobolev Spaces,* Teubner-Texte Math. 31, Teubner, Leipzig, Germany, 1980.

[14] S. Lenhart, V. Protopopescu, and S. Stojanovic, *A Two-Sided Game for Nonlocal Competitive Systems with Control on Source Terms,* IMA Vol. Math. Appl. 53, Springer, New York, 1993, pp. 135–152.

[15] A.W. Leung, *Optimal harvesting-coefficient control of steady-state prey-predator diffusive Volterra–Lotka systems,* Appl. Math. Optim., 31 (1995), pp. 219–241.

[16] A.W. Leung, *Positive solutions for systems of PDE and optimal control,* Nonlinear Anal., 47 (2001), pp. 1345–1356.

[17] A.W. Leung and S. Stojanovic, *Optimal control for elliptic Volterra-Lotka type equations,* J. Math. Anal. Appl., 173 (1993), pp. 603–619.

[18] J.A. Montero, *A uniqueness result for an optimal control problem on a diffusive elliptic Volterra-Lotka type equation,* J. Math. Anal. Appl., 243 (2000), pp. 13–31.

[19] M.A. Pozio and A. Tesei, *Support properties of solutions for a class of degenerate parabolic problems*, Comm. Partial Differential Equations, 12 (1987), pp. 47–75.

[20] S. Stojanovic, *Modeling and minimization of extinction in Volterra-Lotka type equations with free boundaries*, J. Differential Equations, 134 (1997), pp. 320–342.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1578–1603

CONDITIONAL MOMENT GENERATING FUNCTIONS FOR INTEGRALS AND STOCHASTIC INTEGRALS*

C. D. CHARALAMBOUS†, R. J. ELLIOTT‡, AND V. KRISHNAMURTHY§

Abstract. In this paper we present two methods for computing filtered estimates for moments of integrals and stochastic integrals of continuous-time nonlinear systems. The first method utilizes recursive stochastic partial differential equations. The second method utilizes conditional moment generating functions. An application of these methods leads to the discovery of new classes of finite-dimensional filters. For the case of Gaussian systems the recursive computations involve integrations with respect to Gaussian densities, while the moment generating functions involve differentiations of parameter dependent ordinary stochastic differential equations. These filters can be used in Volterra or Wiener chaos expansions and the expectation-maximization algorithm. The latter yields maximum-likelihood estimates for identifying parameters in state space models.

Key words. moment generating functions, finite-dimensional, filters, recursions, expectation-maximization

AMS subject classifications. 93E11, 93E12, 93E10, 60G35

DOI. S036301299833327X

1. Introduction.

Conditional expectations of functionals of systems state processes given noisy observations require, in general, infinite-dimensional computations. To determine whether such conditional expectations are finite-dimensional, it is of interest to derive representations of the conditional distribution.

This paper discusses the following problem. We are given noisy observations $\{y_s; 0 \le s \le t\}$ of the system state process $\{x_s; 0 \le s \le t\}$, and we wish to derive filtered estimates for moments of integrals and stochastic integrals. The underlying mathematical system model can be diverse; for example, it includes continuous-time processes, discrete-time processes, jump point processes, or a combination of these processes. In this paper we focus our attention on continuous-time processes.

Here, our system state process $\{x_s; 0 \le s \le t\}$ and observation process $\{y_s; 0 \le s \le t\}$ are solutions of the Itô stochastic differential equations

$$dx_t = f(t, x_t)dt + \sigma(t, x_t)dw_t, \qquad x(0) \in \mathbb{R}^n, \tag{1.1}$$

$$dy_t = h(t, x_t)dt + \alpha_t dw_t + N_t^{1/2} db_t, \qquad y(0) = 0 \in \mathbb{R}^n, \tag{1.2}$$

in which $\{w_s; 0 \le s \le t\}$ and $\{b_s; 0 \le s \le t\}$, are, respectively, m-dimensional and d-dimensional, independent standard Wiener processes; $x(0)$ is a random variable

*Received by the editors January 28, 1998; accepted for publication (in revised form) April 15, 2003; published electronically November 6, 2003.

http://www.siam.org/journals/sicon/42-5/33327.html

†Department of Electrical and and Computer Engineering, University of Cyprus, 75 Kallipoleos St., P.O. Box 20537, 1678 Nicosia, Cyprus (chadcha@ucy.ac.cy) and School of Information Technology and Engineering, University of Ottawa, 161 Louis Pasteur, Ottawa, ON, Canada, K1N 6N5 (chadcha@site.uottawa.ca), and Adjunct Professor with the Department of Electrical and Computer Engineering, McGill University, Montréal, QB, Canada H3A 2A7. This author's work was supported by the Natural Science and Engineering Research Council of Canada.

‡Haskayne School of Business, University of Calgary, 2500 University Drive NW, Calgary, AB, Canada T2N 1N4 (relliott@ucalgary.ca).

§Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, V6T 1Z4, Canada and Department of Electrical and Electronic Engineering, University of Melbourne, Melbourne, Victoria 3052, Australia (vikramk@ece.ubc.ca).

independent of the Wiener processes. The precise assumptions on the coefficients of our model are stated in section 2.

We are interested in conditional expectations (filtered estimates) of moments of integrals and stochastic integrals

$$L_{0,t}^{\kappa,1} = \left(\int_0^t f^1(s, x_s) ds \right)^\kappa, \quad L_{0,t}^{\kappa,2} = \left(\int_0^t f^2(s, x_s) dw_s \right)^\kappa, \quad L_{0,t}^{\kappa,3} = \left(\int_0^t f^3(s, x_s) db_s \right)^\kappa, \quad \kappa \geq 1, \tag{1.3}$$

for Borel measurable functions $f^1 : [0,T] \times \mathbb{R}^n \to \mathbb{R}$, $f^2 : [0,T] \times \mathbb{R}^n \to (\mathbb{R}^m)'$, $f^3 : [0,T] \times \mathbb{R}^n \to (\mathbb{R}^d)'$, which are continuous in t. Aside from their mathematical value, these estimates are important, for example, in least-squares estimation/filtering, Volterra series expansions of nonlinear realization theory [1], Wiener chaos expansions (of nonlinear filtering) [2], and maximum likelihood estimation through the expectation-maximization (EM) algorithm [3]. For the case $\kappa = 1$, these estimates are important in estimating parameters, a problem which arises in many disciplines, such as signal processing, communications, and control systems.

The first method, Theorem 3.1, utilizes a system of stochastic partial differential equations (SPDEs) that enable us to compute the above estimates recursively. The second method, Theorem 4.5, utilizes conditional moment generating functions for $L_{0,t}^{1,j}, j = 1,2,3$. That is, for a test function $\Phi : \mathbb{R}^n \to \mathbb{R}$, we use measure-valued conditional moment generating functions

$$\widetilde{\beta}_t^{\theta,j}(\Phi) = \widetilde{\mathrm{E}}[\Phi(x_t) \exp\left(\theta L_{0,t}^{1,j}\right) | \mathcal{F}_{0,t}^y], \qquad j = 1,2,3, \qquad \theta = i\omega, \quad i = \sqrt{-1}. \tag{1.4}$$

Therefore, when the unnormalized versions of $\widetilde{\beta}_t^{\theta,j}(\Phi)$ have densities $\beta^{\theta,j}(x,t), j = 1,2,3$, the latter satisfy linear SPDEs. The computation of filtered estimates of moments (1.3) are obtained by simply differentiating the conditional densities with respect to the parameter θ.

For the case of Gaussian system models (i.e., $dx_t = Fx_t dt + Gw_t$, $dy_t = Hx_t dt + N^{\frac{1}{2}} b_t$), we derive filtered estimates for

$$L_{0,t}^{1,1} = \int_0^t x_s' Q x_s ds \quad L_{0,t}^{1,2} = \int_0^t x_s' R dw_s, \quad L_{0,t}^{1,3} = \int_0^t x_s' S db_s. \tag{1.5}$$

Each filtered estimate is propagated by four statistics. Two of these are the conditional mean and error covariances of x_t given $\{y_s; 0 \leq s \leq t\}$ (Kalman filter), while the remaining two are modified versions of the Kalman filter; the latter are driven by the conditional mean and error covariance of the Kalman filter.

In the past, the computation of these filtered estimates was confined to integrals $L_{0,t}^{1,1}$, which are obtained using smoothing operations (e.g., [4]), and certain Lie algebraic techniques applied to Volterra expansions (e.g., [1]). However, for analogous discrete-time systems the filtering estimates in (1.5) are obtained using smoothing operations (e.g., [5]). Recently, conditional expectations for the items in (1.5) were obtained using filtering operations in [6]; the estimates were propagated by five statistics. The techniques in [6], which are different from ours, are only applicable to Gaussian systems, and they are confined to $\kappa = 1$.

2. The Duncan–Mortensen–Zakai (DMZ) equation.

Notation 2.1.

1. "$\prime$" denotes transposition of a matrix;
2. I_k denotes $k \times k$ identity matrices;
3. $(\cdot)_i$ denotes the ith component of a vector and $(\cdot)_{i,j}$ denotes the ijth component of a matrix;
4. $\mathcal{L}(V_1; V_2)$ denotes the space of linear transformations of a vector space V_1 into a vector space V_2;
5. $C^{p,q}_{x,t}(\mathbb{R}^n \times [0,T]) = \{\Phi : \mathbb{R}^n \times [0,T] \to \mathbb{R}^n; \Phi(\cdot,t)$ is "p" times continuously differentiable in "x," and $\Phi(x,\cdot)$ is "q" times continuously differentiable in "t"$\}$;
6. $D_x = [\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \dots, \frac{\partial}{\partial x_n}]', \quad D^2_x = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2}, & \cdots, & \frac{\partial^2}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1}, & \cdots, & \frac{\partial^2}{\partial x_n^2} \end{bmatrix}$;
7. $\Phi : \mathbb{R}^n \to \mathbb{R}$ denotes an arbitrary test function which is $C^2_x(\mathbb{R}^n)$ and has compact support;
8. $\mathrm{E}, \widetilde{\mathrm{E}}$ denote expectations with respect to measures $P, \widetilde{P}$, respectively;
9. $N_t \doteq N_t^{\frac{1}{2}} N_t^{\frac{1}{2}\prime}$.

Assumption 2.2.

1. $f : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n, \sigma : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n, h : [0,T] \times \mathbb{R}^n \to \mathbb{R}^d, T > 0$, are bounded Borel measurable functions;
2. $N : [0,T] \to \mathcal{L}(\mathbb{R}^d; \mathbb{R}^d), \alpha : [0,T] \to \mathcal{L}(\mathbb{R}^n; \mathbb{R}^d), N, \alpha$ are bounded Borel measurable functions, and $\exists \beta_1 > 0, \beta_2 > 0$ such that $N_t \geq \beta_1 I_d \ \forall t \in [0,T]$, $a(t,x) \doteq \sigma(t,x)\sigma(t,x)' \geq \beta_2 I_n \ \forall (t,x) \in [0,T] \times \mathbb{R}^n$;
3. σ is continuous in x, uniformly on compact subsets of $[0,T] \times \mathbb{R}^n$, $\frac{\partial}{\partial x_i}\sigma_{i,j}$ is a bounded measurable function of $(t,x) \in [0,T] \times \mathbb{R}^n$, $1 \leq i,j \leq n$;
4. $|f(t,x) - f(t,z)| + \|\sigma(t,x) - \sigma(t,z)\| \leq k|x - z|$;
5. $x(0)$ has distribution $\Pi_0(dx) = p_0(x)dx$, where $p_0(\cdot) \in L^2(\mathbb{R}^n)$.

The above assumptions, with the exception of statement 4, are assumed to hold throughout the manuscript.

Next, we start with a reference probability measure which is important in deriving certain conditional densities for the filtering problem discussed earlier. Let $(\Omega, \mathcal{F}, P)$ be a reference probability with complete filtration $\{\mathcal{F}_{0,t}; t \in [0,T]\}$, on which we have the following:

(a) $w : [0,T] \times \Omega \to \mathbb{R}^n, b : [0,T] \times \Omega \to \mathbb{R}^d$, which are $\{\mathcal{F}_{0,t}; t \in [0,T]\}$ adapted independent Wiener processes;

(b) $x(0) : \Omega \to \mathbb{R}^n$, an $\mathcal{F}_{0,0}$-measurable random variable, which is independent of $\{w_t, b_t; t \in [0,T]\}$;

(c) processes $\{x_t; t \in [0,T]\}, \{y_t; t \in [0,T]\}$, which (in view of Assumption 2.2) are unique and continuous solutions of the stochastic differential equations

$$dx_t = f(t,x_t)dt - \sigma(t,x_t)\alpha_t' C_t^{-1} h(t,x_t)dt + \sigma(t,x_t)dw_t, \qquad x(0) \in \mathbb{R}^n, \tag{2.1}$$

$$dy_t = \alpha_t dw_t + N_t^{1/2} db_t, \qquad y(0) = 0 \in \mathbb{R}^n, \tag{2.2}$$

where

$$C_t \doteq \alpha_t \alpha_t' + N_t. \tag{2.3}$$

Consider the P-martingale

$$m_t = \int_0^t h'(s,x_s) C_s^{-1} dy_s, \tag{2.4}$$

and introduce the exponential martingale

$$\varepsilon(m_t) = \exp\left(m_t - \frac{1}{2}\langle m, m\rangle_t\right) = \Lambda_{0,t}, \tag{2.5}$$

where $\langle m, m\rangle_t = \int_0^t |C_s^{-1/2} h(s, x_s)|^2 ds$ is the quadratic variation of $\{m_t; t \in [0,T]\}$. By Assumption 2.2, we have $\mathrm{E}[\Lambda_{0,t}] = 1 \forall t \in [0,T]$ (see [7]). Consequently, we define a measure $\widetilde{P}$ through the Radon–Nikodým derivative

$$\Lambda_{0,T} \doteq \mathrm{E}\left[\frac{d\widetilde{P}}{dP}|\mathcal{F}_{0,T}\right] = \varepsilon(m_T). \tag{2.6}$$

Since $\widetilde{P}(\Omega) = \int_\Omega \Lambda_{0,t}(\omega) dP(\omega) = 1 \ \forall t \in [0,T]$, the Girsanov theorem (see [7]) states that $\widetilde{P}$ is a probability measure on $(\Omega, \mathcal{A})$ and that

$$\begin{aligned} \begin{bmatrix} \overline{w}_t \\ \overline{b}_t \end{bmatrix} &= \begin{bmatrix} w_t \\ b_t \end{bmatrix} - \begin{bmatrix} \langle w, m\rangle_t \\ \langle b, m\rangle_t \end{bmatrix} \\ &= \begin{bmatrix} w_t \\ b_t \end{bmatrix} - \begin{bmatrix} \int_0^t \alpha_s' C_s^{-1} h(s, x_s) ds \\ \int_0^t N_s^{1/2,\prime} C_s^{-1} h(s, x_s) ds \end{bmatrix} \end{aligned} \tag{2.7}$$

are independent Wiener processes on $(\Omega, \mathcal{F}, \widetilde{P}; \mathcal{F}_{0,t})$. Substituting (2.7) into (2.1), (2.2), on the new probability space $(\Omega, \mathcal{F}, \widetilde{P}; \mathcal{F}_{0,t})$ we have constructed (weak) solutions $\{x_t; t \in [0,T]\}, \{y_t; t \in [0,T]\}$ of the stochastic equations

$$dx_t = f(t, x_t)dt + \sigma(t, x_t) d\overline{w}_t, \qquad x(0) \in \mathbb{R}^n, \tag{2.8}$$

$$dy_t = h(t, x_t)dt + \alpha_t d\overline{w}_t + N_t^{1/2} d\overline{b}_t, \qquad y(0) = 0 \in \mathbb{R}^d. \tag{2.9}$$

Since $\{\overline{w}_t; t \in [0,T]\}$ and $\{\overline{b}_t; t \in [0,T]\}$ are versions of Wiener processes (which are independent), (2.8), (2.9) constitute our original system model (simply by letting $\overline{w} \to w, \overline{b} \to b$). Note that we may remove the Lipschitz condition Assumption 2.2, statement 4, and employ the martingale approach to construct weak solutions.

Notation 2.3.

1. $\{\mathcal{F}_{0,t}^y; t \in [0,T]\}$ denotes the complete filtration generated by the observations σ-algebra $\sigma\{y_\tau; 0 \le \tau \le t\}$, $\{\mathcal{F}_{0,t}^w; t \in [0,T]\}$ denotes that of $\sigma\{w_\tau; 0 \le \tau \le t\}$, and $\mathcal{F}^{x(0)} = \sigma\{x(0)\}$;
2. The measure-valued process $q_t(\Phi) = \mathrm{E}[\Phi(x_t)\Lambda_{0,t}|\mathcal{F}_{0,t}^y]$ is well defined.

The problem of least-squares filtering is concerned with estimating the conditional mean of x_t given the past and present measurements, i.e., $\mathcal{F}_{0,t}^y$. Thus, the least-squares filtering can be cast in terms of computing conditional expectations $\widetilde{\mathrm{E}}[\Phi(x_t)|\mathcal{F}_{0,t}^y]$.

Lemma 2.4.

1. *A version of Bayes's formula yields*

$$\widetilde{\mathrm{E}}[\Phi(x_t)|\mathcal{F}_{0,t}^y] = \frac{\mathrm{E}[\Phi(x_t)\frac{d\widetilde{P}}{dP}|\mathcal{F}_{0,t}^y]}{\mathrm{E}[\frac{d\widetilde{P}}{dP}|\mathcal{F}_{0,t}^y]} = \frac{q_t(\Phi)}{q_t(1)}. \tag{2.10}$$

2. *If the measure-valued process $q_t(\Phi)$ has an $\mathcal{F}_{0,t}^y$-measurable density function $q : \mathbb{R}^n \times [0,T] \times \Omega \to \mathbb{R}$, then*

$$\widetilde{\mathrm{E}}[\Phi(x_t)|\mathcal{F}_{0,t}^y] = \frac{\int_{\mathbb{R}^n} \Phi(z) q(z,t) dz}{\int_{\mathbb{R}^n} q(z,t) dz}. \tag{2.11}$$

Proof. 1. A version of Bayes' rule yields the equality in (2.10).

2. The proof follows from the existence of the density $q(\cdot)$. □

The existence of the density $q(x,t)$ will follow from the existence and uniqueness of solutions of SPDEs [8, 9, 10], as it will be shown shortly.

We now derive an evolution equation for $q(\cdot)$. Note that $\{\Lambda_{0,t}; t \in [0,T]\}$ is a solution of the stochastic differential equation

$$\Lambda_{0,t} = 1 + \int_0^t \Lambda_{0,s} h'(s,x_s) C_s^{-1} dy_s. \tag{2.12}$$

THEOREM 2.5. *The unnormalized density of the conditional distribution* $\widetilde{P}(x_t \in A|\mathcal{F}_{0,t}^y), A \in \mathcal{B}(\mathbb{R}^n)$ *is* $q(\cdot)$ *and satisfies the SPDE*

$$dq(z,t) = A(t)^* q(z,t)dt + B(t)^* q(z,t)dy_t, \qquad (z,t) \in (0,T] \times \mathbb{R}^n, \tag{2.13}$$

$$q(z,0) = p_0(z), \qquad z \in \mathbb{R}^n, \tag{2.14}$$

where

$$A(t)^*\Phi(x) = \frac{1}{2}\sum_{i,j=1}^n \left(\frac{\partial^2}{\partial x_i \partial x_j} \left((\sigma(t,x)\sigma'(t,x))_{i,j} \Phi(x) \right) \right) - \sum_{i=1}^n \frac{\partial}{\partial x_i}\big(f_i(t,x)\Phi(x)\big), \tag{2.15}$$

$$B_k(t)^*\Phi(x) = \sum_{i=1}^d \left(C_t^{-1}\right)_{i,k} h_i(t,x)\Phi(x) - \sum_{i=1}^n \frac{\partial}{\partial x_i}\left(\left(\sigma(t,x)\alpha_t' C_t^{-1}\right)_{i,k}\Phi(x)\right). \tag{2.16}$$

Proof. Recall that under $P, \{x_t, y_t; t \in [0,T]\}$ are solutions of (2.1), (2.2). Define

$$D_t \doteq I_m - \alpha_t' C_t^{-1} \alpha_t, \tag{2.17}$$

and introduce

$$\widetilde{y}_t = \int_0^t C_s^{-1/2} dy_s, \qquad \widetilde{w}_t = \int_0^t D_s^{-1/2}(dw_s - \alpha_s' C_s^{-1} dy_s). \tag{2.18}$$

Substituting into (2.1) we have

$$\begin{aligned} dx_t = {} & \big(f(t,x_t) - \sigma(t,x_t)\alpha_t' C_t^{-1} h(t,x_t)\big)dt + \sigma(t,x_t) D_t^{1/2} d\widetilde{w}_t \\ & + \sigma(t,x_t)\alpha_t' C_t^{-1/2} d\widetilde{y}_t, \qquad x(0) \in \mathbb{R}^n. \end{aligned} \tag{2.19}$$

Moreover, $\{\widetilde{y}_t; t \in [0,T]\}$ and $\{\widetilde{w}_t; t \in [0,T]\}$ are independent standard Wiener processes, and $\mathcal{F}_{0,t}^y = \widetilde{\mathcal{F}}_{0,t}^y$; that is, no information is gained or lost. By (2.12), (2.18) we deduce

$$\Lambda_{0,t} = 1 + \int_0^t \Lambda_{0,s} h'(s,x_s) C_s^{-1/2} d\widetilde{y}_s. \tag{2.20}$$

By the Itô product rule

$$\begin{aligned} \Phi(x_t)\Lambda_{0,t} = {} & \Phi(x(0)) + \int_0^t \Phi(x_s) d\Lambda_{0,s} + \int_0^t d\Phi(x_s)\Lambda_{0,s} \\ & + \int_0^t d\langle \Phi(x), \Lambda \rangle_s. \end{aligned} \tag{2.21}$$

Since

$$\begin{aligned}\Phi(x_t) &= \Phi(x(0)) + \frac{1}{2}\int_0^t \mathrm{Tr}\big(\sigma(s,x_s)D_s'\sigma'(s,x_s)D_x^2\Phi(x_s)\big)ds \\ &\quad + \frac{1}{2}\mathrm{Tr}\big(\sigma(s,x_s)\alpha_s'C_s^{-1}\alpha_s\sigma'(s,x_s)D_x^2\Phi(x_s)\big)ds \\ &\quad + \int_0^t D_x'\Phi(x_s)\big(f(s,x_s) - \sigma(s,x_s)\alpha_s'C_s^{-1}h(s,x_s)\big)ds \\ &\quad + \int_0^t D_x'\Phi(x_s)\sigma(s,x_s)D_s^{1/2}d\widetilde{w}_s + \int_0^t D_x'\Phi(x_s)\sigma(s,x_s)\alpha_s'C_s^{-1/2}d\widetilde{y}_s, \\ \langle\Phi(x),\Lambda\rangle_t &= \int_0^t \Lambda_{0,s}D_x'\Phi(x_s)\sigma(s,x_s)\alpha_s'C_s^{-1}h(s,x_s)ds,\end{aligned}$$

substituting into (2.21) we have

$$\begin{aligned}\Phi(x_t)\Lambda_{0,t} &= \Phi(x(0)) + \frac{1}{2}\int_0^t \Lambda_{0,s}\mathrm{Tr}\big(\sigma(s,x_s)D_s\sigma'(s,x_s)D_x^2\Phi(x_s)\big)ds \\ &\quad + \frac{1}{2}\int_0^t \Lambda_{0,s}\mathrm{Tr}\big(\sigma(s,x_s)\alpha_s'C_s^{-1}\alpha_s\sigma'(s,x_s)D_x^2\Phi(x_s)\big)ds \\ &\quad + \int_0^t \Lambda_{0,s}D_x'\Phi(x_s)\big(f(s,x_s) - \sigma(s,x_s)\alpha_s'C_s^{-1}h(s,x_s)\big)ds \\ &\quad + \int_0^t \Lambda_{0,s}D_x'\Phi(x_s)\sigma(s,x_s)D_s^{1/2}d\widetilde{w}_s + \int_0^t \Lambda_{0,s}\Phi(x_s)h'(s,x_s)C_s^{-1/2}d\widetilde{y}_s \\ &\quad + \int_0^t \Lambda_{0,s}D_x'\Phi(x_s)\sigma(s,x_s)\alpha_s'C_s^{1/2}d\widetilde{y}_s \\ &\quad + \int_0^t \Lambda_{0,s}D_x'\Phi(x_s)\sigma(s,x_s)\alpha_s'C_s^{-1}h(s,x_s)ds.\end{aligned} \tag{2.22}$$

Conditioning each side of (2.22) on $\mathcal{F}_{0,t}^y$ and then using the mutual independence of $x(0), \{\widetilde{w}_t; t \in [0,T]\}, \{\widetilde{y}_t; t \in [0,T]\}$ (see [11]) and a version of Fubinis theorem [7, 12], we conclude that

$$q_t(\Phi) = q_0(\Phi) + \int_0^t q_s(A(s)\Phi(x))ds + \int_0^t q_s(B(s)\Phi(x))C_s^{-1/2}d\widetilde{y}_s. \tag{2.23}$$

Integrating each term by parts and then substituting $\widetilde{y}_t = \int_0^t C_s^{-1/2}dy_s$ we obtain (2.13), (2.14). □

Next, we employ certain results of variational methods of partial differential equations to show existence and uniqueness of solutions to (2.13) and (2.14).

Introduce the space $H(\mathbb{R}^n) = L^2(\mathbb{R})$ and the Sobolev space $H^1(\mathbb{R}^n)$ defined by

$$H^1(\mathbb{R}^n) = \left\{u \in L^2(\mathbb{R}^n),\ \frac{\partial}{\partial x_i}u \in L^2(\mathbb{R}^n), 1 \le i \le n\right\}.$$

Furnish $H(\mathbb{R}^n), H^1(\mathbb{R}^n)$ with the norm topologies

$$\|u\|_H = \int_{\mathbb{R}^n} |u|^2 dx,\ u \in H(\mathbb{R}^n),$$

$$\|u\|_{H^1} = \left\{\int_{\mathbb{R}^n} |u|^2 dx + \sum_{i=1}^n \int \left|\frac{\partial}{\partial x_i}u\right|^2 dx\right\}^{1/2},\ u \in H^1(\mathbb{R}^n).$$

$H(\mathbb{R}^n)$ and $H^1(\mathbb{R}^n)$ are Hilbert spaces with scalar products defined by

$$(\phi,\psi)_H = \int_{\mathbb{R}^n} \phi\psi dx,\ \phi,\psi \in H(\mathbb{R}^n),$$
$$(\phi,\psi)_{H^1} = \int_{\mathbb{R}^n} \phi\psi dx + \sum_{i=1}^n \int \frac{\partial\phi}{\partial x_i}\frac{\partial\psi}{\partial x_i}dx = (\phi,\psi)_{L^2(\mathbb{R}^n)} + \sum_{i=1}^n \left(\frac{\partial\phi}{\partial x_i}, \frac{\partial\psi}{\partial x_i}\right)_{L^2(\mathbb{R}^n)},\ \phi,\psi \in H^1(\mathbb{R}^n).$$

Let $H^{-1}(\mathbb{R}^n)$ denote the dual of $H^1(\mathbb{R}^n)$ (the space of continuous linear functionals on $H^1(\mathbb{R}^n)$). The norm of elements of $H^{-1}(\mathbb{R}^n)$ is denoted by $\|\cdot\|_*$, and the duality between $H^1(\mathbb{R}^n)$ and $H^{-1}(\mathbb{R}^n)$ is denoted by $\langle\cdot,\cdot\rangle$.

Let

$$B(\cdot)^*u = \begin{bmatrix} B_1(\cdot)^*u \\ \vdots \\ B_d(\cdot)^*u \end{bmatrix},\ u \in H^1(\mathbb{R}^n),$$

and write the adjoint operators of $A(\cdot)^*$ and $B(\cdot)^*$ as

$$\langle u, A(t)^*v\rangle = \langle\ A(t)u, v\rangle = -\frac{1}{2}\sum_{i,j=1}^n \left(a_{i,j}(t,\cdot)\frac{\partial}{\partial x_i}u, \frac{\partial}{\partial x_j}v\right)_{L^2(\mathbb{R}^n)} + \sum_{i=1}^n \left(\tilde{f}_i(t,\cdot)\frac{\partial}{\partial x_i}u, v\right)_{L^2(\mathbb{R}^n)}, \qquad u,v \in H^1(\mathbb{R}^n),$$

where

$$\tilde{f}_i(t,x) = f_i(t,x) - \frac{1}{2}\sum_{j=1}^n \frac{\partial}{\partial x_j}a_{i,j}(t,x)$$

$$\langle u, B(t)^*v\rangle = \langle\ B(t)u, v\rangle = \sum_{i,k=1}^d ((C_t^{-1})_{i,k}h_i(t,\cdot)u, v)_{L^2(\mathbb{R}^n)} + \sum_{k=1}^d\sum_{i=1}^n \left((\sigma(t,\cdot)\alpha_t'C_t^{-1})_{i,k}\frac{\partial}{\partial x_i}u, v\right)_{L^2(\mathbb{R}^n)}, \qquad u,v \in H^1(\mathbb{R}^n).$$

In view of Assumption 2.2, statements 1, 2, 3, and 5, it can be shown that

$$\begin{aligned} &A(\cdot), A(\cdot)^* \in L^\infty((0,T);\mathcal{L}(H^1(\mathbb{R}^n);H^{-1}(\mathbb{R}^n))),\\ &B(\cdot), B(\cdot)^* \in L^\infty((0,T);\mathcal{L}(H^1(\mathbb{R}^n);(L^2(\mathbb{R}^n))^d)). \end{aligned} \tag{2.24}$$

Moreover, $A(t) \in \mathcal{L}(H^1(\mathbb{R}^n);H^{-1}(\mathbb{R}^n))$, $B(t) \in \mathcal{L}(H^1(\mathbb{R}^n);(L^2(\mathbb{R}^n))^d$ satisfy the following coercivity condition. There exist $\lambda_1,\lambda_2 > 0$ such that

$$-2\langle A(t)u,u\rangle + \lambda_1\|u\|^2_{L^2(\mathbb{R}^n)} \geq \lambda_2\|u\|^2_{H^1(\mathbb{R}^n)} + \|Bu\|^2_{(L^2(\mathbb{R}^n))^d}\ \forall u \in H^1(\mathbb{R}^n),\ \forall t \in [0,T]. \tag{2.25}$$

Define the space

$$L_y^2((0,T);H^1) \doteq \{u \in L^2(\Omega,\mathcal{F},P;\ L^2((0,T);H^1);\ a.e.\ on\ [0,T],\ u(t) \in L^2(\Omega,\mathcal{F}_{0,t}^y,P;H^1)\}.$$

LEMMA 2.6. *There exists one and only one solution* $q(\cdot)$ *of* (2.13), (2.14) *in the space*

$$q(\cdot) \in L_y^2((0,T);H^1) \cap L^2(\Omega,\mathcal{F},P;C((0,T);H)).$$

Proof. Assumption 2.2 statements 1, 2, 3, and 5 imply the coercivity condition (2.25), which is then employed to show existence and uniqueness of solutions to (2.13), (2.14) (see [8, 9, 10]). □

The next tool employed in subsequent sections is the concept of fundamental solutions to stochastic differential equations.

DEFINITION 2.7. *A fundamental solution of* (2.13), (2.14) *is an* $\mathcal{F}_{0,t}^y$*-measurable function* $q(z,t;x,s)$, *with* $(z,x) \in \mathbb{R}^n \times \mathbb{R}^n, 0 \le s < t \le T$, *such that the following hold:*

1. $q(\cdot,\cdot;x,s)$ *is a solution of*

(2.26)
$$dq(z,t;x,s) = A(t)^* q(z,t;x,s)dt + B(t)^* q(z,t;x,s)dy_t, \quad 0 < s < t \le T,$$

(2.27)
$$\lim_{t \downarrow s} q(z,t;x,s) = \delta(z-x).$$

2. *For fixed* $(s,x) \in (0,t) \times \mathbb{R}^n$, $q(\cdot,t;x,s) \in C_z^2(\mathbb{R}^n)$.
3. *For* $\varphi : \mathbb{R}^n \to \mathbb{R}$, *which is continuous with compact support,*

$$\lim_{t \downarrow s} \int_{-\infty}^{\infty} q(z,t;x,s)\varphi(x)dx = \varphi(z). \tag{2.28}$$

That is, $\lim_{t \downarrow s} q(z,t;x,s) = \delta(z-x)$ *is a Dirac delta function.*

Unfortunately, Assumption 2.2 is too weak to imply that $q(\cdot,t;x,s) \in C_z^2(\mathbb{R}^n)$. However, if there is no correlation between the state noise and the observation noise (e.g., $\alpha_t = 0\ \forall t \in [0,T]$), and we impose additional smoothness and continuity conditions on (f,σ,h), then by considering the pathwise version of (2.13), (2.14), it can be shown that for each $y \in C([0,T];\mathbb{R}^d)$ there exists a unique solution, which is a fundamental solution [13]. For the correlated case, we have the following result which is found in [9, 10] (see also [14] for alternative conditions).

THEOREM 2.8. *Suppose the coefficient of* A *and* B_k, $k = 1,\dots,d$ *have bounded partial derivatives in* x *of any order. Then*

1. $\{q(z,t;x,s); 0 \le s < t \le T\}$, $(z,x) \in \mathbb{R}^n \times \mathbb{R}^n$, *is a unique fundamental solution of the unnormalized condition density equation* (2.13), (2.14), *and* $q(\cdot,t;x,s) \in C_b^\infty(\mathbb{R}^n)$, $P-a.s. \forall t \in (s,T]$.

2. *A version of the conditional distribution* $\widetilde{P}(x_t \in A|\mathcal{F}_{0,t}^y), A \in \mathcal{B}(\mathbb{R}^n)$, *is*

$$\widetilde{\mathrm{E}}[\Phi(x_t)|\mathcal{F}_{0,t}^y] = \frac{q_t(\Phi)}{q_t(1)} = \frac{\int_{\mathbb{R}^n \times \mathbb{R}^n} \Phi(z)q(z,t;x,0)p_0(x)dxdz}{\int_{\mathbb{R}^n \times \mathbb{R}^n} q(z,t;x,0)p_0(x)dxdz}. \tag{2.29}$$

Proof. 1. This is shown in [10, pp. 227–228].
2. Let $q(z,t;x,s)$ be a solution of (2.26), (2.27); set $\widetilde{q}(z,t) = \int_{\mathbb{R}^n} q(z,t;x,0)p_0(x)dx$. Then

$$\begin{aligned} d\widetilde{q}(z,t) &= \int_{\mathbb{R}^n} dq(z,t;x,0)p_0(x)dx \\ &= \int_{\mathbb{R}^n} A(t)^* q(z,t;x,0)p_0(x)dxdt + \int_{\mathbb{R}^n} B(t)^* q(z,t;x,0)p_0 dx dy_t \\ &= A(t)^*\widetilde{q}(z,t)dt + B(t)^*\widetilde{q}(z,t)dy_t. \end{aligned}$$

This shows that $\widetilde{q}(z,t)$ satisfies (2.13) for $(z,t) \in \mathbb{R}^n \times (0,T]$. Since $\lim_{t\downarrow 0} \widetilde{q}(z,t) = \lim_{t\downarrow 0} \int_{\mathbb{R}^n} q(z,t;x,0)p_0(x)dx = p_0(z)$, we also have (2.14). By Lemma 2.4 we establish (2.29). □

DEFINITION 2.9. *Let $f^1 : [0,T] \times \mathbb{R}^n \to \mathbb{R}, f^2 : [0,T] \times \mathbb{R}^n \to (\mathbb{R}^n)', f^3 : [0,T] \times \mathbb{R}^n \to (\mathbb{R}^d)'$ be Borel measurable and bounded functions.*

1. *The integrals*

$$\begin{aligned} L^{\kappa,1}_{0,t} &= \left(\int_0^t f^1(s,x_s)ds\right)^\kappa, \quad L^{\kappa,2}_{0,t} = \left(\int_0^t f^2(s,x_s)dw_s\right)^\kappa, \\ L^{\kappa,3}_{0,t} &= \left(\int_0^t f^3(s,x_s)db_s\right)^\kappa, \qquad \kappa \geq 1, \end{aligned} \tag{2.30}$$

are well defined.

2. *The measure-valued processes*

$$M^{\kappa,j}_t(\Phi) = \mathrm{E}[\Phi(x_t)\Lambda_{0,t}L^{\kappa,j}_{0,t}|\mathcal{F}^y_{0,t}], \qquad \kappa \geq 0, \quad j = 1,2,3, \tag{2.31}$$

are well defined.

We are interested in filtered estimates of κth moments ($\kappa \geq 1$) of integrals and stochastic integrals. That is, we wish to derive expressions for $\widetilde{\mathrm{E}}[L^{\kappa,j}_{0,t}|\mathcal{F}^y_{0,t}]$. An application of Bayes's theorem yields

$$\widetilde{\mathrm{E}}[L^{\kappa,j}_{0,t}|\mathcal{F}^y_{0,t}] = \frac{\mathrm{E}[\Lambda_{0,t}L^{\kappa,j}_{0,t}|\mathcal{F}^y_{0,t}]}{\mathrm{E}[\Lambda_{0,t}|\mathcal{F}^y_{0,t}]}, \qquad \kappa \geq 1, \quad j = 1,2,3. \tag{2.32}$$

3. Recursive equations. Here we prove that the filtered estimates (2.32) can be expressed in terms of the fundamental solution of the DMZ equation; namely, $q(z,t;x,s), 0 \leq s < t \leq T$, which satisfies (2.13), (2.14). This enables us to conclude that if $q(z,t;x,s)$ is a finite-dimensional statistic, then these filtered estimates can be described in terms of solutions of a finite-number of stochastic differential equations.

THEOREM 3.1. *Suppose $M^{\kappa,j}_t(\cdot)$ have $\mathcal{F}^y_{0,t}$-measurable density functions $M^{\kappa,j} : \mathbb{R}^n \times [0,T] \times \Omega \to \mathbb{R}, j = 1,2,3$.*

Then

$$M^{\kappa,j}(x,t)dx = \mathrm{E}[I_{x_t \in dx}\Lambda_{0,t}L^{\kappa,j}_{0,t}|\mathcal{F}^y_{0,t}], \qquad \kappa \geq 1, \quad j = 1,2,3, \tag{3.1}$$

satisfy the following recursive system of SPDEs:

$$
\begin{aligned}
dM^{\kappa,1}(x,t) &= A(t)^* M^{\kappa,1}(x,t)dt + B(t)^* M^{\kappa,1}(x,t)dy_t \\
&\quad + \kappa f^1(t,x) M^{k-1,1}(x,t)dt, \qquad \kappa \geq 1, \quad (t,x) \,\epsilon\, (0,T] \times \mathbb{R}^n, \qquad (3.2)\\
dM^{\kappa,2}(x,t) &= A(t)^* M^{\kappa,2}(x,t)dt + B(t)^* M^{\kappa,2}(x,t)dy_t \\
&\quad + \frac{1}{2}\kappa(\kappa-1)|f^{2,\prime}(t,x)|^2 M^{\kappa-2,2}(x,t)dt \\
&\quad - \kappa \sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(M^{\kappa-1,2}(x,t)\left(\sigma(t,x) f^{2,\prime}(x,t)\right)_i\right) dt \\
&\quad + \kappa f^2(t,x) M^{\kappa-1,2}(x,t)\alpha_t' C_t^{-1} dy_t, \qquad \kappa \geq 1, \quad (t,x) \,\epsilon\, (0,T] \times \mathbb{R}^n, \qquad (3.3)\\
dM^{\kappa,3}(x,t) &= A(t)^* M^{\kappa,3}(x,t)dt + B(t)^* M^{\kappa,3}(x,t)dy_t \\
&\quad + \frac{1}{2}\kappa(k-1)|C^{1/2} N^{-1/2} f^{3,\prime}(t,x)|^2 M^{\kappa-2,3}(x,t)dt \\
&\quad + \kappa f^3(t,x) M^{\kappa-1,3}(x,t) N^{1/2} C^{-1} dy_t, \qquad \kappa \geq 1, \quad (t,x) \,\epsilon\, (0,T] \times \mathbb{R}^n, \qquad (3.4)
\end{aligned}
$$

where the convention $M^{p,j}(x,t) = 0$ *for* $p < 0$ *is used. The initial conditions are*

$$M^{\kappa,j}(x,0) = 0, \qquad \kappa \geq 1, \quad j = 1,2,3, \tag{3.5}$$

and for $\kappa = 0$

$$M^{0,j}(x,t) = q(x,t), \qquad j = 1,2,3. \tag{3.6}$$

Proof. We shall use induction. Consider (3.2). Now, the case $\kappa = 1$ is easily verified, so it is omitted. Suppose (3.2) holds for $\kappa \to k-1$. We shall show that it also holds for κ. To this end, consider $\Phi(x_t)\Lambda_{0,t} L_{0,t}^{\kappa,1}$, where $\{x_t; t \,\epsilon\, [0,T]\}$ and $\{\Lambda_{0,t}; t \,\epsilon\, [0,T]\}$ are solutions of (2.19), (2.20), respectively. By the Itô product rule

$$L_{0,t}^{\kappa,1} = \kappa \int_0^t L_{0,s}^{\kappa-1,1} f^1(s,x_s)ds, \qquad \kappa \geq 1. \tag{3.7}$$

Employing the Itô product rule once again, we have

$$
\begin{aligned}
\Phi(x_t)\Lambda_{0,t} L_{0,t}^{\kappa,1} &= \int_0^t \Phi(x_s) d(\Lambda_{0,s} L_s^{\kappa,1}) + \int_0^t d\Phi(x_s)\Lambda_{0,s} L_{0,s}^{\kappa,1} \\
&\quad + \int_0^t \langle \Phi(x), \Lambda L^{\kappa,1}\rangle_s. \qquad (3.8)
\end{aligned}
$$

Now, from (3.7), (2.20) we compute

$$
\begin{aligned}
\Lambda_{0,t} L_{0,t}^{\kappa,1} &= \int_0^t \Lambda_{0,s} dL_{0,s}^{\kappa,1} + \int_0^t L_{0,s}^{\kappa,1} d\Lambda_{0,s} + \int_0^t d\langle \Lambda, L^{\kappa,1}\rangle_t \\
&= \kappa \int_0^t f^1(s,x_s)\Lambda_{0,s} L_{0,s}^{\kappa-1,1} ds + \int_0^t \Lambda_{0,s} L_{0,s}^{\kappa,1} h'(s,x_s) C_s^{-1/2} d\widetilde{y}_s. \qquad (3.9)
\end{aligned}
$$

Substituting (3.9) into (3.8) and then proceeding as in the derivation of Theorem

2.5, we obtain

$$\begin{aligned}\Phi(x_s)\Lambda_{0,t}L^{\kappa,1}_{0,t} &= \frac{1}{2}\int_0^t \Lambda_{0,s}L^{\kappa,1}_{0,s}\mathrm{Tr}\left(\sigma(s,x_s)\sigma'(s,x_s)D^2_x\Phi(x_s)\right)ds \\ &+ \int_0^t \Lambda_{0,s}L^{\kappa,1}_{0,s}D'_x\Phi(x_s)\sigma(s,x_s)D^{1/2}_s d\widetilde{w}_s \\ &+ \int_0^t \Lambda_{0,s}L^{\kappa,1}_{0,s}\Phi(x_s)h'(s,x_s)C^{-1/2}_s d\widetilde{y}_s \\ &+ \int_0^t \Lambda_{0,s}L^{\kappa,1}_{0,s}D'_x\Phi(x_s)\sigma(s,x_s)\alpha'_sC^{1/2}_s d\widetilde{y}_s + \kappa\int_0^t \Lambda_{0,s}L^{\kappa-1,1}_{0,s}f^1(s,x_s)ds. \end{aligned} \tag{3.10}$$

Conditioning each side of (3.10) on $\mathcal{F}^y_{0,t}$ using (3.1), and then integrating by parts, we deduce (3.2). When $\kappa = 0, j = 1$, we have $M^{0,1}(x,t)dx = \mathrm{E}[I_{x_t\epsilon dx}\Lambda_{0,t}|\mathcal{F}^y_{0,t}]$, and thus $M^{0,1}(x,t)$ satisfies the DMZ equation.

The derivation (3.3) is done similarly; therefore we shall outline only the important steps. Under measure P,

$$L^{\kappa,2}_{0,t} = \left[\int_0^t f^2(s,x_s)(dw_s - \alpha'_sC^{-1}_s h(s,x_s)ds)\right]^\kappa. \tag{3.11}$$

Substituting $w_t = \int_0^t D^{1/2}_s d\widetilde{w}_s + \int_0^t \alpha'_sC^{-1/2}_s d\widetilde{y}_s$ into (3.11),

$$L^{\kappa,2}_{0,t} = \left[\int_0^t f^2(s,x_s)(D^{1/2}_s d\widetilde{w}_s + \alpha'_sC^{-1/2}_s d\widetilde{y}_s - \alpha'_sC^{-1}_s h(s,x_s)ds)\right]^\kappa. \tag{3.12}$$

By the Itô product rule

$$\begin{aligned} L^{\kappa,2}_{0,t} &= \kappa\int_0^t L^{k-1,2}_{0,s}f^2(s,x_s)(D^{1/2}_s d\widetilde{w}_s + \alpha'_sC^{-1/2}_s d\widetilde{y}_s - \alpha'_sC^{-1}_s h(s,x_s)ds) \\ &+ \frac{1}{2}\kappa(k-1)\int_0^t L^{k-2,2}_{0,s}f^2(s,x_s)D^{1/2}_s D^{1/2,\prime}_s f^{2,\prime}(s,x_s)ds \\ &+ \frac{1}{2}\kappa(k-1)\int_0^t L^{k-2,2}_{0,s}f^2(s,x_s)\alpha'_sC^{-1}_s\alpha_s f^{2,\prime}(s,x_s)ds. \end{aligned} \tag{3.13}$$

Employing the Itô product rule to $\Phi(x_t)\Lambda_{0,t}L^{\kappa,2}_{0,t}$, as in (3.9), (3.10), and then invoking $M^{\kappa,2}(z,t)dx = \mathrm{E}[I_{x_t\epsilon dx}\Lambda_{0,t}L^{\kappa,2}_{0,t}|\mathcal{F}^y_{0,t}]$, after some algebra we derive (3.3), and (3.5) for $j = 2, \kappa \geq 2$. The special case $\kappa = 1, 2$ is done similarly. Also, to derive (3.4), we start with

$$L^{\kappa,3}_{0,t} = \left[\int_0^t f^3(s,x_s)\left(db_s - N^{1/2}_s C^{-1}_s h(s,x_s)ds\right)\right]^\kappa, \tag{3.14}$$

which is defined under measure P, and then we follow the above procedure to obtain (3.4), and (3.5), for $j = 3$. □

Next, we establish existence and uniqueness of the moment processes $M^{\kappa,j}(\cdot), \kappa \geq 1, j = 1, 2, 3$, using the variational methods of SPDEs, similar to Theorem 3.1.

Clearly, (3.2)–(3.4) with their corresponding boundary conditions (3.5), (3.6) are

of the general form

$$M(x,t) = \int_0^t A(s)^* M(x,s)ds + \int_0^t B(s)^* M(x,s)dy_s + \int_0^t \psi(s)ds + \int_0^t \phi(s)dy_s + \int_0^t \eta(s)ds, \tag{3.15}$$

where $\eta(t) \in L_y^2((0,T);H^1), \psi(t) \in L_y^2((0,T);H^{-1}), \phi(t) \in L_y^2((0,T);(L^2(\mathbb{R}^n))^d)$. For example, the fourth right side term of (3.3) belongs to $L_y^2((0,T);(L^2(\mathbb{R}^n))^d)$. Therefore, for finite κ, an application of variational methods of SPDEs (see [8, 9, 10]) implies there exists one and only one solution to (3.15) in the space $M(\cdot) \in L_y^2((0,T);H^1) \cap L_y^2(\Omega,\mathcal{F},P;C((0,T);H))$. Consequently, the moment processes of Theorem 3.1 have unique solutions as well.

Notice that the filtered estimates for $L_{0,t}^{\kappa,j}, \kappa \geq 1, j = 1,2,3$, can be computed from

$$\widetilde{\mathrm{E}}[L_{0,t}^{\kappa,j}|\mathcal{F}_{0,t}^y] = \frac{\int_{\mathbb{R}^n} M^{\kappa,j}(z,t)dz}{\int_{\mathbb{R}^n} q(z,t)dz}, \quad \kappa \geq 1, \quad j = 1,2,3. \tag{3.16}$$

Clearly, if the fundamental solution of the DMZ equation $q(t,t;x,s)$ is finite-dimensional, then according to Lemma 3.2, (3.16) can be computed explicitly in terms of finite numbers of statistics.

LEMMA 3.2. *Suppose the coefficients of $A, B_k,\ k = 1,\ldots,d$, and $f^j,\ 1 \leq j \leq 3$, have bounded partial derivatives in x of any order. Then $M_t^{\kappa,j}(\cdot)$ have $\mathcal{F}_{0,t}^y$-measurable density functions given by*

$$M^{\kappa,1}(z,t) = \kappa \int_0^t \int_{\mathbb{R}^n} f^1(s,x) M^{\kappa-1,1}(x,s) q(z,t;x,s)dxds, \qquad \kappa \geq 1, \tag{3.17}$$

$$\begin{aligned} M^{\kappa,2}(z,t) &= \frac{1}{2}\kappa(\kappa-1)\int_0^t \int_{\mathbb{R}^n} |f^2(s,x)|^2 M^{k-2,2}(x,s) q(z,t;x,s)dxds \\ &\quad - \kappa \int_0^t \int_{\mathbb{R}^n} \sum_{i=1}^n \frac{\partial}{\partial x_i}\left(M^{k-1,2}(x,s)\left(\sigma(s,x)f^{2,\prime}(s,x)\right)_i\right) q(z,t;x,s)dxds \\ &\quad + \kappa \int_0^t \int_{\mathbb{R}^n} f^2(s,x) M^{k-1,2}(x,s)\alpha_s' C_s^{-1} q(z,t;x,s)dxdy_s, \qquad \kappa \geq 1, \end{aligned} \tag{3.18}$$

$$\begin{aligned} M^{\kappa,3}(z,t) &= \frac{1}{2}\kappa(k-1)\int_0^t \int_{\mathbb{R}^n} |C_s^{1/2} N_s^{-1/2} f^3(s,x)|^2 M^{k-2,3}(x,s) q(z,t;x,s)dxds \\ &\quad + \kappa \int_0^t \int_{\mathbb{R}^n} f^3(s,x) M^{k-1,3}(x,s) N^{1/2} C_s^{-1} q(z,t;x,s) C_s^{-1} dxdy_s, \qquad \kappa \geq 1, \end{aligned} \tag{3.19}$$

with the convention $M^{p,j}(x,t) = 0$ for $p < 0$, $j = 1,2,3$.

Proof. Theorem 2.8 establishes the existence and uniqueness of a fundamental solution to the DMZ equation. Let $\widehat{M}^{\kappa,1}(z,t)$ denote the right side of (3.17). For $\kappa = 1$, we have

$$\widehat{M}^{1,1}(z,t) = \int_0^t \int_{\mathbb{R}^n} f^1(x,s) q(x,s) q(z,t;x,s)dxds,$$

because $M^{0,1}(\cdot,\cdot) = q(\cdot,\cdot)$. Then

$$\begin{aligned} d\widehat{M}^{1,1}(z,t) &= f^1(t,z)q(z,t)dt + \int_0^t \int_{\mathbb{R}^n} f^1(s,x)q(x,s)dq(z,t;x,s)dxds \\ &= f^1(t,z)q(z,t)dt + A(t)^* \int_0^t \int_{\mathbb{R}^n} f^1(s,x)q(x,s)q(z,t;x,s)dxdsdt \\ &+ B(t)^* \int_0^t \int_{\mathbb{R}^n} f^1(s,x)q(x,s)q(z,t)dxdsdy_t \\ &= A(t)^*\widehat{M}^{1,1}(z,t)dt + B(t)^*\widehat{M}^{1,1}(z,t)dy_t + f^1(t,z)q(z,t)dt. \end{aligned}$$

Thus, $\widehat{M}^{1,1}(\cdot,\cdot)$ satisfies (3.2); for $t = 0, \widehat{M}^{1,1}(z,0)$, and so (3.17) holds for $\kappa = 1$. Let

$$\widehat{M}^{\kappa,1}(z,t) = \kappa \int_0^t \int_{\mathbb{R}^n} f^1(s,x)\widehat{M}^{k-1,1}(x,s)q(z,t;x,s)dxds \tag{3.20}$$

and assume it satisfies (3.2) for $(t,z)\ \epsilon\ (0,T]\times\mathbb{R}^n$, and (3.5) for $t = 0$. We shall show that

$$\widehat{M}^{k+1,1}(z,t) = (k+1) \int_0^t \int_{\mathbb{R}^n} f^1(s,x)\widehat{M}^{k,1}(x,s)q(z,t;x,s)dxds \tag{3.21}$$

satisfies (3.2), with $k \to k+1$, for $(t,z)\ \epsilon\ (0,T]\times\mathbb{R}^n$. Clearly, $\widehat{M}^{k+1,1}(z,0) = 0$, so (3.5) holds (with $j = 1$). Now,

$$\begin{aligned} d\widehat{M}^{k+1,1}(z,t) &= (k+1)f^1(t,z)\widehat{M}^{k,1}(z,t)dt \\ &\quad + (k+1)\int_0^t \int_{\mathbb{R}^n} f^1(s,x)\widehat{M}^{k,1}(x,s)dq(z,t;x,s)dxds \\ &= (k+1)f^1(t,z)\widehat{M}^{k,1}(z,t)dt \\ &+ (k+1)A^*(t)\int_0^t \int_{\mathbb{R}^n} f^1(s,x)\widehat{M}^{k,1}(x,s)q(z,t;x,s)dxds \\ &+ (k+1)B(t)^*\int_0^t \int_{\mathbb{R}^n} f^1(s,x)\widehat{M}^{k,1}(x,s)q(z,t;x,s)dxdy_s \\ &= (k+1)f^1(t,z)\widehat{M}^{k,1}(z,t)dt + A^*(t)\widehat{M}^{k+1,1}(z,t)dt + B(t)^*\widehat{M}^{k+1,1}(z,t)dy_t. \end{aligned}$$

Hence (3.17) satisfies (3.2), (3.5) with $k \to k+1$. Similarly, one may use induction to verify the representations (3.18), (3.19). □

Next, we introduce an example to demonstrate the computations described in (3.17).

3.1. Specific application. Consider the system

$$\begin{aligned} dx_t &= Fx_t dt + Gdw_t, \qquad x(0)\ \epsilon\ \mathbb{R}^n, \\ dy_t &= Hx_t + N^{\frac{1}{2}}db_t, \qquad y(0) = 0\ \epsilon\ \mathbb{R}^d. \end{aligned}$$

The random variable $x(0)$ is Gaussian. Although the above linear system does not satisfy Assumption 2.2, statements 1, 2, 3, and 5, the fundamental solution of the DMZ equation exists, and it is given by

$$q(z,t;x,s) = \frac{1}{(2\pi)^{n/2}|P_{s,t}|^{1/2}} \exp\left(-\frac{1}{2}|P_{s,t}^{-1/2}(z - r_{s,t}(x))|^2\right) \times \Lambda_{s,t}(x), \tag{3.22}$$

$$dr_{s,t}(x) = \big(F - P_{s,t}H'N^{-1}H\big)r_{s,t}(x)dt + P_{s,t}H'N^{-1}dy_t, \quad r_{s,s}(x) = x, \tag{3.23}$$

$$\dot{P}_{s,t} = FP_{s,t} + P_{s,t}F' - P_{s,t}H'N^{-1}HP_{s,t} + GG', \quad P_{s,s} = 0, \tag{3.24}$$

$$\Lambda_{s,t}(x) = \exp\left(\int_s^t (Hr_{s,\tau})'N^{-1}dy_\tau - \frac{1}{2}\int_s^t |N_\tau^{-\frac{1}{2}}Hr_{s,\tau}(x)|^2 d\tau\right). \tag{3.25}$$

Let

$$r_{s,t}(x) = \Phi_{s,t}x + \beta_{s,t},$$

where

$$\dot{\Phi}_{s,t} = \big(F - P_{s,t}H'N^{-1}H\big)\Phi_{s,t}, \quad \Phi_{s,s} = I_n,$$

$$d\beta_{s,t} = \big(F - P_{s,t}H'N^{-1}H\big)\beta_{s,t}dt + P_{s,t}H'N^{-1}dy_t, \quad \beta_{s,s} = 0.$$

Then

$$\Lambda_{s,t}(x) = \gamma_{s,t}\exp\left(x'\rho_{s,t} - \frac{1}{2}x'S_{s,t}x\right),$$

where

$$\gamma_{s,t} = \exp\left(\int_s^t \beta'_{s,\tau}H'N^{-1}dy_\tau - \frac{1}{2}\int_s^t |N^{-1/2}H\beta_{s,\tau}|^2 d\tau\right),$$

$$S_{s,t} = \int_s^t \Phi'_{s,\tau}H'N^{-1}H\Phi_{s,\tau}d\tau,$$

$$\rho_{s,t} = \int_s^t \Phi'_{s,\tau}H'N^{-1}\big(dy_\tau - H\beta_{s,\tau}d\tau\big).$$

Moreover, the unnormalized conditional density of x_s given $\mathcal{F}_{0,s}^y$ is

$$q(x,s) = \frac{1}{(2\pi)^{n/2}|\Sigma_{0,s}|^{1/2}}\exp\left(-\frac{1}{2}|\Sigma_{0,s}^{-1/2}(x - \widehat{x}_{0,s})|^2\right) \times \widehat{\Lambda}_{0,s},$$

where $\widehat{x}_{0,s}$ is a solution of (3.23) with $\widehat{x}_{0,0} = \xi$, $\Sigma_{0,s}$ is a solution of (3.24) with $\Sigma_{0,0} = \Sigma_0$, and $\widehat{\Lambda}_{0,s}$ is given by (3.25) with $r \to \widehat{x}, P \to \Sigma$. Notice that

$$\int_{\mathbb{R}^n} q(x,s)q(z,t;x,s)dx = \frac{1}{(2\pi)^{n/2}|P_{s,t}|^{1/2}} \times \frac{1}{(2\pi)^{n/2}|\Sigma_{0,s}|^{1/2}} \times \widehat{\Lambda}_{0,s} \times \gamma_{s,t}$$
$$\times \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}|P_{s,t}^{-1/2}\big(z - \Phi_{s,t}x - \beta_{s,t}\big)|^2 - \frac{1}{2}|\Sigma_{0,s}^{-1/2}\big(x - \widehat{x}_{0,s}\big)|^2 + x'\rho_{s,t} - \frac{1}{2}x'S_{s,t}x\right)dx.$$

Therefore, the integral with respect to x is computed by completing the squares. Consequently, we deduce that $\tilde{E}\big[\int_0^t f(x_s)ds|\mathcal{F}_{0,t}^y\big]$ is finite-dimensional computable for large classes of functions $f(x)$ such as $f(x) = x^p, p \geq 1$, p an integer.

4. Conditional moment generating functions. Next we introduce moment generating functions for computing the conditional moments of integrals and stochastic integrals (2.32).

DEFINITION 4.1. *Let* $\theta = i\omega, i = \sqrt{-1}$.

1. *The measure-valued conditional moment generating functions of the stochastic processes* $\{L_{0,t}^{1,j}; t \in [0,T]\}, j = 1,2,3,$ *given by*

$$\widetilde{\beta}_t^{\theta,j}(\Phi) = \widetilde{\mathrm{E}}[\Phi(x_t) \exp\left(\theta L_{0,t}^{1,j}\right) | \mathcal{F}_{0,t}^y], \qquad j = 1,2,3, \tag{4.1}$$

are well defined.

2. *The measure-valued unnormalized conditional moment generating functions of the stochastic processes* $\{L_{0,t}^{1,j}; t \in [0,T]\}, j = 1,2,3,$ *given by*

$$\beta_t^{\theta,j}(\Phi) = \mathrm{E}[\Phi(x_t)\Lambda_{0,t} \exp\left(\theta L_{0,t}^{1,j}\right) | \mathcal{F}_{0,t}^y], \qquad j = 1,2,3, \tag{4.2}$$

are well defined.

LEMMA 4.2. *Suppose* $\beta_t^{\theta,j}(\cdot)$ *have* $\mathcal{F}_{0,t}^y$*-measurable density function* $\beta^{\theta,j} : \mathbb{R}^n \times [0,T] \times \Omega \to \mathbb{R}, j = 1,2,3.$

1. *Then*

$$\widetilde{\mathrm{E}}[\Phi(x_t) \exp\left(\theta L_{0,t}^{1,j}\right) | \mathcal{F}_{0,t}^y] = \frac{\beta_t^{\theta,j}(\Phi)}{q_t(1)} = \frac{\int_{\mathbb{R}^n} \Phi(z)\beta^{\theta,j}(z,t)dz}{\int_{\mathbb{R}^n} q(z,t)dz}, \qquad j = 1,2,3. \tag{4.3}$$

2. *The conditional characteristic functions of the stochastic processes* $\{L_{0,t}^{1,j}; t \in [0,T]\}, j = 1,2,3,$ *are given by*

$$\widetilde{\mathrm{E}}\left[\exp\left(i\omega L_{0,t}^{1,j}\right) | \mathcal{F}_{0,t}^y\right] = \frac{\beta_t^{i\omega,j}(1)}{q_t(1)} = \frac{\int_{\mathbb{R}^n} \beta^{i\omega,j}(z,t)dz}{\int_{\mathbb{R}^n} q(z,t)dz}, \quad j = 1,2,3. \tag{4.4}$$

Proof. The proof is similar to Lemma 2.4. □

THEOREM 4.3. *Suppose* $\beta_t^{\theta,j}(\cdot)$ *have* $\mathcal{F}_{0,t}^y$*-measurable density functions* $\beta^{\theta,j}(\cdot), j = 1,2,3.$

The densities of the measure-valued unnormalized conditional moment generating functions, namely,

$$\beta^{\theta,j}(x,t)dx = \mathrm{E}\left[I_{x_t \in dx}\Lambda_{0,t} \exp\left(\theta L_{0,t}^{1,j}\right) | \mathcal{F}_{0,t}^y\right], \qquad j = 1,2,3, \tag{4.5}$$

satisfy the following system of SPDEs:

$$\begin{aligned} d\beta^{\theta,1}(x,t) &= A(t)^*\beta^{\theta,1}(x,t)dt + B(t)^*\beta^{\theta,1}(x,t)dy_t \\ &\quad + \theta f^1(t,x)\beta^{\theta,1}(x,t)dt, \qquad (t,x) \in (0,T] \times \mathbb{R}^n, \end{aligned} \tag{4.6}$$

$$\begin{aligned} d\beta^{\theta,2}(x,t) &= A(t)^*\beta^{\theta,2}(x,t)dt + B(t)^*\beta^{\theta,2}(x,t)dy_t \\ &\quad + \frac{\theta^2}{2}|f^{2,\prime}(t,x)|^2\beta^{\theta,2}(x,t)dt - \theta\sum_{i=1}^n \frac{\partial}{\partial x_i}\left(\left(\sigma(t,x)f^{2,\prime}(x,t)\right)_i \beta^{\theta,2}(x,t)\right)dt \end{aligned}$$

$$+ \theta f^2(t,x)\beta^{\theta,2}(x,t)\alpha_t' C_t^{-1} dy_t, \qquad (t,x) \in (0,T] \times \mathbb{R}^n, \tag{4.7}$$

$$\begin{aligned} d\beta^{\theta,3}(x,t) &= A(t)^* \beta^{\theta,3}(x,t)dt + B(t)^* \beta^{\theta,3}(x,t)dy_t \\ &\quad + \frac{\theta^2}{2} |C^{1/2} N^{-1/2} f^{3,\prime}(t,x)|^2 \beta^{\theta,3}(x,t)dt \\ &\quad + \theta f^3(t,x)\beta^{\theta,3}(x,t) N^{1/2} C^{-1} dy_t, \qquad (t,x) \in (0,T] \times \mathbb{R}^n. \end{aligned} \tag{4.8}$$

The initial conditions are

$$\beta^{\theta,j}(x,0) = p_0(x), \qquad x \in \mathbb{R}^n, \qquad j = 1,2,3. \tag{4.9}$$

Proof. First, absorb $\exp\left(\theta L_{0,t}^{1,j}\right)$ in the exponential term $\Lambda_{0,t}$ by setting

$$\widehat{\Lambda}_{0,t}^j = \Lambda_{0,t} \exp\left(\theta L_{0,t}^{1,j}\right).$$

Second, apply the Itô product rule as in Theorem 3.1. This derivation is along the lines of information state equations in [13]. □

Equations (4.6), (4.7), (4.8) with their respective boundary conditions (4.9) are of the general form

$$M^\theta(x,t) = p_0(x) + \int_0^t A^\theta(s)^* M(x,s)ds + \int_0^t B^\theta(s)^* M(x,s)dy_s, \tag{4.10}$$

where the operators $A^\theta(t)^*, B^\theta(t)^*$ and their adjoints $A^\theta(t), B^\theta(t)$ depend on θ. Moreover, for sufficiently small $\theta \in \mathbb{R}$, these operators are bounded linear operators as described in (2.24), and there exist $\lambda_1^\theta, \lambda_2^\theta$, which depend on $\theta \in \mathbb{R}$ such that they satisfy the coercivity condition (2.25). Consequently, similar to Lemma 2.6, there exists one and only one solution of (4.6)–(4.9) in the space $\beta^{\theta,j}(\cdot) \in L_y^2((0,T);H^1) \cap L^2(\Omega,\mathcal{F},P;C((0,T);H))$.

LEMMA 4.4. *For $j = 1,2,3$,*

$$E\left[\Phi(x_t)\Lambda_{0,t}\exp\left(\theta L_{0,t}^{1,j}\right)|\mathcal{F}_{0,t}^y\right] = E\left[\Phi(x_t)\Lambda_{0,t}|\mathcal{F}_{0,t}^y\right] + \sum_{\kappa=1}^{\infty} \frac{\theta^\kappa}{\kappa!} E\left[\Phi(x_t)\Lambda_{0,t} L_{0,t}^{\kappa,j}|\mathcal{F}_{0,t}^y\right], \tag{4.11}$$

where the infinite series converges in $L^1(\Omega,\mathcal{F}_{0,t}^y,P)$. Moreover,

$$\beta_t^{\theta,j}(\Phi) = q_t(\Phi) + \sum_{\kappa=1}^{\infty} \frac{\theta^\kappa}{\kappa!} M_t^{\kappa,j}(\Phi), \qquad j = 1,2,3. \tag{4.12}$$

Proof. We shall invoke the following estimate found in [15, p. 353]:

$$\left| e^{\theta x} - \sum_{k=0}^{n} \frac{(\theta x)^k}{k!} \right| \le \min\left\{ \frac{|\theta x|^{n+1}}{(n+1)!}, \frac{2|\theta x|^n}{n!} \right\}, \ \theta \in \mathbb{R}, \ x \in \mathbb{R}.$$

The first right side term is an estimate for $|\theta x|$ small and the second for $|\theta x|$ large. Using the above estimate

$$\begin{aligned}
&\mathrm{E}\Big\{\Big|\mathrm{E}[\Phi(x_t)\Lambda_{0,t}\exp(\theta L^{1,j}_{0,t})|\mathcal{F}^y_{0,t}] - \sum_{k=0}^{n}\frac{\theta^k}{k!}\mathrm{E}[\Phi(x_t)\Lambda_{0,t}L^{k,j}_{0,t}|\mathcal{F}^y_{0,t}]\Big|\Big\}\\
&\le \mathrm{E}\Big\{|\Phi(x_t)|\Lambda_{0,t}\left(\left|\exp(\theta L^{1,j}_{0,t}) - \sum_{k=0}^{n}\frac{\theta^k}{k!}L^{k,j}_{0,t}\right|\right)\Big\}\\
&\le \mathrm{E}\Big\{|\Phi(x_t)|\Lambda_{0,t}\min\left\{\frac{|\theta|^{n+1}L^{n+1,j}_{0,t}}{(n+1)!}, \frac{2|\theta|^n L^{n,j}_{0,t}}{n!}\right\}\Big\}
\end{aligned}$$

(4.13)

$$\le \Big(\mathrm{E}\Big\{|\Phi(x_t)|^2\Lambda^2_{0,t}\Big\}\Big)^{1/2}\left(\mathrm{E}\min\left\{\left(\frac{|\theta|^{n+1}L^{n+1,j}_{0,t}}{(n+1)!}\right)^2, \left(\frac{2|\theta|^n L^{n,j}_{0,t}}{n!}\right)^2\right\}\right)^{1/2}.$$

The first right side term of (4.13) is bounded for any $(t,x)\,\epsilon\,[0,T]\times\mathbb{R}^n$, and the second is bounded because $f^j(t,x),\ j=1,2,3$, are bounded for any $(t,x)\,\epsilon\,[0,T]\times\mathbb{R}^n$.

Moreover, $\lim_{n\to\infty}\left(\frac{\theta^n}{n!}\right)^2\mathrm{E}L^{2n,j}_{0,t}=0$, and therefore in the limit, as $n\to\infty$, the right side of (4.13) converges to zero. Consequently, the following expansion must hold:

$$\mathrm{E}[\Phi(x_t)\Lambda_{0,t}\exp(\theta L^{1,j}_{0,t})|\mathcal{F}^y_{0,t}] = \sum_{k=0}^{\infty}\frac{\theta^k}{k!}\mathrm{E}[\Phi(x_t)\Lambda_{0,t}L^{k,j}_{0,t}|\mathcal{F}^y_{0,t}],$$

which is equivalent to (4.11) and, by Definition 2.9, to (4.12). □

At this stage, we may formally differentiate both sides of (4.12) with respect to θ, and then take the limit as $\theta\to 0$, to obtain relations between $\lim_{\theta\to 0}\frac{d^\kappa}{d\theta^\kappa}\beta^{\theta,j}_t(\Phi)$ and $M^{\kappa,j}_t(\Phi), j=1,2,3$. These results are presented next.

THEOREM 4.5. *We have the following:*

1. $\widetilde{\beta}^{i\omega,j}(1), \beta^{i\omega,j}(1), j=1,2,3$ *have* κ *continuous derivatives with respect to* ω, *w.p.*1.
2.

$$\lim_{\theta\to 0}\frac{d^\kappa}{d\theta^\kappa}\widetilde{\beta}^{\theta,j}_t(\Phi) = \lim_{\theta\to 0}\frac{d^\kappa}{d\theta^\kappa}\frac{\beta^{\theta,j}_t(\Phi)}{q_t(1)} = \widetilde{\mathrm{E}}[\Phi(x_t)L^{\kappa,j}_{0,t}|\mathcal{F}^y_{0,t}] \quad w.p.1, \tag{4.14}$$
$$\theta = i\omega, \quad \kappa\ge 0, \quad j=1,2,3.$$

3.

$$\lim_{\theta\to 0}\frac{d^\kappa}{d\theta^\kappa}\widetilde{\beta}^{\theta,j}_t(1) = \lim_{\theta\to 0}\frac{d^\kappa}{d\theta^\kappa}\frac{\beta^{\theta,j}_t(1)}{q_t(1)} = \widetilde{\mathrm{E}}[L^{\kappa,j}_{0,t}|\mathcal{F}^y_{0,t}] \quad w.p.1, \tag{4.15}$$
$$\theta = i\omega, \quad \kappa\ge 0, \quad j=1,2,3.$$

Proof. Recall that

$$\widetilde{\beta}^{i\omega,j}_t(\Phi) = \widetilde{\mathrm{E}}\big[\Phi(x_t)\exp\big(i\omega L^{1,j}_{0,t}\big)|\mathcal{F}^y_{0,t}\big] = \frac{E\big[\Lambda_{0,t}\exp\big(i\omega L^{1,j}_{0,t}\big)|\mathcal{F}^y_{0,t}\big]}{q_t(1)}.$$

Here $q_t(1)$ is independent of θ. The numerator $\mathrm{E}[\Phi(x_t)\Lambda_{0,t}\exp(i\omega L^{1,j}_{0,t})|\mathcal{F}^y_{0,t}]$ admits the power series expansion of Lemma 4.4, which implies 1, 2, 3. □

4.1. Specific application. Consider the system

$$dx_t = Fx_t dt + Gdw_t, \qquad x(0) \in \mathbb{R}^n,$$
$$dy_t = Hx_t + N^{\frac{1}{2}} db_t, \qquad y(0) = 0,$$
$$f^1(t,x) = \frac{1}{2}x'Qx, \quad f^2(t,x) = x'R, \quad f^3(t,x) = x'S, \qquad Q = Q'.$$

We assume $x(0)$ is a Gaussian random variable.

Suppose F, H are random matrices which we wish to identify or estimate. In [6] an algorithm is presented for estimating these matrices. This involves filtered estimates of the processes $\int_0^t f^1(s,x_s)ds$, $\int_0^t f^2(s,x_s)dw_s, \int_0^t f^3(s,x_s)db_s$. Here we apply Theorem 4.5 to obtain these estimates.

A solution of (2.13), (2.14) is

$$q(x,t) = \frac{1}{(2\pi)^{n/2}|P_t^0|^{1/2}} \exp\left(-\frac{1}{2}|P_t^{0,-1/2}\left(x - \widehat{x}_t^0\right)|^2\right) \times \widehat{\Lambda}_{0,t}^0, \tag{4.16}$$

where $\widehat{x}^0(\cdot), P^0(\cdot), \widehat{\Lambda}^0(\cdot)$ are given by

$$d\widehat{x}_t^0 = F\widehat{x}_t^0 dt + P_t^0 H' N^{-1}\left(dy_t - H\widehat{x}_t^0 dt\right), \qquad \widehat{x}^0(0) = \xi, \tag{4.17}$$

$$\dot{P}_t^0 = FP_t^0 + P_t^\theta F' - P_t^0 H' N^{-1} H P_t^0 + GG', \qquad P^0(0) = P_0, \tag{4.18}$$

$$\widehat{\Lambda}_{0,t}^0 = \exp\left(\int_0^t \left(H\widehat{x}_s^0\right)' N^{-1} dy_s - \frac{1}{2}\int_0^t \left(H\widehat{x}_s^0\right)' N^{-1} H\widehat{x}_s^0 ds\right). \tag{4.19}$$

These computations are easily verified by substitution into the DMZ equation. Recall also that $q(x,t) = M^{0,j}(x,t), j = 1,2,3$.

1. *Computation of* $\widehat{L}_{0,t}^{1,1} = \widetilde{\mathrm{E}}[\frac{1}{2}\int_0^t x_s' Q x_s ds | \mathcal{F}_{0,t}^y]$:

A solution of (4.6), (4.9) is (see, for example, [11, 13])

$$\beta^{\theta,1}(x,t) = \frac{1}{(2\pi)^{n/2}|P_t^\theta|^{1/2}} \exp\left(-\frac{1}{2}|P_t^{\theta,-1/2}\left(x - \widehat{x}_t^\theta\right)|^2\right) \times \widehat{\Lambda}_{0,t}^\theta \times \exp\left(\frac{\theta}{2}\int_0^t \mathrm{Tr}(P_s^\theta Q)ds\right), \tag{4.20}$$

where

$$d\widehat{x}_t^\theta = \left(F + \theta P_t^\theta Q\right)\widehat{x}_t^\theta dt + P_t^\theta H' N^{-1}\left(dy_t - H\widehat{x}_t^\theta dt\right), \qquad \widehat{x}(0) = \xi, \tag{4.21}$$

$$\dot{P}_t^\theta = FP_t^\theta + P_t^\theta F' - P_t^\theta\left(H' N^{-1} H - \theta Q\right) P_t^\theta + GG', \qquad P^\theta(0) = P_0, \tag{4.22}$$

$$\widehat{\Lambda}_{0,t}^\theta = \exp\left(\int_0^t \left(H\widehat{x}_s^\theta\right)' N^{-1} dy_s - \frac{1}{2}\int_0^t \left(H\widehat{x}_s^\theta\right)' N^{-1} H\widehat{x}_s^\theta ds\right). \tag{4.23}$$

In fact, we can show that $\lim_{\theta\to 0} P_t^\theta = P_t^0$, uniformly on compact subsets of $[0,T]$, and $\lim_{\theta\to 0}\widehat{x}_t^\theta = \widehat{x}_t^0$ a.s.

According to Theorem 4.5 we need

$$\frac{d}{d\theta}\frac{\beta_t^{\theta,1}(1)}{\widehat{\Lambda}_{0,t}^0} = \frac{d}{d\theta}\left[\widehat{\Lambda}_{0,t}^\theta\left(\widehat{\Lambda}_{0,t}^0\right)^{-1}\exp\left(\frac{\theta}{2}\int_0^t \mathrm{Tr}(P_s^\theta Q)\right)\right]. \tag{4.24}$$

Let

$$r_t^\theta = \frac{d}{d\theta}\widehat{x}_t^\theta, \qquad \Sigma_t^\theta = \frac{d}{d\theta}P_t^\theta.$$

Then from the differentiability of parameter dependent solutions of stochastic differential equations we know that

$$
\begin{aligned}
r_t^\theta &= \int_0^t \left(F + \theta P_s^\theta Q - P_s^\theta H' N^{-1} H r_t^\theta\right) r_s^\theta ds \\
&\quad + \int_0^t \theta \Sigma_s^\theta Q \widehat{x}_s^\theta ds + \int_0^t \Sigma_s^\theta H' N^{-1} \left(dy_s - H\widehat{x}_s^\theta ds\right) + \int_0^t P_s^\theta Q \widehat{x}_s^\theta ds, \qquad (4.25)\\
\Sigma_t^\theta &= \int_0^t F\Sigma_s^\theta ds + \int_0^t \Sigma_s^\theta F' ds - \int_0^t \Sigma_s^\theta \left(H' N^{-1} H - \theta Q\right) P_s^\theta ds \\
&\quad - \int_0^t P_s^\theta \left(H' N^{-1} H - \theta Q\right) \Sigma_s^\theta ds + \int_0^t P_s^\theta Q P_s^\theta ds \qquad (4.26)
\end{aligned}
$$

are well defined. Similarly as before we have $\lim_{\theta\to 0} r_t^\theta = r_t^0$ (a.s.), $\lim_{\theta\to 0} \Sigma_t^\theta = \Sigma_t^0$, where

$$
\begin{aligned}
r_t^0 &= \int_0^t P_s^0 Q \widehat{x}_s^0 ds + \int_0^t \Sigma_s^0 H' N^{-1} \left(dy_s - H\widehat{x}_s^0 ds\right) \\
&\quad + \int_0^t \left(F - P_s^0 H' N^{-1} H\right) r_s^0 ds, \qquad (4.27)
\end{aligned}
$$

(4.28)

$$
\Sigma_t^0 = \int_0^t \left(F - P^0 H' N^{-1} H\right) \Sigma_s^0 ds + \int_0^t \Sigma_s^0 \left(F - P^0 H' N^{-1} H\right)' ds + \int_0^t P_s^0 Q P_s^0 ds.
$$

Consequently,

$$
\begin{aligned}
&\lim_{\theta\to 0} \frac{d}{d\theta} \left\{ \widehat{\Lambda}_{0,t}^\theta \left(\widehat{\Lambda}_{0,t}^0\right)^{-1} \exp\left(\frac{\theta}{2} \int_0^t \mathrm{Tr}(P_s^\theta Q) ds\right)\right\} \\
&= \lim_{\theta\to 0} \left\{ \left(\int_0^t (Hr_s^\theta)' N^{-1} dy_s - \int_0^t (Hr_s^\theta)' N^{-1} H\widehat{x}_s^\theta ds + \frac{1}{2} \int_0^t \mathrm{Tr}\left(P_s^\theta Q + \theta \Sigma_s^\theta Q\right) ds \right) \right. \\
&\left. \times \widehat{\Lambda}_{0,t}^\theta \left(\widehat{\Lambda}_{0,t}^0\right)^{-1} \exp\left(\frac{\theta}{2} \int_0^t \mathrm{Tr}(P_s^\theta Q) ds\right)\right\} \\
&= \frac{1}{2} \int_0^t \mathrm{Tr}\left(P_s^0 Q\right) ds + \int_0^t \left(Hr_s^0\right)' N^{-1} \left(dy_s - H\widehat{x}_s^0\right) ds. \qquad (4.29)
\end{aligned}
$$

Finally, $\widehat{L}_{0,t}^{1,1} = \widetilde{\mathrm{E}}[\frac{1}{2}\int_0^t x_s' Q x_s ds | \mathcal{F}_{0,t}^y]$ is a solution of the stochastic equation

$$
d\widehat{L}_{0,t}^{1,1} = \frac{1}{2} \mathrm{Tr}\left(P_t^0 Q\right) dt + \left(Hr_t^0\right)' N^{-1} \left(dy_t - H\widehat{x}_t^0 ds\right), \qquad \widehat{L}_{0,t}^{1,1} = 0. \qquad (4.30)
$$

2. *Computation of* $\widehat{L}_{0,t}^{1,2} = \widetilde{\mathrm{E}}[\int_0^t x_s' R dw_s | \mathcal{F}_{0,t}^y]$:
A solution of (4.7), (4.9) is (see [11])

$$
\begin{aligned}
\beta^{\theta,2}(x,t) &= \frac{1}{(2\pi)^{n/2} |P_t^\theta|^{1/2}} \exp\left(-\frac{1}{2} |P_t^{\theta,-1/2} \left(x - \widehat{x}_t^\theta\right)|^2\right) \\
&\quad \times \widehat{\Lambda}_{0,t}^\theta \times \exp\left(\frac{\theta^2}{2} \int_0^t \mathrm{Tr}(P_s^\theta R R') ds\right), \qquad (4.31)
\end{aligned}
$$

where

(4.32)
$$d\widehat{x}_t^\theta = \left(F + \theta^2 P_t^\theta RR' + \theta GR'\right)\widehat{x}_t^\theta dt + P_t^\theta H' N^{-1}\left(dy_t - H\widehat{x}_t^\theta dt\right), \qquad \widehat{x}(0) = \xi,$$

(4.33)
$$\dot{P}_t^\theta = \left(F + \theta GR'\right)P_t^\theta + P_t^\theta\left(F + \theta RG'\right)' - P_t^\theta\left(H'N^{-1}H - \theta^2 RR'\right)P_t^\theta + GG',$$
$$P^\theta(0) = P_0,$$

(4.34)
$$\widehat{\Lambda}_{0,t}^\theta = \exp\left(\int_0^t \left(H\widehat{x}_s^\theta\right)' N^{-1} dy_s - \frac{1}{2}\int_0^t \left(H\widehat{x}_s^\theta\right)' N^{-1} H\widehat{x}_s^\theta ds\right).$$

By Theorem 4.5 we need

$$\frac{d}{d\theta}\frac{\beta_t^{\theta,2}(1)}{\widehat{\Lambda}_{0,t}^0} = \frac{d}{d\theta}\left[\widehat{\Lambda}_{0,t}^\theta\left(\widehat{\Lambda}_{0,t}^0\right)^{-1}\exp\left(\frac{\theta^2}{2}\int_0^t \mathrm{Tr}(P_s^\theta RR')\right)\right]. \tag{4.35}$$

Computing $\lim_{\theta\to 0} r_t^\theta = \lim_{\theta\to 0}\frac{d}{d\theta}\widehat{x}_t^\theta = r_t^0, \lim_{\theta\to 0}\Sigma_t^\theta = \lim_{\theta\to 0}\frac{d}{d\theta}P_t^\theta = P_t^0$, similarly as before, we have

$$\begin{aligned} r_t^0 &= \int_0^t GR'\widehat{x}_s^0 ds + \int_0^t \Sigma_s^0 H' N^{-1}\left(dy_s - H\widehat{x}_s^0 ds\right) \\ &\quad + \int_0^t \left(F - P_s^0 H' N^{-1} H\right) r_s^0 ds, \end{aligned} \tag{4.36}$$

$$\begin{aligned} \Sigma_t^0 &= \int_0^t \left(F - P^0 H' N^{-1} H\right)\Sigma_s^0 ds + \int_0^t \Sigma_s^0\left(F - P^0 H' N^{-1} H\right)' ds \\ &\quad + \int_0^t GR' P_s^0 ds + \int_0^t P_s^0 RG' ds. \end{aligned} \tag{4.37}$$

Hence

$$\begin{aligned} &\lim_{\theta\to 0}\frac{d}{d\theta}\left\{\widehat{\Lambda}_{0,t}^\theta\left(\widehat{\Lambda}_{0,t}^0\right)^{-1}\exp\left(\frac{\theta^2}{2}\int_0^t Tr(P_s^\theta RR')ds\right)\right\} \\ &= \int_0^t \left(Hr_s^0\right)' N^{-1}\left(dy_s - H\widehat{x}_s^0\right) ds. \end{aligned} \tag{4.38}$$

Finally, $\widehat{L}_{0,t}^{1,2} = \widetilde{\mathrm{E}}[\int_0^t x_s' R dw_s | \mathcal{F}_{0,t}^y]$ is a solution of the stochastic equation

$$d\widehat{L}_{0,t}^{1,2} = \left(Hr_t^0\right)' N^{-1}\left(dy_t - H\widehat{x}_t^0 ds\right), \qquad \widehat{L}_{0,t}^{1,2} = 0. \tag{4.39}$$

3. *Computation of* $\widehat{L}_{0,t}^{1,3} = \widetilde{\mathrm{E}}[\int_0^t x_s' S db_s | \mathcal{F}_{0,t}^y]$:
A solution of (4.7), (4.9) is (see [11])

$$\begin{aligned} \beta^{\theta,3}(x,t) &= \frac{1}{(2\pi)^{n/2}|P_t^\theta|^{1/2}}\exp\left(-\frac{1}{2}|P_t^{\theta,-1/2}\left(x - \widehat{x}_t^\theta\right)|^2\right) \\ &\quad \times \widehat{\Lambda}_{0,t}^\theta \times \exp\left(\frac{\theta^2}{2}\int_0^t \mathrm{Tr}(P_s^\theta SS')ds\right), \end{aligned} \tag{4.40}$$

where

$$d\widehat{x}_t^\theta = \left(F + \theta^2 P_t^\theta SS'\right)\widehat{x}_t^\theta dt + P_t^\theta H^{\theta,\prime} N^{-1}\left(dy_t - H^\theta \widehat{x}_t^\theta dt\right), \qquad \widehat{x}(0) = \xi, \tag{4.41}$$

$$\dot{P}_t^\theta = FP_t^\theta + P_t^\theta F' - P_t^\theta\left(H^{\theta,\prime} N^{-1} H^\theta - \theta^2 SS'\right) P_t^\theta + GG', \qquad P^\theta(0) = P_0, \tag{4.42}$$

$$\widehat{\Lambda}_{0,t}^\theta = \exp\left(\int_0^t \left(H^\theta \widehat{x}_s^\theta\right)' N^{-1} dy_s - \frac{1}{2}\int_0^t \left(H^\theta \widehat{x}_s^\theta\right)' N^{-1} H^\theta \widehat{x}_s^\theta ds\right), \tag{4.43}$$

$$H^\theta = H + \theta N^{-1/2,\prime} S'. \tag{4.44}$$

From Theorem 4.5 we need

$$\frac{d}{d\theta}\frac{\beta_t^{\theta,3}(1)}{\widehat{\Lambda}_{0,t}^0} = \frac{d}{d\theta}\left[\widehat{\Lambda}_{0,t}^\theta\left(\widehat{\Lambda}_{0,t}^0\right)^{-1}\exp\left(\frac{\theta^2}{2}\int_0^t \mathrm{Tr}(P_s^\theta SS')\right)ds\right]. \tag{4.45}$$

This can be done as in the previous cases.

Finally, $\widehat{L}_{0,t}^{1,3} = \widetilde{\mathrm{E}}[\int_0^t x_s' S db_s | \mathcal{F}_{0,t}^y]$ is a solution of the stochastic equation

$$\begin{aligned} d\widehat{L}_{0,t}^{1,3} &= \left(Hr_t^0\right)' N^{-1}\left(dy_t - H\widehat{x}_t^0 dt\right) + \left(N^{-\frac{1}{2}} S'\widehat{x}_t^0\right)' N^{-1}\left(dy_t - N^{-\frac{1}{2},\prime} S'\widehat{x}_t^0 dt\right), \\ \widehat{L}_{0,t}^{1,3} &= 0, \end{aligned} \tag{4.46}$$

where

$$\begin{aligned} r_t^0 &= \int_0^t \left(\Sigma_s^0 H' N^{-1} + P_s^0 (SN^{-1/2})' N^{-1}\right)\left(dy_s - H\widehat{x}_s^0 ds\right) \\ &\quad + \int_0^t P_s^0 H' N^{-1}\left(dy_s - (SN^{-1/2})'\widehat{x}_s^0 ds\right) + \int_0^t \left(F - P_s^0 H' N^{-1}\right) r_s^0 ds, \end{aligned} \tag{4.47}$$

$$\begin{aligned} \Sigma_t^0 &= \int_0^t \left(F - P^0 H' N^{-1} H\right)\Sigma_s^0 ds + \int_0^t \Sigma_s^0\left(F - P^0 H' N^{-1} H\right)' ds \\ &\quad - \int_0^t P_s^0\left((SN^{-1/2}) N^{-1} H + (N^{-1}H)'(SN^{-1/2})'\right) P_s^0 ds. \end{aligned} \tag{4.48}$$

□

Remark 4.6. The above methodology can be generalized to joint conditional moment generating functions of $L_{0,t}^{1,j}, L_{0,t}^{1,\ell}, 1 \le j, \ell \le 3$.

5. Applications to nonlinear filtering problems. Both methods introduced in section 3 can be used in Wiener chaos expansions of nonlinear filtering [2] problems.

Consider the nonlinear filtering problem

$$dx_t = f(t, x_t)dt + \sigma(t, x_t)dw_t, \qquad x(0) \in \mathbb{R}^n, \tag{5.1}$$

$$dy_t = h(t, x_t)dt + N_t^{1/2} db_t, \qquad y(0) = 0 \in \mathbb{R}^d. \tag{5.2}$$

Here $\{x_t; t \in [0,T]\}$ and $\{y_t; t \in [0,T]\}$ are the state and observation processes, respectively. Throughout, we assume Assumption 2.2 holds. Similar to section 2, under the reference probability space $(\Omega, \mathcal{A}, P, \mathcal{F}_{0,t})$, processes $\{x_t; t \in [0,T]\}$, $\{y_t; t \in [0,T]\}$, are independent; the former is a solution of (5.1), while the latter is a solution of

$$dy_t = N_t^{1/2} db_t, \qquad y(0) = 0 \in \mathbb{R}^d. \tag{5.3}$$

The Radon–Nikodým derivative is

$$\Lambda_{0,T} \doteq \left[\frac{d\widetilde{P}}{dP}|\mathcal{F}_{0,T}\right] = \exp\left(m_t - \frac{1}{2}\langle m,m\rangle_t\right), \tag{5.4}$$

where $m_t = \int_0^t h'(s,x_s)N_s^{-1}dy_s$. Thus, $\{\Lambda_{0,t}; t \in [0,T]\}$ is a solution of the stochastic differential equation

$$\Lambda_{0,t} = 1 + \int_0^t \Lambda_{0,s}h'(s,x_s)N_s^{-1}dy_s. \tag{5.5}$$

Moreover, if the measured-valued processes $q_t(\Phi) = E[\Phi(x_t)\Lambda_{0,t}|\mathcal{F}_{0,t}^y]$ have a density $q(x,t)$, then

$$dq(z,t) = A(t)^*q(z,t)dt + h'(t,z)q(z,t)N_t^{-1}dy_t, \qquad (z,t) \in [0,T]\times\mathbb{R}^n, \tag{5.6}$$

$$q(z,0) = p_0(z), \qquad z \in \mathbb{R}^n. \tag{5.7}$$

In what follows, we employ some of the recursive systems derived in section 3 to obtain representations for certain asymptotic expansions of $\mathrm{E}[\Phi(x_t)|\mathcal{F}_{0,t}^y]$.

Definition 5.1. *Suppose* $\mathrm{E}[\int_0^T |N_s^{-1/2}h(s,x_s)|^2ds]^p < \infty$. *Then the multiple stochastic integrals*

$$I_t^p[h] = \int_0^t\int_0^{s_1}\cdots\int_0^{s_{p-1}} h'(s_p,x_{s_p})N_{s_p}^{-1}dy_{s_p}h'(s_{p-1},x_{s_{p-1}})N_{s_{p-1}}^{-1}dy_{s_{p-1}} \cdots h'(s_1,x_{s_1})N_{s_1}^{-1}dy_{s_1}, \tag{5.8}$$

$$I_t^{[p]}[h] = \int_0^t\int_0^{s_1}\cdots\int_0^{s_p} \Lambda_{s_{p+1}}h'(s_{p+1},x_{s_{p+1}})N_{s_{p+1}}^{-1}dy_{s_{p+1}}h'(s_p,x_{s_p})N_{s_p}^{-1}dy_{s_p} \cdots h'(s_1,x_{s_1})N_{s_1}^{-1}dy_{s_1} \tag{5.9}$$

are well defined.

Consider the exponential martingale $\{\Lambda_{0,t}; t \in [0,T]\}$. Iterating (5.5) we have

$$\begin{aligned}
\Lambda_{0,t} &= 1 + \int_0^t \Lambda_{0,s}h'(s,x_s)N_s^{-1}dy_s \\
&= 1 + \int_0^t h'(s,x_s)N_s^{-1}dy_s + \int_0^t\int_0^{s_1} \Lambda_{0,s_2}h'(s_2,x_{s_2})N_{s_2}^{-1}dy_{s_2}h'(s_1,x_{s_1})N_{s_1}^{-1}dy_{s_1} \\
&+ \vdots \\
&= 1 + \int_0^t h'(s,x_s)N_s^{-1}dy_s + \int_0^t\int_0^{s_1} \Lambda_{0,s_2}h'(s_2,x_{s_2})N_{s_2}^{-1}dy_{s_2}h'(s_1,x_{s_1})N_{s_1}^{-1}dy_{s_1} \\
&\quad + \cdots + \int_0^t\int_0^{s_1}\cdots\int_0^{s_{p-1}} h'(s_p,x_{s_p})N_{s_p}^{-1}dy_{s_p}h'(s_{p-1},x_{s_{p-1}})N_{s_{p-1}}^{-1}dy_{s_{p-1}}
\end{aligned}$$

$$\cdots h'(s_1,x_{s_1})N_{s_1}^{-1}dy_{s_1} + I_t^{[p]}[h]. \tag{5.10}$$

If we now assume $\mathrm{E}[\Phi^2(x_t)\int_0^T |N_s^{-1}h(s,x_s)|^2ds]^p < \infty$ (which is satisfied by As-

sumption 2.2) and then substitute (5.10) into $q_t^0(\Phi) = \mathrm{E}[\Lambda_{0,t}\Phi(x_t)|\mathcal{F}_{0,t}^y]$ we have

$$\begin{aligned} q_t(\Phi) &= \mathrm{E}[\Lambda_{0,t}\Phi(x_t)|\mathcal{F}_{0,t}^y] = \mathrm{E}[\Phi(x_t)|\mathcal{F}_{0,t}^y] \\ &\quad + \sum_{k=1}^{p} \mathrm{E}\,[\,\Phi(x_t)\int_0^t\int_0^{s_1}\cdots\int_0^{s_{k-1}} h'(s_k,x_{s_k})N_{s_k}^{-1}dy_{s_k}\cdots h'(s_1,x_{s_1})N_{s_1}^{-1}dy_{s_1}|\mathcal{F}_{0,t}^y\,] \\ &\quad + \mathrm{E}\left[\Phi(x_t)I_t^{[p]}[h]|\mathcal{F}_{0,t}^y\right]. \end{aligned} \tag{5.11}$$

Note that under measure P, the processes $\{x_t; t \in [0,T]\}$ and $\{y_t; t \in [0,T]\}$ are independent; therefore $\mathrm{E}[\Phi(x_t)|\mathcal{F}_{0,t}^y] = \mathrm{E}[\Phi(x_t)]$. In addition, the increments $dy_{s_1}, dy_{s_2}, \ldots, dy_{s_k}$ are measurable with respect to $\mathcal{F}_{0,t}^y$; in the scalar case, $d=1$, the second right side term of (5.11) becomes

$$\sum_{k=1}^{p}\int_0^t\int_0^{s_1}\cdots\int_0^{s_{k-1}} \mathrm{E}[\Phi(x_t)h(s_k,x_{s_k})N_{s_k}\cdots h(s_1,x_{s_1})N_{s_1}]dy_{s_k}\cdots dy_{s_1}.$$

Formally, letting $p=\infty$ in (5.11) we derive the full expansion, which is made rigorous in the next theorem.

THEOREM 5.2.

1. *Suppose* $\mathrm{E}[\int_0^t |N_s^{-1/2}h(s,x_s)|^2ds]^p < \infty$ *and* $\mathrm{E}[\Phi(x_t)^2\int_0^t |N_s^{-1/2}h(s,x_s)|^2ds]^p < \infty$. *Then*

$$\widetilde{\mathrm{E}}[\Phi(x_t)|\mathcal{F}_{0,t}^y] = \frac{q_t(\Phi)}{q_t(1)} \tag{5.12}$$
$$= \frac{\mathrm{E}[\Phi(x_t)] + \sum_{k=1}^p \mathrm{E}[\Phi(x_t)I_t^k[h]|\mathcal{F}_{0,t}^y] + \mathrm{E}[\Phi(x_t)I_t^{[p]}[h]|\mathcal{F}_{0,t}^y]}{1 + \sum_{k=1}^p \mathrm{E}[I_t^p[h]|\mathcal{F}_{0,t}^y] + \mathrm{E}[I_t^{[p]}[h]|\mathcal{F}_{0,t}^y]}.$$

2. *Suppose* $\mathrm{E}[\exp\int_0^t |N_s^{-1/2}h(s,x_s)|^2ds] < \infty$ *and*

$$\mathrm{E}[\Phi(x_t)\exp\int_0^t |N_s^{-1/2}h(s,x_s)|^2ds] < \infty.$$

Then

$$\widetilde{\mathrm{E}}[\Phi(x_t)|\mathcal{F}_{0,t}^y] = \frac{q_t(\Phi)}{q_t(1)} = \frac{\mathrm{E}[\Phi(x_t)] + \sum_{k=1}^\infty \mathrm{E}[\Phi(x_t)I_t^k[h]|\mathcal{F}_{0,t}^y]}{1 + \sum_{k=1}^\infty \mathrm{E}[I_t^k[h]|\mathcal{F}_{0,t}^y]} \tag{5.13}$$

and the infinite series of (5.13) *converges in* $L^1(\Omega,\mathcal{F}_{0,t}^y,P)$.

Proof. See [2]. □

5.1. Recursive equations.

DEFINITION 5.3. *Suppose* $\mathrm{E}[\Phi^2(x_t)\int_0^t |N_s^{-1/2}h(s,x_s)|^2]^p < \infty$.

The measure-valued processes

$$M_t^0(\Phi) \doteq \mathrm{E}[\Phi(x_t)] = \int_{\mathbb{R}^n}\Phi(z)p(z,t;x,0)dx, \tag{5.14}$$

$$M_t^k(\Phi) \doteq \mathrm{E}[\Phi(x_t)I_t^k[h]|\mathcal{F}_{0,t}^y], \qquad k \geq 1, \tag{5.15}$$

are well defined.

THEOREM 5.4. *Suppose* $M_t^k(\cdot), k \geq 0$, *have density functions* $M^k(z,t)$.

1. *The density of the distribution* $\widetilde{P}(x_t \in A), A \in \mathcal{B}(\mathbb{R}^n)$, *satisfies the Kolmogorov equation*

$$dp(z,t) = A(t)^* p(z,t)dt, \quad (z,t) \in \mathbb{R}^n \times (0,T]; \qquad \lim_{t\to 0} p(z,t) = p_0(z). \tag{5.16}$$

2. *The densities of* $M_t^k(\cdot), k \geq 1$ *satisfy the following recursive system of SPDEs:*

$$dM^k(z,t) = A(t)^* M^k(z,t)dt + h^*(t,z)M^{k-1}(z,t)N_t^{-1}dy_t, \qquad (z,t) \in \mathbb{R}^n \times [0,T], \tag{5.17}$$

$$M^k(z,0) = 0, \quad z \in \mathbb{R}^n. \tag{5.18}$$

Proof. The distribution of $\{x_t; t \in [0,T]\}$ is the same under measure $\widetilde{P}$ and P. Hence, the density $p(\cdot,\cdot)$ satisfies (5.16).

Now, for $k = 1$ consider

$$I_t^1[h] = \int_0^t h'(s,x_s)N_s^{-1}dy_s. \tag{5.19}$$

By the Itô product rule

$$\begin{aligned}\Phi(x_t)I_t^1[h] = &\int_0^t A(s)^*\Phi(x_s)I_s^1[h]ds \\ &+ \int_0^t D_x'\Phi(x_s)I_s^1[h]\sigma(s,x_s)dw_s \\ &+ \int_0^t \Phi(x_s)h'(s,x_s)N_s^{-1}dy_s.\end{aligned} \tag{5.20}$$

Conditioning both sides of (5.20) on $\mathcal{F}_{0,t}^y$, and then using the independence of $\{w_t; t \in [0,T]\}$ and $\{y_t; t \in [0,T]\}$ (and a version of Fubini's theorem [12]), we have

$$\begin{aligned}M_t^1(\Phi) = &\int_0^t \int_{\mathbb{R}^n} A(s)\Phi(z)M^1(z,s)dzds \\ &+ \int_0^t \int_{\mathbb{R}^n} \Phi(z)h'(s,z)M^0(z,s)N_s^{-1}dzdy_s.\end{aligned} \tag{5.21}$$

Hence (5.17), (5.18) hold for $k = 1$.

Now, for $k = \ell$ consider

$$I_t^\ell[h] = \int_0^t \int_0^{s_1} \cdots \int_0^{s_{\ell-1}} h'(s_\ell, x_{s_\ell})N_{s_\ell}^{-1}dy_{s_\ell} \cdots h'(s,x_{s_1})N_{s_1}^{-1}dy_{s_1}. \tag{5.22}$$

Then

$$dI_t^\ell[h] = I_t^{\ell-1}[h]h'(t,x_t)N_t^{-1}dy_t.$$

By the Itô product rule

$$\begin{aligned}\Phi(x_t)I_t^\ell[h] = &\int_0^t A(s)\Phi(x_s)I_s^\ell[h]ds \\ &+ \int_0^t D_x'\Phi(x_s)I_s^\ell[h]\sigma(s,x_s)dw_s + \int_0^t \Phi(x_s)h'(s,x_s)I_s^{\ell-1}[h]N_s^{-1}dy_s.\end{aligned} \tag{5.23}$$

Similarly as before, conditioning both sides of (5.2) on $\mathcal{F}_{0,t}^y$ we have

$$M_t^\ell(\Phi) = \int_0^t \int_{\mathbb{R}^n} A(s)\Phi(z)M^\ell(z,t)dzds + \int_0^t \int_{\mathbb{R}^n} \Phi(z)h'(s,z)M^{\ell-1}(z,s)N_s^{-1}dzdy_s. \tag{5.24}$$

Hence, (5.17), (5.18) holds for any $k \geq 1$. □

COROLLARY 5.5. *Let* $\{p(z,t;x,s); 0 \leq s < t \leq T\}, (z,x) \in \mathbb{R}^n \times \mathbb{R}^n$ *be the fundamental solution of the density equation* (5.16)*:*

$$dp(z,t;x,s) = A(t)^* p(z,t;x,s)dt, \quad (z,t) \in \mathbb{R}^n \times (0,T]; \quad \lim_{t\to s} P(z,t;x,s) = \delta(z-x). \tag{5.25}$$

Then the solutions of (5.16)–(5.18) *are represented by*

$$M^k(z,t) = \int_0^t \int_{\mathbb{R}^n} h'(s,x)N_s^{-1}M^{k-1}(x,s)p(z,t;x,s)dxdy_s, \qquad k \geq 1, \tag{5.26}$$

$$M^0(z,t) = \int_{\mathbb{R}^n} p(z,t;x,0)p_0(x)dx. \tag{5.27}$$

Proof. Follow the procedures of Theorem 2.5. □

Remark 5.6. Because of the linearity of (5.16), (5.18), any finite expansion of both numerator and denominator of (5.13), say,

$$q_t^\ell(\Phi) = \mathrm{E}[\Phi(x_t)] + \sum_{k=1}^{\ell} \mathrm{E}[\Phi(x_t)I_t^k[h]|\mathcal{F}_{0,t}^y], \qquad \ell \geq 1, \tag{5.28}$$

is a solution of the SPDE

$$dq_t^\ell(\Phi) = q_t^\ell(A(t)\Phi)dt + \sum_{k=1}^{\ell} M_t^{k-1}(h'(t,z)\Phi)N_t^{-1}dy_t. \tag{5.29}$$

From Theorem 5.2, we know that in order to approximate $\widetilde{\mathrm{E}}[\Phi(x_t)|\mathcal{F}_{0,t}^y]$ through a finite series we need to compute $M_t^k(\Phi), 0, \leq k \leq p$. The latter can be computed from the joint-moment generating function of the random processes $\{\int_0^t h'(s,x_s)N^{-1}dy_s; t \in [0,T]\}$ and $\{\int_0^t |h'(s,x_s)|^2 ds; t \in [0,T]\}$.

6. Conclusion. This paper presents two methods for computing conditional moments of integrals and stochastic integrals for general diffusion processes. The first method employs recursive SPDEs; the second method employs conditional moment generating functions. An application of the first method results in new finite-dimensional filters. An application of the second method to the EM algorithm results in a significant reduction in the sufficient statistics required in the computation of the parameters.

REFERENCES

[1] S.I. MARCUS AND A.S. WILLSKY, *Algebraic structure and finite dimensional nonlinear estimation*, SIAM J. Math. Anal., 9 (1978), pp. 312–327.

[2] D. Ocone, *Topics in Nonlinear Filtering Theory*, Ph.D. thesis, M.I.T., Cambridge, MA, 1980.
[3] C. Georghiades and D. Snyder, *The expectation-maximization algorithm for symbol unsynchronized sequence detection*, IEEE Transactions Comm., 39 (1991), pp. 54–61.
[4] D. Ocone, J. Baras, and S. Marcus, *Explicit filters for diffusions with certain nonlinear drifts*, Stochastics, 8 (1982), pp. 1–16.
[5] R. Shumway and D. Stoffer, *An approach to time series smoothing and forecasting using the EM algorithm*, J. Time Ser. Anal., 3 (1982), pp. 253–264.
[6] R.J. Elliott and V. Krishnamurthy, *Exact finite-dimensional filters for maximum likelihood parameter estimation of continuous-time linear Gaussian systems*, SIAM J. Control Optim., 35 (1997), pp. 1908–1923.
[7] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes,* Vol. 1, Springer-Verlag, New York, 1977.
[8] A. Bensoussan, *Stochastic Control of Partially Observable Systems*, Cambridge University Press, Cambridge, UK, 1992.
[9] E. Pardoux, *Spde's and filtering of diffusion processes*, Stochastics, 3 (1979), pp. 127–167.
[10] E. Pardoux, *Equations of Nonlinear Filtering and Applications to Stochastic Control*, Lecture Notes in Math. 972, S. K. Mitter and A. Moro, eds., Springer-Verlag, Berlin, 1982, pp. 208–246.
[11] C. Charalambous and J. Hibey, *Minimum principle for partially observable nonlinear risk-sensitive control problems using measure-valued decompositions*, Stochastics Stochastics Rep., 57 (1996), pp. 247–288.
[12] E. Wong and B. Hajek, *Stochastic Processes in Engineering Systems*, Springer-Verlag, New York, 1985.
[13] C. Charalambous and R. Elliott, *Certain classes of nonlinear partially observable stochastic optimal control problems with explicit control laws equivalent to LEQG/LQG problems*, IEEE Trans. Automat. Control, 42 (1997), pp. 482–497.
[14] H. Kunita, *Stochastic Partial Differential Equations Connected with Nonlinear Filtering*, Lecture Notes in Math. 972, S. K. Mitter and A. Moro, eds., Springer-Verlag, Berlin, 1982, pp. 100–168.
[15] P. Billingsley, *Probability and Measure,* 2nd ed., John Wiley and Sons, New York, 1986.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1604–1622

APPROXIMATE CONTROLLABILITY OF SEMILINEAR DETERMINISTIC AND STOCHASTIC EVOLUTION EQUATIONS IN ABSTRACT SPACES*

NAZIM I. MAHMUDOV†

Abstract. Various sufficient conditions for approximate controllability of linear evolution systems in abstract spaces have been obtained, but approximate controllability of semilinear control systems usually requires some complicated and limited assumptions. In this paper, we show the approximate controllability of the abstract semilinear deterministic and stochastic control systems under the natural assumption that the associated linear control system is approximately controllable. The results are obtained using new properties of symmetric operators (which are proved in this paper), compact semigroups, the Schauder fixed point theorem, and/or the contraction mapping principle.

Key words. symmetric operators, controllability, the Schauder fixed point theorem, the contraction mapping principle, semilinear evolution equations, stochastic evolution equations

AMS subject classifications. 93B, 49E

DOI. S0363012901391688

1. Introduction.

Consider a deterministic or stochastic control system on the finite time interval $I = [0, T]$ with $T > 0$. Let $x(T; x_0, u)$ be its (random or not) state value at time T corresponding to the control $u(\cdot)$ taken from the set of admissible controls U_{ad} and the initial value x_0. Suppose that Z is the state space. Introduce the set

$$R(T; x_0, u) = \{x(T; x_0, u) : u(\cdot) \in U_{ad}\}. \tag{1.1}$$

DEFINITION 1.1. *A control system is said to be approximately controllable on I if $\overline{R(T; x_0, u)} = Z$.*

Controllability of the deterministic systems in finite dimensional spaces has been extensively studied (see [3], [21], and references therein). Several authors (see [1], [2], [3], [7], [8], [9], [14], [15], [16], [17], [18], [19], [20], [21], [22], [26], [28], [33], [34], [35], [36]) studied the concept for systems represented by evolution equations in infinite dimensional spaces. Most of the controllability results for nonlinear infinite dimensional control systems concern the so-called semilinear control systems which consist of a linear and a nonlinear part. Moreover, it should be emphasized that for infinite dimensional systems several concepts of controllability are analyzed.

In section 3, we consider the semilinear evolution system of the form

$$\begin{aligned} &\dot{x}(t) = Ax(t) + Bu(t) + f(t, x(t), u(t)), \quad t \in I = [0, T], \\ &x(0) = x_0, \end{aligned} \tag{1.2}$$

where $f : I \times X \times U \to X$ is a nonlinear operator, X is a separable reflexive Banach space, U is a Hilbert space, A is a linear operator on X, and B is a linear bounded operator from U to X.

*Received by the editors July 2, 2001; accepted for publication (in revised form) April 10, 2003; published electronically November 6, 2003.

http://www.siam.org/journals/sicon/42-5/39168.html

†Department of Mathematics, Eastern Mediterranean University, Gazimagusa, Mersin 10, Turkey (nazim.mahmudov@emu.edu.tr).

A mild solution $x(\cdot)$ of the semilinear evolution system we are considering is defined as a solution of the following integral equation:

$$x(t;x_0,u)=x(t)=S(t)x_0+\int_0^t S(t-s)\left[Bu(s)+f(s,x(s),u(s)\right]ds. \tag{1.3}$$

Any element of $L_2(I,U)$ is called a control. Any solution $x(\cdot;x_0,u)$ of (1.3) is referred to as a state trajectory of the evolution system corresponding to the initial state x_0 and the control $u(\cdot)$.

There have been many papers on the approximate controllability of the semilinear system (1.2) under different conditions; see [21], [36], and references therein.

Yamamoto and Park in [33] considered control systems described by an abstract parabolic differential equation with uniformly bounded nonlinear term and finite dimensional input operator. Naito in [26] investigated the approximate controllability of semilinear control systems under a range condition of the control action operator and an inequality condition on the system parameters. Seidman in [28] considered the reachable set of the abstract equation in a Banach space. Assuming invariance of the reachable set under affine perturbations, "already a fairly strong controllability hypothesis," various conditions are proved under which nonlinear perturbations also leave the reachable set invariant. Li and Yong in [34] studied the same problem assuming the approximate controllability of the associated linear system under arbitrary perturbation in $L_\infty(I,L(X))$.

In modeling physical processes in bounded domains by controlled PDEs, two types of controls, boundary and internal, are typically used. The boundary controls act upon the system from the outside, while the internal controls act in the interior of the system's space domain. Each of these controls can be both distributed, depending on x and t, and lumped, depending on t only. In the beginning of the 1990's, Fabre, Puel, and Zuazua in [14] proved the global approximate controllability for the second order semilinear parabolic equations with nonlinear term $f(t,x,y)$ of sublinear growth in the variable y at infinity. Later, Imanuvilov [19] proved the global exact controllability for the same equation. This result was improved by Fernandez-Cara [17]. On the other hand, for the case of the nonlinear term including ∇y, the approximate controllability was established only recently by Fernandez and Zuazua [16] and Zuazua [36]. The method of these works involves unique continuation and fixed point techniques combined with a variational approach. The lumped controls were studied by Khapalov [20]. The method used in this paper is quite different from the classical fixed point or implicit function arguments.

On the other hand, the approximate controllability for various class of evolution stochastic differential equations was studied by Dubov and Mordukhovich [12], Bashirov and Mahmudov [6], Mahmudov [24], Barbu and Tessitore [4], and Sirbu and Tessitore [30]. To the best knowledge of the author, the controllability of semilinear stochastic evolution systems has not been studied yet.

In this paper, our purpose is to show approximate controllability of the deterministic and stochastic systems under basic assumptions on the system operators. In particular, we assume the approximate controllability of the associated linear system. To prove the main results we develop a constructive approach for approximate controllability for semilinear evolution (deterministic or stochastic) equations. The method is similar to Tikhonov regularization and is based on the new characterization of a symmetric positive operator in terms of strong (weak) convergence of a sequence of (resolvent) operators.

The paper is organized as follows. In section 2, we present some definitions and prove two theorems about the symmetric operators. In particular, Theorem 2.3 gives a characterization of a symmetric positive operator in terms of strong (weak) convergence of a sequence of operators. These results are used in sections 3 and 4 to obtain approximate controllability for the semilinear deterministic and stochastic evolution equations. In section 3, using some constructive control function, we transfer the controllability problem for semilinear systems into a fixed point problem for an appropriate nonlinear operator in a function space. Using the Schauder fixed point theorem, we guarantee the existence of a fixed point of this operator and study an approximate controllability of the system (1.2). In section 4, we study the approximate controllability of the semilinear stochastic evolution systems with uniformly bounded nonlinear term of Lipschitz type. Finally, in section 5, we discuss two examples.

Remark 1.2. Several comments are in order:

1. The result from section 3 can be applied to both distributed and lumped controls.
2. The result obtained in section 3 in some sense extends those in Khapalov [20] that provide the approximate controllability for the one dimensional heat equation to the space
$$W = \left\{ \phi : \phi(\theta) = \sum_{k=1}^{\infty} \alpha_k e_k(\theta), \ \sum_{k=1}^{\infty} \alpha_k c_k^2 < \infty \right\},$$
endowed with the norm $\|\phi\|^2 = \sum_{k=1}^{\infty} \alpha_k c_k^2$, where $\{c_k\}$ is a nonincreasing sequence of specifically defined positive number if $f(x)$ is globally Lipschitz and sublinear.
3. The result of section 4 is new.
4. Combining the methods of this paper with those developed in Lasiecka and Triggiani [22], one may expect the results of this paper to hold for a class of problem with unbounded control operator B.
5. Combining the methods in Zuazua [36] and the techniques of this paper, one may expect finite and approximate controllability of the deterministic and stochastic evolution systems considered here.
6. The analogue of the results obtained in this paper may be proved for exact null controllability.

2. Positive operators. Let Z be a separable reflexive Banach space, and let Z^* stand for its dual space with respect to the continuous pairing $\langle \cdot, \cdot \rangle$. We may assume, without loss of generality, that Z and Z^* are smooth and strictly convex, by virtue of the renorming theorem (see, for example, [5], [34]). In particular, this implies that the duality mapping J of Z into Z^* given by the relations
$$\|J(z)\| = \|z\|, \ \langle J(z), z \rangle = \|z\|^2 \text{ for all } z \in Z$$
is bijective, demicontinuous (i.e., continuous from Z with a strong topology into Z^* with weak topology), and strictly monotonic. Moreover, $J^{-1} : Z^* \to Z$ is also duality mapping.

An operator $\Gamma : Z^* \to Z$ is symmetric if
$$\langle z_1^*, \Gamma z_2^* \rangle = \langle z_2^*, \Gamma z_1^* \rangle$$
for all $z_1^*, z_2^* \in Z^*$. It is easy to see that Γ is linear and continuous (see [32]). Γ is nonnegative if $\langle z^*, \Gamma z^* \rangle \geq 0$ for all $z^* \in Z^*$.

First, we recall the following result about the structure of a class of nonnegative symmetric operators (see [32]).

LEMMA 2.1. *Let Γ be a nonnegative symmetric operator. Then there exists a Hilbert space H and an operator $A \in \mathcal{L}(Z^*, H)$ such that $A^*A = \Gamma$ and $A(Z^*)$ is dense in H. Furthermore, $A^*(H) \subset X$ and $\overline{\Gamma(Z^*)} = \overline{A^*(H)}$.*

Proof. Set $M = \{z^* \in Z^* \mid \langle z^*, \Gamma z^* \rangle = 0\}$. It is obvious that M is closed subspace of Z^*. Let us define the following inner product in a quotient space Z^*/M:

$$(z_1^* + M, z_2^* + M) = \langle z_1^*, \Gamma z_2^* \rangle .$$

It is clear that this inner product is well defined. Let H denote a completion of the obtained pre-Hilbert space. Let A be a natural embedding Z^* to H. We have

$$\|Az^*\|_H^2 = (z^* + M, z^* + M) = \langle z^*, \Gamma z^* \rangle \leq \|\Gamma\| \, \|z^*\|^2 \quad \text{for all } z^* \in Z^*.$$

So A is a continuous operator and $\|A\| \leq \sqrt{\|\Gamma\|}$ and

$$\langle z_1^*, \Gamma z_2^* \rangle = (z_1^* + M, z_2^* + M) = (Az_1^*, Az_2^*) = \langle z_1^*, A^*Az_2^* \rangle .$$

From here $\Gamma = A^*A$. Now let us show that $A^*(H) \subset \overline{\Gamma(Z^*)}$. Indeed, $A^*(A(Z^*)) = \Gamma(Z^*) \subset Z$ and $A^*(H) = A^*(\overline{A(Z^*)})$, and consequently $A^*(H) \subset \overline{\Gamma(Z^*)} \subset Z$. ∎

LEMMA 2.2. *For every $h \in Z$ and $\alpha > 0$ the equation*

$$\alpha z_\alpha + \Gamma J(z_\alpha) = \alpha h \tag{2.1}$$

has a unique solution $z_\alpha = z_\alpha(h) = \alpha \left(\alpha I + \Gamma_0^T J\right)^{-1}(h)$ and

$$\|z_\alpha(h)\| = \|J(z_\alpha(h))\| \leq \|h\| . \tag{2.2}$$

Proof. First, let us consider the following equation in Z^*:

$$\alpha J^{-1}(z^*) + \Gamma z^* = \alpha h. \tag{2.3}$$

Set

$$\mathcal{A}(z^*) = \alpha J^{-1}(z^*) + \Gamma z^* - \alpha h.$$

The operator $\mathcal{A}$ maps Z^* into Z and has the following properties:

1. $\mathcal{A}$ is demicontinuous since J^{-1} is demicontinuous and Γ is linear continuous (so demicontinuous).

2. $\mathcal{A}$ is strictly monotone, i.e., $\langle z_1^* - z_2^*, \mathcal{A}(z_1^*) - \mathcal{A}(z_2^*) \rangle > 0$ for all $z_1^* \neq z_2^*$ in Z^* since J^{-1} is strictly monotonic.

3. There is a $\rho > 0$ such that

$$\langle z^*, \mathcal{A}(z^*) \rangle > 0 \text{ for all } z^* \in Z^* \text{ such that } \|z^*\| > \rho.$$

Indeed, from the inequality

$$\begin{aligned} \langle z^*, \mathcal{A}(z^*) \rangle &= \left\langle z^*, \alpha J^{-1}(z^*) + \Gamma z^* - \alpha h \right\rangle \\ &\geq \alpha(\|z^*\|^2 - \langle z^*, h \rangle) \geq \alpha(\|z^*\|^2 - \|z^*\| \, \|h\|) \\ &= \alpha \left(\|z^*\| - \|h\|\right) \|z^*\| , \end{aligned}$$

property 3 follows.

By the Minty–Browder theorem (see [29, Theorem 2.2]), there exists a unique solution $z^* \in Z^*$ of $\mathcal{A}(z^*) = 0$. Thus (2.3) has a unique solution z^* in Z^*. Then $z_\alpha = J^{-1}(z^*)$ is a unique solution of (2.1). Solving this equation for z_α, we get

$$z_\alpha = \alpha \left(\alpha I + \Gamma_0^T J\right)^{-1}(h).$$

To prove (2.2), let $z_\alpha = z_\alpha(h)$ be the solution of (2.1). Then

$$\begin{aligned} \alpha \langle J(z_\alpha), z_\alpha \rangle + \langle J(z_\alpha), \Gamma J(z_\alpha) \rangle &= \alpha \langle J(z_\alpha), h \rangle, \\ \alpha \|z_\alpha\|^2 + \langle J(z_\alpha), \Gamma J(z_\alpha) \rangle &= \alpha \langle J(z_\alpha), h \rangle, \\ \alpha \|z_\alpha\|^2 \le \alpha \langle J(z_\alpha), h \rangle &\le \alpha \|z_\alpha\| \|h\|, \\ \|z_\alpha\| = \|J(z_\alpha)\| &\le \|h\|. \end{aligned} \tag{2.4}$$

The lemma is proved. □

THEOREM 2.3. *Let Γ be a symmetric operator. Then the following three conditions are equivalent:*

(i) *Γ is positive; that is, $\langle z^*, \Gamma z^* \rangle > 0$ for all nonzero $z^* \in Z^*$.*

(ii) *For all $h \in Z$, $J(z_\alpha(h))$ converges to the zero as $\alpha \to 0^+$ in the weak topology, where $z_\alpha(h) = \alpha(\alpha I + \Gamma J)^{-1}(h)$ is a solution of (2.1).*

(iii) *For all $h \in Z$, $z_\alpha(h) = \alpha(\alpha I + \Gamma J)^{-1}(h)$ converges to the zero as $\alpha \to 0^+$ in the strong topology.*

Proof. (i) $\Rightarrow$ (ii): Suppose that Γ is positive. From (2.2) it follows that we can extract a subsequence (still denoted by z_α) which is weakly convergent; i.e., there exists $\overline{z}^* \in Z^*$ such that

$$\langle J(z_\alpha), z \rangle \to \langle \overline{z}^*, z \rangle \text{ as } \alpha \to 0^+$$

for all $z \in Z$, and since J is bijective there exists $\overline{z} \in Z$ such that

$$\langle J(z_\alpha), z \rangle \to \langle J(\overline{z}), z \rangle \text{ as } \alpha \to 0^+.$$

Then from (2.1) we have

$$\alpha \langle J(\overline{z}), z_\alpha \rangle + \langle J(\overline{z}), \Gamma J(z_\alpha) \rangle = \alpha \langle J(\overline{z}), h \rangle.$$

Taking the limit in the latter equality, we obtain

$$\langle J(\overline{z}), \Gamma J(\overline{z}) \rangle = 0,$$

and, consequently, $J(\overline{z}) = 0$ by positivity of Γ. In fact, we proved that the limit of every weakly convergent subsequence is zero. Thus the sequence $\{J(z_\alpha)\}$ itself converges weakly to zero.

(ii) $\Rightarrow$ (iii): Assume that $J(z_\alpha) \rightharpoonup 0$ as $\alpha \to 0^+$. Then, dividing (2.4) by α and taking the limit, we obtain

$$\lim_{\alpha \to 0^+} \|z_\alpha\|^2 = \lim_{\alpha \to 0^+} \|J(z_\alpha)\|^2 \le \lim_{\alpha \to 0^+} \langle J(z_\alpha), h \rangle = 0.$$

(iii) $\Rightarrow$ (i): Now, assume that for all $h \in Z$, $\lim_{\alpha \to 0^+} \|z_\alpha(h)\| = 0$, but there exists $z^* \neq 0$ such that $\langle z^*, \Gamma z^* \rangle = 0$. Then by Lemma 2.1

$$\langle z^*, \Gamma z^* \rangle = \langle z^*, A^* A z^* \rangle = \|A z^*\|^2 = 0,$$

which implies that $Az^* = 0$ and, consequently, $\Gamma z^* = A^* A z^* = 0$.

On the other hand, as $J : Z \to Z^*$ is bijective, there exists a unique nonzero $z \in Z$ such that $J(z) = z^*$. Then for $h = z$

$$\alpha z + \Gamma J(z) = \alpha h,$$

since $\Gamma J(z) = \Gamma z^* = 0$. Thus $\lim_{\alpha \to 0^+} z_\alpha(h) = h \neq 0$, which leads to a contradiction. □

Remark 2.4. The analogue of this theorem in Hilbert spaces is proved in [23].

THEOREM 2.5. *Let $\Gamma : Z^* \to Z$ be a positive symmetric operator, and let $h : Z \to Z$ be a nonlinear operator. Assume z_α is a solution of the equation*

$$\alpha z_\alpha + \Gamma J(z_\alpha) = \alpha h(z_\alpha) \tag{2.5}$$

and

$$\left\| h(z_\alpha) - \overline{h} \right\| \to 0 \text{ as } \alpha \to 0^+.$$

Then there exists a subsequence of the sequence $\{z_\alpha\}$ converging strongly to zero as $\alpha \to 0^+$.

Proof. From (2.2) and strong convergence of the sequence $\{h(z_\alpha)\}$, it is easy to see that there exists $C > 0$ such that for all $\alpha > 0$

$$\|z_\alpha\| = \|J(z_\alpha)\| \leq \|h(z_\alpha)\| \leq C.$$

Then we can extract a subsequence, still denoted by z_α, such that

$$J(z_\alpha) \rightharpoonup J(\overline{z}_0) \quad \text{as} \quad \alpha \to 0^+$$

for some $\overline{z}_0 \in Z$. Applying $J(\overline{z}_0)$ to (2.5) and taking the limit, we obtain

$$\begin{aligned} \alpha \langle J(\overline{z}_0), z_\alpha \rangle + \langle J(\overline{z}_0), \Gamma J(z_\alpha) \rangle &= \alpha \langle J(\overline{z}_0), h(z_\alpha) \rangle, \\ \lim_{\alpha \to 0^+} \langle J(\overline{z}_0), \Gamma J(z_\alpha) \rangle &= \langle J(\overline{z}_0), \Gamma J(\overline{z}_0) \rangle = 0, \\ J(\overline{z}_0) &= 0. \end{aligned}$$

So, $J(z_\alpha) \rightharpoonup 0$ as $\alpha \to 0^+$. Now applying $J(z_\alpha)$ to (2.5), dividing through by α, and taking the limit, we obtain

$$\begin{aligned} \|z_\alpha\|^2 + \frac{1}{\alpha} \langle J(z_\alpha), \Gamma J(z_\alpha) \rangle &= \langle J(z_\alpha), h(z_\alpha) \rangle, \\ \lim_{\alpha \to 0^+} \|z_\alpha\|^2 &\leq \lim_{\alpha \to 0^+} \langle J(z_\alpha), h(z_\alpha) \rangle \\ &\leq \lim_{\alpha \to 0^+} \left\langle J(z_\alpha), h(z_\alpha) - \overline{h} \right\rangle + \lim_{\alpha \to 0^+} \left\langle J(z_\alpha), \overline{h} \right\rangle = 0. \end{aligned}$$

The proof is complete. □

3. Deterministic systems. We start this section with the following assumptions.

(A1) X is a separable reflexive Banach space, and U is a separable Hilbert space.

(A2) $A : D(A) \subset X \to X$ generates a compact semigroup $S(t), t > 0$, on X.

(A3) The function $f : I \times X \times U \to X$ is continuous, and there exist functions $\lambda_i(\cdot) \in L_1(I, R^+)$ and $\phi_i(\cdot) \in L_1(X \times U, R^+)$, $i = 1, 2, \ldots, q$, such that

$$\|f(t,x,u)\| \le \sum_{i=1}^{q} \lambda_i(t)\,\phi_i(x,u) \quad \text{for all } (t,x,u) \in I \times X \times U.$$

Next, for convenience, let us introduce the following notation:

$$K = \max\{\|S(t)\| : 0 \le t \le T\}, \quad M = \|B\|, \quad \|\lambda_i\| = \int_0^T \lambda_i(s)\,ds,$$
$$k = \max\{1, MK, MKT\},$$
$$a_i = 3kMK^2\|\lambda_i\|, \quad b_i = 3K\|\lambda_i\|, \quad c_i = \max\{a_i, b_i\},$$
$$d_1 = 3kMK(\|x_T\| + K\|x_0\|), \quad d_2 = 3K\|x_0\|, \quad d = \max\{d_1, d_2\}.$$

(A4) For all $\alpha > 0$, $\limsup_{r\to\infty}(r - \sum_{i=1}^{q} \frac{c_i}{\alpha} \sup\{\phi_i(x,u) : \|(x,u)\| \le r\}) = \infty$.

(AB) For every $h \in X$, $z_\alpha(h) = \alpha\left(\alpha I + \Gamma_0^T J\right)^{-1}(h)$ converges to zero as $\alpha \to 0^+$ in strong topology, where

$$L_0^T u := \int_0^T S(T-s)\,Bu(s)\,ds,$$
$$\Gamma_0^T := \int_0^T S(T-s)\,BB^*S^*(T-s)\,ds = L_0^T\left(L_0^T\right)^*,$$

and $z_\alpha(h)$ is a solution of the equation

$$\alpha z_\alpha + \Gamma_0^T J(z_\alpha) = \alpha h.$$

Remark 3.1. By Theorem 2.3, (AB) holds if and only if $\|\left(L_0^T\right)^* z\| > 0$ for all nonzero $z \in X$. In other words, (AB) holds if and only if the corresponding linear system is approximately controllable.

In section 4, it will be shown that the system (1.2) is approximately controllable if for all $\alpha > 0$ there exists a continuous function $(x,u)(\cdot) \in C(I, X \times U)$ such that

$$\begin{cases} u(t) = B^*S^*(T-t)\,J(\left(\alpha I + \Gamma_0^T J\right)^{-1} p(x(\cdot), u(\cdot))), \\ x(t) = S(t)\,x_0 + \int_0^t S(t-s)\,[Bu(s) + f(s, x(s), u(s))]\,ds, \end{cases} \tag{3.1}$$

where

$$p(x(\cdot), u(\cdot)) = x_T - S(T)\,x_0 - \int_0^T S(T-s)\,f(s, x(s), u(s))\,ds.$$

Having noticed this fact, our goal in this section is to find conditions for solvability of (3.1).

For all $\alpha > 0$, define the operator $\mathcal{P}_\alpha$ on $C(I, X \times U)$ as

$$\mathcal{P}_\alpha(x,u) = (z,v), \tag{3.2}$$

where

$$v(t)\,(=v_\alpha(t)) = B^*S^*(T-t)\,J\left((\alpha I+\Gamma_0^T J)^{-1}p(x(\cdot),u(\cdot))\right), \tag{3.3}$$

$$z(t)\,(=z_\alpha(t)) = S(t)\,x_0+\int_0^t S(t-s)\,(Bv_\alpha(s)+f(s,x(s),u(s)))\,ds, \tag{3.4}$$

$$p(x(\cdot),u(\cdot)) = x_T - S(T)\,x_0-\int_0^T S(T-s)\,f(s,x(s),u(s))\,ds.$$

It will be shown that for all $\alpha>0$ the operator $\mathcal{P}_\alpha$ from $C(I,X\times U)$ into itself has a fixed point.

On Banach space $C(I,X\times U)$ introduce a set

$$Y_r := \{(x,u)(\cdot)\in C(I,X\times U)\mid x(0)=x_0,\ \|(x,u)(\cdot)\|\le r\},$$

where r is the positive constant.

THEOREM 3.2. *Assume assumptions* (A1)–(A4) *are satisfied. Then for all* $0<\alpha\le 1$ *the system* (3.1) *has a solution; that is, the operator* $\mathcal{P}_\alpha$ *has a fixed point.*

Proof. The proof of the theorem is long and technical. We split it into two steps.

Step 1. For arbitrary $\alpha>0$ there is a positive constant $r(\alpha)$ such that $\mathcal{P}_\alpha : Y_{r(\alpha)}\to Y_{r(\alpha)}$.

Let

$$\psi_i(r)=\sup\{\phi_i(x,u):\|(x,u)\|\le r\}.$$

By the assumption (A4), for all $\alpha>0$ there exists $r(\alpha)>0$ such that

$$\frac{d}{\alpha}+\sum_{i=1}^q\frac{c_i}{\alpha}\psi_i(r(\alpha))\le r(\alpha).$$

If $(x,u)(\cdot)\in Y_{r(\alpha)}$, from (3.3) and (3.4) we have

$$\begin{aligned}\|v(t)\| &\le \frac{1}{\alpha}MK\left(\|x_T\|+K\|x_0\|+K\int_0^T\sum_{i=1}^q\lambda_i(s)\,\phi_i(x(s),u(s))\,ds\right)\\ &= \frac{1}{\alpha}MK(\|x_T\|+K\|x_0\|)+\frac{1}{\alpha}MK^2\sum_{i=1}^q\|\lambda_i\|\,\psi_i(r(\alpha))\\ &\le \frac{d}{3k\alpha}+\frac{1}{3k}\sum_{i=1}^q\frac{c_i}{\alpha}\psi_i(r(\alpha)) = \frac{1}{3k}\left(\frac{d}{\alpha}+\sum_{i=1}^q\frac{c_i}{\alpha}\psi_i(r(\alpha))\right)\le\frac{r(\alpha)}{3k},\end{aligned}$$

$$\begin{aligned}\|z(t)\| &\le \frac{d}{3}+KMT\|v\|+K\int_0^t\sum_{i=1}^q\lambda_i(s)\,\phi_i(x(s),u(s))\,ds\\ &\le \frac{d}{3}+k\|v\|+\frac{1}{3}\sum_{i=1}^q c_i\psi_i(r(\alpha))\\ &\le \frac{1}{3}\left[d+\sum_{i=1}^q c_i\psi_i(r(\alpha))\right]+k\|v\|\\ &\le \frac{\alpha r(\alpha)}{3}+\frac{r(\alpha)}{3}\le\frac{2r(\alpha)}{3}.\end{aligned}$$

So

$$\|(\mathcal{P}_\alpha(x,u))(t)\| = \|z(t)\| + \|v(t)\| \le r(\alpha).$$

Hence $\mathcal{P}_\alpha$ maps $Y_{r(\alpha)}$ into itself.

Step 2. For all $\alpha > 0$ the operator $\mathcal{P}_\alpha$ maps $Y_{r(\alpha)}$ into a relatively compact subset of $Y_{r(\alpha)}$, and $\mathcal{P}_\alpha$ has a fixed point.

According to the infinite dimensional version of the Ascoli–Arzela theorem, we have to show that

(i) for arbitrary $t \in I$ the set

$$V(t) = \left\{(\mathcal{P}_\alpha(x,u))(t) \mid (x(\cdot), u(\cdot)) \in Y_{r(\alpha)}\right\}$$

is relatively compact;

(ii) for arbitrary $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\|(\mathcal{P}_\alpha(x,u))(t_1) - (\mathcal{P}_\alpha(x,u))(t_2)\| < \varepsilon$$

if $\|x\| \le r$, $|t_1 - t_2| \le \delta$, $t_1, t_2 \in I$.

Let us prove (i). In fact, the case where $t = 0$ is trivial since $V(0) = \{x_0\}$. So let t, $0 < t \le T$, be a fixed real number, and let η be a given real number satisfying $0 < \eta < t$. Define

$$(\mathcal{P}_\alpha^\eta(x,u))(t) = \left[S(\eta) z(t-\eta), B^* S^*(T-t) J\left((\alpha I + \Gamma_0^T J)^{-1} \mathcal{P}(x(\cdot), u(\cdot))\right)\right].$$

Since $S(\eta)$ is compact and $z(t-\eta)$ and $\mathcal{P}(x(\cdot), u(\cdot))$ are bounded on $Y_{r(\alpha)}$, the set

$$V_\eta(t) = \left\{(\mathcal{P}_\alpha^\eta(x,u))(t) \mid (x,u)(\cdot) \in Y_{r(\alpha)}\right\}$$

is a relatively compact set in $X \times U$; that is, we can find a finite set $\{y_i,\ 1 \le i \le m\}$ in $X \times U$ such that

$$V_\eta(t) \subset \bigcup_{i=1}^m N(y_i, \varepsilon/2),$$

where $N(y_i, \varepsilon/2)$ is an open ball in $X \times U$ with the center at y_i of radius $\varepsilon/2$. On the other hand,

$$\|(\mathcal{P}_\alpha(x,u))(t) - (\mathcal{P}_\alpha^\eta(x,u))(t)\| = \left\| \int_{t-\eta}^t S(t-s)\left[Bu(s) + f(s, x(s), u(s))\right] ds \right\|$$

$$\le \frac{1}{\alpha} K^2 M^2 \left(\|x_T\| + K\|x_0\| + K \int_0^T \sum_{i=1}^q \lambda_i(s)\, \psi_i(r(\alpha))\, ds \right) \eta$$

$$+ K \sum_{i=1}^q \int_{t-\eta}^t \lambda_i(s)\, ds \psi_i(r(\alpha)) \le \frac{\varepsilon}{2}.$$

Consequently,

$$V(t) \subset \bigcup_{i=1}^m N(y_i, \varepsilon).$$

Hence, for each $t \in [0,T]$, $V(t)$ is relatively compact in $X \times U$.

Next, we show (ii). We have to show that $V = \left\{ (\mathcal{P}_\alpha (x,u))(\cdot) \mid (x,u)(\cdot) \in Y_{r(\alpha)} \right\}$ is equicontinuous on $[0,T]$. In fact, for $0 < t_1 < t_2 \le T$, we have

$$\begin{aligned}
\|v(t_1) - v(t_2)\| &\le \|B^* S^*(T-t_1) - B^* S^*(T-t_2)\| \\
&\quad \times \frac{1}{\alpha} \left[\|x_T\| + K\|x_0\| + K \int_0^T \sum_{i=1}^q \lambda_i(s)\, \phi_i(x(s), u(s))\, ds \right] \\
&\le \|B^* S^*(T-t_1) - B^* S^*(T-t_2)\| \\
&\quad \times \frac{1}{\alpha} \left[\|x_T\| + K\|x_0\| + K \sum_{i=1}^q \|\lambda_i\|\, \psi_i(r(\alpha)) \right],
\end{aligned}$$

$$\begin{aligned}
\|z(t_1) - z(t_2)\| &\le \|S(t_1) - S(t_2)\|\, \|x_0\| + KM \int_{t_1}^{t_2} \|v(s)\|\, ds \\
&\quad + M \int_0^{t_1} \|S(t_2 - s) - S(t_1 - s)\|\, \|v(s)\|\, ds \\
&\quad + K \int_{t_1}^{t_2} \sum_{i=1}^q \lambda_i(s)\, \phi_i(x(s), u(s))\, ds \\
&\quad + \int_0^{t_1} \|S(t_2 - s) - S(t_1 - s)\| \sum_{i=1}^q \lambda_i(s)\, \phi_i(x(s), u(s))\, ds \qquad (3.5) \\
&\le \|S(t_1) - S(t_2)\|\, \|x_0\| + KM \int_{t_1}^{t_2} \|v(s)\|\, ds \\
&\quad + M \int_0^{t_1} \|S(t_2 - s) - S(t_1 - s)\|\, \|v(s)\|\, ds \\
&\quad + K \sum_{i=1}^q \int_{t_1}^{t_2} \lambda_i(s)\, ds \psi_i(r(\alpha)) \\
&\quad + \sum_{i=1}^q \int_0^{t_1} \|S(t_2 - s) - S(t_1 - s)\|\, \lambda_i(s)\, ds \psi_i(r(\alpha)) \\
&= I_1 + I_2 + I_3 + I_4 + I_5.
\end{aligned}$$

Moreover, for all $(x(\cdot), u(\cdot)) \in Y_{r(\alpha)}$,

$$\begin{aligned}
\|v\| &\le \frac{1}{\alpha} MK \left(\|x_T\| + K\|x_0\| + K \int_0^T \sum_{i=1}^q \lambda_i(s)\, \phi_i(x(s), u(s))\, ds \right) \\
&\le \frac{1}{\alpha} MK \left(\|x_T\| + K\|x_0\| + K \sum_{i=1}^q \|\lambda_i\|\, \psi_i(r(\alpha)) \right).
\end{aligned}$$

Thus the right-hand side of (3.5) does not depend on particular choices of $(x,u)(\cdot)$. It is clear that $I_2 \to 0$ and $I_4 \to 0$ as $t_1 - t_2 \to 0$. Since the semigroup $S(\cdot)$ is compact, $\|S(t_2 - s) - S(t_1 - s)\| \to 0$ as $t_1 - t_2 \to 0$ for arbitrary t, s such that $t - s > 0$. Then $I_1 \to 0$, and by Lebesgue's dominated convergence theorem $I_3 \to 0$ and $I_5 \to 0$ as $t_1 - t_2 \to 0$. So, we obtain the equicontinuity of V. Thus $\mathcal{P}_\alpha \left[Y_{r(\alpha)} \right]$ is equicontinuous and also bounded. By the Ascoli–Arzela theorem, $\mathcal{P}_\alpha \left[Y_{r(\alpha)} \right]$ is

relatively compact in $C(I, X \times U)$. On the other hand, it is easy to see that for all $\alpha > 0, \mathcal{P}_\alpha$ is continuous on $C(I, X \times U)$. Hence, for all $\alpha > 0, \mathcal{P}_\alpha$ is a compact continuous operator on $C(I, X \times U)$. From the Schauder fixed point theorem, $\mathcal{P}_\alpha$ has a fixed point. □

Consider the following linear system with $f(\cdot) \in L_1(I, X)$:

$$z(t; x_0, u) = S(t) x_0 + \int_0^t S(t-s) [Bu(s) + f(s)] ds. \tag{3.6}$$

LEMMA 3.3. *If*

$$p = x_T - S(T) x_0 - \int_0^T S(T-s) f(s) ds$$

and if $u_\alpha(\cdot) \in L_2(I, U)$ *is a control defined by*

$$u_\alpha(t) = B^* S^*(T-t) J \left(\left(\alpha I + \Gamma_0^T J \right)^{-1} p \right), \tag{3.7}$$

then

$$z(T; x_0, u_\alpha) - x_T = -\alpha \left(\alpha I + \Gamma_0^T J \right)^{-1} p \tag{3.8}$$

and

$$\begin{aligned} z(t; x_0, u_\alpha) &= S(t) x_0 + \int_0^t S(t-s) f(s) ds \\ &\quad + \Gamma_0^t S^*(T-t) J \left(\left(\alpha I + \Gamma_0^T J \right)^{-1} p \right), \end{aligned} \tag{3.9}$$

where $x_T \in X$ *and* $\alpha > 0$ *is a parameter.*

Proof. Inserting (3.7) in (3.6), we have

$$\begin{aligned} z(t; x_0, u_\alpha) &= S(t) x_0 + \int_0^t S(t-s) f(s) ds + \int_0^t S(t-s) B u_\alpha(s) ds \\ &= S(t) x_0 + \int_0^t S(t-s) f(s) ds \\ &\quad + \int_0^t S(t-s) B B^* S^*(T-s) J \left(\left(\alpha I + \Gamma_0^T J \right)^{-1} p \right) ds \\ &= S(t) x_0 + \int_0^t S(t-s) f(s) ds \\ &\quad + \Gamma_0^t S^*(T-t) J \left(\left(\alpha I + \Gamma_0^T J \right)^{-1} p \right). \end{aligned}$$

Writing the latter equation for $t = T$ and solving the obtained one for $z(T; x_0, u_\alpha) - x_T$, we obtain (3.8):

$$\begin{aligned} z(T) &= S(T) x_0 + \int_0^T S(T-s) f(s) ds + \Gamma_0^T J \left(\left(\alpha I + \Gamma_0^T J \right)^{-1} p \right) \\ z(T) - x_T &= S(T) x_0 + \int_0^T S(T-s) f(s) ds - x_T \\ &\quad + \left(-\alpha I + \alpha I + \Gamma_0^T J \right) \left(\alpha I + \Gamma_0^T J \right)^{-1} p \\ &= S(T) x_0 + \int_0^T S(T-s) f(s) ds - x_T \end{aligned}$$

$$
\begin{aligned}
&-\alpha\left(\alpha I+\Gamma_0^T J\right)^{-1} p+\left(\alpha I+\Gamma_0^T J\right)\left(\alpha I+\Gamma_0^T J\right)^{-1} p \\
&=-\alpha\left(\alpha+\Gamma_0^T J\right)^{-1} p.
\end{aligned}
$$

The lemma is proved. □

THEOREM 3.4. *Under the conditions* (A1), (A2), (AB)*, and the following:*

(AF) *The function $f: I\times X\times U\to X$ is continuous and uniformly bounded; i.e., there exists $L>0$ such that*

$$\|f(t,x,u)\|\le L \quad \text{for all } (t,x,u)\in I\times X\times U,$$

the system (1.2) *is approximately controllable.*

Proof. It is obvious that the conditions (A3) and (A4) follow from (AF).

Let $\left(x_\alpha^*,u_\alpha^*\right)(\cdot)$ be a fixed point of $\mathcal{P}_\alpha$ in $Y_{r(\alpha)}$. Then $x_\alpha^*(\cdot)$ is a mild solution of (1.2) on $[0,T]$ under the control

$$u_\alpha^*(t)=B^*S^*(T-t)J\left(\left(\alpha I+\Gamma_0^T J\right)^{-1} p\left(x_\alpha^*(\cdot),u_\alpha^*(\cdot)\right)\right)$$

and satisfies the following equality:

$$x_\alpha^*(T)=x_T-\alpha\left(\alpha I+\Gamma_0^T J\right)^{-1} p\left(x_\alpha^*(\cdot),u_\alpha^*(\cdot)\right).$$

In other words, by Lemma 2.2, $z_\alpha=x_\alpha^*(T)-x_T$ is a solution of the equation

$$\alpha z_\alpha+\Gamma_0^T J(z_\alpha)=\alpha h_\alpha$$

with

$$h_\alpha=-p\left(x_\alpha^*(\cdot),u_\alpha^*(\cdot)\right)=S(T)x_0+\int_0^T S(T-s)f\left(s,x_\alpha^*(s),u_\alpha^*(s)\right)ds-x_T.$$

By (AF)

$$\int_0^T\left\|f\left(s,x_\alpha^*(s),u_\alpha^*(s)\right)\right\|^2 ds\le L^2T,$$

and, consequently, the sequence $\left\{f\left(\cdot,x_\alpha^*(\cdot),u_\alpha^*(\cdot)\right)\right\}$ is bounded and belongs to $L_2(I,X)$. Then there is a subsequence still denoted by $\left\{f\left(\cdot,x_\alpha^*(\cdot),u_\alpha^*(\cdot)\right)\right\}$ that weakly converges to, say, $f(\cdot)$ in $L_2(I,X)$. Then by Corollary 3.3 from [34], we obtain

$$
\begin{aligned}
\|h_\alpha-h\| &=\left\|\int_0^T S(T-s)\left[f\left(s,x_\alpha^*(s),u_\alpha^*(s)\right)-f(s)\right]ds\right\| \\
&\le \sup_{0\le t\le T}\left\|\int_0^t S(t-s)\left[f\left(s,x_\alpha^*(s),u_\alpha^*(s)\right)-f(s)\right]ds\right\|\to 0,
\end{aligned}
$$

where

$$h=S(T)x_0+\int_0^T S(T-s)f(s)\,ds-x_T$$

as $\alpha\to 0^+$ because of the compactness of an operator $g(\cdot)\to\int_0^{\cdot} S(\cdot-s)g(s)\,ds: L_2(I,X)\to C(I,X)$. Then by Theorem 2.5

$$\left\|x_\alpha^*(T)-x_T\right\|=\|z_\alpha\|\to 0$$

as $\alpha\to 0^+$. This gives the approximate controllability. The theorem is proved. □

4. Stochastic systems. In this section, we examine the approximate controllability of the following stochastic semilinear control system in a Hilbert space X:

$$\begin{aligned} dx(t) &= [Ax(t) + Bu(t) + f(t, x(t), u(t))]dt + g(t, x(t), u(t))dw(t), \\ x(0) &= x_0, \quad t \in I = [0, T]. \end{aligned} \tag{4.1}$$

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t \uparrow \subset \mathcal{F}, t \geq 0\}, P)$ denote a complete probability space equipped with a family of nondecreasing subsigma algebras. Let $\mathbf{E}\{\cdot\}$ denote the integration with respect to the measure P. All random processes considered in the paper will be assumed to be strongly $\mathcal{F}_t$-progressively measurable processes unless stated otherwise. Let E be a separable Hilbert space, and let $\{w(t), t \geq 0\}$ be a Wiener process with values in E with covariance operator Q, where Q is a positive nuclear operator in E. We assume that there exist a complete orthonormal system $\{e_k\}$ in E, a bounded sequence of nonnegative real numbers λ_k such that $Qe_k = \lambda_k e_k$, $k = 1, 2, \ldots$, and a sequence $\{\beta_k\}$ of independent Brownian motions such that

$$\langle w(t), e\rangle = \sum_{k=1}^{\infty} \sqrt{\lambda_k}\, \langle e_k, e\rangle\, \beta_k(t), \ e \in E, \ t \geq 0.$$

Further, we assume that $\mathcal{F}_t$ is generated by $\{w(s) : 0 \leq s \leq t\}$. Let $L_2^0 := L_2(Q^{1/2}E, X)$ be the space of all Hilbert–Schmidt operators from $Q^{1/2}E$ to X. The space L_2^0 is a separable Hilbert space, equipped with the norm $\|\Psi\|_Q^2 = tr[\Psi Q \Psi^*]$. We use $L_p(\mathcal{F}, X)$, $1 \leq p \leq \infty$, to denote the Banach space of strongly $\mathcal{F}$-measurable, X-valued random variables satisfying $\mathbf{E}\|x\|_X^2 < \infty$. Since for each $t \geq 0$ the subsigma algebras $\mathcal{F}_t$ are complete, $L_p(\mathcal{F}_t, X)$ are closed subspaces of $L_p(\mathcal{F}, X)$, and hence they are also Banach spaces. Similarly, $L_p^{\mathcal{F}}(I, X)$ will denote the Banach space of $\mathcal{F}_t$-progressively measurable random processes defined on I, taking values from X satisfying $\mathbf{E}\int_I \|x(t)\|_X^2\, dt < \infty$. $C(I, L_2(\Omega, \mathcal{F}, P, X))$ is the Banach space of continuous maps from $I = [0, T]$ into $L_2(\Omega, \mathcal{F}, P, X)$ satisfying the condition $\sup_{t\in I} \mathbf{E}\|x(t)\|^2 < \infty$. $\mathcal{H}_2$ is the closed subspace of $C(I, L_2(\Omega, \mathcal{F}, P, X))$ consisting of measurable and $\mathcal{F}_t$-adapted processes $x(t)$. Then $\mathcal{H}_2$ is a Banach space with the norm topology given by $\|x\|_{\mathcal{H}_2} = (\sup_{t\in I} \mathbf{E}\|x(t)\|^2)^{1/2}$. Further notation will be introduced as required.

Concerning the operators A, B, f, and g, we assume the following hypotheses:

(H1) The operator A generates a compact semigroup $S(\cdot)$.

(H2) The functions $f : I \times X \times U \to X$ and $g : I \times X \times U \to L_2^0$ are continuous, and there exists a constant $L > 0$ such that

$$\begin{aligned} \|f(t, x, u) - f(t, y, u)\| + \|g(t, x, u) - g(t, y, u)\|_Q &\leq L\|x - y\|, \\ \|f(t, x)\| + \|g(t, x)\|_Q &\leq L \ \text{ for all } t \in I, \ x, y \in X. \end{aligned}$$

(H3) For every $0 \leq r < T$, the operator $\alpha\left(\alpha I + \Gamma_r^T\right)^{-1}$ converges to zero as $\alpha \to 0^+$ in the strong operator topology; that is, the linear deterministic system that corresponds to (4.1) is approximately controllable on every $[r, T]$, $0 \leq r < T$ (or on every $[0, t]$, $0 < t \leq T$).

It is clear that under these conditions the system (4.1) admits a mild solution $x(\cdot) \in \mathcal{H}_2$ for any $x_0 \in X$, $u(\cdot) \in L_2^{\mathcal{F}}(I, U)$; see Da Prato and Zabczyk [11].

For any $\alpha > 0$ define the operator

$$\mathcal{P}_\alpha(y, v) = (z, w),$$

where

$$
\begin{aligned}
z(t) &= S(t)x_0 + \int_0^t S(t-r)Bw(r)\,dr + \int_0^t S(t-r)f(r,y(r),v(r))\,dr \\
&\quad + \int_0^t S(t-r)\sigma(r,y(r),v(r))\,dw(r), \\
w(t) &= B^*S^*(T-t)\left[\left(\alpha+\Gamma_0^T\right)^{-1}(\mathbf{E}h - S(T)x_0) + \int_0^t \left(\alpha+\Gamma_r^T\right)^{-1}\varphi(r)\,dw(r)\right] \\
&\quad - B^*S^*(T-t)\int_0^t \left(\alpha+\Gamma_r^T\right)^{-1} S(T-r)f(r,y(r),v(r))\,dr \\
&\quad - B^*S^*(T-t)\int_0^t \left(\alpha+\Gamma_r^T\right)^{-1} S(T-r)\sigma(r,y(r),v(r))\,dw(r),
\end{aligned}
$$

and $\varphi(\cdot) \in L_2^{\mathcal{F}}\left(I, L_2^0\right)$ from the representation $h = \mathbf{E}h + \int_0^T \varphi(s)\,dw(s)$ of $h \in L_2(\mathcal{F}_T, X)$; see [24].

It will be shown that the system (4.1) is approximately controllable if for all $\alpha > 0$ there exists a fixed point of the operator $\mathcal{P}_\alpha$. To show that $\mathcal{P}_\alpha$ has a fixed point, we employ the contraction mapping principle in $\mathcal{H}_2$.

THEOREM 4.1. *For all $\alpha > 0$ the operator $\mathcal{P}_\alpha$ has a unique fixed point in $\mathcal{H}_2$.*

Proof. It is obvious that $\mathcal{P}_\alpha$ maps $\mathcal{H}_2$ into itself. Show that for every $\alpha > 0$ there exists $n > 0$ such that $\mathcal{P}_\alpha^n$ is a contraction mapping. To do this, let $\mathcal{P}_\alpha(y_1, v_1) = (z_1, w_1)$ and $\mathcal{P}_\alpha(y_2, v_2) = (z_2, w_2)$. Then for each $\alpha > 0$ there exists $L(\alpha)$ such that

$$
\mathbf{E}\left\|(z_1,w_1)(t) - (z_2,w_2)(t)\right\|^2 \le tL(\alpha)\left\|(y_1,v_1) - (y_2,v_2)\right\|_{\mathcal{H}_2}^2.
$$

Evidently

$$
\mathbf{E}\left\|\mathcal{P}_\alpha^2(y_1,v_1)(t) - \mathcal{P}_\alpha^2(y_2,v_2)(t)\right\|^2 \le \frac{t^2}{2!}L^2(\alpha)\left\|(y_1,v_1) - (y_2,v_2)\right\|_{\mathcal{H}_2}^2.
$$

By mathematical induction,

$$
\begin{aligned}
&\mathbf{E}\left\|\mathcal{P}_\alpha^n(y_1,v_1)(t) - \mathcal{P}_\alpha^n(y_2,v_2)(t)\right\|^2 \\
&\le L(\alpha)\int_0^t E\left\|\mathcal{P}_\alpha^{n-1}(y_1,v_1)(s) - \mathcal{P}_\alpha^{n-1}(y_2,v_2)(s)\right\|^2 ds \\
&\le \frac{t^n}{n!}L^n(\alpha)\left\|(y_1,v_1) - (y_2,v_2)\right\|_{\mathcal{H}_2}^2.
\end{aligned}
$$

Thus

$$
\sup_{t\in[0,T]} \mathbf{E}\left\|\mathcal{P}_\alpha^n(y_1,v_1)(t) - \mathcal{P}_\alpha^n(y_2,v_2)(t)\right\|^2 \le \frac{(L(\alpha)t)^n}{n!}\left\|(y_1,v_1) - (y_2,v_2)\right\|_{\mathcal{H}_2}^2.
$$

For any fixed $\alpha > 0$ there exists n such that $\frac{(L(\alpha)t)^n}{n!} < 1$. This results that $\mathcal{P}_\alpha^n$ is a contraction mapping for sufficiently large n. Then by the contraction mapping principle the operator $\mathcal{P}_\alpha$ has a unique fixed point in $\mathcal{H}_2$. The theorem is proved. □

THEOREM 4.2. *Assume hypotheses $(H1), (H2)$, and $(H3)$ are satisfied. Then the system (4.1) is approximately controllable.*

Proof. Let $(x_\alpha^*, u_\alpha^*)(\cdot)$ be a fixed point of $\mathcal{P}_\alpha$ in $\mathcal{H}_2$. Using the Fubini theorem, one can show that any fixed point of $\mathcal{P}_\alpha$ satisfies

$$\begin{aligned} x_\alpha^*(T) &= h - \alpha\left(\alpha I + \Gamma_0^T\right)^{-1}(\mathbf{E}h - S(T)x_0) - \int_0^T \alpha\left(\alpha I + \Gamma_r^T\right)^{-1}\varphi(r)\,dw(r) \\ &\quad + \int_0^T \alpha\left(\alpha I + \Gamma_r^T\right)^{-1} S(T-r) f(r, x_\alpha^*(r), u_\alpha^*(r))dr \\ &\quad + \int_0^T \alpha\left(\alpha I + \Gamma_r^T\right)^{-1} S(T-r)\sigma(r, x_\alpha^*(r), u_\alpha^*(r))dw(r). \end{aligned} \tag{4.2}$$

By (H2)

$$\|f(s, x_\alpha^*(s), u_\alpha^*(s))\|^2 + \|g(s, x_\alpha^*(s), u_\alpha^*(s))\|_Q^2 \le L$$

in $I \times \Omega$. Then there is a subsequence, still denoted by $\{f(s, x_\alpha^*(s)), g(s, x_\alpha^*(s))\}$, weakly converging to, say, $(f(s,\omega), g(s,\omega))$ in $X \times L_2^0$. The compactness of $S(t), t > 0$, implies that

$$\begin{aligned} &S(T-s) f(s, x_\alpha^*(s), u_\alpha^*(s)) \to S(T-s) f(s), \\ &S(T-s) g(s, x_\alpha^*(s), u_\alpha^*(s)) \to S(T-s) g(s) \text{ in } I \times \Omega. \end{aligned}$$

On the other hand,

$$\|S(T-s) f(s, x_\alpha^*(s), u_\alpha^*(s))\|^2 + \|S(T-s) g(s, x_\alpha^*(s), u_\alpha^*(s))\|_Q^2 \le K^2 L \text{ in } I \times \Omega.$$

Thus by the Lebesgue dominated convergence theorem

$$\begin{aligned} &\mathbf{E}\int_0^T \|S(T-s)[f(s, x_\alpha^*(s), u_\alpha^*(s)) - f(s)]\|^2 ds \\ &+\mathbf{E}\int_0^T \|S(T-s)[g(s, x_\alpha^*(s), u_\alpha^*(s)) - g(s)]\|^2 ds \to 0 \text{ as } \alpha \to 0^+. \end{aligned}$$

Then having in mind $\left\|\alpha R(\alpha, \Gamma_r^T)\right\|^2 \le 1$ and $\alpha R(\alpha, \Gamma_r^T) \to 0$ in strong operator topology for all $0 \le r < T$, from (4.2) we obtain

$$\begin{aligned} \sqrt{\mathbf{E}\|x_\alpha^*(T) - h\|^2} &\le \left\|\alpha R(\alpha, \Gamma_0^T)(\mathbf{E}h - S(T)x_0)\right\| + \sqrt{\mathbf{E}\int_0^T \|\alpha R(\alpha, \Gamma_r^T)\|^2 \|\varphi(r)\|^2 dr} \\ &+ \sqrt{T\mathbf{E}\int_0^T \|\alpha R(\alpha, \Gamma_r^T)\|^2 \|S(T-r)[f(s, x_\alpha^*(r), u_\alpha^*(r)) - f(r)]\|^2 dr} \\ &+ \sqrt{\mathbf{E}\int_0^T \|\alpha R(\alpha, \Gamma_r^T)\|^2 \|S(T-r)[g(r, x_\alpha^*(r), u_\alpha^*(r)) - g(r)]\|^2 dr} \\ &\to 0 \end{aligned}$$

as $\alpha \to 0^+$. This gives the approximate controllability. The theorem is proved. ∎

5. Applications.

Example 1. Consider a control system governed by the semilinear heat equation

$$\begin{aligned} dx(t,\theta) &= [x_{\theta\theta}(t,\theta) + Bu(t,\theta) + f(t,x(t,\theta))] + dw(t), \\ x(t,0) &= x(t,\pi) = 0, \quad 0 \le t \le T, \ \ 0 < \theta < \pi. \end{aligned} \tag{5.1}$$

Let $X = L_2[0,\pi]$, and let $A : X \to X$ be an operator defined by

$$Az = z''$$

with domain

$$D(A) = \{z \in X \mid z, z' \ \text{ are absolutely continuous } \ z'' \in X, \ z(0) = z(\pi) = 0\}.$$

Then

$$Az = \sum_{n=1}^{\infty} \left(-n^2\right)(z, e_n)\, e_n(\theta), \quad z \in D(A),$$

where $e_n(\theta) = \sqrt{2/\pi}\sin n\theta$, $0 \le x \le \pi$, $n = 1, 2, \ldots$. It is known that A generates an analytic semigroup $S(t), t > 0$, in X and is given by

$$S(t)z = \sum_{n=1}^{\infty} e^{-n^2 t}(z, e_n)\, e_n(\theta), \quad z \in X.$$

Now define an infinite dimensional space

$$U = \left\{ u = \sum_{n=2}^{\infty} u_n e_n(\theta) \,\middle|\, \sum_{n=2}^{\infty} u_n^2 < \infty \right\}$$

with a norm defined by $\|u\| = \left(\sum_{n=2}^{\infty} u_n^2\right)^{1/2}$ and a linear continuous mapping B from U to X as follows:

$$Bu = 2u_2 e_1(\theta) + \sum_{n=2}^{\infty} u_n e_n(\theta).$$

It is obvious that for $u(t,\theta,\omega) = \sum_{n=2}^{\infty} u_n(t,\omega)\, e_n(\theta) \in L_2^{\mathcal{F}}(I,U)$,

$$Bu(t) = 2u_2(t)\, e_1(\theta) + \sum_{n=2}^{\infty} u_n(t)\, e_n(\theta) \in L_2^{\mathcal{F}}(I,X).$$

Moreover,

$$B^* v = (2v_1 + v_2)\, e_2(\theta) + \sum_{n=3}^{\infty} v_n e_n(\theta),$$

$$B^* S^*(t)x = \left(2x_1 e^{-t} + x_2 e^{-4t}\right) e_2(\theta) + \sum_{n=3}^{\infty} x_n e^{-n^2 t} e_n(\theta)$$

for $v = \sum_{n=1}^{\infty} v_n e_n(\theta)$ and $x = \sum_{n=1}^{\infty} x_n e_n(\theta)$. Let

$$\|B^* S^*(t)x\| = 0, \ \ t \in [0,T];$$

it follows that

$$\left\|2x_1e^{-t}+x_2e^{-4t}\right\|^2+\sum_{n=3}^{\infty}\|x_ne^{-n^2t}\|^2=0,\quad t\in[0,T],$$
$$\Rightarrow x_n=0,\quad n=1,2,\cdots\Rightarrow x=0.$$

Thus the deterministic linear system corresponding to (5.1) is approximately controllable on $[0,T]$ for every T, and by Theorem 4.2, the system (5.1) is approximately controllable on $[0,T]$ provided that f satisfies the assumptions (H2).

Example 2. We consider the following homogeneous Dirichlet problem for the semilinear one dimensional heat equation. Let $X=L_2[0,\pi]$, $U=L_2[0,T]$, and $A=d^2/d\theta^2$ with

$$D(A)=\{y\in X\mid y,dy/d\theta \text{ are absolutely continuous and } d^2y/d\theta^2\in X \text{ and } y(0)=y(\pi)=0\}.$$

Put $e_n(\theta)=(2/\pi)^{1/2}\sin n\theta$, $0\le\theta\le\pi$, $n=1,2,\dots$; then $\{e_n(\theta),\ n=1,2,\dots\}$ is an orthogonal basis for X and e_n is the eigenfunction corresponding to the eigenvalue $\lambda_n=n^2$ of the operator A, $n=1,2,\dots$.

Consider a control system governed by the semilinear heat equation

$$\begin{aligned}\frac{\partial x(t,\theta)}{\partial t}&=\frac{\partial^2x(t,\theta)}{\partial\theta^2}+f(x(t,\theta))+\chi_{(l_1,l_2)}(\theta)\,u(t),\\ z(t,0)&=z(t,\pi)=0,\quad 0\le t\le T,\\ z(0,\theta)&=z_0(\theta),\qquad 0\le\theta\le\pi,\end{aligned}\tag{5.2}$$

where $\chi_{(l_1,l_2)}(\theta)$ is the characteristic function of a given subinterval $(l_1,l_2)\subset(0,\pi)$. Now we can define the bounded linear operator $B:R\to L_2(0,\pi)$ by $(Bu)(t)=\chi_{(l_1,l_2)}(\theta)\,u(t)$, and the nonlinear operator f on X is assumed to satisfy (AF).

In [20], Khapalov showed that if $f(z)$ is globally Lipschitz and sublinear, the above control system is approximately controllable in

$$W=\left\{\phi:\phi(\theta)=\sum_{k=1}^{\infty}\alpha_ke_k(\theta),\ \sum_{k=1}^{\infty}\alpha_kc_k^2<\infty\right\},$$

endowed with the norm $\|\phi\|^2=\sum_{k=1}^{\infty}\alpha_kc_k^2$, where $\{c_k\}$ is a nonincreasing sequence of specifically defined positive numbers. Note that $L_2[0,\pi]$ is continuously embedded into W.

If $l_1\pm l_2$ is an irrational number, then the linear system corresponding to (5.2) is approximately controllable, and by Theorem 3.4, the system (5.2) is approximately controllable.

Acknowledgments. The author is grateful to the two anonymous referees and the corresponding editor for suggestions which led to the improvement of the paper and for pointing out [8] and to Professor Agamirza Bashirov of the Eastern Mediterranean University for fruitful discussions.

REFERENCES

[1] N. U. Ahmed and S. Kerbal, *On approximate controllability for semilinear systems*, Nonlinear World, 2 (1995), pp. 53–67.

[2] K. Balachandran, P. Balasubramaniam, and J. P. Dauer, *Controllability of nonlinear integrodifferential systems in Banach space*, J. Optim. Theory Appl., 84 (1995), pp. 83–91.
[3] K. Balachandran and J. P. Dauer, *Controllability of nonlinear systems via fixed-point theorems,* J. Optim. Theory Appl., 53 (1987), pp. 345–352.
[4] V. Barbu and G. Tessitore, *Considerations on the controllability of stochastic linear heat equations*, in Stochastic Partial Differential Equations and Applications, Lecture Notes in Pure and Appl. Math. 227, G. Da Prato and L. Tubaro, eds., Marcel Dekker, New York, 2002, pp. 39–51.
[5] V. Barbu and Th. Precupanu, *Convexity and Optimization in Banach Spaces*, 2nd ed., Mathematics and Its Applications (East European Series) 10, D. Reidel Publishing Co., Dordrecht, The Netherlands, 1986.
[6] A. E. Bashirov and N. I. Mahmudov, *On concepts of controllability for linear deterministic and stochastic systems*, SIAM J. Control Optim., 37 (1999), pp. 1808–1821.
[7] W. Bian, *Approximate controllability for semilinear systems*, Acta Math. Hungar., 81 (1998), pp. 41–57.
[8] W. C. Chewning and T. I. Seidman, *A convergent scheme for boundary control of the heat equation*, SIAM J. Control Optim., 15 (1977), pp. 64–72.
[9] E. N. Chuckwu and S. M. Lenhart, *Controllability questions for nonlinear systems in abstract spaces*, J. Optim. Theory Appl., 68 (1991), pp. 437–462.
[10] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1992.
[11] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Encyclopedia Math. Appl. 44, Cambridge University Press, Cambridge, UK, 1992.
[12] M. A. Dubov and B. S. Mordukhovich, *Theory of controllability of linear stochastic systems*, Differ. Eq., 14 (1978), pp. 1609–1612.
[13] V. N. Do, *Controllability of semilinear systems*, J. Optim. Theory Appl., 65 (1990), pp. 41–52.
[14] C. Fabre, J. P. Puel, and E. Zuazua, *Approximate controllability of the semilinear heat equation*, Proc. Roy. Soc. Edinburgh Sect. A 125 (1995), pp. 31–61.
[15] H. O. Fattorini, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.
[16] L. A. Fernandez and E. Zuazua, *Approximate controllability for semilinear heat equation involving gradient terms*, J. Optim. Theory Appl., 101 (1999), pp. 307–328.
[17] E. Fernandez-Cara, *Null controllability of the semilinear heat equation*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 87–103.
[18] R. K. George, *Approximate controllability of nonautonomous semilinear systems*, Nonlinear Anal., 24 (1995), pp. 1377–1393.
[19] O. Yu. Imanuvilov, *Boundary controllability of parabolic equations*, Sb. Math., 186 (1995), pp. 879–900.
[20] A. Yu. Khapalov, *Approximate controllability properties of the semilinear heat equation with lumped controls*, Int. J. Appl. Math. Comput. Sci., 9 (1999), pp. 751–765.
[21] J. Klamka, *Schauder's fixed-point theorem in nonlinear controllability problems*, Control Cybernet., 29 (2000), pp. 153–165.
[22] I. Lasiecka and R. Triggiani, *Control theory for partial differential equations: Continuous and approximation theories* II, in Abstract Hyperbolic-Like Systems over a Finite Time Horizon, Encyclopedia Math. Appl. 75, Cambridge University Press, Cambridge, UK, 2000, pp. i–xii, 645–1067, I1–I4.
[23] N. I. Mahmudov, *Controllability of linear stochastic systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 724–731.
[24] N. I. Mahmudov, *Controllability of linear stochastic systems in Hilbert spaces*, J. Math. Anal. Appl., 259 (2001), pp. 64–82.
[25] S. Nakagiri and M. Yamamoto, *Controllability and observability of linear retarded systems in Banach space*, Internat. J. Control, 49 (1989), pp. 1489–1504.
[26] K. Naito, *Controllability of semilinear control systems dominated by the linear part*, SIAM J. Control Optim., 25 (1987), pp. 715–722.
[27] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[28] T. I. Seidman, *Invariance of the reachable set under nonlinear perturbations*, SIAM J. Control Optim., 25 (1987), pp. 1173–1191.
[29] R. E. Showalter, *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*, AMS, Providence, RI, 1997.
[30] M. Sirbu and G. Tessitore, *Null controllability of an infinite dimensional SDE with state and control-dependent noise*, Systems Control Lett., 44 (2001), pp. 385–394.

[31] R. Triggiani, *Controllability and observability in Banach space with bounded operators*, SIAM J. Control, 13 (1975), pp. 462–491.

[32] H. H. Vahaniya, V. I. Tarieladze, and C. A. Chobanian, *Probability Distributions in Banach Spaces*, Nauka, Moscow, 1985 (in Russian).

[33] M. Yamamoto and J. Y. Park, *Controllability for parabolic equations with uniformly bounded nonlinear terms*, J. Optim. Theory Appl., 66 (1990), pp. 515–532.

[34] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser Boston, Boston, 1995.

[35] H. X. Zhou, *Approximate controllability for a class of semilinear abstract equations*, SIAM J. Control Optim., 21 (1983), pp. 551–565.

[36] E. Zuazua, *Controllability of partial differential equations and its semi-discrete approximations*, Discrete Contin. Dyn. Syst., 8 (2002), pp. 469–513.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1623–1642

VERIFICATION THEOREMS FOR HAMILTON–JACOBI–BELLMAN EQUATIONS*

MAURO GARAVELLO†

Abstract. We study an optimal control problem in Bolza form, and we consider the value function associated to this problem. We prove two verification theorems which ensure that, if a function W satisfies some suitable weak continuity assumptions and a Hamilton–Jacobi–Bellman inequality outside a countably $\mathcal{H}^n$-rectifiable set, then it is lower than or equal to the value function. These results can be used for optimal synthesis approach.

Key words. verification theorem, optimal control, HJB equation, value function, viscosity solution

AMS subject classifications. 49K15, 93C15, 49L25

DOI. S0363012902392688

1. Introduction.

In this paper we consider a control system of the type

$$\dot{x} = f(t, x, u), \qquad u \in U, \tag{1.1}$$

where $x \in \mathbb{R}^n$ is the state, $U \subset \mathbb{R}^q$ is the control space, and f is the controlled dynamic. Given a target $S \subset \mathbb{R}^n$, a running cost $L(t, x, u)$, a final cost $\psi(t, x)$, and an initial condition (t_0, x_0), we consider the optimal control problem in Bolza form consisting in minimizing the integral of L summed with the value of ψ at final points for trajectories that start at x_0 at time t_0 and reach the target S. We define in the usual way the value function $V(t_0, x_0)$ to be the infimum of the problem with initial condition (t_0, x_0). It is well known that, under special conditions on the data of the problem, V satisfies the dynamic programming principle. This is the key point in order to prove that V is a viscosity solution of the Hamilton–Jacobi–Bellman (HJB) equation; see [2]. In general, it may happen that V is not the unique viscosity solution for this equation. With some more requirements on the Hamiltonian and on the regularity of the value function, V becomes the unique viscosity solution to the HJB equation; see [2].

Therefore, given a function W with suitable properties, it is possible to determine if W coincides with the value function, checking if it is a viscosity solution to the HJB equation. This type of theorem, called a verification theorem, is useful, for example, when a candidate value function is produced by means of the construction of a synthesis [19]. It is then natural to ask for minimal conditions under which a function W coincides with the value function. If we know that W was obtained via a synthesis, then the inequality $W \geq V$ is granted by construction, and thus we make this assumption. Then, for W to coincide with the value function, we prove it is sufficient that, outside a rectifiable set of codimension one, W is differentiable and it satisfies an HJB inequality in the classical sense. Moreover, we make use of only some weak continuity assumptions, already used in [19] to prove optimality of a regular extremal synthesis; see Theorems 5.2 and 6.1 for details. A first result in this direction can be found in [12], where the HJB inequality is asked outside a locally

*Received by the editors May 8, 2002; accepted for publication (in revised form) May 4, 2003; published electronically November 6, 2003.
http://www.siam.org/journals/sicon/42-5/39268.html
†SISSA–ISAS, Via Beirut, 2–4, 34014 Trieste, Italy (mgarav@sissa.it).

finite collection of regular manifolds of positive codimension (under more restrictive continuity assumptions). Notice that, for an optimal control problem, if the value function is also semiconcave, it is differentiable outside a countably $\mathcal{H}^n$-rectifiable set; see [1, 9, 21].

We start considering the main assumptions for the problem and present two technical lemmas, one of which deals with the cardinality of the intersections between admissible trajectories and a countably $\mathcal{H}^n$-rectifiable set, while the other gives some conditions to ensure the monotonicity of a real valued function. Also, we state, without proofs, two propositions dealing with the properties of the solution to (1.1) and, in particular, dealing with existence, uniqueness, and continuous dependence on data.

Then, in section 3, we recall briefly the synthesis approach and various results available in the literature for comparison. Some examples of regular optimal synthesis, to which our main results are applicable, are given.

The first case we treat is the problem of finite time. We define a value function as the infimum, over all admissible trajectories reaching the target in finite time. The main result of this part is Theorem 5.2, which permits to verify if the function W is lower than or equal to the value function.

Next, we consider the infinite time problem. In this case the value function (6.1) is defined as the infimum of the cost functional over all admissible trajectories reaching the target in infinite time. The main result of this section is Theorem 6.1, which gives sufficient conditions on the function W to ensure the inequality $W \leq V$, where V is the value function. In this case, for a technical reason, we consider a suitable neighborhood S_1 of the target S, and we suppose that the final cost ψ is defined on S_1 in order to give sense to the limit in the definition of the value function (6.1). As a corollary of Theorems 5.2 and 6.1, we can treat a mixed case (see also [18]), considering at the same time the trajectories reaching the target both in finite time and in infinite time.

A key ingredient for Theorems 5.2 and 6.1 is the positiveness of the Lagrangian L, in order to prevent some bad phenomena such as the permanence of the system for an arbitrary interval of times in a region where L is negative making the value function equal to $-\infty$, as we see in Example 4. More precisely, it is not necessary to suppose L positive in the whole space, but some relaxed assumptions can be made, as we see in Remark 4.

This paper ends with an appendix, where we give the definition of a noncontinuous viscosity solution as in [2] and we state Theorem A.7, which ensures that, under suitable assumptions, the value functions (2.4) and (6.1) are viscosity solutions to the HJB equation.

2. Preliminaries. We consider a control system

$$\dot{x}(t) = f(t, x(t), u(t)), \quad (t, x) \in \Omega, \quad u(t) \in U, \tag{2.1}$$

where the following hold:

(A-1) Ω is an open and connected subset of $\mathbb{R} \times \mathbb{R}^n$.

(A-2) U is a nonempty subset of $\mathbb{R}^q$ for some $q \geq 1$, $q \in \mathbb{N}$.

(A-3) $\mathcal{U} = L^p(\mathbb{R}; U)$ with $1 \leq p < +\infty$ is the set of admissible controls.

(A-4) $f : \Omega \times U \to \mathbb{R}^n$ is measurable in t, continuous in (x, u), differentiable in x, and, for each $u \in U$, $D_x f(\cdot, \cdot, u)$, bounded on compact sets. Moreover, there exists $\varphi_1 : \mathbb{R} \to \mathbb{R}^+$ integrable, and for every K, compact subset of Ω, there exist a modulus

of continuity ω_K and a constant $L_K > 0$ such that, if $(t,x) \in K$ and $(t,y) \in K$, then for all u

$$\begin{cases} |f(t,x,u) - f(t,y,u)| \le \omega_K(|x-y|), \\ (f(t,x,u) - f(t,y,u)) \cdot (x-y) \le L_K |x-y|^2, \\ |f(t,x,u)| \le L_K(\varphi_1(t) + |u|^p). \end{cases} \tag{2.2}$$

We consider a function $L : \Omega \times U \to \mathbb{R}$ and assume the following:

(A-5) L is measurable in t and continuous in (x,u). Moreover, there exist $\varphi_2 : \mathbb{R} \to \mathbb{R}^+$ integrable and, for every $R \ge 0$, $C_R \ge 0$ such that

$$|L(t,x,u)| \le C_R(\varphi_2(t) + |u|^p), \quad |(t,x)| \le R. \tag{2.3}$$

In this paper we indicate with $x(\cdot\,; u, t_0, x_0)$ the solution to (2.1) such that $x(t_0;u,t_0,x_0) = x_0$. Define the value function

$$V(t_0,x_0) := \inf_{\substack{u \in \mathcal{U} \\ (T,x(T;u,t_0,x_0)) \in S}} \left\{ \int_{t_0}^{T} L(s, x(s;u,t_0,x_0), u(s)) ds + \psi(T, x(T;u,t_0,x_0)) \right\}, \tag{2.4}$$

where S, the target, is a closed subset of $\mathbb{R} \times \mathbb{R}^n$ contained in Ω, and $\psi : S \to \mathbb{R}$ is the final cost. We recall the following definition.

DEFINITION 2.1. *A subset A of Ω is a* countably $\mathcal{H}^n$-rectifiable set *if there exist A_1 and A_2 such that $A = A_1 \cup A_2$, A_1 is a finite or countable union of connected $\mathcal{C}^1$ submanifolds of positive codimension, and $\mathcal{H}^n(A_2) = 0$, where $\mathcal{H}^k$ is the k-dimensional Hausdorff measure.*

3. Examples of syntheses. In the next sections, we give sufficient conditions for a candidate value function W to coincide with V. Besides some regularity conditions, we require an HJB inequality outside a countably $\mathcal{H}^n$-rectifiable set. This regularity is shared by every function W obtained from a regular synthesis; thus it can be used to prove the optimality of the synthesis itself. In this section, we give various examples to which Theorem 5.2 is applicable. First, we need some definitions.

DEFINITION 3.1. *A synthesis Γ is a collection $\{(x_{(\bar t,\bar y)}(\cdot), u_{(\bar t,\bar y)})\}_{(\bar t,\bar y)\in\Omega}$ such that $x_{(\bar t,\bar y)}(\cdot) = x(\cdot; u_{(\bar t,\bar y)}, \bar t, \bar y) : [\bar t, \tau(\bar t,\bar y)] \to \mathbb{R}^n$, $u_{(\bar t,\bar y)} \in \mathcal{U}$, for every $(\bar t,\bar y) \in \Omega$, $x_{(\bar t,\bar y)}(\tau(\bar t,\bar y)) \in S$, and for every $t \in [\bar t, \tau(\bar t,\bar y)]$*

$$u_{(t,x_{(\bar t,\bar y)}(t))}(s) = u_{(\bar t,\bar y)}(s+t) \quad a.e.$$

and

$$x_{(t,x_{(\bar t,\bar y)}(t))}(\cdot) = x_{(\bar t,\bar y)}(\cdot + t).$$

DEFINITION 3.2. *A synthesis Γ is optimal if every $u_{(\bar t,\bar y)}$ is an optimal control.*

There is a standard method in geometric control theory to construct an optimal synthesis; see [4]. This consists of four steps: (1) using the Pontryagin maximum principle and other geometric tools to study the properties of optimal trajectories, (2) determining a finite dimensional family $\mathcal{F}$ of extremal trajectories sufficient for optimality, i.e., such that for each initial data $(\bar t,\bar y) \in \Omega$ there exists an element of $\mathcal{F}$ solving the corresponding optimal control problem, (3) constructing a synthesis formed by extremal trajectories, and (4) proving its optimality. In many cases, for autonomous systems, it happens that the extremal synthesis is associated to a feedback

$u : \mathbb{R}^n \to U$ that is smooth on each stratum of a stratification; see [19] for details. Roughly speaking, a stratification is a locally finite collection of disjoint regular submanifolds of various dimensions that is a partition and such that the boundary of each manifold is a union of manifolds of higher codimensions. In this case, the synthesis is called regular in the sense of Boltyanskii–Brunovský; see [3, 8, 19].

Step (4) of the geometric control approach can thus be obtained in essentially two ways: either by using the regularity of the synthesis (see [19]) or by proving that the candidate value function W associated to the synthesis coincides with V. The latter is exploited in [12] for a continuous W, defined on a subset of $\mathbb{R}^n$, that is differentiable and satisfies the HJB equation outside a locally finite union of smooth submanifolds of positive codimension. Then the optimality is granted for initial points for which all admissible trajectories remain in the domain of W. A mild generalization is obtained in [5], where trajectories can exit the domain of W, but the boundary of the domain of W is a level set of W itself. Another approach is one of nonsmooth analysis, using the various verification theorems that can be proved; see, for example, [20].

Our main results (see Theorems 5.2 and 6.1) generalize previous results in the following way:

1. As in [5], we assume that W can be defined on a subset and the boundary of its domain is a level curve of W.

2. We ask W to be differentiable and satisfy the HJB equation only outside a countably $\mathcal{H}^n$-rectifiable set.

3. W is only lower semicontinuous (satisfying other weak continuity assumptions).

A direct comparison with results of nonsmooth analysis is difficult. However, we point out that the value function fails in general to be locally Lipschitz continuous (see Example 1) for regular synthesis. In case of locally Lipschitz regularity, our result is a consequence of those obtained by nonsmooth analysis methods; see, for example, [10, 20].

We now give some examples to illustrate the applicability of our results. A whole class of examples can be found in [6, 17]. The first example shows a typical regular synthesis with a not locally Lipschitz continuous value function. In the second, the value function is not continuous, and it is differentiable only outside a countably $\mathcal{H}^n$-rectifiable set. The last example shows the well-known Fuller phenomenon. In this case, optimal trajectories have an infinite number of switchings, and the methods of Boltyanskii–Brunovský do not work (while the method developed in [19] does work in the case of the Fuller phenomenon).

Example 1. Let $x \in \mathbb{R}$ and $u \in [-1, 1]$. Consider the control system

$$\ddot{x} + x = u$$

and the problem of reaching the origin in minimum time. If we define $x_1 = x$ and $x_2 = \dot{x}$, we obtain the following first-order system:

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = -x_1 + u. \end{cases} \tag{3.1}$$

Every optimal trajectory is a bang-bang trajectory, i.e., formed by arcs corresponding to control $+1$ or -1. The synthesis is illustrated in Figure 3.1. There are some "switching curves":

(i) all semicircles of radius 1 contained in $\{(x_1, x_2) : x_2 \le 0\}$ and centered at $(2n + 1, 0)$, with $n \in \mathbb{N} \setminus \{0\}$;

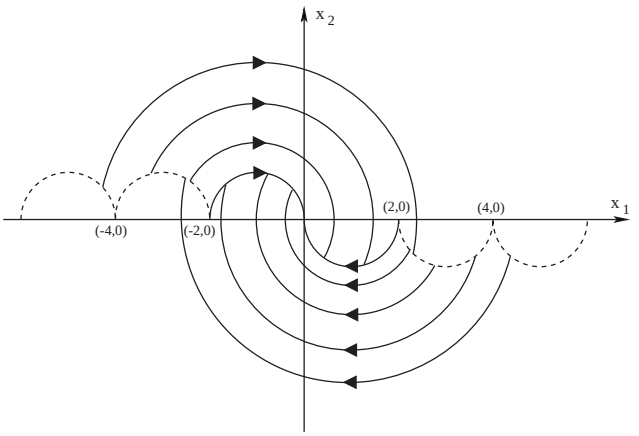


FIG. 3.1. *Synthesis of system* (3.1).

(ii) all semicircles of radius 1 contained in $\{(x_1, x_2) : x_2 \geq 0\}$ and centered at $(-2n-1, 0)$, with $n \in \mathbb{N} \setminus \{0\}$.

Optimal trajectories switch along these curves, i.e., change control from $+1$ to -1 or vice versa. Let $\gamma^{\pm}$ be the trajectory that switches at points $(\pm 2, 0)$ (defined, say, on $[-\infty, 0]$). Then the value function is not locally Lipschitz continuous at any point of supp $(\gamma^{\pm})$, but it satisfies all the hypotheses of Theorem 5.2.

Example 2. Let $\Omega = \mathbb{R}^2$, $f \equiv 0$, and $L \equiv 1$. Consider the target

$$S = \{(t, x) : x \neq 0, t = \sin(1/x)\} \cup \{x = 0, -1 \leq t \leq 1\} \cup \{t \geq 1\}$$

and the final cost ψ constantly equal to 0. The value function for this problem is given by

$$V(t,x) = \begin{cases} \sin(1/x) - t & \text{if} \quad x \neq 0, t \leq \sin(1/x), \\ 1 - t & \text{if} \quad x \neq 0, \sin(1/x) < t < 1, \\ 0 & \text{if} \quad t \geq 1, \\ -1 - t & \text{if} \quad x = 0, t \leq -1, \\ 0 & \text{if} \quad x = 0, -1 < t < 1. \end{cases}$$

This function satisfies all the hypotheses of Theorem 5.2, and clearly it is not continuous. Moreover, it is differentiable outside a countably $\mathcal{H}^n$-rectifiable set A, which is not a locally finite union of regular manifolds.

Example 3 (Fuller phenomenon). Let us consider the system

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = u, \end{cases}$$

with $|u| \leq 1$, $\Omega = \mathbb{R} \times \mathbb{R}^2$, $S = \mathbb{R} \times \{0\}$, $\psi \equiv 0$, and $L(t, x_1, x_2, u) = x_1^2$. This problem is well known in the literature; see, for example, [22]. Every optimal trajectory is composed by an infinite number of bang-bang arcs, while the time for reaching the origin of $\mathbb{R}^2$ is finite. There are two switching curves ζ^+ and ζ^- which separate $\mathbb{R}^2$ into two regions Z^+ and Z^-, where the optimal trajectory uses, respectively, the controls $u = +1$ and $u = -1$; see Figure 3.2. The value function of this problem satisfies all the hypotheses of Theorem 5.2.

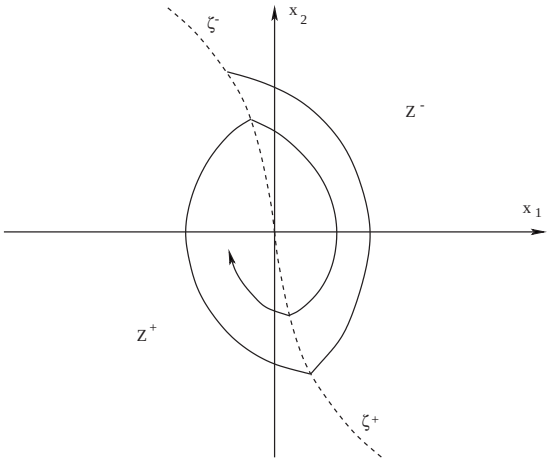


FIG. 3.2. *Synthesis of Fuller phenomenon.*

4. Some useful results. We start by recalling without proofs some classical results about ODEs.

PROPOSITION 4.1 (local existence and uniqueness). *Assume* (A-1)–(A-4). *With fixed* $u \in \mathcal{U}$ *and* $(t_0, x_0) \in \Omega$, *there exist* $\delta > 0$ *and a unique absolutely continuous function* $x(\cdot; u, t_0, x_0) : [t_0, t_0 + \delta] \to \mathbb{R}^n$ *solution to* (2.1).

PROPOSITION 4.2 (continuous dependence on data). *Assume* (A-1)–(A-4). *Let* $(t_0, x_0) \in \Omega$, $(t_0, x_n) \in \Omega$ *for every* $n \in \mathbb{N}$ *and* $u \in \mathcal{U}$, $u_n \in \mathcal{U}$ *for every* $n \in \mathbb{N}$. *Let us suppose that there exists a time* $T > t_0$ *such that* $x(\cdot; u, t_0, x_0)$ *and* $x(\cdot; u_n, t_0, x_n)$ *are defined in* $[t_0, T]$. *If* $x_n \to x_0$ *and* $u_n \to u$ *in the strong topology of* $L^p([t_0, T]; U)$ *as* $n \to +\infty$, *then* $x(\cdot; u_n, t_0, x_n) \to x(\cdot; u, t_0, x_0)$ *uniformly in* $[t_0, T]$ *as* $n \to +\infty$.

Now, we present two technical lemmas used to prove the theorems of the next sections.

LEMMA 4.3. *Fix an element* $\omega \in U$, $t' < t''$ *and* $x \in \mathbb{R}^n$ *with* $(t'', x) \in \Omega$. *Assume that there exists* $\mathcal{W}$, *an open neighborhood of* x *in* $\mathbb{R}^n$, *such that* $\zeta^y(\cdot)$, *the solution to* $\dot{\zeta}^y(t) = f(t, \zeta^y(t), \omega)$ *with* $\zeta^y(t'') = y$, *is defined on* $[t', t'']$ *for any* $y \in \mathcal{W}$ *and* $(t, \zeta^y(t)) \in \Omega$ *for all* $t \in [t', t'']$. *Let* A *be a countable* $\mathcal{H}^n$*-rectifiable set. Then for a.e.* $y \in \mathcal{W}$ *the set* $B^y := \{t \in [t', t''] : (t, \zeta^y(t)) \in A\}$ *is finite or countable.*

This lemma is a slight generalization of a result proved in Theorem 2.14 of [19], since here we consider the trajectory coupled with time.

Proof. We can write $A = A_1 \cup A_2$, where $A_1 = \cup_j M_j$ and $\{M_j\}_{j \in J}$ is a finite or countable family of connected submanifolds of $\mathbb{R}^{n+1}$ of codimension $d_j > 0$, and $\mathcal{H}^n(A_2) = 0$. After replacing each M_j by a finite or countable family of open submanifolds of M_j, we may assume that the M_j are embedded. Define $\widetilde{\mathcal{W}} :=]t', t''[\times \mathcal{W}$, and let Φ be the map $\widetilde{\mathcal{W}} \ni (t, y) \mapsto (t, \zeta^y(t)) \in \Omega$. The Jacobian of Φ is

$$\mathbf{J}\Phi = \left(\begin{array}{c|c} 1 & 0 \cdots\cdots 0 \\ \hline \mathbf{b} & \mathbf{V}^{\zeta}(t; t', \mathbf{Id}) \end{array}\right), \tag{4.1}$$

where $\mathbf{b}$ is the column vector $f(t, \zeta^y(t), \omega)$ and $\mathbf{V}^\zeta(t; t', \mathbf{Id})$ is the fundamental matrix solution to the linear system

$$\dot{v}(t) = -D_x f(t, \zeta^y(-t + t' + t''), w) \cdot v(t) \tag{4.2}$$

such that $\mathbf{V}^\zeta(t'; t', \mathbf{Id}) = \mathbf{Id}$. So the determinant of $\mathbf{J}\Phi$ is equal to the determinant of $\mathbf{V}^\zeta(t; t', \mathbf{Id})$, which is equal to $\exp \int_{t'}^t \operatorname{tr}(-D_x f(s, \zeta^y(-s+t'+t''), \omega))ds$, by Liouville's theorem (see [14]). In particular, $\det(\mathbf{J}\Phi)$ is strictly positive for any $t \in [t', t'']$. Moreover, by (A-4) $\operatorname{tr}(-D_x f)$ is bounded on compact sets and then there exist $c > 0$, $C > 0$ such that $0 < c \leq \det(\mathbf{J}\Phi) \leq C$.

So Φ is a Lipschitz diffeomorphism. In particular, we have $\mathcal{H}^n(\Phi^{-1}(A_2)) = 0$. Now for each j consider $\widetilde{M}_j := \Phi^{-1}(M_j)$. It is an embedded submanifold of codimension $d_j > 0$. Let $\Pi : \widetilde{\mathcal{W}} \to \mathcal{W}$ be the canonical projection. Consider the set S_j consisting of the points $s \in \widetilde{M}_j$ such that $\Pi|_{\widetilde{M}_j}$ is not regular. Thus, by Sard's theorem, $\mathcal{L}^n(\Pi(S_j)) = 0$. Moreover, $\mathcal{H}^n(\Pi(\Phi^{-1}(A_2))) = 0$. So the set $\mathcal{B} := \Pi(\Phi^{-1}(A_2)) \cup (\bigcup_j \Pi(S_j))$ has Lebesgue measure 0 in $\mathbb{R}^n$.

Let $y \in \mathcal{W} \setminus \mathcal{B}$. Then $(t, \zeta^y(t)) \notin A_2$ if $t' < t < t''$. To obtain the thesis, it is sufficient to show that, for each j, the set $E_j = \{t \in]t', t''[: (t, \zeta^y(t)) \in M_j\}$ is at most countable. Fix j and suppose $t \in E_j$. $\widetilde{M}_j$ has codimension $d_j > 0$, so the dimension ν_j of $\widetilde{M}_j$ is less than or equal to n. Since $y \notin \mathcal{B}$, the map $d\Pi(t, y) : T_{(t,y)}\widetilde{M}_j \to \mathbb{R}^n$ is onto; thus $\nu_j = n$ and $d\Pi(t, y)$ is injective. Obviously $d\Pi(t, y)(\frac{\partial}{\partial t}) = 0$, so $\frac{\partial}{\partial t} \notin T_{(t,y)}\widetilde{M}_j$ and, consequently, $(\tilde{t}, y) \notin \widetilde{M}_j$ if $0 < |\tilde{t} - t| \leq \varepsilon$ for $\varepsilon > 0$ sufficiently small. Therefore, t is an isolated point of E_j, and so the lemma is proved. □

LEMMA 4.4. *Let g be a real valued function on a compact interval $[a, b]$. Assume that there exists a finite or countable subset E of $[a, b]$ with the following properties:*

(i) $\liminf_{h \downarrow 0} \frac{g(x+h)-g(x)}{h} \geq 0$ *for all* $x \in [a, b[\setminus E$,
(ii) $\liminf_{h \downarrow 0} g(x + h) \geq g(x)$ *for all* $x \in [a, b[$,
(iii) $\liminf_{h \downarrow 0} g(x - h) \leq g(x)$ *for all* $x \in]a, b]$.

Then $g(b) \geq g(a)$.

For a proof of this, lemma, see [19, Lemma B.1].

5. Problem with finite time. We indicate with ∂Q the topological boundary of an arbitrary $Q \subseteq \mathbb{R} \times \mathbb{R}^n$. Before stating the theorem, we need the following definition.

DEFINITION 5.1. *Suppose that we have a time-varying Lipschitz-continuous vector field X on $\mathbb{R}^n$ and $W : \Omega \to \mathbb{R} \cup \{\pm\infty\}$. We say that W has the* no downward jumps *(NDJ) property along X if for any $[a, b] \ni t \mapsto \gamma(t)$, solution to $\dot{\gamma}(t) = X(t, \gamma(t))$ such that $(t, \gamma(t)) \in \Omega$ for all $t \in [a, b]$, we have $\liminf_{h \downarrow 0} W(t - h, \gamma(t - h)) \leq W(t, \gamma(t))$ whenever $t \in]a, b]$.*

THEOREM 5.2. *Suppose* (A-1)–(A-5) *hold. Let $Q \subseteq \Omega$ be an open subset containing S. Let $W : \overline{Q} \to \mathbb{R}$ be a lower semicontinuous function such that the following hold:*

(i) *W has the NDJ property along every time-varying vector field of the type $f(t, x, u)$ with $u \in U$ fixed and for each t*

$$\text{ess–liminf}_{y \to x} W(t, y) \leq W(t, x).$$

(ii) $W \le \psi$ *on* S.
(iii) *At every point* $(t,x) \in \partial Q$ *one has*

$$W(t,x) = \sup_{(s,y)\in Q} W(s,y).$$

(iv) *There exists a countably* $\mathcal{H}^n$*-rectifiable set* $A \subseteq \Omega$ *such that* W *is differentiable on* $Q \setminus A$ *and satisfies*

$$W_s(s,y) + \inf_{\omega\in U} \{W_y(s,y)\cdot f(s,y,\omega) + L(s,y,\omega)\} \ge 0 \quad \text{on } Q\setminus A.$$

(v) $L \ge 0$.
Then $W \le V$ *on* Q. *If* $Q = \Omega$, *we can drop hypotheses* (iii) *and* (v).

Proof. Suppose by contradiction that there exists $(t_0,x_0)\in Q$ such that $W(t_0,x_0) > V(t_0,x_0)$. In particular, $V(t_0,x_0) < +\infty$. First, let us consider the case $V(t_0,x_0) > -\infty$. So we can find $\varepsilon > 0$, $\delta > 0$ such that

$$V(t_0,x_0) \le W(t_0,x_0) - 2\varepsilon \tag{5.1}$$

and, by the lower semicontinuity of W,

$$|x - x_0| < \delta \quad \Rightarrow \quad W(t_0,x) > V(t_0,x_0) + \varepsilon. \tag{5.2}$$

We can find $u^* \in \mathcal{U}$ such that $x^*(\cdot) := x(\cdot; u^*, t_0, x_0)$ satisfies $(T, x^*(T)) \in S$ and

$$\int_{t_0}^{T} L(s,x^*(s),u^*(s))ds + \psi(T,x^*(T)) \le V(t_0,x_0) + \frac{\varepsilon}{2}. \tag{5.3}$$

Moreover, for every $l \in \mathbb{N}$ there exists $u_l \in \mathcal{U}$ such that $\|u_l - u^*\|_{L^p([t_0,T])} \le \frac{1}{l}$, and u_l is piecewise constant and left continuous. By [7, Theorem IV.9], there exist a subsequence of $(u_l)_l$, denoted again by $(u_l)_l$, and a function $h \in L^p([t_0,T])$ such that $|u_l| \le h$ a.e. and u_l converges to u^* a.e. as $l \to +\infty$. Hence, if we denote by $x_l(\cdot)$ the trajectory $x(\cdot; u_l, T, x^*(T))$, for l sufficiently big we have (see Proposition 4.2)

$$|x_l(t) - x^*(t)| < \frac{\delta}{2} \qquad \forall t \in [t_0,T] \tag{5.4}$$

and

$$\left| \int_{t_0}^{T} [L(s,x_l(s),u_l(s)) - L(s,x^*(s),u^*(s))]ds \right| \le \frac{\varepsilon}{2}. \tag{5.5}$$

Fix l such that (5.4) and (5.5) hold and an interval $]t',t'']$ such that $u_l(t) \equiv \omega$ on $]t',t'']$. Suppose that $(t,x_l(t)) \in Q$ for all $t \in [t',t'']$. Let $\zeta^y(t)$ be the trajectory associated to the constant control ω such that $\zeta^y(t'') = y$. By the fact that $d(\partial Q, \{(t,x_l(t)) : t \in [t',t'']\}) > 0$, we can find an open neighborhood $\mathcal{W}$ of $x_l(t'')$ in $\mathbb{R}^n$ such that $(t'',y) \in Q$ for all $y \in \mathcal{W}$ and $\{(t,\zeta^y(t)) : t \in [t',t'']\} \subseteq Q$ for all $y \in \mathcal{W}$. By Lemma 4.3, we have that for a.e. $y \in \mathcal{W}$ the set $B^y := \{t \in [t',t''] : (t,\zeta^y(t)) \in A\}$ is at most countable.

Therefore, since for every fixed t ess–liminf$_{y\to x}W(t,y) \le W(t,x)$, then for every $\delta_j \to 0$, $\delta_j > 0$ there exists a sequence $(y_j^l)_j \in \mathbb{N}$ such that $y_j^l \to x_l(t'')$, $W(t'', y_j^l) \le W(t'', x_l(t'')) + \delta_j$, and $B^{y_j^l}$ is at most countable. Consider the following function defined on $[t', t'']$:

$$\varphi_j^l(t) := W(t, \zeta^{y_j^l}(t)) + \int_{t'}^{t} L(s, \zeta^{y_j^l}(s), \omega)ds.$$

By the choice of y_j^l and the hypotheses (iv), φ_j^l is differentiable a.e. with a nonnegative derivative. By the lower semicontinuity of W and the NDJ condition, it follows that φ_j^l verifies the hypotheses of Lemma 4.4 and so $\varphi_j^l(t') \le \varphi_j^l(t'')$. Thus

$$W(t', \zeta^{y_j^l}(t')) \le W(t'', \zeta^{y_j^l}(t'')) + \int_{t'}^{t''} L(s, \zeta^{y_j^l}(s), \omega)ds. \tag{5.6}$$

Now, using the fact that $\zeta^{y_j^l}(t'') = y_j^l$, we obtain

$$\begin{aligned} W(t', \zeta^{y_j^l}(t')) &\le W(t'', y_j^l) + \int_{t'}^{t''} L(s, \zeta^{y_j^l}(s), \omega)ds \\ &\le W(t'', x_l(t'')) + \delta_j + \int_{t'}^{t''} L(s, \zeta^{y_j^l}(s), \omega)ds. \end{aligned} \tag{5.7}$$

By Proposition 4.2, $\zeta^{y_j^l}(\cdot) \to x_l(\cdot)$ as $j \to +\infty$, and so by the Lebesgue theorem and the lower semicontinuity of W, passing to the limit as $j \to +\infty$, we obtain

$$W(t', x_l(t')) \le W(t'', x_l(t'')) + \int_{t'}^{t''} L(s, x_l(s), \omega)ds. \tag{5.8}$$

First consider the case $\{(t, x_l(t)) : t \in [t_0, T]\} \subseteq Q$. Summing (5.8) over each interval on which u_l is constant, we have

$$W(t_0, x_l(t_0)) \le W(T, x_l(T)) + \int_{t_0}^{T} L(s, x_l(s), u_l(s))ds. \tag{5.9}$$

Now, $x_l(T) = x^*(T)$ by definition, and so, using (5.2)–(5.5) and (ii),

$$\begin{aligned} W(t_0, x_l(t_0)) &\le W(T, x^*(T)) + \int_{t_0}^{T} L(s, x_l(s), u_l(s))ds \\ &\le \psi(T, x^*(T)) + \int_{t_0}^{T} L(s, x_l(s), u_l(s))ds \\ &\le V(t_0, x_0) + \frac{\varepsilon}{2} - \int_{t_0}^{T} L(s, x^*(s), u^*(s))ds \\ &\quad + \int_{t_0}^{T} L(s, x_l(s), u_l(s))ds \\ &\le V(t_0, x_0) + \varepsilon < W(t_0, x_l(t_0)). \end{aligned}$$

This is a contradiction.

Suppose now $\{(t, x_l(t)) : t \in [t_0, T]\} \not\subseteq Q$. Define

$$\hat{\tau} := \inf \{t \leq T : (s, x_l(s)) \in Q \quad \forall s \in [t, T]\}. \tag{5.10}$$

In particular, $(\hat{\tau}, x_l(\hat{\tau})) \in \partial Q$. Using the same argument to pass from (5.8) to (5.9), we obtain that for every $\tau > \hat{\tau}$

$$W(\tau, x_l(\tau)) \leq W(T, x^*(T)) + \int_\tau^T L(s, x_l(s), u_l(s))ds, \tag{5.11}$$

and so, using (ii) and (5.3),

$$\begin{aligned} W(\tau, x_l(\tau)) &\leq \psi(T, x^*(T)) + \int_\tau^T L(s, x_l(s), u_l(s))ds \\ &\leq V(t_0, x_0) + \frac{\varepsilon}{2} - \int_{t_0}^T L(s, x^*(s), u^*(s))ds \\ &\quad + \int_\tau^T L(s, x_l(s), u_l(s))ds. \end{aligned} \tag{5.12}$$

Using (5.1), (5.5), and (v), we obtain for all $\tau > \hat{\tau}$

$$W(\tau, x_l(\tau)) \leq V(t_0, x_0) + \varepsilon \leq W(t_0, x_0) - \varepsilon. \tag{5.13}$$

Passing to the liminf as $\tau \to \hat{\tau}$ and using the lower semicontinuity of W, we conclude

$$W(\hat{\tau}, x_l(\hat{\tau})) \leq W(t_0, x_0) - \varepsilon, \tag{5.14}$$

and so by (iii)

$$W(t_0, x_0) \leq \sup_{(t,x)\in Q} W(t, x) \leq W(t_0, x_0) - \varepsilon, \tag{5.15}$$

which is a contradiction.

Now, we have to treat the case $V(t_0, x_0) = -\infty$. Since $W(t_0, x_0) > -\infty$ and W is lower semicontinuous, we may find two constants $M > 1$ and $\delta > 0$ such that

$$W(t_0, x) > -M$$

for every x so that $|x - x_0| < \delta$. Moreover, we can find $u^* \in \mathcal{U}$ such that $x^*(\cdot) := x(\cdot; u^*, t_0, x_0)$ satisfies $(T, x^*(T)) \in S$ and

$$\int_{t_0}^T L(s, x^*(s), u^*(s))ds + \psi(T, x^*(T)) \leq -2M.$$

With the same arguments in the first part of the proof, we may find a control $u_l \in \mathcal{U}$ piecewise constant and left continuous such that, if we denote by $x_l(\cdot)$ the trajectory $x(\cdot; u_l, T, x^*(T))$,

$$|x_l(t) - x^*(t)| < \frac{\delta}{2} \quad \forall t \in [t_0, T]$$

and

$$\left| \int_{t_0}^{T} [L(s, x_l(s), u_l(s)) - L(s, x^*(s), u^*(s))] ds \right| \leq 1.$$

Repeating the same calculations as before, we obtain that

$$\begin{aligned} -M &\leq W(T, x^*(T)) + \int_{t_0}^{T} L(s, x_l(s), u_l(s)) ds \\ &\leq \psi(T, x^*(T)) + \int_{t_0}^{T} L(s, x^*(s), u^*(s)) ds + 1 \\ &\leq -2M + 1, \end{aligned}$$

which gives $M \leq 1$, a contradiction.

This concludes the proof of the theorem. □

COROLLARY 5.3. *Let us suppose that W satisfies all the hypotheses of the previous theorem. If, moreover, $W \geq V$, then $W = V$.*

Remark 1. If W is produced by a synthesis procedure, the inequality $W \geq V$ always holds, and so if W satisfies all the hypotheses of Theorem 5.2, then W coincides with the value function.

Using the same techniques as those of the previous theorem, we can prove a corollary for value functions generated by approximated syntheses and give a bound of the error thus produced.

COROLLARY 5.4. *Suppose* (A-1)–(A-5) *hold. Let $Q \subseteq \Omega$ be an open subset containing S. Let $W : \overline{Q} \to \mathbb{R}$ be a lower semicontinuous function verifying the NDJ property along every time-varying vector field of the type $f(t, x, u)$ with $u \in U$ fixed. Moreover, we assume that, for each t,* ess–$\liminf_{y \to x} W(t, y) \leq W(t, x)$ *and that there exist $\varepsilon > 0$ and $g \in L^1(\mathbb{R})$, $g \geq 0$, such that the following hold:*

(i) *$W \leq \psi + \varepsilon$ on S.*

(ii) *At every point $(t, x) \in \partial Q$ one has*

$$W(t, x) = \sup_{(s,y) \in Q} W(s, y).$$

(iii) *There exists a countably $\mathcal{H}^n$-rectifiable set $A \subseteq \Omega$ such that W is differentiable on $Q \setminus A$ and satisfies*

$$W_s(s, y) + \inf_{\omega \in U} \{W_y(s, y) \cdot f(s, y, \omega) + L(s, y, \omega)\} \geq -\varepsilon g(s) \quad \text{on } Q \setminus A.$$

(iv) *$L \geq -\varepsilon g$.*

Then $W \leq V + \varepsilon(1 + \|g\|_1)$ on Q.

Proof. Note that $L(t, x, u) + \varepsilon g(t) \geq 0$, and so

$$\begin{aligned} W(t_0, x_0) &\leq \inf_{\substack{u \in \mathcal{U} \\ (T, x(T; u, t_0, x_0)) \in S}} \left\{ \int_{t_0}^{T} L(s, x(s; u, t_0, x_0), u(s)) ds + \psi(T, x(T; u, t_0, x_0)) \right\} \\ &\leq V(t_0, x_0) + \varepsilon(1 + \|g\|_{L^1}). \end{aligned}$$ □

Remark 2. Notice that the value function of an optimal control problem has the NDJ property along every possible direction as a consequence of the dynamic

programming principle. Indeed, for every $(t,y) \in \Omega \setminus S$ and for every admissible control $u \in \mathcal{U}$ (in particular, for every control $\omega\chi_I$, where $\omega \in U$ and I is a bounded interval), the function

$$h \mapsto \int_t^{t+h} L(s, x(s;u,t,y), u(s))ds + V(t+h, x(t+h;u,t,y))$$

is nondecreasing for $h \in [0,\delta]$ and δ small enough.

Instead, the hypothesis

$$\text{ess–liminf}_{y\to x} W(t,y) \le W(t,x)$$

for each t fixed says that for every $\varepsilon > 0$ there exists a subset $V \subseteq \{y \in \mathbb{R}^n : |y-x| \le \varepsilon\}$ of strictly positive Lebesgue measure such that

$$\inf_{y\in V} W(t,y) \le W(t,x).$$

So, if we consider a set $V_1 \subseteq \mathbb{R}^n$ of zero Lebesgue measure with x as a cluster point, the set $V \setminus V_1$ has a strictly positive Lebesgue measure. In the proof of Theorem 5.2, this fact is used to avoid the points y for which B^y is not countable. Moreover, this hypothesis, coupled with the lower semicontinuity of W, gives the following:

- For each t,

$$W(t,x) = \liminf_{y\to x} W(t,y) = \text{ess–liminf}_{y\to x} W(t,y).$$

Remark 3. Hypothesis (iii) of Theorem 5.2 says that, in the case $Q \neq \Omega$, the boundary of Q must be a level set of the function W. We can relax the same hypothesis in the following way:

- At every point $(t,x) \in \partial Q$ one has

$$\liminf_{\substack{\tau\to t, y\to x \\ (\tau,y)\in Q}} W(\tau,y) \ge \sup_{(s,y)\in Q} W(s,y).$$

Thus the conclusion of the theorem remains valid. Moreover, if we define with $R(t,x)$ the set of points reachable with an admissible control from (t,x), the previous condition can be replaced by

$$\inf_{(s,y)\in R(t,x)\cap\partial Q} W(s,y) \ge W(t,x),$$

and the conclusion still holds.

The hypothesis of the positiveness of L is almost optimal as the next example shows. However, the Lagrangian L may be negative on some region if trajectories cannot stay for too long in such a region, and one can relax the assumption (v) as shown in Remark 4.

Example 4. Consider the system $\dot{x} = u$, $U = [-1,1]$ and $\mathcal{U} = L^1(\mathbb{R};U)$, $\Omega = \mathbb{R}^2$, $S = \mathbb{R} \times \{0\}$, $Q = \mathbb{R}\times]-1,1[$ with the Lagrangian $L(t,x,u) = u^2 + x^4 - 6x^3 + 7x^2$ (see Figure 5.1) and $\psi \equiv 0$ on S. Since the Lagrangian is negative in a region where the system can stay for an arbitrary interval of times, clearly the value function for this problem is equal to $-\infty$. If $W \equiv C$ on $\overline{Q}$ with C a negative constant, then W verifies all the hypotheses of the Theorem 5.2 but (v). In fact (i), (ii), and (iii) are obvious, while (iv) holds because L is positive on Q and W is differentiable on Q. So

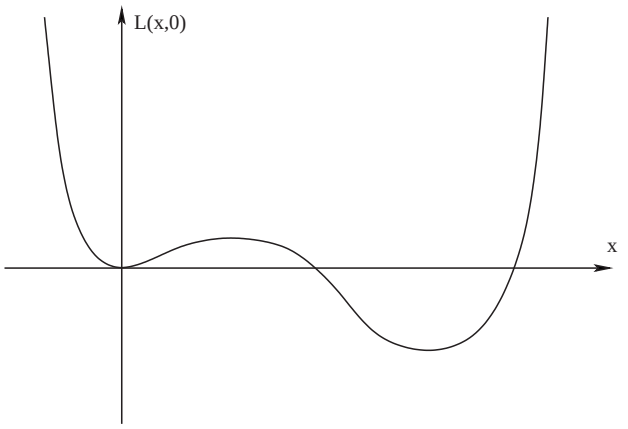


FIG. 5.1. $L(x,0)$ *of Example* 4.

there exist infinitely many functions W defined on $\overline{Q}$ verifying all the hypotheses of Theorem 5.2 except (v), which are not lower than or equal to the value function V.

Remark 4. If one wants to eliminate hypothesis (v) from the previous theorem, one may assume one of the following conditions:

(a) Fix $\varepsilon > 0$ and $(\bar{t},\bar{x}) \in Q$. We call $x_\varepsilon : [\bar{t},T] \to \mathbb{R}^n$ an ε–quasi-optimal trajectory (ε–q.o.t.) for $(\bar{t},\bar{x})$ if

(a.1) $\exists\, u_\varepsilon \in \mathcal{U}$ such that $\dot{x}_\varepsilon(s) = f(s, x_\varepsilon(s), u_\varepsilon(s))$ for a.e. $s \in [\bar{t},T]$,

(a.2) $x_\varepsilon(\bar{t}) = \bar{x}$,

(a.3) $(T, x_\varepsilon(T)) \in S$,

(a.4) $V(\bar{t},\bar{x}) + \varepsilon \geq \int_{\bar{t}}^{T} L(s, x_\varepsilon(s), u_\varepsilon(s))ds + \psi(T, x_\varepsilon(T))$.

Now define Q_1 as the set of points $(\bar{t},\bar{x}) \in Q$ such that, for every $\varepsilon > 0$, there exists x_ε, an ε–q.o.t. for $(\bar{t},\bar{x})$, satisfying $(s, x_\varepsilon(s)) \in Q$ for any $s \in [\bar{t},T]$. What we need is that $L \geq 0$ in $\Omega \backslash Q_1$. In fact, under this assumption, we may suppose that $(s, x(s)) \in \Omega \backslash Q_1$ for every $s \in [t_0, \hat{\tau}[$, where x is the trajectory defined in the proof of Theorem 5.2 and the time $\hat{\tau}$ is defined in (5.10). So the integral $\int_{t_0}^{\hat{\tau}} L(s, x(s), u(s))ds$ is positive. Otherwise, we can assume $Q_1 = Q$.

(b) We can also use a hypothesis similar to one given in [15, 16]. For any $(\bar{t},\bar{x}) \in \Omega$ and $u \in \mathcal{U}$, let $x_{\bar{t},\bar{x}}(\cdot;u) := x(\cdot;u,\bar{t},\bar{x})$ be the solution to (2.1) associated to the control u. Consider the set P consisting of those points $(\bar{t},\bar{x})$ of Q such that

$$\int_{\bar{t}}^{T} L(s, x_{\bar{t},\bar{x}}(s;u), u(s))ds \geq 0 \qquad \forall T > t, \quad \forall u \in \mathcal{U}.$$

We have to suppose that, if $(\bar{t},\bar{x}) \in Q \setminus [P \cup S]$, there exist a bounded and open set B, $(\bar{t},\bar{x}) \in B \subseteq Q$, $B \cap S = \emptyset$, so that $\partial B \subseteq Q$, and a positive number M strictly less than

$$\inf_{u\in\mathcal{U}} \left\{ T > 0 : d((\bar{t}+T, x_{\bar{t},\bar{x}}(\bar{t}+T;u)), \partial B) \leq d((\bar{t},\bar{x}), \partial B)/2 \right\}$$

such that, for all $u \in \mathcal{U}$, $(M+\bar{t}, x_{\bar{t},\bar{x}}(M+\bar{t};u)) \in Q \cap P$ and

$$\int_{\bar{t}}^{\bar{t}+M} L(s, x_{\bar{t},\bar{x}}(s), u(s))ds \geq 0,$$

and this allows us to conclude the proof of Theorem 5.2 without using $L \geq 0$ on the whole space.

Example 5. Consider the system $\dot{x} = u$, $U = [-1, 1]$, $\mathcal{U} = L^1(\mathbb{R}; U)$, $\Omega = \mathbb{R}^+ \times \mathbb{R}$, $S = \mathbb{R}^+ \times \{0\}$, $Q = \mathbb{R}^+ \times]-1, 1[$, $\psi = 0$ on S and the Lagrangian defined by

$$L(t,x,u) := \begin{cases} u^2 + x^2 & \text{if} \quad x \leq 1, \\ (u^2+1)(2-x) + (x-1)(u^2 + Ct) & \text{if} \quad 1 < x < 2, \\ u^2 + x^2 - 6x + 8 + Ct & \text{if} \quad x \geq 2. \end{cases}$$

It is clear that this Lagrangian, for C sufficiently big, satisfies the conditions (a) and (b) of the previous remark, even if it is not positive outside Q.

Remark 5. We can relax hypotheses (iii) and (v) with the following:

(iii') the boundary ∂Q is a level set of W;

(v') $L \geq 0$ on $\Omega \setminus Q$.

With these hypotheses, we can obtain an inequality of type (5.8) for each interval where the couple time-trajectory is in Q, and then, using (iii'), (v'), the lower semicontinuity of W, and the NDJ property, we can obtain (5.9).

6. Problem with infinite time. In this section we consider the control system (2.1) and assume that (A-1)–(A-5) hold with $0 \leq C_R \leq C$ for some $C > 0$ and every $R > 0$. Moreover, we suppose that the target S is a closed subset of $\mathbb{R} \times \mathbb{R}^n$ which satisfies the following structural property:

$(*)$ For any $T > 0$, there exists $(t, x) \in S$ with $t \geq T$.

Let S_1 be an open neighborhood of S contained in Ω. Assume that the final cost ψ is defined on S_1 and, if $d((t, x(t; u, t_0, x_0)), S) \to 0$ as $t \to +\infty$, then the trajectory $x(\cdot; u, t_0, x_0)$ is definitively in S_1; that is,

$(**)$ $\exists\, T > t$ such that $(s, x(s; u, t_0, x_0)) \in S_1$ for all $s \geq T$.

Define the value function

(6.1)
$$V(t_0, x_0) := \inf_{\substack{u \in \mathcal{U} \\ d((t,x(t;u,t_0,x_0)),S) \to 0 \\ \text{as } t \to +\infty}} \left\{ \int_{t_0}^{+\infty} L(s, x(s; u, t_0, x_0), u(s)) ds + \limsup_{t \to +\infty} \psi(t, x(t; u, t_0, x_0)) \right\}.$$

In other words, we consider only the trajectories that approach the target S in infinite time. Notice that this condition does not imply that $(T, x(T)) \notin S$ for any $T \geq t_0$.

Remark 6. The introduction of an open neighborhood of the target S is due to a technical reason and precisely to the fact that it is necessary to compare the candidate value function to the final cost near the target. Notice that in the following theorem the set Q must contain S_1. For example we consider $\Omega = \mathbb{R}^+ \times \mathbb{R}$, $S = \mathbb{R}^+ \times \{0\}$, $Q = \{(t, x) : t > 0, x < 1/t\}$, and $S_1 = \{(t, x) : t > 0, x < 3/t\}$. If $(t, 2/t)$ with $t > 0$ is a trajectory, then it is definitely in S_1, but it is never in Q.

THEOREM 6.1. *Let $Q \subseteq \Omega$ be an open subset containing S_1. Let $W : \overline{Q} \to \mathbb{R}$ be a lower semicontinuous function such that the following hold:*

(i) *W has the NDJ property along every time-varying vector field of the type $f(t, x, u)$ with $u \in U$ fixed, and for each t,*

$$\text{ess–liminf}_{y \to x} W(t, y) \leq W(t, x).$$

(ii) *$W \leq \psi$ on S_1.*

(iii) *At every point* $(t,x)\in\partial Q$ *one has*

$$W(t,x)=\sup_{(s,y)\in Q} W(s,y).$$

(iv) *There exists a countable* $\mathcal{H}^n$*-rectifiable set* $A\subseteq\Omega$ *such that* W *is differentiable in* $Q\setminus A$ *and satisfies*

$$W_s(s,y)+\inf_{\omega\in U}\{W_y(s,y)\cdot f(s,y,\omega)+L(s,y,\omega)\}\geq 0 \quad \textit{in } Q\setminus A.$$

(v) $L\geq 0$.

Then $W\leq V$ *on* Q. *If* $Q=\Omega$, *we can drop hypotheses* (iii) *and* (v).

Proof. Suppose by contradiction that there exists $(t_0,x_0)\in Q$ such that $W(t_0,x_0)>V(t_0,x_0)$. In particular, $V(t_0,x_0)<+\infty$. First, let us consider the case $V(t_0,x_0)>-\infty$. As in the first part of the proof of Theorem 5.2, we can find $\varepsilon>0$ and $\delta>0$ such that the following hold:

$$V(t_0,x_0)\leq W(t_0,x_0)-2\varepsilon, \tag{6.2}$$

$$|x-x_0|<\delta \quad\Rightarrow\quad W(t_0,x)>V(t_0,x_0)+\frac{3\varepsilon}{2}\ . \tag{6.3}$$

We can choose $u^*\in\mathcal{U}$, with the property that the trajectory $(t,x^*(t))$ approaches the target when $t\to+\infty$, such that

$$\int_{t_0}^{+\infty} L(s,x^*(s),u^*(s))ds+\limsup_{t\to+\infty}\psi(t,x^*(t))\leq V(t_0,x_0)+\frac{\varepsilon}{2}\ , \tag{6.4}$$

where $x^*(\cdot)$ is the trajectory corresponding to the control u^* such that $x^*(t_0)=x_0$.

Consider, now, a strictly increasing sequence of times $T_j>t_0$ converging to $+\infty$. We may suppose that $(t,x^*(t))\in Q$ for every $t\geq T_1$. Fix $j\in\mathbb{N}$. For every $l\in\mathbb{N}$, there exists $u_j^l\in\mathcal{U}$ piecewise constant and left continuous such that $\|u_j^l-u^*\|_{L^p([t_0,T_j])}\leq\frac{1}{l}$. So, by [7, Theorem IV.9], we can extract a subsequence of $(u_j^l)_l$, denoted again with $(u_j^l)_l$, and we can find a function $h_j\in L^p([t_0,T_j])$ such that $|u_j^l|\leq h_j$ a.e. for every $l\in\mathbb{N}$ and $u_j^l\to u^*$ for a.e. $t\in[t_0,T_j]$ as $l\to+\infty$. Thus, denoting with $x_j^l(\cdot)$ the trajectory $x(\cdot;u_j^l,T_j,x^*(T_j))$, for l sufficiently big we have (see Proposition 4.2)

$$\left|x_j^l(t)-x^*(t)\right|\leq\frac{\delta}{2}\qquad\forall t\in[t_0,T_j] \tag{6.5}$$

and then

$$\left|\int_{t_0}^{T_j}\left[L(s,x_j^l(s),u_j^l(s))-L(s,x^*(s),u^*(s))\right]ds\right|\leq\frac{\varepsilon}{2}. \tag{6.6}$$

Now, fix $l\in\mathbb{N}$ such that (6.5) and (6.6) hold. First, let us suppose that $\{(t,x_j^l(t)):t\in[t_0,T_j]\}\subseteq Q$. So, using Lemma 4.3, Lemma 4.4, the same arguments as in the proof of Theorem 5.2, and (6.6), we conclude

$$\begin{aligned}W(t_0,x_j^l(t_0))&\leq W(T_j,x_j^l(T_j))+\int_{t_0}^{T_j}L(s,x_j^l(s),u_j^l(s))ds\\&\leq W(T_j,x^*(T_j))+\int_{t_0}^{T_j}L(s,x^*(s),u^*(s))ds+\frac{\varepsilon}{2}.\end{aligned}$$

Using (6.3) and (6.5), we have

$$V(t_0,x_0)+\frac{3\varepsilon}{2}<W(T_j,x^*(T_j))+\int_{t_0}^{T_j}L(s,x^*(s),u^*(s))ds+\frac{\varepsilon}{2}. \tag{6.7}$$

Now consider the other case and precisely $\{(t,x_j^l(t)):t\in[t_0,T_j]\}\not\subseteq Q$. Define

$$\tau_j^l:=\inf\left\{t\geq t_0:(s,x_j^l(s))\in Q\ \ \forall s\in[t,T_j]\right\}. \tag{6.8}$$

Given $\tau_j^l<t<T_j$,

$$W(t,x_j^l(t))\leq W(T_j,x_j^l(T_j))+\int_t^{T_j}L(s,x_j^l(s),u_j^l(s))ds. \tag{6.9}$$

Considering the fact that $(t,x_j^l(t))\to(\tau_j^l,x_j^l(\tau_j^l))$ as $t\to\tau_j^l$, $(\tau_j^l,x_j^l(\tau_j^l))\in\partial Q$ and (iii), we obtain

$$W(t_0,x_0)\leq W(T_j,x^*(T_j))+\int_{\tau_j^l}^{T_j}L(s,x_j^l(s),u_j^l(s))ds. \tag{6.10}$$

We can now use the hypothesis (v), (6.3), and (6.6) in order to have

$$\begin{aligned}V(t_0,x_0)+\frac{3\varepsilon}{2}&<W(t_0,x_0)\\ &\leq W(T_j,x^*(T_j))+\int_{t_0}^{T_j}L(s,x^*(s),u^*(s))ds+\frac{\varepsilon}{2}.\end{aligned} \tag{6.11}$$

In all cases we have that, for every $j\in\mathbb{N}$,

$$V(t_0,x_0)+\frac{3\varepsilon}{2}<W(T_j,x^*(T_j))+\int_{t_0}^{T_j}L(s,x^*(s),u^*(s))ds+\frac{\varepsilon}{2}. \tag{6.12}$$

So, applying the limsup as $j\to+\infty$, we get

$$\begin{aligned}V(t_0,x_0)+\frac{3\varepsilon}{2}&\leq\limsup_{j\to+\infty}W(T_j,x^*(T_j))+\int_{t_0}^{+\infty}L(s,x^*(s),u^*(s))ds+\frac{\varepsilon}{2}\\ &\leq\limsup_{t\to+\infty}W(t,x^*(t))+\int_{t_0}^{+\infty}L(s,x^*(s),u^*(s))ds+\frac{\varepsilon}{2}.\end{aligned}$$

For t sufficiently big, $(t,x^*(t))\in S_1$, and so, using (ii) and (6.4),

$$\begin{aligned}V(t_0,x_0)+\frac{3\varepsilon}{2}&\leq\limsup_{t\to+\infty}\psi(t,x^*(t))+\int_{t_0}^{+\infty}L(s,x^*(s),u^*(s))ds+\frac{\varepsilon}{2}\\ &\leq V(t_0,x_0)+\varepsilon,\end{aligned}$$

which implies

$$V(t_0,x_0)\leq V(t_0,x_0)-\frac{\varepsilon}{2},$$

which is a contradiction.

It remains the case that $V(t_0, x_0) = -\infty$. Since $W(t_0, x_0) > -\infty$ and W is lower semicontinuous, we may find two constants $M > 1$ and $\delta > 0$ such that

$$W(t_0, x) > -M$$

for every x so that $|x - x_0| < \delta$. Moreover, we can find $u^* \in \mathcal{U}$ such that $x^*(\cdot) := x(\cdot; u^*, t_0, x_0)$ approaches the target when $t \to +\infty$ and

$$\int_{t_0}^{+\infty} L(s, x^*(s), u^*(s))ds + \limsup_{t\to+\infty} \psi(t, x^*(t)) \le -2M.$$

Consider a strictly increasing sequence of times $T_j > t_0$ converging to $+\infty$ and repeat the previous arguments in order to find a control $u_j^l \in \mathcal{U}$ piecewise constant, left continuous, and such that, if $x_j^l(\cdot) := x(\cdot; u_j^l, T_j, x^*(T_j))$,

$$\left|x_j^l(t) - x^*(t)\right| \le \frac{\delta}{2} \quad \forall t \in [t_0, T_j]$$

and

$$\left|\int_{t_0}^{T_j} [L(s, x_j^l(s), u_j^l(s)) - L(s, x^*(s), u^*(s))]ds\right| \le 1.$$

Proceeding as before, we obtain that

$$\begin{aligned} -M &\le W(T_j, x^*(T_j)) + \int_{t_0}^{T_j} L(s, x_j^l(s), u_j^l(s))ds \\ &\le W(T_j, x^*(T_j)) + \int_{t_0}^{T_j} L(s, x^*(s), u^*(s))ds + 1 \end{aligned}$$

for every $j \in \mathbb{N}$. Passing to the limit, we have

$$\begin{aligned} -M &\le \limsup_{j\to+\infty} W(T_j, x^*(T_j)) + \int_{t_0}^{+\infty} L(s, x^*(s), u^*(s))ds + 1 \\ &\le \limsup_{t\to+\infty} W(t, x^*(t)) + \int_{t_0}^{+\infty} L(s, x^*(s), u^*(s))ds + 1 \\ &\le \limsup_{t\to+\infty} \psi(t, x^*(t)) + \int_{t_0}^{+\infty} L(s, x^*(s), u^*(s))ds + 1 \\ &\le -2M + 1, \end{aligned}$$

which gives $M \le 1$, a contradiction.

So the theorem is proved. □

COROLLARY 6.2. *Let W satisfy all the hypotheses of the previous theorem and, moreover, $W \ge V$, where V is defined in* (6.1). *Then W coincides with the value function.*

Remark 7. In Theorem 6.1 the condition (ii) can be relaxed in the following way:

$$\limsup_{t\to+\infty} W(t, x(t)) \le \limsup_{t\to+\infty} \psi(t, x(t))$$

for every $x(\cdot)$ solution to (2.1) such that $d((t, x(t)), S) \to 0$ as $t \to +\infty$.

So, if one wants to minimize a Lagrangian cost without final cost, the condition becomes

$$\limsup_{t\to+\infty} W(t,x(t)) \le 0$$

for every $x(\cdot)$ with the above property.

Remark 8. If we assume that there exists $\eta > 0$ such that $S + B(0,\eta) \subseteq S_1$, where $B(0,\eta)$ is the ball in $\mathbb{R}^{n+1}$ centered in 0 with radius η, then hypothesis $(**)$ obviously holds. In fact, suppose $d((t,x(t;u,t_0,x_0)),S) \to 0$ as $t \to +\infty$. Then there exists $T > 0$ such that $d((s,x(s;u,t_0,x_0)),S) < \frac{\eta}{2}$ for all $s \ge T$. So we can choose an element $(t(s),y(s)) \in S$ in order to have $d((s,x(s;u,t_0,x_0)),(t(s),y(s))) < \frac{\eta}{2}$ for all $s \ge T$. So the points $(s,x(s;u,t_0,x_0)) \in S + B(0,\eta) \subseteq S_1$ for every $s \ge T$.

Remark 9. We obtain a generalization of Theorems 5.2 and 6.1 considering the same problem (2.1) with assumptions (A-1)–(A-4), but we accept at the same time all the trajectories that hit the target in finite time or that tend to the target in infinite time. Obviously an analogous theorem to Theorems 5.2 and 6.1 holds.

Remark 10. Also, in this case we can substitute hypothesis (iii) of Theorem 6.1 in an analogous way as in Remark 3. Moreover, we can eliminate hypothesis (v) of Theorem 6.1 in the same way as in Remark 4.

Appendix. Viscosity solutions and value functions. This appendix is intended to recall the notion of viscosity sub- and supersolution and to state some known properties of the value function. Proofs are analogous to those of [2].

Let Ω_1 be an open subset of $\mathbb{R}\times\mathbb{R}^n$. We need the following definitions.

DEFINITION A.1. *Let $f : A \to \overline{\mathbb{R}}$ be a function where A is an open subset of $\mathbb{R}^l$ for some $l \in \mathbb{N}\setminus\{0\}$. The lower semicontinuous envelope f_* and the upper semicontinuous envelope f^* of f are defined by*

$$f_*(x) := \lim_{r\to 0^+} \inf\{f(y) : y \in A, |y-x| \le r\},$$
$$f^*(x) := \lim_{r\to 0^+} \sup\{f(y) : y \in A, |y-x| \le r\}.$$

PROPOSITION A.2. *The lower semicontinuous (resp., upper semicontinuous) envelope of a function f is a lower semicontinuous (resp., upper semicontinuous) function. More precisely, it is the greatest (resp., least) lower semicontinuous (resp., upper semicontinuous) function less than or equal to (resp., greater than or equal to) f. Moreover, f is continuous if and only if $f_* = f^*$.*

DEFINITION A.3. *We say that a lower semicontinuous function $V : \Omega_1 \to \overline{\mathbb{R}}$ is a viscosity supersolution to $F(t,x,D_tV,D_xV) = 0$ in Ω_1 if, for any $\varphi \in \mathcal{C}^1(\Omega_1)$ and for any $(t_0,x_0) \in \Omega_1$ point of local minimum for $V-\varphi$, one has*

$$F^*(t_0,x_0,D_t\varphi(t_0,x_0),D_x\varphi(t_0,x_0)) \ge 0.$$

DEFINITION A.4. *We say that an upper semicontinuous function $V : \Omega_1 \to \overline{\mathbb{R}}$ is a viscosity subsolution to $F(t,x,D_tV,D_xV) = 0$ in Ω_1 if, for any $\varphi \in \mathcal{C}^1(\Omega_1)$ and for any $(t_0,x_0) \in \Omega_1$ point of local maximum for $V-\varphi$, one has*

$$F_*(t_0,x_0,D_t\varphi(t_0,x_0),D_x\varphi(t_0,x_0)) \le 0.$$

DEFINITION A.5. *We say that a function $V : \Omega_1 \to \overline{\mathbb{R}}$ is a viscosity solution to $F(t,x,D_tV,D_xV) = 0$ in Ω_1 if V_* is a viscosity supersolution and V^* is a viscosity subsolution to the equation.*

Remark 11. Note that the notion of viscosity solution is not bilateral in the sense that the sets of viscosity solutions to $F = 0$ and $-F = 0$ in general are different.

Let us consider the following hypotheses:

(H-1) The functions f and L are continuous in all the variables.

(H-2) U is a bounded set.

We have the following.

PROPOSITION A.6. *Let us assume* (A-1)–(A-5) *and* (H-1)–(H-2). *Then the value function V defined in* (2.4) *satisfies the dynamic programming principle, that is,*

$$V(t_0,x_0) = \inf_{\substack{u\in\mathcal{U}\\(T,x(T;u,t_0,x_0))\in S}} \left\{ \int_{t_0}^{T_1} L(s,x(s;u,t_0,x_0),u(s))ds + V(T_1,x(T_1;u,t_0,x_0)) \right\}$$

for every $(t_0,x_0) \in \Omega \setminus S$ and for every T_1 less than the minimum time to reach the target.

An analogous proposition holds for the value function V defined in (6.1).

Let us now state without proof the result that ensures that the value function is a viscosity solution to an HJB equation.

THEOREM A.7. *Let us assume* (A-1)–(A-5) *and* (H-1)–(H-2). *Then the value functions* (2.4) *and* (6.1) *are viscosity solutions of*

$$-V_s(t,x) - \inf_{\omega\in U} \{f(t,x,\omega)\cdot V_y(t,x) + L(t,x,\omega)\} = 0 \quad \text{in } \Omega \setminus S.$$

Acknowledgments. The author wishes to thank Professor B. Piccoli for having proposed to him the study of this problem and for his useful advice, and the referees for their suggestions for improvement.

REFERENCES

[1] G. ALBERTI, L. AMBROSIO, AND P. CANNARSA, *On the singularities of convex functions*, Manuscripta Math., 76 (1992), pp. 421–435.

[2] M. BARDI AND I. CAPUZZO-DOLCETTA, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*, Birkhäuser Boston, Boston, 1997.

[3] V. G. BOLTYANSKII, *Sufficient conditions for optimality and the justification of the dynamic programming method*, SIAM J. Control, 4 (1966), pp. 326–361.

[4] U. BOSCAIN AND B. PICCOLI, *Geometric control approach to synthesis theory. Control theory and its applications*, Rend. Sem. Mat. Univ. Politec. Torino, 56 (1998), pp. 53–68.

[5] A. BRESSAN, *Lecture Notes on the Mathematical Theory of Control*, SISSA, Trieste, Italy, 1994.

[6] A. BRESSAN AND B. PICCOLI, *A generic classification of time-optimal planar stabilizing feedbacks*, SIAM J. Control Optim., 36 (1998), pp. 12–32.

[7] H. BREZIS, *Analyse fonctionnelle: Théorie et applications*, Masson, Paris, 1987.

[8] P. BRUNOVSKÝ, *Existence of regular syntheses for general problems*, J. Differential Equations, 38 (1980), pp. 317–343.

[9] P. CANNARSA, A. MENNUCCI, AND C. SINESTRARI, *Regularity results for solutions of a class of Hamilton–Jacobi equations*, Arch. Ration. Mech. Anal., 140 (1997), pp. 197–223.

[10] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Canad. Math. Soc. Ser. Monogr. Adv. Texts, Wiley, New York, 1983.

[11] L. C. EVANS AND R. F. GARIEPY, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC, Boca Raton, FL, 1992.

[12] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[13] G. B. FOLLAND, *Real Analysis: Modern Techniques and Their Applications*, J. Wiley and Sons, New York, 1984.

[14] P. HARTMAN, *Ordinary Differential Equations*, S. H. Hartman, Baltimore, 1973.

[15] M. MALISOFF, *On the Bellman equation for control problems with exit times and unbounded cost functionals*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 23–28.

[16] M. Malisoff and H. J. Sussmann, *Further results on the Bellman equation for optimal control problems with exit times and nonnegative instantaneous costs*, in Proceedings of the 39th Annual IEEE Conference on Decision and Control (Sidney, Australia), IEEE Control Systems Society, Piscataway, NJ, 2000.
[17] B. Piccoli, *Classification of generic singularities for the planar time-optimal synthesis*, SIAM J. Control Optim., 34 (1996), pp. 1914–1946.
[18] B. Piccoli, *Infinite time regular synthesis*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 381–405.
[19] B. Piccoli and H. J. Sussmann, *Regular synthesis and sufficiency conditions for optimality*, SIAM J. Control Optim., 39 (2000), pp. 359–410.
[20] R. Vinter, *Optimal Control*, Birkhäuser Boston, Boston, 2000.
[21] L. Zajicek, *On the points of multiplicity of monotone operators*, Comment. Math. Univ. Carolin., 19 (1978), pp. 179–189.
[22] M. I. Zelikin and V. F. Borisov, *Theory of Chattering Control with Applications to Astronautics, Robotics, Economics, and Engineering*, Birkhäuser Boston, Boston, 1994.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1643–1660

BOUNDARY STABILIZATION FOR A HYBRID SYSTEM OF VISCOELASTICITY*

QIONG ZHANG† AND FALUN HUANG‡

Abstract. In this paper, we study the boundary feedback stabilization of a hybrid system that is composed of a viscoelastic thin plate with one part of its edge clamped and the remaining free part attached to a viscoelastic rigid body. By adopting the frequency domain method, we prove that the boundary feedback controls together with the dissipation induced by the memory effect are strong enough to secure the result of the exponential stability of energy. We also reach the result of the exponential stability of the hybrid system on the domain with corners without any extra hypotheses on the regularity of solutions.

Key words. hybrid system, viscoelasticity, C_0 contraction semigroup, exponential stability, nonsmooth domain

AMS subject classifications. 35Q72, 73F15, 73D30, 73H10, 73K10, 93C20, 93D15

DOI. S0363012901389913

1. Introduction.

In recent years, much attention has been paid to the topic of control and stabilization of hybrid (or coupled) systems, in which the dynamics of systems are related to possible rigid bodies (see, e.g., [14, 23, 26, 28, 29, 31, 32, 33] and references therein). However, little attention was paid to the hybrid systems of special materials. In this paper, we consider a linear hybrid model that is composed of a viscoelastic thin plate clamped on one part of its boundary and along the other free part rimmed with a viscoelastic flange that has mass and bending moment inertia. Mathematically speaking, the vibration u of the plate is described by a viscoelastic plate equation with two dynamical viscoelastic boundary conditions:

$$
\begin{cases}
u''(t) + \Delta^2 u(t) + \displaystyle\int_0^\infty D'(s)\Delta^2 u(t-s)ds = 0 & \text{in } \Omega,\\
u(t) = \partial_\nu u(t) = 0 & \text{on } \Gamma_0,\\
J\partial_\nu u''(t) + B_1 u(t) + \displaystyle\int_0^\infty D'(s)B_1 u(t-s)ds = -\partial_\nu u'(t) & \text{on } \Gamma_1,\\
\rho u''(t) - B_2 u(t) - \displaystyle\int_0^\infty D'(s)B_2 u(t-s)ds = -u'(t) & \text{on } \Gamma_1,\\
u(0) = u_0,\ u'(0) = u_1 & \text{in } \Omega,\\
u(-s) = \theta(s), \quad 0 < s < \infty & \text{in } \Omega,
\end{cases}
\tag{1.1}
$$

*Received by the editors May 29, 2001; accepted for publication (in revised form) April 4, 2003; published electronically November 6, 2003. This work was supported by the National Nature Science Foundation of China.
http://www.siam.org/journals/sicon/42-5/38991.html
†Institute of Systems Science, Academia Sinica, Beijing, 100080, China (q.zhang@amss.ac.cn).
‡Department of Mathematics, Sichuan University, Chengdu, 610064, Sichuan, China.

where B_1, B_2 denote the boundary operators associated with the plate equation:

$$B_1 u = \Delta u + (1-\mu)\left(2\nu_1\nu_2 \frac{\partial^2 u}{\partial x_1 \partial x_2} - \nu_1^2 \frac{\partial^2 u}{\partial x_2^2} - \nu_2^2 \frac{\partial^2 u}{\partial x_1^2}\right),$$

$$B_2 u = \partial_\nu \Delta u + (1-\mu)\partial_\tau \left[\left(\nu_1^2 - \nu_2^2\right) \frac{\partial^2 u}{\partial x_1 \partial x_2} + \nu_1\nu_2 \left(\frac{\partial^2 u}{\partial x_2^2} - \frac{\partial^2 u}{\partial x_1^2}\right)\right].$$

$\nu = (\nu_1, \nu_2)$ is the unit outer normal vector and $\tau = (-\nu_2, \nu_1)$ is the unit tangent vector. $\Omega \subset R^2$ is a bounded open domain with smooth or nonsmooth boundary $\Gamma = \Gamma_0 \bigcup \Gamma_1$. $0 < \mu < \frac{1}{2}$ is the Poisson ratio of elasticity, $\rho > 0$ is the linear boundary density, and $J > 0$ is the bending moment of inertia per unit length of the boundary. $D(\cdot)$ denotes the relaxation function. θ is the specified "history," and u_0, u_1 are initial data. $-\partial_\nu u'(t)$ and $-u'(t)$ are the boundary feedback controls applied on the free boundary part Γ_1.

As far as the wave equation with viscoelastic damping is considered, Dafermos first proved that the energy of the system tends to zero asymptotically under the Dirichlet boundary conditions [6, 7]. Day [8] obtained an explicit rate at which the energy decays to zero. Desch and Miller [9, 10] and Hannsgen and Wheeler [15] provided the results on the exponential stability. Fabiano and Lazzari [11] investigated the three-dimensional viscoelastic system and obtained the result on the exponential stability. In [25] (or [27]), the authors introduced an abstract frame to study the viscoelastic system and obtained a spectrum determined growth rate property. Lagnese [18] proved the exponential stability of the viscoelastic Kirchhoff plate. In this paper, we will show that the feedback controls $-\partial_\nu u'(t)$ and $-u'(t)$(noncompact) just applied on the rimmed part of the boundary of the plate are sufficient to provide the uniformly exponential energy decay of the viscoelastic model on smooth and nonsmooth domains.

A common way to prove the exponential stability for the mechanical systems was given by Huang in [16] (or lately, Weiss in [34]). If $T(t)$ is a bounded C_0 semigroup with generator A, and if the resolvent of A is bounded on the imaginary axis, then $T(t)$ is exponentially stable. Such a frequency domain method was used in [32], where an explicit representation of the resolvent operator is needed, as well as in [4], where it is combined with the energy multiplier technique [17, 18] and where the "geometric conditions" are necessary. In this paper, we adopt the indirect contradiction argument of the frequency domain method to avoid the requirement for any explicit knowledge of the resolvent of A. Moreover, due to the dissipation induced by the memory effect, our approach enables us to get the desired exponential stability of the system (1.1) on smooth or nonsmooth domains even if dispensing all geometric restrictions that are routinely imposed when the multiplier technique is adopted.

In the special case of domain with corners, one of the major differences between this and the case of smooth domain in studying the stabilization problem is that additional energy terms are contributed by the twisting moment at corner points, and, consequently, extra feedback controls are required to treat the "corner effects" [3, 4]. Another difference is that the solutions lose regularity because of the presence of corners [2, 13]. Then neither the energy multipliers technique, which needs the high regularity of solutions, nor the microlocal analysis method [1, 20], which could be used only on the domain with sufficiently smooth boundary, can be applied in this situation. We introduce several proper spaces and operators to skip this difficulty, and the result of exponential stability of system (1.1) on the nonsmooth domain could be achieved through the frequency domain method.

In the next section, we introduce the viscoelastic hybrid system on the domain with sufficiently smooth boundary and prove the well-posedness through the classical semigroup theory. In section 3, we derive the uniformly exponential decay of energy of system (1.1) on the domain with sufficiently smooth boundary. In section 4, domain with corners is considered. By introducing proper pointwise constraints on the corner points, we get the well-posedness and the exponential stability of the system.

2. The well-posedness. In this section, we deal with the system (1.1) described above when the boundary of domain Ω is sufficiently smooth. Assume Γ is C^2-smooth and its partition satisfies $\overline{\Gamma}_0 \bigcap \overline{\Gamma}_1 = \emptyset$, the clamped part Γ_0 is not empty and has positive boundary measure, and the rimmed part Γ_1 is relatively open in Γ. Let the relaxation function $D(\cdot)$ satisfy the following basic conditions:

(H1) $D(t) \in C^2[0,\infty)$.

(H2) $D(t) > 0$, $D'(t) < 0$, $D''(t) \geq 0$.

(H3) $D_\infty > 0$.

Condition (H2) means that the memory is strictly decreasing, the rate of the loss of the memory is increasing, and $D_\infty \doteq D(\infty)$, $D'_\infty \doteq D'(\infty)$ exist, $D_\infty \geq 0$, $D'_\infty = 0$. (H3) implies that the material is viscoelastic solid. We shall suppose that $D(0) = 1$ without affecting the result.

We rewrite system (1.1) as follows:

$$
\begin{cases}
u''(t) + D_\infty \Delta^2 u(t) + \displaystyle\int_0^\infty D'(s)\Delta^2 w(t,s)ds = 0 & \text{in } \Omega,\\
w(t,s) = u(t-s) - u(t), \quad 0 < s < \infty, & \text{in } \Omega,\\
u(t) = \partial_\nu u(t) = 0 & \text{on } \Gamma_0,\\
J\partial_\nu u''(t) + D_\infty B_1 u(t) + \displaystyle\int_0^\infty D'(s)B_1 w(t,s)ds = -\partial_\nu u'(t) & \text{on } \Gamma_1,\\
\rho u''(t) - D_\infty B_2 u(t) - \displaystyle\int_0^\infty D'(s)B_2 w(t,s)ds = -u'(t) & \text{on } \Gamma_1,\\
u(0) = u_0,\ u'(0) = u_1 & \text{in } \Omega,\\
w(0,s) = \theta(s) - u_0, \quad 0 < s < \infty, & \text{in } \Omega.
\end{cases} \tag{2.1}
$$

Define the associated energy of (2.1) as

$$
\begin{aligned}
E(t) \doteq \frac{1}{2}\Bigg[& D_\infty a(u(t)) + \int_\Omega |u'(t)|^2 dx + \int_0^\infty |D'(s)| a(u(t-s)\\
& - u(t))ds + \int_{\Gamma_1} (J|\partial_\nu u'(t)|^2 + \rho|u'(t)|^2)d\Gamma\Bigg],
\end{aligned} \tag{2.2}
$$

where $a(u) = a(u,u)$, and $a(u,v)$ is the following sesquilinear form:

$$
\begin{aligned}
a(u,v) = \int_\Omega \Bigg[& \frac{\partial^2 u}{\partial x_1^2}\overline{\frac{\partial^2 v}{\partial x_1^2}} + \frac{\partial^2 u}{\partial x_2^2}\overline{\frac{\partial^2 v}{\partial x_2^2}} + \mu\left(\frac{\partial^2 u}{\partial x_1^2}\overline{\frac{\partial^2 v}{\partial x_2^2}} + \frac{\partial^2 u}{\partial x_2^2}\overline{\frac{\partial^2 v}{\partial x_1^2}}\right)\\
& + 2(1-\mu)\frac{\partial^2 u}{\partial x_1 \partial x_2}\overline{\frac{\partial^2 v}{\partial x_1 \partial x_2}}\Bigg] dx.
\end{aligned}
$$

A straightforward calculation gives that

$$\frac{d}{dt}E(t) = -\frac{1}{2}\int_0^\infty D''(s)a(u(t-s)-u(t))ds - \int_{\Gamma_1}(|u'(t)|^2 + |\partial_\nu u'(t)|^2)d\Gamma \le 0.$$

Therefore, the energy of the system (2.1) decreases on $[0,\infty)$.

We at first formulate system (2.1) as a first-order evolution equation on a certain Hilbert space. Set

$$V^m = \{u \in H^m(\Omega) \ : \ u|_{\Gamma_0} = \partial_\nu u|_{\Gamma_0} = 0\}, \quad m \ge 2 \text{ is an integer.}$$

When $m = 2$, we define the equivalent norm on V^2 as

$$\|u\|_{V^2} = [D_\infty a(u)]^{\frac{1}{2}}.$$

It is clear that $a(u,v)$ is well-posed on V^2 since Γ_0 is nonempty and has positive boundary measure. Let the "history space" $L^2([0,\infty),|D'|,V^2)$ consist of V^2-valued functions $w(\cdot)$ on $[0,\infty)$ for which

$$\|w(\cdot)\|_{L^2([0,\infty),|D'|,V^2)} = \left[\int_0^\infty |D'(s)|a(w(s))ds\right]^{\frac{1}{2}} < \infty.$$

Set the Hilbert space

$$H = V^2 \times L^2(\Omega) \times L^2([0,\infty),|D'|,V^2) \times L^2_J(\Gamma_1) \times L^2_\rho(\Gamma_1)$$

with the energy norm

$$\begin{aligned}\|(u,v,w(\cdot),\xi,\eta)\|_H = [&\|u\|^2_{V^2} + \|v\|^2_{L^2(\Omega)} + \|w(\cdot)\|^2_{L^2([0,\infty),|D'|,V^2)} \\ &+ \|\xi\|^2_{L^2_J(\Gamma_1)} + \|\eta\|^2_{L^2_\rho(\Gamma_1)}]^{\frac{1}{2}},\end{aligned}$$

where $\|\xi\|^2_{L^2_J(\Gamma_1)} = J\int_{\Gamma_1}|\xi|^2 d\Gamma$, and $\|\eta\|^2_{L^2_\rho(\Gamma_1)} = \rho\int_{\Gamma_1}|\eta|^2 d\Gamma$.

For further analysis, the following integration-by-parts formula is needed:

$$\int_\Omega \Delta^2 u\bar{v}dx = a(u,v) + \int_{\Gamma_1} B_2 u\bar{v}d\Gamma - \int_{\Gamma_1} B_1 u\overline{\partial_\nu v}d\Gamma \quad \forall u \in W,\ v \in V^2, \tag{2.3}$$

where

$$W = \{u \in V^2 \ : \ \Delta^2 u \in L^2(\Omega), B_1 u, B_2 u \in L^2(\Gamma_1)\}.$$

Now we define an unbounded linear operator A on H as

$$\begin{aligned}D(A) = \{&z = (u,v,w(\cdot),\xi,\eta) \in H \ : \ D_\infty u + Lw(\cdot) \in W, v \in V^2, \\ &w(\cdot) \in C^1([0,\infty),|D'|,V^2), w(0) = 0, \xi = \partial_\nu v|_{\Gamma_1}, \eta = v|_{\Gamma_1}\}, \\ A(u,v,w(\cdot),\xi,\eta) = \Big(&v, -D_\infty\Delta^2 u - \Delta^2 Lw(\cdot), -\frac{\partial}{\partial s}w(\cdot) - v, -\frac{1}{J}(D_\infty B_1 u \\ &+ B_1 Lw(\cdot) + \xi), \frac{1}{\rho}(D_\infty B_2 u + B_2 Lw(\cdot) - \eta)\Big),\end{aligned}$$

where

$$C^1([0,\infty),|D'|,V^2)=\left\{w(\cdot)\in L^2([0,\infty),|D'|,V^2):\ \frac{\partial}{\partial s}w(\cdot)\in L^2([0,\infty),|D'|,V^2)\right\},$$

and the operator L is

$$L:\ L^2([0,\infty),|D'|,V^m)\to V^m,\quad Lw(\cdot)=\int_0^\infty D'(s)w(s)ds.$$

Let $z(t)=(u(t),u'(t),u(t-s)-u(t),\partial_\nu u'(t)|_{\Gamma_1},u'(t)|_{\Gamma_1})$. Then the system (2.1) can be formulated as

$$z'(t)=Az(t),\qquad z(0)=z_0\in H. \tag{2.4}$$

THEOREM 2.1. *Assume that the relaxation function $D(\cdot)$ satisfies* (H1), (H2), *and* (H3). *Then the operator A generates a C_0 contraction semigroup e^{tA} on H.*

Proof. For $z=(u,v,w(\cdot),\xi,\eta)\in D(A)$, we have

$$\begin{aligned}&\Re e(Az,z)_H\\ &=\ -\frac{1}{2}\int_0^\infty D''(s)a(w(s))ds-\int_{\Gamma_1}(|\partial_\nu v|^2+|v|^2)d\Gamma\le 0.\end{aligned} \tag{2.5}$$

Thus, A is dissipative. On the other hand, for any $f=(f_1,f_2,f_3(\cdot),f_4,f_5)\in H$, solve the equation

$$(I-A)z=f,\quad z=(u,v,w(\cdot),\xi,\eta)\in D(A). \tag{2.6}$$

This implies

$$\begin{cases}u-v=f_1,\\ v+D_\infty\Delta^2u+\Delta^2Lw(\cdot)=f_2,\\ \dfrac{\partial}{\partial s}w(s)+v+w(s)=f_3(s),\\ (J+1)\xi+D_\infty B_1u+B_1Lw(\cdot)=Jf_4,\\ (\rho+1)\eta-D_\infty B_2u-B_2Lw(\cdot)=\rho f_5.\end{cases}$$

Eliminating v and using $\xi=\partial_\nu v|\Gamma_1$, $\eta=v|\Gamma_1$, we obtain

$$u+D_\infty\Delta^2u+\Delta^2Lw(\cdot)=f_1+f_2, \tag{2.7}$$

$$w(s)=-(1-e^{-s})(u-f_1)+\int_0^s e^{\sigma-s}f_3(\sigma)d\sigma, \tag{2.8}$$

$$(J+1)\partial_\nu u+D_\infty B_1u+B_1Lw(\cdot)=Jf_4+(J+1)\partial_\nu f_1, \tag{2.9}$$

$$(\rho+1)u-D_\infty B_2u-B_2Lw(\cdot)=\rho f_5+(\rho+1)f_1. \tag{2.10}$$

It follows from (2.8) that

$$Lw(\cdot)=\int_0^\infty D'(s)w(s)ds=X(u-f_1)+Y,$$

where

$$X=\int_0^\infty D'(s)(e^{-s}-1)ds=-D_\infty+\int_0^\infty D(s)e^{-s}ds$$

and

$$Y = \int_0^\infty \int_0^s D'(s) e^{\sigma - s} f_3(\sigma) d\sigma ds.$$

Using the formula (2.3), we deduce that (2.7) and (2.9)–(2.10) are equivalent to the following equality:

$$\begin{aligned} &\int_\Omega u\overline{\phi} dx + (D_\infty + X)a(u,\phi) + (J+1)\int_{\Gamma_1} \partial_\nu u \overline{\partial_\nu \phi} d\Gamma \\ &+ (\rho+1)\int_{\Gamma_1} u\overline{\phi} d\Gamma = \int_\Omega (f_1 + f_2)\overline{\phi} dx + a(Xf_1 - Y, \phi) \\ &+ \int_{\Gamma_1} [(Jf_4 + (J+1)\partial_\nu f_1)\overline{\partial_\nu \phi} + (\rho f_5 + (\rho+1)f_1)\overline{\phi}] d\Gamma \quad \forall \phi \in V^2. \end{aligned} \tag{2.11}$$

Due to the Lax–Milgram theorem, (2.11) admits a unique solution $u \in V^2$ for any given $f \in H$. Therefore, it follows from (2.7)–(2.10) that the range of $I - A$ is H. By Theorem 4.6 in [30], we have $\overline{D(A)} = H$. The generation of C_0 contraction semigroup now follows from the Lumer–Phillips theorem. □

Now the following well-posedness result is obvious.

COROLLARY 2.2. *Suppose that the relaxation function $D(\cdot)$ satisfies hypotheses* (H1), (H2), *and* (H3). *Then we have the following:*

(i) *For any initial value $z_0 \in H$, system* (2.1) *has a unique mild solution satisfying*

$$z(t) = (u(t), u'(t), u(t-s) - u(t), \partial_\nu u'(t)|_{\Gamma_1}, u'(t)|_{\Gamma_1}) \in C([0,\infty), H).$$

Moreover,

$$\|z(t)\|_H \le \|z_0\|_H \quad \forall t \ge 0.$$

(ii) *For any initial value $z_0 \in D(A)$, system* (2.1) *has a unique classical solution satisfying*

$$z(t) \in C^1([0,\infty), H) \bigcap C([0,\infty), D(A)).$$

3. Exponential stability. We first recall Huang's frequency domain theorem [16].

LEMMA 3.1. *Let e^{tA} be a C_0 semigroup in Hilbert space H, and there exists a position constant number M such that $\|e^{tA}\| \le M (t \ge 0)$. Then e^{tA} is exponentially stable if and only if $\{\lambda \in C : \Re e\lambda = 0\} \subset \rho(A)$ and $\sup\{\|(\lambda - A)^{-1}\|_H : \Re e\lambda = 0\} < \infty$.*

We will use the following properties of the relaxation function.

LEMMA 3.2. *Assume that the function $D(\cdot)$ satisfies* (H1)–(H3). *Then for any $\epsilon > 0$, there exists a constant $\delta > 0$ such that*

$$\inf_{\{\beta \in R\,:\,|\beta| \ge \epsilon > 0\}} \int_0^\infty |D'(s)||e^{-i\beta s} - 1|^2 ds \ge \delta. \tag{3.1}$$

Proof. Through direct computation, we have

$$\begin{aligned}\Pi(\beta) &\doteq \int_0^\infty |D'(s)||e^{-i\beta s}-1|^2 ds \\ &= 2\int_0^\infty |D'(s)|(1-\cos\beta s)ds \\ &= 2\int_0^\infty |D'(s)|ds - \int_0^\infty |D'(s)|(e^{-i\beta s}+e^{i\beta s})ds.\end{aligned} \tag{3.2}$$

Applying the Riemann–Lebesgue lemma [12] yields that there exists $\kappa > \epsilon$ such that

$$\left|\int_0^\infty |D'(s)|(e^{-i\beta s}+e^{i\beta s})ds\right| \le \int_0^\infty |D'(s)|ds, \quad |\beta| > \kappa.$$

Therefore,

$$\Pi(\beta) \ge \int_0^\infty |D'(s)|ds, \quad |\beta| > \kappa. \tag{3.3}$$

Furthermore, notice that Π is a positive continuous function on compact subset $\{\beta \in R : \epsilon \le |\beta| \le \kappa\}$. Thus there exists $\delta_0 > 0$ such that

$$\Pi(\beta) \ge \delta_0, \quad \epsilon \le |\beta| \le \kappa. \tag{3.4}$$

Set $\delta \doteq \min\{\delta_0, \int_0^\infty |D'(s)|ds\}$. We deduce from (3.3) and (3.4) that

$$\Pi(\beta) \ge \delta > 0, \quad \epsilon \le |\beta| \le \infty. \qquad \square$$

Now we turn to our main result in this section. A further assumption about the relaxation function is set.

(H4) There exists a constant $k > 0$ such that $D''(s) + kD'(s) \ge 0$ for all $s \ge 0$.

THEOREM 3.3. *Let the relaxation function $D(\cdot)$ satisfy hypotheses* (H1)–(H4). *Then*

$$0 \in \rho(A).$$

Proof. For any $f = (f_1, f_2, f_3(\cdot), f_4, f_5)$, we consider the equation

$$Az = f, \quad z = (u, v, w(\cdot), \xi, \eta),$$

which implies

$$v = f_1, \tag{3.5}$$

$$D_\infty \Delta^2 u + \Delta^2 L w(\cdot) = -f_2, \tag{3.6}$$

$$\frac{\partial}{\partial s} w(s) + v = -f_3(s), \tag{3.7}$$

$$D_\infty B_1 u + B_1 L w(\cdot) + \xi = -J f_4, \tag{3.8}$$

$$D_\infty B_2 u + B_2 L w(\cdot) - \eta = \rho f_5. \tag{3.9}$$

It follows from (3.5) that

$$\xi = \partial_\nu v|_{\Gamma_1} = \partial_\nu f_1|_{\Gamma_1}, \qquad \eta = v|_{\Gamma_1} = f_1|_{\Gamma_1}. \tag{3.10}$$

Furthermore, we can deduce from (3.5) and (3.7) that

$$w(s) = -\int_0^s f_3(\sigma)d\sigma - s f_1. \tag{3.11}$$

It is clear that $\frac{\partial}{\partial s}w(\cdot) \in L^2([0,\infty), |D'|, V^2)$. And from assumption (H4), we have $-D'(s) \le -D'(0)e^{-ks}$ for all $s \ge 0$. Thus $w(\cdot) \in L^2([0,\infty), |D'|, V^2)$ by (3.11). Finally, we have from (3.6), (3.8) and (3.9) that $D_\infty u + Lw(\cdot) \in W$, and for all $\phi \in V$,

$$\begin{aligned} & D_\infty a(u,\phi) \\ = & -\int_\Omega f_2\overline{\phi}dx + a\left(L\left(\int_0^s f_3(\sigma)d\sigma + sf_1\right),\phi\right) \\ & -\int_{\Gamma_1}(Jf_4\overline{\partial_\nu\phi} + \rho f_5\overline{\phi})d\Gamma - \int_{\Gamma_1}(\partial_\nu f_1\overline{\partial_\nu\phi} + f_1\overline{\phi})d\Gamma. \end{aligned} \tag{3.12}$$

Applying the Lax–Milgram theorem yields that there exists a unique $u \in V^2$ satisfying (3.12) for all fixed $f \in H$. Combining (3.10)–(3.11), we obtain that $Az = f$ admits a unique solution $z \in D(A)$ for all $f \in H$. Thus A^{-1} exists and is bounded by the closed graph theorem. □

THEOREM 3.4. *Suppose the boundary* Γ *is* C^2*-smooth and the relaxation function* $D(\cdot)$ *satisfies* (H1)–(H4). *Then*

(i) $\{i\omega : \omega \in R\} \subset \rho(A)$;

(ii) e^{tA} *is an exponentially stable* C_0 *semigroup; i.e., there exist two positive constants* M *and* ϖ *such that*

$$E(t) \le Me^{-\varpi t} \quad \forall t \ge 0.$$

Proof. From Lemma 3.1 and Theorem 3.3, we will prove (i) and (ii) at the same time if we can verify the following condition: there exists $r > 0$ such that

$$\inf\{\|(i\omega - A)z\|_H : z \in D(A), \|z\|_H = 1, \omega \in R\} \ge r. \tag{3.13}$$

Suppose (3.13) is not true. By the continuity of the resolvent and the resonance theorem, there exist a sequence $\omega_n \in R$ and a sequence of vectors $z_n = (u_n, v_n, w_n(\cdot), \xi_n, \eta_n) \in D(A)$ $(n = 1, 2, \ldots)$ such that

$$\|z_n\|_H = 1, \quad n = 1, 2, \ldots, \tag{3.14}$$

$$\lim_{n\to\infty} \|(i\omega_n - A)z_n\|_H = 0. \tag{3.15}$$

Moreover, it follows from (3.15) that $\lim_{n\to\infty} \Re e(Az_n, z_n)_H = 0$. Consequently,

$$\lim_{n\to\infty}\int_{\Gamma_1}|\partial_\nu v_n|^2 d\Gamma = 0, \qquad \lim_{n\to\infty}\int_{\Gamma_1}|v_n|^2 d\Gamma = 0, \tag{3.16}$$

and

$$\lim_{n\to\infty}\int_0^\infty D''(s)a(w_n(s))ds = 0. \tag{3.17}$$

By the assumption (H4), we can get from (3.17) that

$$\begin{aligned} \lim_{n\to\infty}\|w_n(\cdot)\|^2_{L^2([0,\infty),|D'|,V^2)} &= -\lim_{n\to\infty}\int_0^\infty D'(s)a(w_n(s))ds \\ &\le \frac{1}{k}\lim_{n\to\infty}\int_0^\infty D''(s)a(w_n(s))ds = 0. \end{aligned} \tag{3.18}$$

Furthermore, in view of (3.15), we have

$$g_n \doteq i\omega_n u_n - v_n \longrightarrow 0 \qquad \text{in } V^2, \tag{3.19}$$

$$h_n(s) \doteq i\omega_n w_n(s) + \frac{\partial}{\partial s} w_n(s) + v_n \longrightarrow 0 \qquad \text{in } L^2([0,\infty), |D'|, V^2). \tag{3.20}$$

Substituting (3.19) into (3.20) yields

$$\frac{\partial}{\partial s} w_n(s) + i\omega_n w_n(s) - h_n(s) + i\omega_n u_n - g_n \longrightarrow 0 \qquad \text{in} \quad L^2([0,\infty), |D'|, V^2).$$

Through direct computation, we obtain

$$w_n(s) = (e^{-i\omega_n s} - 1)\left(u_n - \frac{1}{i\omega_n} g_n\right) + \int_0^s e^{-i\omega_n(s-\tau)} h_n(\tau) d\tau \text{ in } \quad L^2([0,\infty), |D'|, V^2).$$

Thus,

$$\begin{aligned} &\left(\int_0^\infty |D'(s)||e^{-i\omega_n s} - 1|^2 ds\right) a(u_n) \\ \leq\ & \|w_n(\cdot)\|^2_{L^2([0,\infty),|D'|,V^2)} + \frac{1}{|\omega_n|^2}\left(\int_0^\infty |D'(s)||e^{-i\omega_n s} - 1|^2 ds\right) a(g_n) \\ &+ \left\| \int_0^{\cdot} e^{-i\omega_n(\cdot - \tau)} h_n(\tau) d\tau \right\|^2_{L^2([0,\infty),|D'|,V^2)}. \end{aligned} \tag{3.21}$$

By Lemma 3.2 and (3.18), we can deduce from (3.21) that

$$\lim_{n\to\infty} a(u_n) = 0. \tag{3.22}$$

On the other hand, substituting (3.16) and (3.18) into (3.14) yields

$$D_\infty a(u_n) + \int_\Omega |v_n|^2 dx \longrightarrow 1. \tag{3.23}$$

Assume that $\varrho \doteq \lim_{n\to\infty} \int_\Omega |v_n|^2 dx$, $\varrho \neq 0$. Then $\lim_{n\to\infty} D_\infty a(u_n) = 1 - \varrho$. Taking the inner product of (3.19) with v_n in $L^2(\Omega)$, we get

$$\lim_{n\to\infty} \int_\Omega |v_n|^2 dx = -\lim_{n\to\infty} i\omega_n \int_\Omega v_n \overline{u_n} dx. \tag{3.24}$$

Also, we take the inner product of (3.19) with v_n in $L^2(\Gamma_1)$ to get

$$\lim_{n\to\infty} \int_{\Gamma_1} |v_n|^2 d\Gamma = -\lim_{n\to\infty} i\omega_n \int_{\Gamma_1} v_n \overline{u_n} d\Gamma.$$

Using (3.16) yields

$$\lim_{n\to\infty} i\omega_n \int_{\Gamma_1} v_n \overline{u_n} d\Gamma = 0. \tag{3.25}$$

Similarly,

$$\lim_{n\to\infty} i\omega_n \int_{\Gamma_1} \partial_\nu v_n \overline{\partial_\nu u_n} d\Gamma = 0. \tag{3.26}$$

Furthermore, it follows from (3.15) that

$$\begin{aligned} &i\omega_n v_n + D_\infty \Delta^2 u_n + \Delta^2 L w_n(\cdot) \longrightarrow 0 && \text{in } L^2(\Omega),\\ &(i\omega_n J + 1)\partial_\nu v_n + D_\infty B_1 u_n + B_1 L w_n(\cdot) \longrightarrow 0 && \text{in } L^2(\Gamma_1),\\ &(i\omega_n \rho + 1) v_n - D_\infty B_2 u_n - B_2 L w_n(\cdot) \longrightarrow 0 && \text{in } L^2(\Gamma_1). \end{aligned} \tag{3.27}$$

Consequently,

$$\begin{aligned} &i\omega_n \int_\Omega v_n \overline{u_n} dx + D_\infty a(u_n) + a(L w_n(\cdot), u_n)\\ &+ \int_{\Gamma_1} [i\omega_n (J\partial_\nu v_n \overline{\partial_\nu u_n} + \rho v_n \overline{u_n}) + (\partial_\nu v_n \overline{\partial_\nu u_n} + v_n \overline{u_n})] d\Gamma \longrightarrow 0. \end{aligned} \tag{3.28}$$

Substituting (3.16), (3.17) and (3.24)–(3.26) into (3.28), we obtain

$$\lim_{n\to\infty} D_\infty a(u_n) = \lim_{n\to\infty} \int_\Omega |v_n|^2 dx - \lim_{n\to\infty} a(L w_n(\cdot), u_n). \tag{3.29}$$

From the Hölder inequality, we have [18]

$$\begin{aligned} &\left| \int_\Omega L(\partial_{ij} w(s)) \overline{L(\partial_{kl} w(s))} dx \right|\\ = &\left| \int_\Omega \int_0^\infty D'(s) \partial_{ij} w(s) ds \overline{\int_0^\infty D'(s) \partial_{kl} w(s) ds} dx \right|\\ \le &\int_\Omega \left| \int_0^\infty D'(s) \partial_{ij} w(s) ds \right| \left| \int_0^\infty D'(s) \partial_{kl} w(s) ds \right| dx\\ \le &\int_0^\infty |D'(s)| ds \int_\Omega \left(\int_0^\infty |D'(s)| |\partial_{ij} w(s)|^2 ds \right)^{\frac{1}{2}} \left(\int_0^\infty |D'(s)| |\partial_{kl} w(s)|^2 ds \right)^{\frac{1}{2}} dx\\ \le &\frac{1}{2}(1 - D_\infty) \left(\int_\Omega \int_0^\infty |D'(s)| |\partial_{ij} w(s)|^2 ds dx + \int_\Omega \int_0^\infty |D'(s)| |\partial_{kl} w(s)|^2 ds dx \right), \end{aligned}$$

where $\partial_{ij} = \frac{\partial^2}{\partial x_i \partial x_j}, i, j, k, l = 1, 2.$
Thus there exists a constant $C > 0$ such that

$$a(Lw(\cdot)) \le C \|w(\cdot)\|^2_{L^2([0,\infty), |D'|, V^2)}. \tag{3.30}$$

From (3.29), (3.30) and (3.18), we can assert that

$$\lim_{n\to\infty} a(u_n) = \frac{\varrho}{D_\infty} = \frac{1}{2D_\infty} \neq 0.$$

This is in contradiction to (3.22). The proof is completed. □

4. Stabilization on nonsmooth domain. When the viscoelastic plate is posed on a domain with corners, the analysis of well-posedness and stabilization for system (1.1) is somehow different from that in sections 2 and 3 and therefore must be further discussed. Mathematically, the basic difference of the four-order elliptic boundary problem between the domain without corners and the domain with corners is the integration-by-parts formula. Since the corners are present, the extra energy terms

appear. In order to take the corner effects into account, the integration-by-parts formula needs to be amended. Hence the additional feedback control is also needed to keep the dissipation of the system. On the other hand, it is well known that the solutions of the elliptic boundary problems on the nonsmooth domain lose the regularity (see, e.g., [2, 13]). Therefore, we can apply neither the energy multiplier technique [17, 18] that requires high regularity, $u \in H^{\frac{7}{2}+\epsilon}(\Omega)$ $(0 < \epsilon \leq \frac{1}{2})$, nor the method of microlocal analysis [1, 20] that demands the boundary is C^∞-smooth. Here, we introduce several proper spaces and operators. With the help of the frequency domain method, we get around the technical difficulties due to the emergence of corners and reach the desired exponential stability result without any extra assumptions on the regularity of solutions.

Before the analysis, we formulate the condition about the domain with corners.

(H5) Ω is a bounded open connected convex domain in R^2. Its boundary $\partial\Omega$ is C^2-smooth everywhere except at finite corner points $P \doteq \{P_i : i = 1, 2, \ldots, l\}$. $\partial\Omega$ is the union of two connected subsets, Γ_0 and Γ_1, where Γ_0 is not empty and has positive boundary measure, Γ_1 is relatively open in Γ, and Γ_0 and Γ_1 either are disjoint or share two common end points which are corner points.

LEMMA 4.1 (see [3]). *Let Ω satisfy* (H5). *Let $\partial\Omega$ be parameterized in a counterclockwise sense. Then, for sufficiently smooth functions u and v and for the sesquilinear form $a(\cdot,\cdot)$ in section* 2, *we have*

$$\begin{aligned}\int_\Omega \Delta^2 u\overline{v}dx = a(u,v) + \int_{\Gamma_1} (B_2 u\overline{v} - B_1 u\overline{\partial_\nu v})d\Gamma \\ + \sum_{i=1}^{l} [M_t(u)(P_i)]\overline{v(P_i)},\end{aligned} \tag{4.1}$$

where B_1 and B_2 are given as in section 1,

$$[M_t(u)(P_i)] = M(u)(P_i^+) - M(u)(P_i^-)$$

is the jump of $M_t(u)$ across P_i in the direction of increasing arc length, and

$$M(u) = (1-\mu)\left[(\nu_1^2 - \nu_2^2)\frac{\partial^2 u}{\partial x_1 \partial x_2} + \nu_1\nu_2\left(\frac{\partial^2 u}{\partial x_2^2} - \frac{\partial^2 u}{\partial x_1^2}\right)\right]$$

is the twisting moment.

Comparing (2.3) with (4.1), we notice that the sum in (4.1) denotes the work done by l corner forces $[M_t(u)(P_i)](i = 1, 2, \ldots, l)$, acting through l corner displacements $v(P_i)$ [3, 4].

For sufficiently smooth function $u(t)$, we can naturally derive the following boundary feedback law:

$$\begin{cases} u(t) = \partial_\nu u(t) = 0 & \text{on } \Gamma_0, \\ J\partial_\nu u''(t) + B_1 u(t) + \displaystyle\int_0^\infty D'(s)B_1 u(t-s)ds = -\partial_\nu u'(t) & \text{on } \Gamma_1/P, \\ \rho u''(t) - B_2 u(t) - \displaystyle\int_0^\infty D'(s)B_2 u(t-s)ds = -u'(t) & \text{on } \Gamma_1/P \end{cases} \tag{4.2}$$

and

$$[M_t(u)(t,P_i)] + \int_0^\infty D'(s)[M_t(u)(t-s,P_i)]ds = u'(t,P_i) \quad \text{for } P_i \in \Gamma/\overline{\Gamma}_0. \tag{4.3}$$

Then a straight computation gives that for sufficiently smooth functions u and u', the energy of the system on the domain with corners, still defined by (2.2), is nonincreasing:

$$\begin{aligned}\frac{d}{dt}E(t) = -\frac{1}{2}\int_0^\infty D''(s)a(u(t-s)-u(t))ds - \int_{\Gamma_1}(|u'(t)|^2 \\ +|\partial_\nu u'(t)|^2)d\Gamma - \sum_{P_i\in\Gamma/\overline{\Gamma}_0}|u'(t,P_i)|^2 \le 0.\end{aligned} \tag{4.4}$$

Remark 4.2. More precisely, for sufficiently smooth functions u and u', we can deduce the following boundary conditions at corner points [4]:

$$[M_t(u)(t,P_i)] + \int_0^\infty D'(s)[M_t(u)(t-s,P_i)]ds = \begin{cases} u'(t,P_i) & \text{for } P_i\in\Gamma/\overline{\Gamma}_0, \\ 0 & \text{for } P_i\in\overline{\Gamma}_0. \end{cases}$$

Remark 4.3. Boundary conditions (4.2) and (4.3) require that u and u' are sufficiently smooth. In fact, in order for the pointwise limits of the twisting moments $M_t(u)(P_i^+)$, $M_t(u)(P_i^-)$ and the pointwise values $u'(P_i)$ to exist ($i=1,2,\dots,l$), the sufficient conditions are

$$\begin{cases} u\in C^{2,\alpha_1}(\Gamma), \quad 0<\alpha_1<1, \\ u'\in C^{0,\alpha_2}(\Gamma), \quad 0<\alpha_2<1. \end{cases}$$

If $u\in V^{\frac{7}{2}+\varepsilon}$, $0<\varepsilon\le\frac{1}{2}$, and $u'\in V^2$, there would be no problem in (4.2) and (4.3) from the imbedding theorem on the domain with Lipschitz boundary [13]. However, since the boundary Γ contains corners, the classical regularity results for elliptic boundary problems may no longer be valid [2, 13].

We are now in a position to determine the C_0 semigroup and its infinitesimal generator corresponding to the following system:

$$\begin{cases} u''+D_\infty\Delta^2 u+\displaystyle\int_0^\infty D'(s)\Delta^2 w(s)ds=0 & \text{in } \Omega, \\ w(t,s)=u(t-s)-u(t), \quad 0<s<\infty, & \text{in } \Omega, \\ u=\partial_\nu u=0 & \text{on } \Gamma_0, \\ J\partial_\nu u''+D_\infty B_1 u+\displaystyle\int_0^\infty D'(s)B_1 w(s)ds=-\partial_\nu u' & \text{on } \Gamma_1/P, \\ \rho u''-D_\infty B_2 u-\displaystyle\int_0^\infty D'(s)B_2 w(s)ds=-u' & \text{on } \Gamma_1/P, \\ D_\infty[M_t(u)(P_i)]+\displaystyle\int_0^\infty D'(s)[M_t(w)(s,P_i)]ds=u'(P_i) & \text{for } P_i\in\Gamma/\overline{\Gamma}_0, \\ u(0)=u_0, \quad u'(0)=u_1 & \text{in } \Omega, \\ w(0,s)=\theta(s)-u_0, \quad 0<s<\infty, & \text{in } \Omega. \end{cases} \tag{4.5}$$

To overcome the difficulties due to the loss of the regularity of solution for system (4.5), we introduce several proper spaces and operators. First, we define two

sesquilinear forms:

$$c(u,\hat{u}) = \int_\Omega u\overline{\hat{u}}dx + J\int_{\Gamma_1} \partial_\nu u\overline{\partial_\nu \hat{u}}d\Gamma + \rho\int_{\Gamma_1} u\overline{\hat{u}}d\Gamma \quad \forall u,\hat{u}\in V^2;$$

$$b(u,\hat{u}) = \int_{\Gamma_1} \partial_\nu u\overline{\hat{u}}d\Gamma + \int_{\Gamma_1} u\overline{\hat{u}}d\Gamma + \sum_{i=1}^{l} u(P_i)\overline{\hat{u}(P_i)} \quad \forall u,\hat{u}\in V^2.$$

The completion of the space V^2 normed by $c^{\frac{1}{2}}(\cdot,\cdot)$ is denoted by $\mathcal{H}$. Furthermore, we set

$$\hat{\mathcal{H}} = L^2(\Omega)\times L^2_J(\Gamma_1)\times L^2_\rho(\Gamma_1),$$
$$\|(u_1,u_2,u_3)\|_{\hat{\mathcal{H}}} = [\|u_1\|^2_{V^2} + \|u_2\|^2_{L^2_J(\Gamma_1)} + \|u_3\|^2_{L^2_\rho(\Gamma_1)}]^{\frac{1}{2}}.$$

Then we know from [24] that $\mathcal{H}$ is isomorphic and isometric to $\hat{\mathcal{H}}$.

Remark 4.4. In fact, define the operator J as $Ju = (u, \partial_\nu u|_{\Gamma_1}, u|_{\Gamma_1})$. Then J is the isometrical isomorphism of $\mathcal{H}$ onto $\hat{\mathcal{H}}$.

Let $(V^2)^*$ be the dual of V^2 pivotal to $\mathcal{H}$. It is clear that $V^2 \hookrightarrow \mathcal{H} = \mathcal{H}^* \hookrightarrow (V^2)^*$, where "$\hookrightarrow$" denotes the continuous dense injection. We define the operators as[1]

$$\mathcal{C}: \mathcal{H}\to\mathcal{H}^*, \qquad \langle \mathcal{C}u,\hat{u}\rangle_{\mathcal{H}\times\mathcal{H}} = c(u,\hat{u}) \quad \forall u,\hat{u}\in\mathcal{H};$$
$$\mathcal{A}: V^2\to(V^2)^*, \quad \langle \mathcal{A}u,\hat{u}\rangle_{(V^2)^*\times V^2} = (u,\hat{u})_{V^2} = D_\infty a(u,\hat{u}) \quad \forall u,\hat{u}\in V^2;$$
$$\mathcal{B}: V^2\to(V^2)^*, \quad \langle \mathcal{B}u,\hat{u}\rangle_{(V^2)^*\times V^2} = b(u,\hat{u}) \quad \forall u,\hat{u}\in V^2.$$

It is clear that $\mathcal{C}$ is the isomorphism of $\mathcal{H}$ and $\mathcal{A}$ is the isomorphism of V^2 onto $(V^2)^*$. $\mathcal{B}$ is a symmetric nonnegative operator.

Then we write (4.5) as follows:

$$\mathcal{C}u''(t) + \mathcal{A}u(t) + \frac{1}{D_\infty}\mathcal{A}L(u(t-\cdot)-u(t)) + \mathcal{B}u'(t) = 0 \qquad \text{in } (V^2)^*. \tag{4.6}$$

Let us set the Hilbert space

$$H_c = V^2\times\mathcal{H}\times L^2([0,\infty),|D'|,V^2)$$
$$= V^2\times\hat{\mathcal{H}}\times L^2([0,\infty),|D'|,V^2)$$

with the energy norm

$$\|(u,v,w(\cdot))\|_{H_c} = [\|u\|^2_{V^2} + \|v\|^2_{\mathcal{H}} + \|w(\cdot)\|^2_{L^2([0,\infty),|D'|,V^2)}]^{\frac{1}{2}}$$
$$= [\|u\|^2_{V^2} + \|v\|^2_{L^2(\Omega)} + \|w(\cdot)\|^2_{L^2([0,\infty),|D'|,V^2)}$$
$$+ \|\partial_\nu v\|^2_{L^2_J(\Gamma_1)} + \|v\|^2_{L^2_\rho(\Gamma_1)}]^{\frac{1}{2}}.$$

Set $(\mathcal{A}_w w)(s) = \mathcal{A}(w(s))$. Then $\mathcal{A}_w$ is the isomorphism of $L^2([0,\infty),|D'|,V^2)$ onto $(L^2([0,\infty),|D'|,V^2))^* = L^2([0,\infty),|D'|,(V^2)^*)$. Note that H_c^*, the dual of H_c,

[1] Here we denote the value of u^* at u by $\langle u^*,u\rangle$.

is $(V^2)^* \times \mathcal{H} \times L^2([0,\infty), |D'|, (V^2)^*)$. Now we define two operators as

$$C = \begin{pmatrix} \mathcal{A} & 0 & 0 \\ 0 & \mathcal{C} & 0 \\ 0 & 0 & \mathcal{A}_w \end{pmatrix} \text{ is the isomorphism of } H_c \text{ onto } H_c^*;$$

$$A_c = \begin{pmatrix} 0 & \mathcal{A} & 0 \\ -\mathcal{A} & -\mathcal{B} & -\dfrac{1}{D_\infty}\mathcal{A}L \\ 0 & -\mathcal{A}_w & -\mathcal{A}_w \dfrac{\partial}{\partial s} \end{pmatrix} : D(A_c) \to H_c^*,$$

$$D(A_c) = \Big\{ u, v, w(\cdot)) \in H_c \,:\, v \in V^2, \mathcal{A}u + \mathcal{B}v + \frac{1}{D_\infty}\mathcal{A}Lw(\cdot) \in \mathcal{H},$$
$$w(0) = 0, w(\cdot) \in C^1([0,\infty), |D'|, V^2) \Big\}.$$

It is obvious that $C^{-1}A_c : D(A_c) \to H_c$. Assume $z(t) \doteq (u(t), u'(t), u(t-s) - u(t))$. Then we can formulate (4.5) as

$$z'(t) = C^{-1}A_c z(t), \qquad z(0) = z_0 \in H_c. \tag{4.7}$$

The existence of the unique solution of system (4.5) on the nonsmooth domain is guaranteed by the following theorem.

THEOREM 4.5. *Let the domain Ω and the relaxation function $D(\cdot)$ satisfy* (H5) *and* (H1)–(H3)*, respectively. Then the unbounded operator $C^{-1}A_c$ generates a C_0 contraction semigroup $e^{tC^{-1}A_c}$ on H_c.*

Proof. For $z = (u, v, w(\cdot)) \in D(A_c)$, we have that

$$\begin{aligned} \Re e(C^{-1}A_c z, z)_{H_c} &= (v,u)_{V^2} - c\left(\mathcal{C}^{-1}\mathcal{A}u + \mathcal{C}^{-1}\mathcal{B}v + \frac{1}{D_\infty}\mathcal{C}^{-1}\mathcal{A}Lw(\cdot), v\right) \\ &\quad - \left(v + \frac{\partial}{\partial s}w(\cdot), w(\cdot)\right)_{L^2([0,\infty),|D'|,V^2)} \\ &= -b(v,v) - \int_0^\infty |D'(s)| a\left(\frac{\partial}{\partial s}w(s), w(s)\right) ds \\ &= -\frac{1}{2}\int_0^\infty D''(s)a(w(s))ds - \int_{\Gamma_1}(|\partial_\nu v|^2 + |v|^2)d\Gamma \\ &\quad - \sum_{P_i \in \Gamma/\overline{\Gamma}_0} |v(P_i)|^2 \le 0. \end{aligned} \tag{4.8}$$

Thus $C^{-1}A_c$ is dissipative. To prove the range of $I - C^{-1}A_c$ is H_c, we let $f = (f_1, f_2, f_3(\cdot)) \in H_c$ and consider the resolvent equation $(I - C^{-1}A_c)z = f$, $z = (u, v, w(\cdot)) \in D(C^{-1}A_c)$. It is equivalent to

$$\begin{cases} u - v = f_1 \in V^2, \\ v + \mathcal{C}^{-1}\mathcal{A}u + \mathcal{C}^{-1}\mathcal{B}v + \dfrac{1}{D_\infty}\mathcal{C}^{-1}\mathcal{A}Lw(\cdot) = f_2 \in \mathcal{H}, \\ w(\cdot) + v + \dfrac{\partial}{\partial s}w(\cdot) = f_3(\cdot) \in L^2([0,\infty), |D'|, V^2). \end{cases}$$

Therefore,

$$v = u - f_1 \in V^2, \tag{4.9}$$

$$u + \mathcal{C}^{-1}\mathcal{A}u + \mathcal{C}^{-1}\mathcal{B}u + \frac{1}{D_\infty}\mathcal{C}^{-1}\mathcal{A}Lw(\cdot) = f_1 + \mathcal{C}^{-1}\mathcal{B}f_1 + f_2 \in \mathcal{H}, \tag{4.10}$$

$$w(s) = -(1 - e^{-s})(u - f_1) + \int_0^s e^{\sigma - s} f_3(\sigma)d\sigma \in L^2([0,\infty), |D'|, V^2). \tag{4.11}$$

From (4.10), we have

$$\begin{aligned} & c(u,\phi) + D_\infty a(u,\phi) + b(u,\phi) + a(Lw(\cdot),\phi) \\ = \;& c(f_1 + f_2, \phi) + b(f_1,\phi) \quad \forall \phi \in V^2. \end{aligned} \tag{4.12}$$

By the same argument as that in Theorem 2.1, we know that (4.11) and (4.12) admit a unique solution $u \in V^2$ for any $f \in H_c$. Combining this with (4.9) and (4.11), we get that the range of $I - C^{-1}A_c$ is H_c. Thus $C^{-1}A_c$ is m-dissipative. □

By an approach similar to that employed in Theorem 3.3, we can get the following result.

THEOREM 4.6. *Suppose that the relaxation function $D(\cdot)$ satisfies hypotheses* (H1)–(H4) *and the domain Ω satisfies* (H5). *Then we have*

$$0 \in \rho(C^{-1}A_c).$$

THEOREM 4.7. *Suppose the relaxation function $D(\cdot)$ satisfies* (H1)–(H4) *and the nonsmooth domain Ω satisfies* (H5). *Then there exist positive constants M_c and ϖ_c such that the energy $E(t)$ of the system* (4.5), *still defined by* (2.2), *satisfies*

$$E(t) \le M_c e^{-\varpi_c t} \qquad \forall t \ge 0.$$

Proof. From Lemma 3.1, we need only to show

$$\sup\{\|(i\omega - C^{-1}A_c)^{-1}\|_{H_c} \;:\; \omega \in R\} < \infty.$$

If it fails, there are a sequence $\omega_n \in R$ and a sequence of vectors $z_n = (u_n, v_n, w_n(\cdot)) \in D(A_c)$ such that

$$\|z_n\|_{H_c} = 1, \quad n = 1, 2, \ldots, \tag{4.13}$$

$$\lim_{n\to\infty} \|(i\omega_n - C^{-1}A_c)z_n\|_{H_c} = 0. \tag{4.14}$$

It follows from (4.14) that $\lim_{n\to\infty} \Re e(C^{-1}A_c z_n, z_n)_{H_c} = 0$. Thus

$$\lim_{n\to\infty} \int_0^\infty D''(s)a(w_n(s))ds = 0, \tag{4.15}$$

$$\lim_{n\to\infty} \int_{\Gamma_1} |\partial_\nu v_n|^2 d\Gamma = \lim_{n\to\infty} \int_{\Gamma_1} |v_n|^2 d\Gamma = 0, \tag{4.16}$$

$$\lim_{n\to\infty} v_n(P_i) = 0, \quad i = 1, \ldots, l. \tag{4.17}$$

From the hypothesis (H4), (4.15) yields

$$\lim_{n\to\infty} \|w_n(\cdot)\|_{L^2([0,\infty),|D'|,V^2)} = 0. \tag{4.18}$$

Moreover, by (4.14),

$$i\omega_n u_n - v_n \longrightarrow 0 \quad \text{in } V^2; \tag{4.19}$$

$$i\omega_n v_n + \mathcal{C}^{-1}\mathcal{A}u_n + \mathcal{C}^{-1}\mathcal{B}v_n + \frac{1}{D_\infty}\mathcal{C}^{-1}\mathcal{A}Lw_n(\cdot) \longrightarrow 0 \quad \text{in } \mathcal{H}; \tag{4.20}$$

$$i\omega_n w_n(\cdot) + v_n + \frac{\partial}{\partial s}w_n(\cdot) \longrightarrow 0 \quad \text{in } L^2([0,\infty), |D'|, V^2). \tag{4.21}$$

Using the same argument as that in Theorem 3.4, we can obtain from (4.18) and (4.21) that

$$\lim_{n\to\infty} a(u_n) = 0. \tag{4.22}$$

On the other hand, under the assumption $\varrho \doteq \lim_{n\to\infty}\int_\Omega |v_n|^2 dx$, we can deduce from (4.13), (4.16) and (4.18) that $\lim_{n\to\infty} D_\infty a(u_n) = 1 - \lim_{n\to\infty}\int_\Omega |v_n|^2 dx = 1 - \varrho$. As in the proof of Theorem 3.4, it follows from (4.19) that

$$\lim_{n\to\infty} -i\omega_n \int_\Omega v_n\overline{u_n}dx = \lim_{n\to\infty}\int_\Omega |v_n|^2 dx = \varrho \tag{4.23}$$

and

$$\lim_{n\to\infty} i\omega_n \int_{\Gamma_1} \partial_\nu v_n \overline{\partial_\nu u_n} d\Gamma = \lim_{n\to\infty} i\omega_n \int_{\Gamma_1} v_n\overline{u_n}d\Gamma = 0. \tag{4.24}$$

Moreover, (4.20) yields

$$\begin{aligned} 0 \longleftarrow\ & i\omega_n c(v_n, u_n) + c\left(\mathcal{C}^{-1}\mathcal{A}u_n + \mathcal{C}^{-1}\mathcal{B}v_n + \frac{1}{D_\infty}\mathcal{C}^{-1}\mathcal{A}Lw_n(\cdot), u_n\right) \\ =\ & i\omega_n\left[\int_\Omega v_n\overline{u_n}dx + \int_{\Gamma_1}(J\partial_\nu v_n\overline{u_n} + \rho v_n\overline{u_n})d\Gamma\right] + D_\infty a(u_n) \\ & + b(v_n, u_n) + a(Lw_n(\cdot), u_n). \end{aligned} \tag{4.25}$$

Notice that $|u(P_j)| \le C_j$ ($C_j > 0$ are constants, $j = 1, 2, \ldots$) since $u \in C^{0,\alpha}(\Gamma_1)$, $0 < \alpha < 1$. Therefore, substituting (4.16)–(4.18), (4.23) and (4.24) into (4.25), we obtain

$$\lim_{n\to\infty} D_\infty a(u_n) = \varrho.$$

Therefore, $\varrho = \frac{1}{2}$. This is in contradiction to (4.22). □

Acknowledgments. We would like to thank Professors Kangsheng Liu and Zhuangyi Liu for their helpful discussions. We also appreciate the suggestions of the referees.

REFERENCES

[1] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary,* SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[2] R. Bey, J.-P. Loheac, and M. Moussaoui, *Singularities of the solution of a mixed problem for a general second order elliptic equation and boundary stabilization of the wave equation,* J. Math. Pures. Appl. (9), 78 (1999), pp. 1043–1067.
[3] G. Chen, M. P. Coleman, and Z. Ding, *Some corner effects on the loss of self adjointness and the non-excitation of vibration for thin plates and shells,* Quart. J. Mech. Appl. Math., 51 (1998), pp. 213–239.
[4] G. Chen, M. P. Coleman, and K. S. Liu, *Boundary stabilization of Donnell's shallow circular cylindrical shell,* J. Sound Vibration, 209 (1998), pp. 265–298.
[5] S. Chen, K. Liu, and Z. Liu, *Spectrum and stability for elastic systems with global or local Kelvin–Voigt damping,* SIAM J. Appl. Math., 59 (1998), pp. 651–668.
[6] C. M. Dafermos, *Asymptotic stability in viscoelasticity,* Arch. Rational Mech. Anal., 37 (1970), pp. 297–308.
[7] C. M. Dafermos, *An abstract Volterra equation with applications to linear viscoelasticity,* J. Differential Equations, 7 (1970), pp. 554–569.
[8] A. Day, *The decay of energy in a viscoelastic body,* Matematika, 27 (1980), pp. 268–286.
[9] W. Desch and R. K. Miller, *Exponential stabilization of Volterra integrodifferential equations in Hilbert space,* J. Differential Equations, 79 (1987), pp. 366–389.
[10] W. Desch and R. K. Miller, *Exponential stabilization of Volterra integral equations with singular kernels,* J. Integral Equations Appl., 1 (1988), pp. 397–433.
[11] R. H. Fabiano and B. Lazzari, *On the existence and asymptotic stability of solutions for linearly viscoelastic solids,* Arch. Rational Mech. Anal., 116 (1991), pp. 139–152.
[12] G. B. Folland, *Real Analysis, Modern Techniques and Their Applications,* John Wiley and Sons, New York, 1984.
[13] P. Grisvard, *Elliptic Problems in Nonsmooth Domains,* Pitman, London, 1985.
[14] S. Hansen and E. Zuazua, *Exact controllability and stabilization of a vibrating string with an interior point mass,* SIAM J. Control Optim., 33 (1995), pp. 1357–1391.
[15] K. B. Hannsgen and R. L. Wheeler, *Viscoelastic and boundary feedback damping: Precise energy decay rates when creep modes are dominant,* J. Integral Equations Appl., 2 (1990), pp. 495–527.
[16] F. L. Huang, *Characteristic conditions for exponential stability of linear dynamical systems in Hilbert spaces,* Ann. Differential Equations, 1 (1985), pp. 43–56.
[17] V. Komornik, *Exact Controllability and Stabilization: The Multiplier Method,* Masson, Paris, John Wiley and Sons, Chichester, 1994.
[18] J. E. Lagnese, *Boundary Stabilization of Thin Plates,* SIAM Stud. Appl. Math. 10, SIAM, Philadelphia, 1989.
[19] J. E. Lagnese and J. L. Lions, *Modeling, Analysis and Control of Thin Plates,* Masson, Paris, 1988.
[20] I. Lasiecka, *Sharp trace estimates of solutions to Kirchhoff and Euler-Bernoulli equations,* Appl. Math. Optim., 28 (1993), pp. 277–306.
[21] J. L. Lions, *Exact controllability, stabilization and perturbations for distributed systems,* SIAM Rev., 30 (1988), pp. 1–68.
[22] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications,* Vol. 1, Springer-Verlag, New York, 1972.
[23] W. Littman and L. Markus, *Stabilization of a hybrid system of elasticity by feedback boundary damping,* Ann. Mat. Pura Appl. (4), 152 (1998), pp. 281–330.
[24] K. S. Liu, *Local Boundary Control and Stabilization for Distributed Parameter Systems,* Ph.D. thesis, Fudan University, People's Republic of China, 1991 (in Chinese).
[25] K. S. Liu and Z. Y. Liu, *On the type of C_0-semigroup associated with the abstract linear viscoelastic system,* Z. Angew. Math. Phys., 47 (1996), pp. 1–15.
[26] K. S. Liu and Z. Y. Liu, *Boundary stabilization of nonhomogenous beam with rotatory inertia at the tip,* J. Comput. Appl. Math., 114 (2000), pp. 1–10.
[27] Z. Y. Liu and S. M. Zheng, *Semigroups Associated With Dissipative Systems,* CRC Press, Boca Raton, FL, 1999.
[28] L. Markus and Y. You, *Dynamical boundary control for elastic plates of general shape,* SIAM J. Control Optim., 31 (1993), pp. 983–992.
[29] S. Micu and E. Zuazua, *Asymptotics for the spectrum of a fluid/structure hybrid system arising in the control of noise,* SIAM J. Math. Anal., 29 (1998), pp. 967–1001.
[30] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations,* Springer-Verlag, New York, 1983.
[31] B. Rao, *Stabilization of elastic plates with dynamical boundary control,* SIAM J. Control Optim., 36 (1998), pp. 148–163.

[32] R. Rebarber, *Exponential stability of coupled beams with dissipative joints: A frequency domain approach,* SIAM J. Control Optim., 33 (1995), pp. 1–28.
[33] D. L. Russell, *A general framework for the study of indirect damping mechanisms in elastic systems,* J. Math. Anal. Appl., 173 (1993), pp. 339–358.
[34] G. Weiss, *Weak L^p-stability of a linear semigroup on a Hilbert space implies exponential stability,* J. Differential Equations, 76 (1988), pp. 269–285.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1661–1680

CONTROLLABILITY TO THE TRAJECTORIES OF PHASE-FIELD MODELS BY ONE CONTROL FORCE*

F. AMMAR KHODJA†, A. BENABDALLAH†, C. DUPAIX†, AND I. KOSTIN†

Abstract. In this article, we study the controllability to the trajectories of 2×2 nonlinear parabolic systems for control forces acting on a single equation of the system. This result, which in particular applies to Caginalp's phase-field model, actually extends those obtained for the semilinear heat equations. The proof relies on Kakutani's fixed point theorem and makes use of an observability estimate for the associated linearized system.

Key words. nonlinear parabolic systems, controllability

AMS subject classifications. 35K50, 35K60, 93C20, 93B05

DOI. S0363012902417826

1. Introduction.

The aim of this article is the study of controllability for phase-field models when the control force acts on a single equation of the system. The models that we consider here are generalizations of Caginalp's phase-field model [9] (see also [8]) in its enthalpy formulation. For given time $T > 0$ and bounded domain $\Omega \subset \mathbb{R}^N$ $(1 \leq N < 6)$ with smooth boundary $\partial\Omega$, it reads as follows:

$$\begin{cases} \phi_t = \Delta\phi - h(\phi) + u, & \\ u_t = D\Delta u - \Delta\phi \ + f & \text{in } (0,T) \times \Omega = Q_T, \\ \phi = u = 0 & \text{on } (0,T) \times \partial\Omega = \Sigma_T, \\ \phi(0) = \phi_0,\ u(0) = u_0 & \text{in } \Omega, \end{cases} \tag{1}$$

where $D > 0$ is a constant and $h : \mathbb{R} \to \mathbb{R}$ is a C^1 function (locally Lipschitz continuous is actually sufficient). We assume that $f = \chi_\omega g$, where $g \in L^2(Q_T)$ and χ_ω is the characteristic function of a nonempty open and fixed set $\omega \Subset \Omega$ (namely $\overline{\omega} \subset \Omega$). The unknown functions ϕ and u can, respectively, be interpreted as a phase parameter and an enthalpy.

As we will see below, a part of the results here obtained depend on the behavior of the function h near $\pm\infty$. However, this hypothesis on h is not a restriction from the phase transition point of view. In fact, besides some monotonicity property of h, the main requirement for the phase-field models is that at least for $|s| \leq 1$, $h(s) = c_1 H'(s) - c_2 s$, where c_1, c_2 are positive constants and H is a symmetric double-well potential having two local minima at $s = \pm 1$. Although this model is known to be inconsistent with the second law of thermodynamics, it has been proved to be quite useful since many other models of the phase transition phenomena can be derived from Caginalp's model by taking suitable limits with respect to the parameters of the system (i.e., D, c_1, and c_2).

In this work we are interested in the controllability of such systems and we recall, therefore, that system (1) is said to be *controllable to the trajectories at time T if for any initial data (ϕ_0, u_0) there exists a control f such that the corresponding solution*

*Received by the editors November 14, 2002; accepted for publication (in revised form) April 2, 2003; published electronically November 6, 2003.
http://www.siam.org/journals/sicon/42-5/41782.html

†Université de Franche-Comté, Département de Mathématiques, CNRS-UMR 6623, 16 Route de Gray, 25030 Besançon Cedex, France (ammar@descartes.univ-fcomte.fr, assia@descartes.univ-fcomte.fr, dupaix@descartes.univ-fcomte.fr, kostin@descartes.univ-fcomte.fr).

(ϕ, u) of (1) *is defined on* $[0, T]$ *and satisfies*

$$(\phi(T,.), u(T,.)) = (\phi^*(T,.), u^*(T,.)) \text{ a.e. in } \Omega, \tag{2}$$

where (ϕ^*, u^*) *is any bounded solution of* (1) *defined on* $[0, T]$ *associated with the data* ϕ_0^*, u_0^*, *and* f^*.

System (1) *is locally controllable to the trajectories at time* T *if there exists a constant* $r > 0$ *such that for any initial data* (ϕ_0, u_0) *satisfying* $\|\phi_0 - \phi_0^*\| + \|u_0 - u_0^*\| < r$ *the solution* (ϕ, u) *of system* (1) *satisfies* (2).

The functional spaces will be precised in the forthcoming sections.

Considering $(\phi - \phi^*, u - u^*)$ as a new unknown, the problem is reduced to drive the solution at time T to the state $(0, 0)$. Therefore, referring to the linear case, this kind of controllability is often called *null-controllability*. Notice that in the finite dimensional case and for linear problems, null, exact, and approximate controllability are equivalent (see, for instance, [21]).

The originality of our approach is to control 2×2 nonlinear systems by *acting (locally in space) on a single equation* of it. From this point of view and to our knowledge, our results are the first ones in this direction. Controlling a system with a minimum number of forces, or by forces satisfying an algebraic or a differential (or any other type) relation, is a common problem in the control theory (see, for instance, [19, Chapter V, p. 322] for other types of systems). Other approaches deal with dynamic controls which, roughly speaking, means that the control itself obeys a dynamic (see [1] and the references therein). From this last point of view, the variable u of system (1) can be seen as a dynamic control with respect to the variable ϕ.

Null-controllability of linear and semilinear heat equations has been extensively studied in recent years. The main ingredients, therefore, are global estimates of Carleman type for linear parabolic equations with an additional fixed point argument for the nonlinear case. The appropriate version of the Carleman estimates was established, in their most general form, by Imanuvilov [17] and Fursikov and Imanuvilov [16] (see also Lebeau and Robbiano [20] for the linear heat equation). These estimates were first used to prove the controllability to the trajectories of semilinear heat equations with globally Lipschitz continuous nonlinearities (see [16]) and then to extend this property to superlinear heat equations (see, for instance, [12], [5], [14]). In contrast with this collection of results, few works deal with the controllability of linear or nonlinear parabolic systems. In [10] De Teresa considers a semilinear heat equation coupled with an adjoint problem. The specific form of the system allows the proof of controllability to the trajectories in the case of globally Lipschitz continuous nonlinearities. However, in the case of system (1) such a proof does not work since the two equations are coupled in a very different way. Recently, Anita and Barbu [3] have considered a 2×2 reaction-diffusion system with a bilinear nonlinearity. They proved the local controllability to the stationary solutions of the system by two control forces acting on both equations of the system. Their result seems to be the first one dealing with the controllability to the trajectories for semilinear parabolic systems. More recently, Barbu [6] has considered the phase-field model (1) with a cubic nonlinearity and in an (equivalent) temperature formulation. He has proved the local controllability to the stationary solutions by *two control* forces localized on the same subdomain. At this stage let us point out that *the one control force result* cannot be obtained as a simple generalization of Barbu's result. It relies on the construction (using a multiplier technique) of a suitable functional Λ with suitable weights (see Lemma 3.4 and (34)). This is one of the crucial points of the proof of our results. It

also appears to be useful for the study of the null-controllability by a single force of abstract general linear "parabolic" systems of two equations [2]. Besides the sufficient conditions of controllability, in [2] we also construct a counterexample showing that if the coupling operators are "too compact," null-controllability fails to hold.

The sketch of the proof of our results is by now well known and was adopted for a scalar semilinear parabolic equation by several authors (see [16], [12], [5], [14]). The idea is to prove an observability estimate for the "linearized" system and then to use a fixed point theorem. The main difficulty is to prove the observability estimate for the "linearized" system corresponding to *the control by a single force*. To achieve this goal we first establish a global Carleman-type estimate for a linear parabolic system. This implies the controllability by two forces. In a second step we obtain an observability estimate which implies the controllability of the linear system by a single force. The fixed point method we use makes it necessary to construct a sufficiently regular control which induces by parabolic regularity a solution in a suitable space (here $L^\infty(Q_T)$). This explains the need for a "refined" observability estimate (see Lemma 3.4).

The paper is organized as follows. In the second section, by admitting an essential lemma (see Lemma 2.3), we state and prove the main result of this article (Theorem 2.1). We prove this lemma in the fourth section, after having shown in the third section the crucial observability estimate for a linear adjoint problem (see Lemma 3.4).

We conclude this section with the following two remarks.

1. Since the diffusion coefficient D in (1) changes only the constants in the proofs, we assume for simplicity that $D = 1$. However, the dependence of these constants on D (and on c_1 and c_2 for Caginalp's model) is explicit and may be of interest from the point of view of the theory of phase transitions.

2. Also for the sake of simplicity, we consider only the homogeneous Dirichlet boundary conditions. However, all the results seem to still be valid for the homogeneous Neumann boundary conditions.

2. The main result. To simplify, we will work with a function h satisfying

$$h \in C^1(R), \quad h(0) = 0. \tag{3}$$

With this assumption, system (1) admits $(0, 0)$ as a global solution associated with null initial data and control.

Let $1 \le N < 6$ and $\frac{N+2}{2} < q_N < 2\frac{N+2}{N-2}$ if $N \ge 3$, and $q_N \in (2, +\infty)$ if $N = 1, 2$ (even if the physical case reduces to $N = 1, 2, 3$). We use the following notation:

$$W_q^{2,1}(Q_T) = \{\zeta \in L^q(Q_T);\ D_t^r D_x^s \zeta \in L^q(Q_T); 2r + s \le 2\},$$

and $W^{m,p}(\Omega)$ is the standard Sobolev space. With the assumption on q_N, we have the following embeddings [18]:

$$W_{q_N}^{2,1}(Q_T) \hookrightarrow L^\infty(Q_T), \quad W^{2(1-\frac{1}{q_N}),q_N}(\Omega) \hookrightarrow L^\infty(\Omega). \tag{4}$$

Our main result is the following theorem.

THEOREM 2.1. *Assume that h satisfies* (3). *Let $T > 0$, $1 \le N < 6$, and let (ϕ^*, u^*) be a globally defined and bounded solution of* (1) *associated with the data $\phi_0^*, u_0^* \in L^2(\Omega)$ and $g^* \in L^2(Q_T)$.*

(i) Controllability to the trajectories. *If h satisfies*

$$\lim_{|s|\to+\infty} \frac{h(s)}{|s|\ \ln^{3/2}(1 + |s|)} = 0, \tag{5}$$

then $\forall \phi_0, u_0 \in L^2(\Omega)$ *with* $\phi_0 - \phi^*, u_0 - u_0^* \in H_0^1(\Omega) \cap W^{2(1-\frac{1}{q_N}),q_N}(\Omega)$, *one can find* $g \in L^2(Q_T)$ *with* $g - g^* \in L^{q_N}(Q_T)$ *such that there exists* (ϕ_g, u_g) *solution of* (1) *with* $\phi_g - \phi^*, u_g - u^* \in W_{q_N}^{2,1}(Q_T)$ *and satisfying*

$$(\phi_g - \phi^*)(T) = 0, \ \ (u_g - u^*)(T) = 0.$$

(ii) Local controllability to the trajectories. *There is* $\rho > 0$ *such that if* $\phi_0 - \phi^*, u_0 - u_0^* \in H_0^1(\Omega) \cap W^{2(1-\frac{1}{q_N}),q_N}(\Omega)$ *with* $\|(\phi_0 - \phi^*, u_0 - u_0^*)\|_{L^\infty(\Omega)} \leq \rho$, *one can find* $g - g^* \in L^{q_N}(Q_T)$ *such that there exists* (ϕ_g, u_g) *solution of* (1) *with* $\phi_g - \phi^*, u_g - u^* \in W_{q_N}^{2,1}(Q_T)$ *and satisfying*

$$(\phi_g - \phi^*)(T) = 0, \ \ (u_g - u^*)(T) = 0.$$

Remark 2.2. This result implies the one of Barbu [6, Theorem 1, p. 364].

Proof of Theorem 2.1. For the sake of simplicity, we assume that $\phi_0^* = u_0^* = 0$ and $g^* = 0$ so that $(\phi^*, u^*) = (0, 0)$, and we introduce the function:

$$\nu(s) = \begin{cases} \frac{h(s)}{s} & \text{if } s \neq 0, \\ h'(0) & \text{if } s = 0. \end{cases}$$

For $R > 0$, we set

$$K_R = \left\{ z \in L^\infty(Q_T); \ \|z\|_{L^\infty(Q_T)} \leq R \right\},$$

and let $z \in K_R$.

Consider the "linearized" version of (1) with $a = -\nu(z)$:

$$\begin{cases} \psi_t = \Delta\psi + a\psi + v & \text{in } Q_T, \\ v_t = \Delta v - \Delta\psi + \chi_\omega g & \text{in } Q_T, \\ \psi = v = 0 & \text{on } \Sigma_T, \\ \psi(0,.) = \psi_0, \, v(0,.) = v_0 & \text{in } \Omega. \end{cases} \tag{6}$$

Note that if $(\phi^*, u^*) \neq (0, 0)$, it suffices to consider the function

$$\widetilde{\nu} \ : \ Q_T \times \mathbb{R} \to \mathbb{R},$$
$$\widetilde{\nu}(t, x, s) = \begin{cases} \frac{h(s+\phi^*) - h(\phi^*)}{s} & \text{if } s \neq 0, \\ h'(\phi^*) & \text{if } s = 0 \end{cases}$$

and to study (6) with $a(t, x) = -\widetilde{\nu}(t, x, z)$. The crucial point is the following result.

LEMMA 2.3. *Let* $X_0 = (\psi_0, v_0) \in (H_0^1(\Omega) \cap W^{2(1-\frac{1}{q_N}),q_N}(\Omega))^2$. *For any* $T > 0$, *there exists* $g \in L^{q_N}(Q_T)$ *such that the associated solution* ψ_g, v_g *of* (6) *is in* $L^2(0, T; H_0^1(\Omega)) \cap W_{q_N}^{2,1}(Q_T)$ *and, moreover,*

$$(\psi_g, v_g)(T) = 0 \ \ \textit{a.e. } \Omega, \tag{7}$$

$$\|\chi_\omega g\|^2_{L^{q_N}(Q_T)} \leq C_T \ \|X_0\|^2_{L^2(\Omega)}, \tag{8}$$

with

$$C_T = \exp\left(C \left(1 + \frac{1}{T} + (1 + \|a\|_\infty) T + \|a\|_\infty^{2/3} \right) \right).$$

The next sections are devoted to the proof of this lemma. With this result in hand, let us continue the proof of Theorem 2.1.

For each $z \in K_R$, we apply Lemma 2.3 and consider the set $F(z) \subset L^2(Q_T)$ of the first components ψ_g of all the solutions $\psi_g, v_g \in L^2(0,T;H_0^1(\Omega)) \cap W^{2,1}_{q_N}(Q_T)$ associated with any control $g \in L^{q_N}(Q_T)$ such that $(\psi_g, v_g)(T) = 0$ a.e. sΩ and $\|\chi_\omega g\|^2_{L^{q_N}(Q_T)} \leq C_T \|X_0\|^2_{L^2(\Omega)}$. The set is a nonempty closed convex subset of $L^2(Q_T)$.

Let us prove that for $R > 0$ sufficiently large we have $F(K_R) \subset K_R$. To do this, we first show that

$$\|\psi_g\|^2_{L^\infty(Q_T)} \leq \|(\psi_g, v_g)\|^2_{L^\infty(Q_T)} \leq C_T \|X_0\|^2_{L^\infty(\Omega)}. \tag{9}$$

Indeed, using the fundamental matrix $\Gamma = (\Gamma_{ij})_{1\leq i,j\leq 2}$ associated with the parabolic operator

$$L_0 = \begin{pmatrix} \partial_t - \Delta & 0 \\ \Delta & \partial_t - \Delta \end{pmatrix}$$

(see [18, p. 620]), with homogeneous Dirichlet boundary conditions, we can write down the solution of (6) with the notation $X_g = (\psi_g, v_g)$

$$\begin{aligned} X_g(t,x) = &\int_\Omega \Gamma(t,0,x,y) X_0(y)\, dy \\ &+ \int_0^t \int_\Omega \Gamma(t,\tau,x,y)\, (a\psi_g + v_g, \chi_\omega g)\, (\tau,y) dy\, d\tau. \end{aligned}$$

The entries Γ_{ij} of Γ satisfy for $1 \leq i,j \leq 2$

$$\begin{gathered} \left|D_t^k D_x^s \Gamma_{ij}\right| \leq C(t-\tau)^{-\frac{N+2k+s}{2}} \exp\left(-C\frac{|x-y|^2}{t-\tau}\right), \\ x, y \in \Omega,\ 0 < \tau < t < T, \quad 2s + k = 0, 1, 2 \end{gathered} \tag{10}$$

(notice that in this special situation, $\Gamma_{12} = 0$). The positive constant C depends only on Ω and $1 \leq i,j \leq 2$. We will assume, without loss of generality, that it is the same constant. Set

$$\Gamma_0(t,x) = t^{-\frac{N}{2}} \exp\left(-C\frac{|x|^2}{t}\right), \quad (t,x) \in (0,\infty) \times \mathbb{R}^N.$$

We know by [18, Theorem 10.4] that $X_g \in W^{2,1}_{q_N}(Q_T)$ and the embeddings (4) allow us to write

$$\begin{aligned} \|X_g(t)\|_{L^\infty(\Omega)} \leq C \Bigg(&\|\Gamma_0(t,.) * |X_0|\|_{L^\infty(\Omega)} + \left\|\Gamma_0 \underset{t,x}{*} \chi_\omega\, |g|\right\|_{L^\infty(\Omega)} \\ &+ (1 + \|a\|_\infty) \left\|\Gamma_0 \underset{t,x}{*} |X_g|\right\|_{L^\infty(\Omega)} \Bigg), \end{aligned} \tag{11}$$

where $(f \underset{t,x}{*} g)(t,x) = \int_0^t d\tau \int_\Omega f(t-\tau, x-y) dy$. Now, we first have

$$\begin{aligned} \|\Gamma_0(t,.) * |X_0|\|_{L^\infty(\Omega)} &\leq \|\Gamma_0(t,.)\|_{L^1(\Omega)} \|X_0\|_{L^\infty(\Omega)} \\ &\leq C \|X_0\|_{L^\infty(\Omega)}, \end{aligned}$$

since for any $t>0$, $\|\Gamma_0(t,.)\|_{L^1(\Omega)} \leq C$ (a constant depending only on the dimension N) and, in the same way,

$$\left\| \Gamma_0 \underset{t,x}{*} |X_g| \right\|_{L^\infty(\Omega)} \leq C \int_0^t \|X_g(\tau)\|_{L^\infty(\Omega)} \, d\tau.$$

From Young's inequality (see [7]), we get with $\frac{1}{p}+\frac{1}{q_N}=1$ and the condition on q_N:

$$\begin{aligned} \left\| \Gamma_0 \underset{t,x}{*} \chi_\omega |g| \right\|_{L^\infty(\Omega)} &\leq \left\| \Gamma_0 \underset{t,x}{*} \chi_\omega |g| \right\|_{L^\infty(Q_t)} \\ &\leq \|\Gamma_0\|_{L^p(Q_t)} \|\chi_\omega g\|_{L^{q_N}(Q_t)} \\ &\leq C T^{-\frac{N+2}{2q_N}+1} \|\chi_\omega g\|_{L^{q_N}(Q_T)}. \end{aligned}$$

These three last inequalities transform (11) into

$$\begin{aligned} \|X_g(t)\|_{L^\infty(\Omega)} \leq C \Bigg(& \|X_0\|_{L^\infty(\Omega)} + T^{-\frac{N+2}{2q_N}+1} \|\chi_\omega g\|_{L^{q_N}(Q_T)} \\ & + \left(1+\|a\|_{L^\infty(Q_T)}\right) \int_0^t \|X_g(\tau)\|_{L^\infty(\Omega)} \, d\tau \Bigg), \end{aligned}$$

and from Gronwall's inequality we get

$$\|X_g\|_{L^\infty(Q_T)} \leq C e^{C\left(1+\|a\|_{L^\infty(Q_T)}\right)T} \left(\|X_0\|_{L^\infty(\Omega)} + T^{-\frac{N+2}{2q_N}+1} \|\chi_\omega g\|_{L^{q_N}(Q_T)} \right), \tag{12}$$

and (9) follows from (12) and Lemma 2.3.

Now, from (5) it follows that for any $\eta>0$ there exists $C_\eta = C_\eta(h) > 0$ such that

$$|\nu(s)|^{2/3} \leq C_\eta + \eta \ln(1+|s|) \ \forall s \in \mathbb{R} \tag{13}$$

and

$$\begin{aligned} \|\psi_g\|^2_{L^\infty(Q_T)} &\leq \exp\left(C \left(1 + \frac{1}{T} + (1+\|\nu(z)\|_{L^\infty(Q_T)})T + \|\nu(z)\|^{2/3}_{L^\infty(Q_T)} \right) \right) \|X_0\|^2_{L^\infty(\Omega)} \\ &\leq \exp\Bigg(C \Bigg(1 + \frac{1}{T} + (1+[C_\eta + \eta \ln(1+R)]^{3/2})T \\ &\quad + C_\eta + \eta \ln(1+R) \Bigg) \Bigg) \|X_0\|^2_{L^\infty(\Omega)}. \end{aligned}$$

Choosing $T := T(R,\eta) = [C_\eta + \eta \ln(1+R)]^{-1}$, we get for R sufficiently large

$$\begin{aligned} \|\psi_g\|^2_{L^\infty(Q_T)} &\leq \exp\left(C \left(1 + C_\eta + \eta \ln(1+R) \right) \right) \|X_0\|^2_{L^\infty(\Omega)} \\ &\leq (1+R)^{\eta C} \exp\left(C \left(1 + C_\eta \right) \right) \|X_0\|^2_{L^\infty(\Omega)}. \end{aligned}$$

Choosing $\eta = \frac{1}{2C}$ yields

$$\|\psi_g\|^2_{L^\infty(Q_T)} \leq C(1+R)^{1/2} \|X_0\|^2_{L^\infty(\Omega)},$$

and clearly this will imply, for R sufficiently large, that

$$\|\psi_g\|^2_{L^\infty(Q_T)} \leq R.$$

It follows that $F(K_R) \subset K_R$. Parabolic regularity implies that $F(K_R)$ is relatively compact in $L^2(Q_T)$ and exactly as in [5], and F is semicontinuous using again [18, Theorem 10.4]. Applying the Kakutani fixed point theorem (see, for instance, [4]) in the space $L^2(Q_T)$, we deduce that there is at least one $z \in L^\infty(Q_T)$ such that $z \in F(z)$. Therefore, there is at least one pair (ψ_g, v_g) satisfying the first claim of Theorem 2.1. Actually, we have proved this first claim for any $T = T(R,\eta)$, but, clearly, this implies the same result for any $T > T(R,\eta)$: in this case, we choose a control defined on $(0, T(R,\eta))$ which gives a solution satisfying (7) at $T = T(R,\eta)$ and extend it by 0 to the whole interval $(0,T)$.

The second claim of Theorem 2.1 is obtained starting from (9) and choosing X_0 so that $C_T \|X_0\|^2_{L^\infty(\Omega)} \le R$. This ends the proof of Theorem 2.1. ∎

3. Observability estimate. The main result of this section is the proof of the observability estimate (32) in Lemma 3.4 which allows us to build controls satisfying estimate (8) of Lemma 2.3.

Following [15], let us introduce some notation. Let $\omega' \Subset \omega$ be a subdomain of ω and let β be a $C^2(\overline{\Omega})$ function such that

$$\min\left\{|\nabla\beta(x)|\,,\ x \in \overline{\Omega\backslash\omega'}\right\} > 0 \quad \text{and} \quad \frac{\partial\beta}{\partial n} \le 0 \ \text{ on } \partial\Omega, \tag{14}$$

where n denotes the outward unit normal to $\partial\Omega$. Moreover, we can always assume that β satisfies

$$\min\left\{\beta(x),\ x \in \overline{\Omega}\right\} \ge \max\left(\frac{3}{4}\|\beta\|_{L^\infty(\Omega)}, \ln(3)\right). \tag{15}$$

Finally, we introduce the following functions with parameters $\lambda > 0$ and $\tau > 0$:

$$\rho(t,x) := \frac{e^{\lambda\beta(x)}}{t(T-t)}, \qquad (t,x) \in Q_T, \tag{16}$$

$$\alpha(t,x) := \tau\frac{e^{\frac{4}{3}\lambda\|\beta\|_{L^\infty(\Omega)}} - e^{\lambda\beta(x)}}{t(T-t)}, \qquad (t,x) \in Q_T. \tag{17}$$

Then the following result holds (Carleman estimate).

THEOREM 3.1 (see [15, Theorem 7.1, p. 288]). *There exist $\lambda_0 > 0$, $\tau_0 > 0$ and a positive constant C such that $\forall\lambda \ge \lambda_0$, $\forall\tau \ge \tau_0$, and $\forall s \ge -3$ the inequality*

$$\int_{Q_T}\left(\frac{1}{\lambda}|z_t|^2 + \left|D_x^2 z\right|^2 + \lambda\tau^2\rho^2|\nabla z|^2 + \lambda^4\tau^4\rho^4 z^2\right)\rho^{2s-1}e^{-2\alpha}dxdt$$

$$\le C\left(\tau\int_{Q_T}|z_t \pm \Delta z|^2\rho^{2s}e^{-2\alpha}dxdt + \lambda^4\tau^4\int_0^T\int_{\omega'} z^2\rho^{2s+3}e^{-2\alpha}dxdt\right) \tag{18}$$

holds for any function z satisfying an homogeneous Dirichlet condition and such that the right-hand side of (18) *is finite. Moreover, the constants C and λ_0 depend only on Ω and ω'. The constant τ_0 is of the form*

$$\tau_0 = c_0(\Omega,\omega')(T + T^2).$$

The explicit dependence in time of the constants is not given in [15]. We refer to [13], where the above formula for τ_0 is obtained.

In what follows, the symbol C will stand for various constants independent of T and a.

The adjoint problem associated with (6) is

$$\begin{cases} \varphi_t = \Delta\varphi + a\ \varphi - \Delta w & \text{in } (0,T)\times\Omega = Q_T, \\ w_t = \Delta w + \varphi & \text{in } Q_T, \\ \varphi = w = 0 & \text{on } (0,T)\times\partial\Omega = \Sigma_T, \\ \varphi(0) = \varphi_0,\ w(0) = w_0, & \text{in } \Omega. \end{cases} \tag{19}$$

Let us introduce the following notation: For given λ and τ as in Theorem 3.1, we set $\delta = \tau\rho$ and consider the functional:

$$I(s,z) = \int_{Q_T} \left(\frac{1}{\lambda} |z_t|^2 + |\Delta z|^2 + \lambda\delta^2 |\nabla z|^2 + \lambda^4\delta^4 z^2 \right) \delta^{2s-1} e^{-2\alpha} dxdt. \tag{20}$$

LEMMA 3.2. *Let $\lambda_0 > 1$, $\tau_1 = C(T + (1 + \|a\|^{2/3}_{L^\infty(Q_T)})T^2)$, C being the constant given in Theorem* 3.1. *Then $\forall\lambda \geq \lambda_0$, $\forall\tau \geq \tau_1$, and $\forall s \geq -3/2$, the solution (φ, w) of* (19) *satisfies the estimate:*

$$I\left(s - \frac{3}{2}, \varphi\right) + I(s,w) \leq C(1+\lambda^4) \int_0^T \int_{\omega'} \left(\varphi^2\delta^{2s} + w^2\delta^{2s+3}\right) e^{-2\alpha} dxdt. \tag{21}$$

As a consequence, we get

$$I(-3,\varphi) + I\left(-\frac{3}{2}, w\right) \leq C(1+\lambda^4) \int_0^T \int_{\omega'} \left(\varphi^2 + w^2\right) e^{-2\alpha} dxdt. \tag{22}$$

Remark 3.3. The estimate (22) is sufficient to prove controllability by two forces as in Barbu [6].

Proof. Applying Theorem 3.1 to the first equation of problem (6) and multipyling (18) by τ^{2s-1}, we obtain

$$\begin{aligned} I(s,\varphi) \leq C \Bigg(& \int_{Q_T} \left(|\Delta w|^2 + |a\varphi|^2 \right) \delta^{2s} e^{-2\alpha}\, dx\, dt \\ & + \lambda^4 \int_0^T \int_{\omega'} \varphi^2 \delta^{2s+3} e^{-2\alpha}\, dx\, dt \Bigg). \end{aligned} \tag{23}$$

Theorem 3.1 applied to the second equation of (6) yields

$$\begin{aligned} \int_{Q_T} |\Delta w|^2 \delta^{2s} e^{-2\alpha} dxdt \leq\ & I(s+\frac{1}{2}, w) \\ & \leq C \Bigg(\int_{Q_T} \varphi^2 \delta^{2s+1} e^{-2\alpha} dxdt \\ & \quad + \lambda^4 \int_0^T \int_{\omega'} w^2 \delta^{2s+4} e^{-2\alpha} dxdt \Bigg). \end{aligned}$$

Inserting the latter estimate in (23), we get

$$\begin{aligned} I(s,\varphi) \leq C \Bigg(& \int_{Q_T} \left(\varphi^2\delta^{2s+1} + |a\varphi|^2 \delta^{2s} \right) e^{-2\alpha} dxdt \\ & + \lambda^4 \int_0^T \int_{\omega'} \left(\varphi^2\delta^{2s+3} + w^2\delta^{2s+4}\right) e^{-2\alpha} dxdt \Bigg). \end{aligned} \tag{24}$$

Observe that

$$\begin{aligned}&\int_{Q_T}\left(\delta^{2s+1}\varphi^2+\delta^{2s}\left|a\varphi\right|^2\right)e^{-2\alpha}dxdt\\&\leq\int_{Q_T}\left(\delta^{2s+1}\varphi^2+\left\|a\right\|^2_{L^\infty(Q_T)}\,\delta^{2s}\left|\varphi\right|^2\right)e^{-2\alpha}dxdt.\end{aligned}\tag{25}$$

Thus, we get

$$\begin{aligned}I(s,\varphi)\leq C\Bigg(&\int_{Q_T}\left(\delta^{2s+1}\varphi^2+\left\|a\right\|^2_{L^\infty(Q_T)}\,\delta^{2s}\left|\varphi\right|^2\right)e^{-2\alpha}dxdt\\&+\lambda^4\int_0^T\int_{\omega'}\left(\varphi^2\delta^{2s+3}+w^2\delta^{2s+4}\right)e^{-2\alpha}dxdt\Bigg).\end{aligned}\tag{26}$$

In order to get rid of the first integral at the right-hand side of inequality (26), we transfer it to the left to obtain

$$\begin{aligned}&\frac{1}{2}\int_{Q_T}\lambda^4\tau^4\rho^{2s+3}\left|\varphi\right|^2e^{-2\alpha}dxdt\\&-C\tau\int_{Q_T}\left(\tau\rho+\left\|a\right\|^2_{L^\infty(Q_T)}\right)\rho^{2s}\left|\varphi\right|^2e^{-2\alpha}dxdt\\&=\int_{Q_T}\left(\frac{1}{2}\lambda^4\tau^3\rho^3-C\tau\rho-C\left\|a\right\|^2_{L^\infty(Q_T)}\right)\tau\rho^{2s}\left|\varphi\right|^2e^{-2\alpha}dxdt\\&\geq\int_{Q_T}\left(\frac{1}{4}\lambda^4\tau^3\rho^3-C\left\|a\right\|^2_{L^\infty(Q_T)}\right)\tau\rho^{2s}\left|\varphi\right|^2e^{-2\alpha}dxdt,\end{aligned}\tag{27}$$

provided (see (16)) $\tau\geq C\dfrac{T^2}{\lambda^2}$.

Now notice that

$$\frac{1}{4}\lambda^4\tau^3\rho^3-C\left\|a\right\|^2_{L^\infty(Q_T)}\geq 2^4\frac{\lambda^4}{T^6}\tau^3-C\left\|a\right\|^2_{L^\infty(Q_T)}\geq 0,$$

provided

$$\begin{aligned}\tau&\geq C\frac{T^2}{(2\lambda)^{4/3}}\left\|a\right\|^{2/3}_{L^\infty(Q_T)}\\&\geq CT^2\left\|a\right\|^{2/3}_{L^\infty(Q_T)}.\end{aligned}\tag{28}$$

Taking into account these computations, (26) becomes

$$I(s,\varphi)\leq C\lambda^4\int_0^T\int_{\omega'}\left(\varphi^2\delta^{2s+3}+w^2\delta^{2s+4}\right)e^{-2\alpha}dxdt,\tag{29}$$

provided $\tau\geq\tau_1$.

Similarly, for the second equation of problem (6), we obtain for $s\geq -3/2$ and using (29)

$$
\begin{aligned}
I(s,w) &\leq C\left(\int_{Q_T}\varphi^2\delta^{2s}e^{-2\alpha}dxdt+\lambda^4\int_0^T\int_{\omega'}w^2\delta^{2s+3}e^{-2\alpha}dxdt\right)\\
&\leq C\left(\frac{1}{\lambda^4}I\left(s-\frac{3}{2},\varphi\right)+\lambda^4\int_0^T\int_{\omega'}w^2\delta^{2s+3}e^{-2\alpha}dxdt\right)\\
&\leq C\int_0^T\int_{\omega'}\left(\varphi^2\delta^{2s}+w^2\delta^{2s+1}+\lambda^4w^2\delta^{2s+3}\right)e^{-2\alpha}dxdt\\
&\leq C\int_0^T\int_{\omega'}\left(\varphi^2\delta^{2s}+\lambda^4w^2\delta^{2s+3}\right)e^{-2\alpha}dxdt, \qquad (30)
\end{aligned}
$$

the last inequality being obtained by noting that $\delta \geq 1$ for τ sufficiently large.

Adding up inequalities (29) and (30), we get (21).

The consequence (22) follows from $\delta \geq 1$ for τ sufficiently large and by taking $s=-\frac{3}{2}$ □

We are now ready to state our crucial lemma.

LEMMA 3.4. *Under the assumptions of Lemma* 3.2, $\forall r \in [0,2)$ *there exists a constant* $C=C_r$ *such that*

$$
\int_0^T\int_{\omega'}\left(\varphi^2+w^2\right)\ e^{-2\alpha}\,dx\,dt \leq C\left(\ 1+\frac{\tau^8}{T^{14}}+\|a\|^6_{L^\infty(Q_T)}\right)\int_0^T\int_{\omega}\ e^{-r\alpha}w^2\,dx\,dt. \qquad (31)
$$

As an immediate consequence, it follows that

$$
I(-3,\varphi)+I\left(-\frac{3}{2},w\right) \leq C\left(\ 1+\frac{\tau^8}{T^{14}}+\|a\|^6_{L^\infty(Q_T)}\right)\int_0^T\int_{\omega}\ e^{-r\alpha}w^2\,dx\,dt. \qquad (32)
$$

Remark 3.5. We will see later that it is important to be able to choose $r>1$. Notice that r cannot be equal to 2: this is the "cost" of the control by a single force.

Proof. The main idea is to estimate $\int_0^T\int_{\omega'}\varphi^2\ e^{-2\alpha}\,dx\,dt$ by $\int_0^T\int_{\omega}\ e^{-r\alpha}w^2\,dx\,dt$ for some $r\in[0,2)$ using the second equation of (19). To do this, we first localize the system in space, multiply the second equation by $-\beta_0 e^{-2\alpha}\eta\varphi$, and manage the "bad" terms appearing (see $\Lambda(t)$ in (34) and our paper [2] for the construction of this function in an abstract setting).

Let $\xi\in C^\infty(R^n)$ be a truncation function satisfying

$$
\begin{cases}
\xi(x)=1 & \forall x\in\omega',\\
0<\xi(x)\leq 1 & \forall x\in\omega'',\\
\xi(x)=0 & \forall x\in\mathbb{R}^n\backslash\omega'',
\end{cases} \qquad (33)
$$

where $\omega' \Subset \omega'' \Subset \omega \Subset \Omega$. We introduce the function $\eta:=\xi^6$. For real numbers $\beta_0,\beta_1,p,q>0$, which will be chosen below, set

$$
\Lambda(t)=\int_\Omega\left(e^{-p\alpha}\eta^{4/3}\varphi^2-\beta_0e^{-2\alpha}\eta\varphi w+\beta_1e^{-q\alpha}\eta^{2/3}w^2\right)dx. \qquad (34)
$$

Differentiating Λ with respect to t and replacing φ_t and w_t by their expressions given

by (19) we obtain

$$
\begin{aligned}
\Lambda' &= \int_\Omega \left(-pe^{-p\alpha}\eta^{4/3}\varphi^2 + 2\beta_0 e^{-2\alpha}\eta\varphi w - \beta_1 q\, e^{-q\alpha}\eta^{2/3}w^2\right)\alpha_t\, dx \\
&\quad + \int_\Omega 2e^{-p\alpha}\eta^{4/3}\varphi(\Delta\varphi + a\varphi - \Delta w)\, dx \\
&\quad -\beta_0 \int_\Omega e^{-2\alpha}\eta\left[\varphi(\Delta w + \varphi) + w(\Delta\varphi + a\varphi - \Delta w)\right] dx \\
&\quad + \int_\Omega 2\beta_1\, e^{-q\alpha}\eta^{2/3} w(\Delta w + \varphi) dx.
\end{aligned} \tag{35}
$$

We have $\Lambda(0) = \Lambda(T) = 0$, and therefore the integration of (35) over $(0,T)$ yields

$$
\begin{aligned}
\beta_0 \int_{Q_T} e^{-2\alpha}\eta\varphi^2 dx &= \int_{Q_T} \Big\{ \left(-p\alpha_t + 2a\right) e^{-p\alpha}\eta^{4/3}\varphi^2 - \beta_1\, q\alpha_t\, e^{-q\alpha}\eta^{2/3}w^2 \\
&\qquad + \left(\beta_0\ (2\alpha_t - a)\, e^{-2\alpha}\eta + 2\beta_1\, e^{-q\alpha}\eta^{2/3}\right)\varphi\, w\Big\}\, dx\, dt \\
&\quad +2\int_{Q_T} e^{-p\alpha}\eta^{4/3}\,\varphi\,\Delta\varphi\, dx\, dt \\
&\quad -\int_{Q_T}\left(2e^{-p\alpha}\eta^{4/3} + \beta_0 e^{-2\alpha}\eta\right)\varphi\,\Delta w\, dx\, dt \\
&\quad -\int_{Q_T} \beta_0\, e^{-2\alpha}\eta\, w\,\Delta\varphi\, dx dt \\
&\quad +\int_{Q_T}\left(\beta_0\, e^{-2\alpha}\eta + 2\beta_1 e^{-q\alpha}\eta^{2/3}\right) w\,\Delta w\, dx\, dt \\
&= J_1 + J_2 + J_3 + J_4 + J_5.
\end{aligned} \tag{36}
$$

Now we estimate each of the five terms $J_1, \dots, J_5$.

For J_1, one has

$$
\begin{aligned}
J_1 &= \int_{Q_T} \left\{\left(-p\alpha_t + 2a\right) e^{-p\alpha}\eta^{4/3}\varphi^2 - \beta_1\, q\alpha_t\, e^{-q\alpha}\eta^{2/3}w^2\right. \\
&\qquad \left. + \left(\beta_0\ (2\alpha_t - a)\, e^{-2\alpha}\eta + 2\beta_1\, e^{-q\alpha}\eta^{2/3}\right)\varphi\, w\right\}\, dx\, dt.
\end{aligned}
$$

In order to estimate $\beta_0\ (2\alpha_t - a)\, e^{-2\alpha}\eta\,\varphi\, w$ in terms of $e^{-2\alpha}\eta\varphi^2$, we need to bound $(2\alpha_t - a)^2\, e^{-2\alpha}\eta^{2/3}w^2$. So, since $\alpha_t \notin L^\infty(Q_T)$, we introduce $r \in [0,2)$ and write $e^{-2\alpha} = e^{-(2-r)\alpha}e^{-r\alpha}$. Assuming that

$$
p > 2,\ q > 1 + \frac{r}{2}, \tag{37}
$$

we get, with $\beta_1 \geq 1$,

$$
\begin{aligned}
J_1 &\leq \left(\frac{1}{2} + \left\|\left(-p\alpha_t + 2a\right) e^{-(p-2)\alpha}\eta^{1/3}\right\|_{L^\infty(Q_T)}\right)\int_{Q_T} e^{-2\alpha}\eta\varphi^2\, dx\, dt \\
&\quad + \left\{\frac{1}{2}\left\|\beta_0\ (2\alpha_t - a)\, e^{-(1-\frac{r}{2})\alpha}\eta^{1/3} + 2\beta_1\, e^{-(q-1-\frac{r}{2})\alpha}\right\|^2_{L^\infty(Q_T)}\right. \\
&\quad \left. + \beta_1 q\left\|\alpha_t\, e^{-(q-r)\alpha}\eta^{1/3}\right\|_{L^\infty(Q_T)}\right\}\int_{Q_T}\eta^{1/3}e^{-r\alpha}w^2 dx\, dt. \\
&\leq C\left[\left(1 + \left\|\alpha_t e^{-(p-2)\alpha}\right\|_{L^\infty(Q_T)} + \|a\|^2_{L^\infty(Q_T)}\right)\int_{Q_T} e^{-2\alpha}\eta\varphi^2\, dx\, dt\right.
\end{aligned}
$$

$$+\left(\left(\|a\|^2_{L^\infty(Q_T)}+\left\|\alpha_t\, e^{-(1-\frac{r}{2})\alpha}\right\|_{L^\infty(Q_T)}\right)\beta_0^2 + \left(1+\left\|\alpha_t\, e^{-(q-r)\alpha}\right\|_{L^\infty(Q_T)}\right)\beta_1^2\right)\int_{Q_T}\eta^{1/3}e^{-r\alpha}w^2dxdt\Bigg],$$

where $C=C(p,q,\|\eta\|_{L^\infty(\Omega)})$. Now, assuming that $\tau\geq\tau_1$ and $\beta_1\geq 1$, one obtains

$$J_1\leq C\left[\left(1+\frac{\tau}{T^3}+\|a\|^2_{L^\infty(Q_T)}\right)\int_{Q_T}e^{-2\alpha}\eta\varphi^2dxdt + \left(1+\|a\|^2_{L^\infty(Q_T)}+\frac{\tau^2}{T^6}\right)\left(\beta_1^2+\beta_0^2\right)\int_{Q_T}\eta^{1/3}e^{-r\alpha}w^2dxdt\right]. \tag{38}$$

Concerning J_2,

$$\begin{aligned}J_2&=2\int_{Q_T}e^{-p\alpha}\eta^{4/3}\,\varphi\,\Delta\varphi\,dx\,dt\\&=-2\int_{Q_T}e^{-p\alpha}\eta^{4/3}\,|\nabla\varphi|^2\,dx\,dt-2\int_{Q_T}\varphi\,\nabla\varphi.\nabla\left(e^{-p\alpha}\eta^{4/3}\right)\,dx\,dt\\&=-2\int_{Q_T}e^{-p\alpha}\eta^{4/3}\,|\nabla\varphi|^2\,dx\,dt+\int_{Q_T}\varphi^2.\Delta\left(e^{-p\alpha}\eta^{4/3}\right)\,dx\,dt\\&=-2\int_{Q_T}e^{-p\alpha}\eta^{4/3}\,|\nabla\varphi|^2\,dx\,dt\\&\quad+\int_{Q_T}\left(e^{-2\alpha}\eta\varphi^2\right).e^{2\alpha}\eta^{-1}\Delta\left(e^{-p\alpha}\eta^{4/3}\right)\,dx\,dt.\end{aligned}$$

We have

$$\Delta\left(e^{-p\alpha}\eta^l\right)=e^{-p\alpha}\left[\left(p^2\,|\nabla\alpha|^2-p\Delta\alpha\right)\eta^l+l\left(\Delta\eta-2p\nabla\alpha.\nabla\eta\right)\eta^{l-1}+p(p-1)\,|\nabla\eta|^2\,\eta^{l-2}\right].$$

So

$$\begin{aligned}e^{2\alpha}\eta^{-1}\Delta\left(e^{-p\alpha}\eta^{4/3}\right)=e^{-(p-2)\alpha}\Big[&\left(p^2\,|\nabla\alpha|^2-p\Delta\alpha\right)\eta^{1/3}\\&+\frac{4}{3}\left(\Delta\eta-2p\nabla\alpha.\nabla\eta\right)\eta^{-2/3}\\&+p(p-1)\,|\nabla\eta|^2\,\eta^{-5/3}\Big].\end{aligned}$$

Note that

$$\frac{\nabla\eta}{\eta^{5/6}}=6\nabla\xi\in L^\infty(\Omega),\quad \frac{\nabla\eta}{\eta^{2/3}}=6\xi\nabla\xi\in L^\infty(\Omega),$$
$$\frac{\Delta\eta}{\eta^{2/3}}=30\,|\nabla\xi|^2+6\xi\Delta\xi\in L^\infty(\Omega).$$

It follows from this last computation and $p>2$ that

$$\left\|e^{2\alpha}\eta^{-1}\Delta\left(e^{-p\alpha}\eta^{4/3}\right)\right\|_{L^\infty(Q_T)}\leq C\left(1+\frac{\tau^2}{T^4}\right),$$

where $C = C(p, \|\eta\|_{L^\infty(\Omega)})$. Coming back to J_2, we get

$$J_2 \leq -2\int_{Q_T} e^{-p\alpha}\eta^{4/3}\ |\nabla\varphi|^2\ dx\,dt + C\left(1+\frac{\tau^2}{T^4}\right)\int_{Q_T} e^{-2\alpha}\eta\varphi^2\ dx\,dt. \tag{39}$$

We now estimate $J_3 + J_4$. We have

$$\begin{aligned}
J_3 + J_4 &= -\int_{Q_T} \beta_0 e^{-2\alpha}\eta\left(\varphi\,\Delta w + w\,\Delta\varphi\right)\ dx\,dt - \int_{Q_T} 2e^{-p\alpha}\eta^{4/3}\varphi\,\Delta w dxdt \\
&= -\int_{Q_T} \beta_0 e^{-2\alpha}\eta\left(\Delta\left(\varphi\, w\right) - 2\nabla\varphi.\nabla w\right)\ dx\,dt \\
&\quad + \int_{Q_T} 2e^{-p\alpha}\eta^{4/3}\nabla\varphi\,.\nabla w dxdt \\
&\quad + \int_{Q_T} 2\varphi\nabla(e^{-p\alpha}\eta^{4/3}).\nabla w\,dxdt \\
&= -\beta_0\int_{Q_T} \Delta\left(e^{-2\alpha}\eta\right)\varphi\, w\,dx\,dt \\
&\quad +2\int_{Q_T}\left(\beta_0 e^{-2\alpha}\eta + e^{-p\alpha}\eta^{4/3}\right)\nabla\varphi\,.\nabla w dxdt \\
&\quad + \int_{Q_T} 2\varphi\nabla(e^{-p\alpha}\eta^{4/3}).\nabla w\,dxdt.
\end{aligned}$$

Proceeding as previously, thanks to the assumption $r < 2$, we first get

$$\begin{aligned}
\left|\beta_0\int_{Q_T}\Delta\left(e^{-2\alpha}\eta\right)\varphi\, w\,dx\,dt\right| \leq C\Bigg(&\int_{Q_T} e^{-2\alpha}\eta\varphi^2\ dx\,dt \\
&+ \beta_0^2\left(1+\frac{\tau^4}{T^8}\right)\int_{Q_T}\eta^{1/3}e^{-r\alpha}w^2 dxdt\Bigg).
\end{aligned}$$

In the same way, with $p > 2$ and $\eta^{-1/6}\nabla\eta \in L^\infty(\Omega)$,

$$\begin{aligned}
\left|\int_{Q_T} 2\varphi\nabla(e^{-p\alpha}\eta^{4/3}).\nabla w\,dxdt\right| \leq C\Bigg(&\int_{Q_T} e^{-2\alpha}\eta\varphi^2\ dx\,dt \\
&+\left(1+\frac{\tau^2}{T^4}\right)\int_{Q_T}\eta^{2/3}e^{-2(p-1)\alpha}\left|\nabla w\right|^2 dxdt\Bigg).
\end{aligned}$$

It appears that

$$\begin{aligned}
J_3 + J_4 \leq C\Bigg(&\int_{Q_T} e^{-2\alpha}\eta\varphi^2\ dx\,dt + \beta_0^2\left(1+\frac{\tau^4}{T^8}\right)\int_{Q_T}\eta^{1/3}e^{-r\alpha}w^2 dxdt \\
&+\left(1+\frac{\tau^2}{T^4}\right)\int_{Q_T}\eta^{2/3}e^{-2(p-1)\alpha}\left|\nabla w\right|^2 dxdt\Bigg) \\
&+2\int_{Q_T}\left(\beta_0 e^{-2\alpha}\eta + e^{-p\alpha}\eta^{4/3}\right)\nabla\varphi\,.\nabla w dxdt.
\end{aligned} \tag{40}$$

Finally, we estimate J_5.

$$J_5 = \int_{Q_T}\left(\beta_0\, e^{-2\alpha}\eta + 2\beta_1 e^{-q\alpha}\eta^{2/3}\right)\, w\,\Delta w\,dx\,dt$$

$$= -\int_{Q_T} \left(\beta_0\, e^{-2\alpha}\eta + 2\beta_1 e^{-q\alpha}\eta^{2/3}\right) |\nabla w\,|^2\, dx\, dt$$
$$+\frac{1}{2}\int_{Q_T} \Delta\left(\beta_0\, e^{-2\alpha}\eta + 2\beta_1 e^{-q\alpha}\eta^{2/3}\right) w^{\,2} dx\, dt.$$

Again, in the same way, we get using the condition $q > r$ and the definition of η:

$$\frac{1}{2}\left|\int_{Q_T} \Delta\left(\beta_0\, e^{-2\alpha}\eta + 2\beta_1 e^{-q\alpha}\eta^{2/3}\right) w^{\,2} dx\, dt\right| \leq C\beta_1\left(1+\frac{\tau^2}{T^4}\right)\int_{Q_T} \eta^{1/3} e^{-r\alpha} w^2 dx dt.$$

Thus

$$J_5 \leq -\int_{Q_T} \left(\beta_0\, e^{-2\alpha}\eta + 2\beta_1 e^{-q\alpha}\eta^{2/3}\right) |\nabla w\,|^2\, dx\, dt$$
$$+C\beta_1\left(1+\frac{\tau^2}{T^4}\right)\int_{Q_T} \eta^{1/3} e^{-r\alpha} w^2 dx dt. \tag{41}$$

Fix $\beta_0 = 2C(1+\frac{\tau^2}{T^4} + \|a\|^2_{L^\infty(Q_T)})$. Then from (38), (39), (40), and (41), we get with $\tau \geq \tau_0$, conditions (37) and $\beta_1 \geq 1$:

$$\int_{Q_T} e^{-2\alpha}\eta\varphi^2 dx \leq 2C\left(1+\|a\|^2_{L^\infty(Q_T)} + \frac{\tau^4}{T^8}\right)\frac{\left(\beta_1^2+\beta_0^2\right)}{\beta_0}\int_{Q_T} \eta^{1/3} e^{-r\alpha} w^2 dx dt$$
$$-\frac{4}{\beta_0}\int_{Q_T} e^{-p\alpha}\eta^{4/3}\, |\nabla\varphi|^2\, dx\, dt$$
$$-\frac{2}{\beta_0}\int_{Q_T} \Big(\beta_0\, e^{-2\alpha}\eta + 2\beta_1 e^{-q\alpha}\eta^{2/3}$$
$$-C\left(1+\frac{\tau^2}{T^4}\right)\eta^{2/3} e^{-2(p-1)\alpha}\Big) |\nabla w|^2\, dx dt$$
$$+\frac{4}{\beta_0}\int_{Q_T} \left(\beta_0 e^{-2\alpha}\eta + e^{-p\alpha}\eta^{4/3}\right)\nabla\varphi\,.\nabla w\, dx dt.$$

Consider the three last terms in the right-hand side of the previous inequality. Assume moreover that

$$\frac{1}{2}q + 1 < p < 4 - q. \tag{42}$$

For β_1 sufficiently large, we have

$$\left(\beta_0 e^{-2\alpha}\eta + e^{-p\alpha}\eta^{4/3}\right)^2 \leq 2e^{-p\alpha}\eta^{4/3}\Big(\beta_0\, e^{-2\alpha}\eta + 2\beta_1 e^{-q\alpha}\eta^{2/3}$$
$$-C\left(1+\frac{\tau^2}{T^4}\right)\eta^{2/3} e^{-2(p-1)\alpha}\Big) \text{ on } Q_T.$$

Here is the proof. Indeed, it is sufficient to prove

$$\frac{1}{4}\left(\beta_0 e^{-(2-\frac{p+q}{2})\alpha} + e^{-\frac{p-q}{2}\alpha}\eta^{1/3}\right)^2 + \frac{1}{2}C\left(1+\frac{\tau^2}{T^4}\right) e^{-(2p-q-2)\alpha} \leq \beta_1 \text{ on } Q_T.$$

Taking into account (42), this last estimate is true if, for instance,

$$\beta_1 \geq C\left(\left(\beta_0\ +1\right)^2 + 1 + \frac{\tau^2}{T^4}\right)$$

with a large constant $C = C(\Omega, \omega, \omega', \eta)$. To summarize, we have

$$\int_{Q_T} e^{-2\alpha}\eta\varphi^2 dxdt \le C\frac{\left(1+\|a\|^2_{L^\infty(Q_T)}+\frac{\tau^4}{T^8}\right)\left(\beta_1^2+\beta_0^2\right)}{\beta_0}\int_{Q_T}\eta^{1/3}e^{-r\alpha}w^2dxdt,$$

and all the computations are valid with the following conditions:

$$\begin{aligned} &r<2,\ p>2,\ q>1+\frac{r}{2},\ \frac{1}{2}q+1<p<4-q,\\ &\beta_0 = 2C\left(1+\frac{\tau^2}{T^4}+\|a\|^2_{L^\infty(Q_T)}\right),\\ &\beta_1 \ge C\left(\left(\beta_0\ +1\right)^2+1+\frac{\tau^2}{T^4}\right),\\ &\tau\ge\tau_1. \end{aligned} \tag{43}$$

It is clear that there is a nonempty set of (p, q, r) verifying (43): for instance, $(2+\frac{1}{16}, 2-\frac{1}{8}, \frac{3}{2})$ satisfies this condition. With maybe a modified constant C, we get the following final estimate:

$$\int_{Q_{T,\omega'}} e^{-2\alpha}\varphi^2 dxdt \le C\left(\ 1+\frac{\tau^8}{T^{14}}+\|a\|^6_{L^\infty(Q_T)}\right)\int_{Q_{T,\omega}} e^{-r\alpha}w^2dxdt,$$

where $Q_{T,\omega} = (0,T)\times\omega$. This final estimate ends the proof of the lemma. ∎

4. Proof of Lemma 2.3. For $r\in(1,2)$ and $\varepsilon>0$, we define

$$J_\varepsilon(g) = \frac{1}{2}\int_{Q_T} e^{r\alpha}g^2dxdt+\frac{1}{2\varepsilon}\left\|(\psi,v)(T)\right\|^2_{L^2(\Omega)},$$

where $g\in L^2(Q_T)$ and (ψ, v) is the associated solution of (6). Introduce also the dual functional (see [11]):

$$J^*_\varepsilon(Y_0) = \frac{1}{2}\int_0^T\int_\omega e^{-r\alpha}w^2dxdt+\frac{\varepsilon}{2}\left\|Y_0\right\|^2_{L^2(\Omega)}+\int_\Omega Y(0).X_0dx,$$

where $X_0 = (\psi_0, v_0)\in H^1_0(\Omega)\times H^1_0(\Omega)$ is the fixed initial data of (6) and $Y = (\varphi, w)$ is the solution of the backward linear system with initial data $Y_0 = (\varphi_0, w_0)\in L^2(\Omega)\times L^2(\Omega)$

$$\begin{cases} -\varphi_t = \Delta\varphi + a\ \varphi - \Delta w & \text{in } (0,T)\times\Omega = Q_T,\\ -w_t = \Delta w+\varphi & \text{in } Q_T,\\ \varphi = w = 0 & \text{on } (0,T)\times\partial\Omega = \Sigma_T,\\ \varphi(T)=\varphi_0,\ w(T)=w_0, & \text{in } \Omega. \end{cases} \tag{44}$$

It is easy to prove that the minimization problems $\min_g J_\varepsilon(g)$ and $\min_{Y_0} J^*_\varepsilon(Y_0)$ have both exactly one solution $(g_\varepsilon, Y_{0\varepsilon})$ and, moreover, by the maximum principle (or see, for instance, [11])

$$g_\varepsilon = \chi_\omega e^{-r\alpha}w_\varepsilon \text{ on } Q_T;\ \ Y_{0\varepsilon} = -\frac{1}{\varepsilon}(\psi_\varepsilon, v_\varepsilon)(T) \text{ on } \Omega, \tag{45}$$

where $(\psi_\varepsilon, v_\varepsilon)$ (resp., $(\varphi_\varepsilon, w_\varepsilon)$) is the solution of (6) (resp., (44)) associated with g_ε (resp., $Y_{0\varepsilon}$). Since $J^*_\varepsilon(Y_{0\varepsilon}) \leq 0$, we get

(46)
$$\frac{1}{2}\int_0^T \int_\omega e^{-r\alpha} w_\varepsilon^2 dxdt + \frac{1}{2\varepsilon} \|(\psi_\varepsilon, v_\varepsilon)(T)\|^2_{L^2(\Omega)} \leq \|(\varphi_\varepsilon, w_\varepsilon)(0)\|_{L^2(\Omega)} \cdot \|X_0\|_{L^2(\Omega)}.$$

To obtain a uniform estimate, we will need the following result.

LEMMA 4.1. *For $r \in (0,2)$, any solution pair of* (44) *satisfies the estimate*

$$\|(\varphi, w)(0)\|^2_{L^2(\Omega)} \leq C_T \int_0^T \int_\omega e^{-r\alpha} w^2 dxdt, \tag{47}$$

with

$$C_T = \exp\left(C\left(1 + \frac{1}{T} + (1 + \|a\|_{L^\infty(Q_T)})T + \|a\|^{2/3}_{L^\infty(Q_T)}\right)\right).$$

We will prove this lemma at the end of this section. From (47) and (47), we get $\forall \varepsilon > 0$

$$\frac{1}{2}\int_0^T \int_\omega e^{-r\alpha} w_\varepsilon^2 dxdt + \frac{1}{2\varepsilon} \|(\psi_\varepsilon, v_\varepsilon)(T)\|^2_{L^2(\Omega)} \leq C_T \|X_0\|^2_{L^2(\Omega)}. \tag{48}$$

We obtain from this last estimate a control in $L^2(Q_T)$; however, as it appeared clearly in the proof of Theorem 2.1, it is not sufficient. So we will prove that our control is in $L^{q_N}(Q_T)$.

We introduce $\zeta_\varepsilon = e^{-r\alpha} w_\varepsilon$. It satisfies by (44)

$$\begin{cases} (\zeta_\varepsilon)_t + \Delta\zeta_\varepsilon = f_\varepsilon & \text{in } (0,T)\times\Omega = Q_T, \\ \zeta_\varepsilon = 0 & \text{on } (0,T)\times\partial\Omega = \Sigma_T, \\ \zeta_\varepsilon(T) = 0 & \text{in } \Omega, \end{cases}$$

with

$$f_\varepsilon = -2r\nabla\alpha.\left(e^{-r\alpha}\nabla w_\varepsilon\right) + \left(\Delta\left(e^{-r\alpha}\right) - \left(e^{-r\alpha}\right)_t\right) w_\varepsilon - e^{-r\alpha}\varphi_\varepsilon.$$

By parabolic regularity, we have

$$\|\zeta_\varepsilon\|_{W_2^{2,1}(Q_T)} \leq C \|f_\varepsilon\|_{L^2(Q_T)}. \tag{49}$$

On the other hand, setting

$$I_1 = \int_{Q_T} e^{-2r\alpha}\varphi_\varepsilon^2 \, dx\, dt$$

we have, using (32) in Lemma 3.4,

$$\begin{aligned} I_1 &= \int_{Q_T} \left(\delta^3 e^{-2(r-1)\alpha}\right)\left(\delta^{-3} e^{-2\alpha}\varphi_\varepsilon^2\right) dx\, dt \\ &\leq \left\|\delta^3 e^{-2(r-1)\alpha}\right\|_{L^\infty(Q_T)} \int_{Q_T} \delta^{-3} e^{-2\alpha}\varphi_\varepsilon^2 \, dx\, dt \\ &\leq C\left\|\delta^3 e^{-2(r-1)\alpha}\right\|_{L^\infty(Q_T)} \left(1 + \frac{\tau^8}{T^{14}} + \|a\|^6_{L^\infty(Q_T)}\right) \int_0^T \int_\omega e^{-r\alpha} w_\varepsilon^2 dxdt \\ &\leq C\left(1 + \frac{\tau^8}{T^{14}} + \|a\|^6_{L^\infty(Q_T)}\right) \int_0^T \int_\omega e^{-r\alpha} w_\varepsilon^2 dxdt. \end{aligned}$$

Notice that $\left\|\delta^3 e^{-2(r-1)\alpha}\right\|_{L^\infty(Q_T)}$ is finite since we have assumed $r>1$ and the same remark holds in what follows. In the same way, setting

$$I_2 = \int_{Q_T} \left|\nabla\alpha.\left(e^{-r\,\alpha}\nabla w_\varepsilon\right)\right|^2 dx\,dt,$$

$$\begin{aligned} I_2 &\le \int_{Q_T} \left(|\nabla\alpha|^2\,\delta^2 e^{-2(r-1)\alpha}\right)\left(\delta^{-2}e^{-2\alpha}\,|\nabla w_\varepsilon|^2\right) dx\,dt \\ &\le \left\||\nabla\alpha|^2\,\delta^2 e^{-2(r-1)\alpha}\right\|_{L^\infty(Q_T)} C\left(1+\frac{\tau^8}{T^{14}}+\|a\|^6_{L^\infty(Q_T)}\right)\int_0^T\int_\omega e^{-r\alpha}w_\varepsilon^2 dxdt \\ &\le C\left(1+\frac{\tau^8}{T^{14}}+\|a\|^6_{L^\infty(Q_T)}\right)\int_0^T\int_\omega e^{-r\alpha}w_\varepsilon^2 dxdt, \end{aligned}$$

and, if

$$I_3 = \int_{Q_T} \left|\left(\Delta\left(e^{-r\,\alpha}\right)-\left(e^{-r\,\alpha}\right)_t\right)\,w_\varepsilon\right|^2 dx\,dt,$$

$$\begin{aligned} I_3 &= \int_{Q_T}\left|-r\Delta\alpha+r^2\,|\nabla\alpha|^2+r\alpha_t\right|^2 e^{-2r\alpha}w_\varepsilon^2 dx\,dt \\ &= \int_{Q_T}\left(\left|-r\Delta\alpha+r^2\,|\nabla\alpha|^2+r\alpha_t\right|^2 e^{-2(r-1)\alpha}\right)\left(e^{-2\alpha}w_\varepsilon^2\right)dx\,dt \\ &\le C\left\|\left|-r\Delta\alpha+r^2\,|\nabla\alpha|^2+r\alpha_t\right|^2 e^{-2(r-1)\alpha}\right\|_{L^\infty(Q_T)} \\ &\quad\left(1+\frac{\tau^8}{T^{14}}+\|a\|^6_{L^\infty(Q_T)}\right)\int_0^T\int_\omega e^{-r\alpha}w_\varepsilon^2 dxdt \\ &\le C\left(1+\frac{\tau^8}{T^{14}}+\|a\|^6_{L^\infty(Q_T)}\right)\int_0^T\int_\omega e^{-r\alpha}w_\varepsilon^2 dxdt. \end{aligned}$$

It follows from these three last inequalities and (49) that

$$\|\zeta_\varepsilon\|^2_{W_2^{2,1}(Q_T)} \le C\left(1+\frac{\tau^8}{T^{14}}+\|a\|^6_{L^\infty(Q_T)}\right)\int_0^T\int_\omega e^{-r\alpha}w_\varepsilon^2 dxdt.$$

Now, by the embedding $W_2^{2,1}(Q_T)\hookrightarrow L^{q_N}(Q_T)$ (see, for instance, [18, Lemma 3.2, p. 80])

$$\|\zeta_\varepsilon\|^2_{L^{q_N}(Q_T)} \le C\left(1+\frac{\tau^8}{T^{14}}+\|a\|^6_{L^\infty(Q_T)}\right)\int_0^T\int_\omega e^{-r\alpha}w_\varepsilon^2 dxdt.$$

Going back to our control, we get, using (48),

$$\begin{aligned} \|g_\varepsilon\|^2_{L^{q_N}(Q_T)} &= \|\chi_\omega\zeta_\varepsilon\|^2_{L^{q_N}(Q_T)} \\ &\le C\left(1+\frac{\tau^8}{T^{14}}+\|a\|^6_{L^\infty(Q_T)}\right)\int_0^T\int_\omega e^{-r\alpha}w_\varepsilon^2 dxdt \\ &\le C_T\ \|X_0\|^2_{L^2(\Omega)}. \end{aligned} \tag{50}$$

From (50) and [18, Theorem 10.4, p. 621], it follows, at least for a subsequence, that for $\varepsilon \to 0$

$$\begin{aligned} g_\varepsilon &\rightharpoonup g && \text{weakly in } L^{q_N}(Q_T), \\ (\psi_\varepsilon, v_\varepsilon) &\rightharpoonup (\psi, v) && \text{weakly in } L^2(0,T;H_0^1(\Omega)) \cap W^{2,1}_{q_N}(Q_T), \end{aligned}$$

and $((\psi, v), g)$ satisfy (6) with $(\psi, v)(T) = 0$ and

$$\|\chi_\omega g\|^2_{L^{q_N}(Q_T)} \leq C_T \ \|X_0\|^2_{L^2(\Omega)}. \qquad \square$$

To complete the proof, it remains to show Lemma 4.1.

Proof of Lemma 4.1. According to Lemma 3.4

$$\int_{T/4}^{3T/4} \int_\Omega \left(\delta^{-3} |\varphi|^2 + |w|^2\right) e^{-2\alpha} dx dt \leq C \left(1 + \frac{\tau^8}{T^{14}} + \|a\|^6_{L^\infty(Q_T)}\right) \int_0^T \int_\omega e^{-r\alpha} w^2 \, dx \, dt, \tag{51}$$

provided $\tau \geq \tau_1$. Let us estimate $\delta^{-3} e^{-2\alpha}$ from below on $(T/4, 3T/4) \times \Omega$:

$$\begin{aligned} \delta^{-3}(t,x) e^{-2\alpha(t,x)} &= \tau^{-3} t^3 (T-t)^3 e^{-3\lambda\beta(x)} \exp\left(-2\tau \frac{e^{\frac{4}{3}\lambda\|\beta\|_{L^\infty(\Omega)}} - e^{\lambda\beta(x)}}{t(T-t)}\right) \\ &\geq \tau^{-3} e^{-3\lambda\|\beta\|_{L^\infty(\Omega)}} t^3 (T-t)^3 \exp\left(-2\tau \frac{e^{\frac{4}{3}\lambda\|\beta\|_{L^\infty(\Omega)}} - e^{\lambda \ln 3}}{t(T-t)}\right) \\ &\geq C\tau^{-3} T^6 \exp\left(-\frac{C\tau}{T^2}\right) \\ &\geq C\tau^{-3} \exp\left(-\frac{C\tau}{T^2}\right). \end{aligned}$$

Inserting this inequality into (51), we get

$$\begin{aligned} &\int_{T/4}^{3T/4} \int_\Omega \left(|\varphi|^2 + |w|^2\right) dx \, dt \\ &\leq C\tau^3 \exp\left(\frac{C\tau}{T^2}\right) \left(1 + \frac{\tau^8}{T^{14}} + \|a\|^6_{L^\infty(Q_T)}\right) \int_0^T \int_\omega e^{-r\alpha} w^2 \, dx \, dt. \end{aligned} \tag{52}$$

If (φ, w) satisfies (44), then for $m = \|a\|_{L^\infty(Q_T)} + \frac{1}{2}$ we have

$$\begin{aligned} &\frac{d}{dt}\left(e^{-2m(T-t)}\left(|\varphi(t)|_2^2 + |w(t)|_2^2\right)\right) \\ &= 2me^{-2m(T-t)}\left(|\varphi(t)|_2^2 + |w(t)|_2^2\right) + 2e^{-2m(T-t)}\left(\int_\Omega (ww_t + \varphi\varphi_t)(t)dx\right), \end{aligned} \tag{53}$$

or, together with the Cauchy–Schwarz and Young inequalities,

$$\begin{aligned} &\frac{d}{dt}\left(e^{-2m(T-t)}\left(\|\varphi(t)\|^2_{L^2(\Omega)} + \|w(t)\|^2_{L^2(\Omega)}\right)\right) \\ &\geq 2e^{-2m(T-t)}\left\{\left(m - \frac{1}{2}\right)\|w(t)\|^2_{L^2(\Omega)} + \left(m - \|a\|_{L^\infty(Q_T)} - \frac{1}{2}\right)\|\varphi(t)\|^2_{L^2(\Omega)}\right\} \geq 0. \end{aligned} \tag{54}$$

The integration of (54) over $[T/4, 3T/4]$ yields

$$\frac{T}{2}e^{-2mT}\left(\|\varphi(0)\|^2_{L^2(\Omega)}+\|w(0)\|^2_{L^2(\Omega)}\right)\le\int_{\frac{T}{4}}^{\frac{3T}{4}}\int_\Omega e^{-2m(T-t)}\left(\varphi^2+w^2\right)dx\,dt$$
$$\le e^{-m\frac{T}{2}}\int_{\frac{T}{4}}^{\frac{3T}{4}}\int_\Omega\left(\varphi^2+w^2\right)dx\,dt.$$

Using (52) and setting $\tau=C(T+(1+\|a\|^{2/3}_{L^\infty(Q_T)})T^2)$ with C sufficiently large, we finally arrive at

$$\|\varphi(0)\|^2_{L^2(\Omega)}+\|w(0)\|^2_{L^2(\Omega)}$$
$$\le\frac{2}{T}e^{m\frac{3T}{2}}\int_{\frac{T}{4}}^{\frac{3T}{4}}\int_\Omega\left(\varphi^2+w^2\right)dxdt$$
$$\le\exp\left(C\left(1+\frac{1}{T}+(1+\|a\|_{L^\infty(Q_T)})T+\|a\|^{2/3}_{L^\infty(Q_T)}\right)\right)\int_0^T\int_\omega e^{-r\alpha}w^2dxdt,$$

which is the desired estimate. □

REFERENCES

[1] F. Ammar Khodja and A. Benabdallah, *Sufficient conditions for uniform stabilization of second order equations by dynamical controllers*, Dynam. Contin. Discrete Impuls. Systems, 7 (2000), pp. 207–222.

[2] F. Ammar Khodja, A. Benabdallah, C. Dupaix, and I. Kostin, *Null-Controllability of Some Systems of Parabolic Type by One Control Force*, Prépublications du Laboratoire de Mathématiques de Besançon 2003/08, 2003.

[3] S. Anita and V. Barbu, *Local exact controllability of a reaction-diffusion system*, Differential Integral Equations, 14 (2001), pp. 577–587.

[4] J. P. Aubin and A. Cellina, *Differential Inclusions. Set-Valued Map and Viability Theory*, Springer-Verlag, Berlin, 1984.

[5] V. Barbu, *Exact controllability of the superlinear heat equation*, Appl. Math. Optim., 42 (2000), pp. 73–89.

[6] V. Barbu, *Local controllability of the phase field system*, Nonlinear Anal., 50 (2002), pp. 363–372.

[7] H. Brezis, *Analyse Fonctionnelle. Théorie et Applications*, Masson, Paris, 1983.

[8] M. Brokate and J. Sprekels, *Hysteresis and Phase Transitions*, Appl. Math. Sci. 121, Springer-Verlag, New York, 1996.

[9] G. Caginalp, *An analysis of a phase field model for a free boundary*, Arch. Rational Mech. Anal., 92 (1986), pp. 205–245.

[10] L. De Teresa, *Insensitizing controls for a semilinear heat equation*, Comm. Partial Differential Equations, 25 (2000), pp. 39–72.

[11] I. Ekeland and R. Temam, *Analyse Convexe et Problèmes Variationnels*, Dunod, Gauthier-Villars, Paris, 1974.

[12] E. Fernández-Cara, *Null controllability of the semilinear heat equation*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 87–103.

[13] E. Fernández-Cara and E. Zuazua, *The cost of approximate controllability for heat equations: The linear case*, Adv. Differential Equations, 5 (2000), pp. 465–514.

[14] E. Fernández-Cara and E. Zuazua, *Null and approximate controllability for weakly blowing up semilinear heat equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 583–616.

[15] A. Fursikov, *Optimal Control of Distributed Systems. Theory and Applications.* Transl. Math. Monogr., 187, AMS, Providence, RI, 2000.

[16] A. Fursikov and O. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Ser. 34, Seoul National University, Korea, 1996.

[17] O. Yu. Imanuvilov, *Boundary controllability of parabolic equations*, Uspekhi Mat. Navk, 48 (1993), pp. 211–212 (in Russian); Russian Math. Survey, 48 (1993), pp. 192–194 (in English).

[18] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type,* Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.
[19] J.-L. Lions, *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués*, Vol. 1, RMA 8, Masson, Paris, 1988.
[20] G. Lebeau and L. Robbiano, *Contrôle exact de l'équation de la chaleur*, Comm. Partial Differential Equations, 20 (1995), pp. 335–356.
[21] J. Zabczyk, *Mathematical Control Theory: An Introduction*, Birkhäuser, Boston, Cambridge, MA, 1992.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1681–1702

RICCATI EQUATIONS FOR STABLE WELL-POSED LINEAR SYSTEMS: THE GENERIC CASE*

RUTH F. CURTAIN†

Abstract. Under the generic assumption that zero is in the resolvent set of the generator, we show that the optimal control problem for a stable well-posed linear system is equivalent to a control problem for its reciprocal system which has bounded generating operators. Consequently, the operator X that defines the optimal cost satisfies a Riccati equation with bounded operators. Previous results needed various regularity assumptions to obtain X as a solution to a Riccati equation resembling that in the finite-dimensional theory.

Key words. Riccati equations, well-posed linear systems, weakly regular linear systems, Popov function, spectral factorizations, optimal control

AMS subject classifications. 49N10, 93B05, 93C25

DOI. 10.1137/S0363012901399362

1. Introduction.

A key to solving a number of problems in systems and control is the existence of a solution to a Riccati equation which is closely related to a spectral factorization problem and to a linear quadratic optimal control problem. More concretely, consider the finite-dimensional linear system

$$\dot{x}(t) = Ax(t) + Bu(t),\ x(0) = x_0,\ \ y(t) = Cx(t) \tag{1.1}$$

and the cost functional

$$J(x_0, u) = \int_0^\infty \left\langle \begin{bmatrix} Q & N^* \\ N & R \end{bmatrix} \begin{bmatrix} y(t) \\ u(t) \end{bmatrix}, \begin{bmatrix} y(t) \\ u(t) \end{bmatrix} \right\rangle \mathrm{d}t, \tag{1.2}$$

where A, B, C, R, N, Q are matrices of appropriate dimensions, and $R = R^* > 0, Q = Q^*$.

The *optimal control problem* is to find the input function $u^{opt} \in \mathbf{L}_2(0, \infty; \mathbb{C}^m)$ that minimizes $J(x_0, u)$, where x, y are given by (1.1), and to compute this minimum. The corresponding *Popov function* is given by

$$\Pi(i\omega) = R + N\mathbf{G}(i\omega) + \mathbf{G}(i\omega)^* N^* + \mathbf{G}(i\omega)^* Q \mathbf{G}(i\omega), \tag{1.3}$$

where $\mathbf{G}(i\omega) = C(i\omega I - A)^{-1}B$ and $\omega \in \mathbb{R}$.

We consider the case corresponding to a coercive Popov function: $\Pi(i\omega) \geq \epsilon I$ for some $\epsilon > 0$ and all $\omega \in \mathbb{R}$. Then it is known that the control problem has a minimum given by

$$J(x_0, u^{opt}) = \langle Xx_0, x_0 \rangle, \tag{1.4}$$

where X is the minimal self-adjoint matrix solution of the associated *Riccati equation*

$$XA + A^*X + C^*QC = (XB + C^*N^*)R^{-1}(B^*X + NC). \tag{1.5}$$

*Received by the editors December 10, 2001; accepted for publication (in revised form) May 6, 2003; published electronically November 14, 2003.
http://www.siam.org/journals/sicon/42-5/39936.html
†Department of Mathematics, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands (R.F.Curtain@math.rug.nl).

Moreover, $u^{opt}(t) = -R^{-1}(B^*X + NC)x(t) \in \mathbf{L}_2(0,\infty;\mathbb{C}^m)$ (the closed-loop system need not be stable). For each matrix solution X of (1.5) we obtain a spectral factorization of Π:

$$\Pi(i\omega) = \mathbf{\Xi}(i\omega)^*\mathbf{\Xi}(i\omega), \tag{1.6}$$

$$\mathbf{\Xi}(s) = W[I + R^{-1}(B^*X + NC)(sI - A)^{-1}B], \tag{1.7}$$

where $W^*W = R$. If we choose W to be invertible, then $\mathbf{\Xi}$ has the inverse

$$[I - R^{-1}(B^*X + NC)(sI - A + BR^{-1}(B^*X + NC))^{-1}B]W^{-1}.$$

Both $\mathbf{\Xi}$ and its inverse are in $\mathbf{H}_\infty$.

In this paper, we investigate some extensions of the above theory to the class of infinite-dimensional systems known as *stable well-posed linear systems*. This is a large class of systems, including many systems described by partial differential equations with boundary control and point observations, as well as by delay equations with delayed observations and control action (see Weiss [18], [17], [16], [19], [20]).

Staffans [11] and Weiss and Weiss [22] solved the optimal control problem (1.1)–(1.2) for the subclass of stable weakly regular linear systems using a spectral factorization approach under the coercivity assumption $\Pi(i\omega) \geq \varepsilon I$ for some $\varepsilon > 0$ and almost all $\omega \in \mathbb{R}$. In [14], [12] Staffans solved the problem for the unstable well-posed class, and a complete theory of more general optimal control problems and Riccati equations can be found in Mikkola [9]. The solution for the stable case is obtained in terms of a spectral factorization like (1.6). In its full generality the theory becomes very technical. Although it is possible to give nice conditions for the existence of a solution to the optimal control problem, and the minimum cost has the form (1.4), it is not always possible to obtain a Riccati equation like (1.5). In Weiss and Weiss [22], as well as the weak regularity assumption, they need to assume that the spectral factor $\mathbf{\Xi}$ in (1.6) is regular. In general, this is hard to check, but the following sufficient conditions are known:

1. B is less than maximally unbounded ($\alpha(B) < \frac{1}{2}$ in (2.18)) in Weiss and Curtain [21] .
2. B is infinite-time admissible, $\mathcal{Y}$ is finite-dimensional, and C is bounded in Weiss and Weiss [22].
3. The input-output map is a convolution with an operator-valued L_1-function plus a constant feedthrough and delays (see Staffans [13] and Mikkola [9]).

Even if the spectral factor is regular, the general Riccati equation may have a different factor to R in (1.5) which is rather disconcerting and impossible to compute in general. One sufficient condition for R to be in (1.5) is condition 2 above. While examples are given in [22] and Weiss and Zwart [23] which show that these anomalies can actually occur, it is curious that they have never appeared in the alternative partial differential equation approach to Riccati equations (see Lasiecka and Triggiani [7]).

In this paper, we consider the optimal control problem for a stable well-posed linear system and we attempt to clarify the technicalities under the extra assumption that zero is in the resolvent set of the infinitesimal generator of the weakly regular linear system. This ensures the existence of its *reciprocal system* $A^{-1}, A^{-1}B, -CA^{-1}$, $\mathbf{G}(0)$, where A, B, C are the generating operators of the original well-posed linear system and $\mathbf{G}$ is its transfer function. Note that the generating operators of the reciprocal system are all bounded! Interesting connections between well-posed systems and their reciprocal systems are established in section 3. In particular, they have the

same controllability and observability gramians. The concept of reciprocal systems can be traced back to Livsic in his book [8]. As reported in [3], although this concept has been implicitly used, for example, in studying controllability, it has taken a long while for the reciprocal system approach to be used as a tool for finding solutions to Riccati equations for infinite-dimensional systems. Earlier results include Callier and Grabowski [1], [2]), where they assume scalar control and observation operators, exponential stability of A, and strong stability of its inverse, and Curtain [4], where the control operator is assumed to be bounded. Our results show that none of these strong assumptions is necessary. In section 2 we give the relevant background on well-posed and weakly regular linear systems, and in section 4 we first introduce the optimal control problem for a stable well-posed system and review the known results from [22] and [11]. Under the generic assumption $0 \in \rho(A)$, we show that the optimal control problem for the well-posed system is equivalent to one for the reciprocal system. The minimum cost is given by (1.4), where X is the minimal solution of the following reciprocal Riccati equation with bounded operators:

$$\begin{aligned} &A^{-*}X + XA^{-1} + A^{-*}C^*QCA^{-1} \\ &= (B^*A^{-*}X - N_-CA^{-1})^*R_-^{-1}(B^*A^{-*}X - N_-CA^{-1}), \end{aligned}$$

where $N_- = N + G_0^*Q$, $R_- = R + NG_0 + G_0^*N^* + G_0^*QG_0$, and $G_0 = \mathbf{G}(0)$. Moreover, the spectral factor $\mathbf{\Xi}$ is given by $\mathbf{\Xi}(s) = D_-^\xi + C_-^\xi(\frac{1}{s}I - A^{-1})^{-1}A^{-1}$. This represents a complete solution to the problem without making any regularity assumptions, and it is our main new contribution. However, it is instructive to compare our solution with those in the existing literature. In particular, we give new necessary and sufficient conditions for $\mathbf{\Xi}$ to be (weakly) regular; namely, Σ is (weakly) regular and the following limit exists in the (weak) strong topology:

$$\lim_{\lambda\to\infty} B^*A^{-*}X\lambda(\lambda I - A)^{-1}B.$$

If, in addition, $B^*_{\Lambda_w}X(\lambda I - A)^{-1}B$ has the weak limit $V \in \mathcal{L}(\mathcal{U})$ as $\lambda \to \infty$, then $(D^\xi)^*D^\xi = R + V$. This formula was previously obtained in [14]. We also obtain a new result: if the degree of unboundedness of B is $< \frac{1}{2}$ (B is less than maximally unbounded), then $\mathbf{\Xi}$ is uniformly regular with an invertible feedthrough operator and X also satisfies the Riccati equation (4.12) given in [22].

Finally, we remark that the term "generic" is not misplaced. Analogous results may be obtained by replacing the assumption that $0 \in \rho(A)$ by the assumption that $i\omega \in \rho(A)$ for some real ω. In this case, the relevant reciprocal system is $\Sigma(A_\omega^{-1}, A_\omega^{-1}B, -CA_\omega^{-1}, \mathbf{G}(i\omega))$. While there are generators (e.g., the shift operator) that do not satisfy the above condition, most do.

2. Preliminaries. Since our results are based on the theory of *well-posed linear systems* and *weakly regular linear systems*, we review the relevant theory from Weiss [16], [17], [18], [19], [20], Weiss and Curtain [21] and Weiss and Weiss [22]. We begin with some notation.

DEFINITION 2.1. *Let $\mathcal{Z}_1, \mathcal{Z}_2$ be Hilbert spaces, $\mathcal{B}$ a Banach space, and $\Omega \subset \mathbb{R}$.*

- $\mathbf{L}_2(\Omega, \mathcal{Z}_1)$ *is the space of Lebesgue-measurable, square integrable, $\mathcal{Z}_1$-valued functions on Ω.*
- $\mathbf{L}_2^{loc}(0,\infty;\mathcal{Z}_1)$ *is the space of Lebesgue-measurable functions from $[0,\infty)$ to $\mathcal{Z}_1$, which are square integrable on $[0,\tau]$ for every $\tau > 0$, with the topology determined by the seminorms $\|\cdot\|_{\mathbf{L}_2[0,\tau]}$.*

- $\mathbf{H}_2(\mathcal{Z}_1)$ *is the space of holomorphic* $\mathcal{Z}_1$*-valued functions on the open right half-plane* $\mathbb{C}^+$ *which are uniformly square integrable on vertical lines.*
- $\mathbf{H}_\infty(\mathcal{B})$ *is the space of bounded, holomorphic,* $\mathcal{B}$*-valued functions on* $\mathbb{C}^+$*.*
- $\mathbf{L}_\infty(\mathcal{L}(\mathcal{Z}_1,\mathcal{Z}_2))$ *is the space of essentially bounded, weakly measurable,* $\mathcal{L}(\mathcal{Z}_1, \mathcal{Z}_2)$*-valued functions on the imaginary axis.*

Scalar-valued function spaces will be denoted $\mathbf{H}_2, \mathbf{H}_\infty, \mathbf{L}_\infty$, etc. For simplicity, we suppose that all Hilbert spaces are separable. Let $\mathcal{W}$ be any such Hilbert space. We denote the right shift by τ on $\mathbf{L}_2^{loc}(0,\infty;\mathcal{W})$ by $\mathbf{S}_\tau$, i.e.,

$$(\mathbf{S}_\tau w)(t) = \begin{cases} 0, & 0 \le t < \tau, \\ w(t-\tau), & t \ge \tau. \end{cases}$$

An operator $\mathbb{F}$ on $\mathbf{L}_2^{loc}(0,\infty;\mathcal{W})$ is called *shift-invariant* if $\mathbb{F}\,\mathbf{S}_\tau = \mathbf{S}_\tau\mathbb{F}$ for all $\tau > 0$. $\mathbf{P}_\tau$ denotes the projection of $\mathbf{L}_2^{loc}(0,\infty;\mathcal{W})$ into $\mathbf{L}_2(0,\infty;\mathcal{W})$ by truncation, defined for $w \in \mathbf{L}_2^{loc}(0,\infty;\mathcal{W})$ by

$$(\mathbf{P}_\tau w)(t) = \begin{cases} w(t), & 0 \le t < \tau, \\ 0, & t \ge \tau. \end{cases}$$

For $w_1, w_2 \in \mathbf{L}_2^{loc}(0,\infty;\mathcal{W})$ and $\tau \ge 0$, the τ*-concatenation* of w_1 and w_2, denoted $w_1 \underset{\tau}{\diamond} w_2$, is defined by

$$\left(w_1 \underset{\tau}{\diamond} w_2\right)(t) = \mathbf{P}_\tau w_1 + \mathbf{S}_\tau w_2.$$

We now define well-posed linear systems on Hilbert spaces.

DEFINITION 2.2. *Let* $\mathcal{U}, \mathcal{X}, \mathcal{Y}$ *be given Hilbert spaces. A* well-posed linear system *on* $\mathcal{U}$*,* $\mathcal{X}$*, and* $\mathcal{Y}$ *is a quadruple* $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$*, where*

(i) $\mathbb{T} = (\mathbb{T}_t)_{t\ge 0}$ *is a strongly continuous semigroup of bounded linear operators on* $\mathcal{X}$*;*

(ii) $\Phi = (\Phi_t)_{t\ge 0}$ *is a family of bounded linear operators from* $\mathbf{L}_2(0,\infty;\mathcal{U})$ *to* $\mathcal{X}$ *such that*

$$\Phi_{\tau+t}\left(u \underset{\tau}{\diamond} v\right) = \mathbb{T}_t\Phi_\tau u + \Phi_t v$$

for any $u, v \in \mathbf{L}_2(0,\infty;\mathcal{U})$ *and any* $\tau, t \ge 0$*;*

(iii) Ψ *is a continuous linear operator from* $\mathcal{X}$ *to* $\mathbf{L}_2^{loc}(0,\infty;\mathcal{Y})$ *such that for any* $x \in \mathcal{X}$ *and* $\tau > 0$*,*

$$\Psi x = \Psi x \underset{\tau}{\diamond} \Psi\mathbb{T}_\tau x;$$

(iv) $\mathbb{F}$ *is a continuous linear operator from* $\mathbf{L}_2(0,\infty;\mathcal{U})$ *to* $\mathbf{L}_2^{loc}(0,\infty;\mathcal{Y})$ *such that for any* $u, v \in \mathbf{L}_2(0,\infty;\mathcal{U})$*,*

$$\mathbb{F}\left(u \underset{\tau}{\diamond} v\right) = \mathbb{F}u \underset{\tau}{\diamond} (\Psi\Phi_\tau u + \mathbb{F}v).$$

$\mathcal{U}$ *is the* input space*,* $\mathcal{X}$ *is the* state-space*, and* $\mathcal{Y}$ *is the* output space*.*

If A is the generator of the strongly continuous semigroup $\mathbb{T}$ on $\mathcal{X}$, we denote by $\mathcal{X}_1$ the space $\mathcal{D}(A)$ with the norm $\|z\|_1 = \|(\beta I - A)z\|$, where $\beta \in \rho(A)$, and $\mathcal{X}_{-1}$ is the completion of $\mathcal{X}$ with respect to the norm $\|z\|_1 = \|(\beta I - A)^{-1}z\|$. The choice of β is unimportant, since different choices produce equivalent norms. Consequences of

the definition are the existence of certain operators A, B, C. Assumption (i) implies the existence of the infinitesimal generator $A \in \mathcal{L}(\mathcal{X}_1, \mathcal{X})$ of $\mathbb{T}$. Assumptions (i) and (ii) above imply the existence of a unique $B \in \mathcal{L}(\mathcal{U}, \mathcal{X}_{-1})$, called the *control operator* of Σ, such that for all $t \geq 0$,

$$\Phi_t u = \int_0^t \mathbb{T}_{t-\sigma} B u(\sigma)\, \mathrm{d}\sigma. \tag{2.1}$$

The fact that $\Phi_t u \in \mathcal{X}$ means that B is an *admissible control operator* for $\mathbb{T}$ (i.e., $\Phi_t \in \mathcal{L}(\mathbf{L}_2(0,t;\mathcal{U}), \mathcal{X})$). From (2.1) we see that $\Phi_t u$ depends only on $\mathbf{P}_t u$, and so Φ_t has a natural extension to $\mathbf{L}_2^{loc}(0,\infty;\mathcal{U})$. B is called *infinite-time admissible* for $\mathbb{T}$ if for all $u \in \mathcal{L}(\mathbf{L}_2(0,\infty;\mathcal{U}))$,

$$\sup_{0\leq t<\infty} \left\| \int_0^t \mathbb{T}_\sigma B u(\sigma)\, \mathrm{d}\sigma \right\|_{\mathcal{X}} < \infty. \tag{2.2}$$

In this case, we can define the *extended input map* $\tilde{\Phi} \in \mathcal{L}(\mathbf{L}_2(0,\infty;\mathcal{U}), \mathcal{X})$ by

$$\tilde{\Phi} v = \lim_{T\to\infty} \int_0^T \mathbb{T}_\sigma B v(\sigma)\, \mathrm{d}\sigma. \tag{2.3}$$

If $x_0 \in \mathcal{X}$ is the initial state of Σ and $u \in \mathbf{L}_2^{loc}(0,\infty;\mathcal{U})$ is its input function, then the *state trajectory* of Σ, $x : [0,\infty) \to \mathcal{X}$ is defined by

$$x(t) = \mathbb{T}_t x_0 + \Phi_t u \tag{2.4}$$

for all $t \geq 0$. The function x is continuous in $\mathcal{X}$ and it satisfies the differential equation

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{2.5}$$

in the strong sense in $\mathcal{X}_{-1}$. The function x is the unique solution of (2.5) satisfying the initial condition $x(0) = x_0$. If u has the Laplace transform $\hat{u}$ and $x_0 = 0$, then x has the Laplace transform $\hat{x}(s) = (sI - A)^{-1} B\hat{u}(s)$ for all s with $\mathrm{Re}(s)$ sufficiently large. The operator Ψ in Definition 2.2 is called the *extended output map* of Σ. More generally, any operator Ψ which satisfies assumption (iii) in Definition 2.2 is called an *extended output map* for $\mathbb{T}$. For every such Ψ there exists a unique $C \in \mathcal{L}(\mathcal{X}_1, \mathcal{Y})$ called the *observation operator* of Ψ, such that

$$(\Psi x_0)(t) = C\mathbb{T}_t x_0 \tag{2.6}$$

for every $x_0 \in \mathcal{X}_1$ and every $t \geq 0$. This C determines Ψ, since $\mathcal{X}_1$ is dense in $\mathcal{X}$. The function $y_0 = \Psi x_0$ has a Laplace transform $\hat{y}_0$, and we have $\hat{y}_0(s) = C(sI - A)^{-1} x_0$ for all $x_0 \in \mathcal{X}$ and for $\mathrm{Re}(s)$ sufficiently large. If Ψ is *bounded*, i.e., $\Psi \in \mathcal{L}(\mathcal{X}, \mathbf{L}_2(0,\infty;\mathcal{Y}))$, then we say that C is *infinite-time admissible*.

In Grabowski [5] it is shown that C is infinite-time admissible if and only if the following Lyapunov equation has a self-adjoint nonnegative solution $L \in \mathcal{L}(X)$ for all $z, x \in D(A)$:

$$\langle Az, Lx\rangle + \langle z, LAx\rangle = -\langle Cz, Cx\rangle. \tag{2.7}$$

The *observability gramian* $L_C = \Psi^*\Psi$ is the minimal solution of (2.7) (see Hansen and Weiss [6], where the dual controllability gramian is treated).

The Λ-*extension* of C is defined by

$$C_\Lambda z = \lim_{\lambda \to +\infty} C\lambda(\lambda I - A)^{-1} z. \tag{2.8}$$

The domain $\mathcal{D}(C_\Lambda)$ consists of those $z \in \mathcal{X}$ for which the above limit exists (λ is real). If we replace C by C_Λ in (2.6), then it holds for all $x_0 \in \mathcal{X}$ and almost every $t \geq 0$. The operator C_{Λ_w}, the *weak* Λ-*extension* of C, is defined by

$$C_{\Lambda_w} z = \text{weak} \lim_{\lambda \to +\infty} C\lambda(\lambda I - A)^{-1} z. \tag{2.9}$$

The domain of C_{Λ_w} consists of those $z \in \mathcal{X}$ for which the above limit exists. C_{Λ_w} is an extension of C_Λ and they are equal if $\mathcal{Y}$ is finite-dimensional. Note that for all $x_0 \in \mathcal{X}$ and almost all $t \geq 0$,

$$(\Psi x_0)(t) = C_{\Lambda_w} \mathbb{T}_t x_0 = C_\Lambda \mathbb{T}_t x_0. \tag{2.10}$$

The operator $\mathbb{F}$ in Definition 2.1 is called the *extended input-output map* of Σ. $\mathbb{F}$ is shift-invariant, which implies that $\mathbb{F}$ is *causal*:

$$\mathbf{P}_\tau \mathbb{F} = \mathbf{P}_\tau \mathbb{F} \mathbf{P}_\tau \quad \text{for all} \quad \tau \geq 0. \tag{2.11}$$

Using (2.11), we can extend $\mathbb{F}$ continuously to $\mathbf{L}_2^{loc}(0, \infty; \mathcal{U})$. If $\mathbb{T}_t$ is exponentially stable, then $\mathbb{F} \in \mathcal{L}(\mathbf{L}_2(0, \infty; \mathcal{U}), \mathbf{L}_2(0, \infty; \mathcal{Y}))$ and B and C are infinite-time admissible. We can represent $\mathbb{F}$ via the *transfer function* $\mathbf{G}$ of Σ, which is a bounded analytic $\mathcal{L}(\mathcal{U}, \mathcal{Y})$-valued function on some right half-plane in $\mathbb{C}$. We do not distinguish between two transfer functions defined on different right half-planes if one is a restriction of the other. The connection between $\mathbb{F}$ and $\mathbf{G}$ is as follows: if $u \in \mathbf{L}_2(0, \infty; \mathcal{U})$, then $y = \mathbb{F}u$ has a Laplace transform $\hat{y}$ and, for $\text{Re}(s)$ sufficiently large,

$$\hat{y}(s) = \mathbf{G}(s)\hat{u}(s), \tag{2.12}$$

and $\mathbb{F} \in \mathcal{L}(\mathbf{L}_2(0, \infty; \mathcal{U}), \mathbf{L}_2(0, \infty; \mathcal{Y}))$ if and only if $\mathbf{G} \in \mathbf{H}_\infty(\mathcal{L}(\mathcal{U}, \mathcal{Y}))$. For s, β in some right half-plane, the relationship with the generating operators is given by

$$\begin{aligned} \mathbf{G}(s) - \mathbf{G}(\beta) &= C[(sI - A)^{-1} - (\beta I - A)^{-1}]B \\ &= (\beta - s)C(sI - A)^{-1}(\beta I - A)^{-1}B. \end{aligned} \tag{2.13}$$

In fact, a necessary and sufficient condition for A, B, C to generate a well-posed system is that B and C are admissible for $T(\cdot)$ and the right-hand side of (2.13) be uniformly bounded on some right half-plane for one $\beta \in \rho(A)$. The formula (2.13) can be extended to all s, β in the resolvent set of A, and it is used as the definition of the transfer function on page 10 in Staffans and Weiss [15]. If $\rho(A)$ is connected, this agrees with the analytic continuation to $\rho(A)$, but otherwise these two extensions may differ. To avoid confusion we call this extension the *characteristic function* and denote it by $\mathfrak{G}$. Under our assumptions in section 4, the two concepts are consistent.

LEMMA 2.3. *Suppose that the well-posed linear system Σ has the generating operators A, B, C and the transfer function $\mathbf{G}$. If C is an infinite-time admissible observation operator, then $\mathbf{G}$ has an analytic extension on $\mathbb{C}_0^+$ and $\mathbf{G} = \mathfrak{G}$ in $\rho(A) \cap \mathbb{C}_0^+$.*

Proof. Now $\Psi \in \mathcal{L}(\mathcal{X}, \mathbf{L}_2(0, \infty; \mathcal{Y}))$ and so its Laplace transform is such that $\hat{\Psi}x \in \mathbf{H}_2(\mathcal{Y})$ for $x \in \mathcal{X}$ and from (2.10) $\hat{\Psi}(s)x = C(sI - A)^{-1}x$ for s in some right

half-plane. For $x \in D(A)$ we also have $\hat{\Psi}(s)(sI - A)x = Cx$, and since both sides are analytic on $\mathbb{C}_0^+$, this extends to $s \in \mathbb{C}_0^+$. So we obtain

$$\hat{\Psi}(s)x = C(sI - A)^{-1}x \quad \text{for} \quad s \in D(A) \cap \mathbb{C}_0^+, \tag{2.14}$$

and since $D(A)$ is dense in $\mathcal{X}$, this extends to $x \in \mathcal{X}$. Now we fix a number $\beta \in \rho(A)$ in some right half-plane and rewrite (2.13) as follows:

$$\mathbf{G}(s) - \mathbf{G}(\beta) = (\beta - s)\hat{\Psi}(s)(\beta I - A)^{-1}B. \tag{2.15}$$

This shows that $\mathbf{G}$ has an extension to an analytic function on $\mathbb{C}_0^+$. Similarly, for $z \in \mathbb{C}_0^+$, we have

$$\mathbf{G}(z) - \mathbf{G}(\beta) = (\beta - z)\hat{\Psi}(z)(\beta I - A)^{-1}B, \tag{2.16}$$

and subtracting (2.16) from (2.15) gives

$$\begin{aligned}\mathbf{G}(s) - \mathbf{G}(z) &= \beta[\hat{\Psi}(s) - \hat{\Psi}(z)](\beta I - A)^{-1}B - [s\hat{\Psi}(s) - z\hat{\Psi}(z)](\beta I - A)^{-1}B \\ &= (z - s)\hat{\Psi}(s)(zI - A)^{-1}B \quad \text{for} \quad s, z \in \rho(A) \cap \mathbb{C}_0^+ \\ &= (z - s)C(sI - A)^{-1}(zI - A)^{-1}B, s, z \in \rho(A) \cap \mathbb{C}_0^+,\end{aligned}$$

where we have used (2.14). Thus for $s, z \in D(A) \cap \mathbb{C}_0^+$, we obtain

$$\mathbf{G}(s) - \mathbf{G}(z) = \mathfrak{G}(s) - \mathfrak{G}(z). \qquad \square$$

Clearly, the above lemma has a dual version with B an infinite-time admissible control operator.

If $\mathbf{G}$ is a bounded, analytic, $\mathcal{L}(\mathcal{U}, \mathcal{Y})$-valued function defined on some right half-plane in $\mathbb{C}$, then a *realization* of $\mathbf{G}$ is a well-posed linear system Σ whose transfer function is $\mathbf{G}$.

In order to obtain nice state-space formulas we need to assume a regularity condition.

DEFINITION 2.4. *The system Σ (or its transfer function $\mathbf{G}$) is called* weakly regular *if the following limit exists in $\mathcal{Y}$ for all $v \in \mathcal{U}$:*

$$\textit{weak} \lim_{\lambda \to +\infty} \mathbf{G}(\lambda)v = Dv. \tag{2.17}$$

Note that λ in the above is real. Σ (or $\mathbf{G}$) is called *regular* if the limit (2.17) exists in the norm topology of $\mathcal{Y}$ and uniformly regular if the limit exists in the operator norm topology. The operator $D \in \mathcal{L}(\mathcal{U}, \mathcal{Y})$ is called the *feedthrough operator* of Σ (or of $\mathbf{G}$). A sufficient condition for weak regularity of a well-posed linear system is that $C_{\Lambda_w}(\beta I - A)^{-1}B$ exists for some β in the resolvent set of A. If $\mathcal{Y}$ is finite-dimensional, then weak regularity equals regularity, and if B is bounded, Σ is uniformly regular. Other conditions for regularity can be expressed in terms of the degree of unboundedness of B or of C. If B is an admissible control operator for $\mathbb{T}$, its *degree of unboundedness*, denoted $\alpha(B)$, is the infimum of those $\alpha \geq 0$ for which there exist positive constants δ, ω such that

$$\|(\lambda I - A)^{-1}B\|_{\mathcal{L}(\mathcal{U},\mathcal{X})} \leq \frac{\delta}{\lambda^{1-\alpha}} \quad \text{for all } \lambda \in (\omega, \infty). \tag{2.18}$$

It is known that $\alpha(B) \leq \frac{1}{2}$, and if B is bounded, then $\alpha(B) = 0$. The *degree of unboundedness of an admissible observation operator* C is the degree of unboundedness of the admissible control operator C^* for $\mathbb{T}^*$. We reformulate the recent result from Proposition 4.2 in [21].

LEMMA 2.5. *Let B and C be admissible control and observation operators for $T(\cdot)$. If either B or C has a degree of unboundedness less than $\frac{1}{2}$, then A, B, C are the generating operators of a uniformly regular system.*

One advantage of considering (weakly) regular linear systems is that the formulas for $\mathbf{G}$ and $\mathbb{F}$ are the same as the finite-dimensional ones with C_{Λ_w} replacing C (for well-posed systems they are rather complicated).

THEOREM 2.6. *If Σ is weakly regular, then the following hold:*

(i) $\mathbf{G}(s) = C_{\Lambda_w}(sI - A)^{-1}B + D$ *for sufficiently large* Re s.

(ii) $\mathbb{F} : \mathbf{L}_2^{loc}(0, \infty; \mathcal{U}) \to \mathbf{L}_2^{loc}(0, \infty; \mathcal{Y})$ *is given by*

$$(\mathbb{F}u)(t) = C_{\Lambda_w} \int_0^t \mathbb{T}_{t-\sigma} B u(\sigma)\, \mathrm{d}\sigma + Du(t)$$

for almost all $t \geq 0$.

(iii) *If x is the state trajectory of Σ corresponding to the initial state $x_0 \in \mathcal{X}$ and the input function $u \in \mathbf{L}_2^{loc}(0, \infty; \mathcal{U})$, then $x(t)$ given by (2.4) satisfies $x(t) \in \mathcal{D}(C_{\Lambda_w})$ for almost every $t \geq 0$, and the output function of Σ, $y = \Psi x_0 + \mathbb{F}u$ satisfies*

$$y(t) = C_{\Lambda_w} x(t) + Du(t) \quad \textit{for almost all } t \geq 0.$$

If Σ is regular, then C_Λ may replace C_{Λ_w} in the above. A, B, C, D are called the generating operators *of Σ.*

In this paper the operator $B^*_{\Lambda_w}$ occurs frequently, where

$$B^*_{\Lambda_w} z = \text{weak} \lim_{\lambda \to \infty} B^* \lambda (\lambda I - A^*)^{-1} z, \tag{2.19}$$

and its domain consists of all $z \in \mathcal{X}$ for which the weak limit makes sense.

The following proposition about inverses of regular linear systems follows from results in sections 4 and 7 in [19] by considering unity feedback of $I - \mathbf{G}$.

PROPOSITION 2.7. *Let Σ be a regular linear system with generating operators A, B, C, D and transfer function $\mathbf{G}$. Then D is invertible if and only if $\mathbf{G}^{-1}$ is regular. In this case $\mathbf{G}^{-1}$ is the transfer function of the regular linear system Σ^{inv} with generating operators $A^{inv}, B^{inv}, C^{inv}, D^{inv}$ given by*

$$\begin{aligned} A^{inv}x &= (A - BD^{-1}C_\Lambda)x \quad \textit{for all } x \in D(A^{inv}), \\ D(A^{inv}) &= \{x \in D(C_\Lambda) \mid (A - BD^{-1}C_\Lambda)x \in \mathcal{X}\}, \\ C^{inv}x &= -D^{-1}C_\Lambda x \quad \textit{for all } x \in \mathcal{D}(A^{inv}), \\ B^{inv} &= BD^{-1}, D^{inv} = D^{-1}. \end{aligned}$$

Under the assumptions of this proposition, we call Σ^{inv} the inverse of Σ.

3. Reciprocal systems. In this section we consider a well-posed linear system Σ with the generating operators A, B, C, transfer function $\mathbf{G}$, and characteristic function $\mathfrak{G}$, and we assume that $0 \in \rho(A)$. Using the definition of the transfer function from (2.13), it is clear that for all $s \in \rho(A)$

$$\begin{aligned} \mathfrak{G}(s) - \mathfrak{G}(0) &= C[(sI - A)^{-1} + A^{-1}]B \\ &= -CA^{-1}\left(\frac{1}{s}I - A^{-1}\right)^{-1} A^{-1}B. \end{aligned} \tag{3.1}$$

Note that $A^{-1}, A^{-1}B, CA^{-1}$ are all bounded operators, and so they generate a regular linear system with feedthrough operator zero, and (3.1) holds for all $s \in \rho(A)$. This motivates the following definition.

DEFINITION 3.1. *Let* $\mathbf{G}$ *be the transfer function of a well-posed linear system* Σ *with generating operators* A, B, C *with* $0 \in \rho(A)$. *We call the regular linear system* Σ_- *with generating operators* $A^{-1}, A^{-1}B, -CA^{-1}, \mathfrak{G}(0)$ *the* reciprocal system corresponding to Σ. *We denote its transfer function and characteristic functions by* $\mathbf{G}_-$ *and* $\mathfrak{G}_-$, *respectively.*

In the next lemma we prove some interesting connections between well-posed linear systems and their reciprocal systems.

LEMMA 3.2. *Under the assumptions of Definition* 3.1 *and denoting the group with generator* A^{-1} *by* $\mathbb{T}_-$, *the following hold:*

1. $\mathfrak{G}_-(\frac{1}{s}) = \mathfrak{G}(s)$ *for all* $s \in \rho(A)$.
2. C *is an infinite-time admissible observation operator for* $\mathbb{T}$ *if and only if* CA^{-1} *is an infinite-time admissible operator for* $\mathbb{T}_-$. *In this case, their observability gramians are equal and are the minimal solutions of their respective Lyapunov equations.*
3. B *is an infinite-time admissible control operator for* $\mathbb{T}$ *if and only if* $A^{-1}B$ *is an infinite-time admissible control operator for* $\mathbb{T}_-$. *In this case, their controllability gramians are equal and are the minimal solutions of their respective Lyapunov equations.*
4. *If* C *is an infinite-time admissible observation operator for* $\mathbb{T}$, *then* $\mathbf{G}(s) = \mathbf{G}_-(\frac{1}{s})$ *for* $s \in \mathbb{C}_0^+$. *Moreover,* $\mathbf{G} \in \mathbf{H}_\infty(\mathcal{L}(\mathcal{U}, \mathcal{Y}))$ *if and only if* $\mathbf{G}_- \in \mathbf{H}_\infty(\mathcal{L}(\mathcal{U}, \mathcal{Y}))$.

Proof.

1. This follows from (3.1).
2. From Grabowski [5] we know that C is an infinite-time admissible operator for $\mathbb{T}$ if and only if Lyapunov equation (2.7) has a self-adjoint nonnegative solution $L \in \mathcal{L}(\mathcal{X})$. An analogous statement holds for CA^{-1} and the following Lyapunov equation with values in $\mathcal{X}$:

$$A^{-*}L_C + L_C A^{-1} = -A^{-*}C^*CA^{-1}. \tag{3.2}$$

   It is clear that L solves (2.7) if and only if it solves (3.2). Moreover, if C is infinite-time admissible, the observability gramian of Σ is $\Psi^*\Psi$, and it is the minimal solution of (2.7). An analogous statement holds for the reciprocal system and this proves the claim.
3. This is a dual statement to the infinite-time observability (see Hansen and Weiss [6]).
4. Now, from part 2, both C and CA^{-1} are infinite-time admissible, and so from Lemma 2.3 we have

$$\mathbf{G}(s) = \mathfrak{G}(s) \quad \text{for} \quad s \in \rho(A) \cap \mathbb{C}_0^+ \quad \text{and}$$

$$\mathbf{G}_-(s) = \mathfrak{G}_-(s) \quad \text{for} \quad s \in \rho(A^{-1}) \cap \mathbb{C}_0^+.$$

   This together with (3.1) gives

$$\mathbf{G}(s) = \mathbf{G}_-\left(\frac{1}{s}\right) \quad \text{for} \quad s \in \rho(A) \cap \mathbb{C}_0^+.$$

However, from Lemma 2.3 both $\mathbf{G}$ and $\mathbf{G}_-$ are analytic on $\mathbb{C}_0^+$, and so the above equality holds on $\mathbb{C}_0^+$. □

We now relate the reciprocals of inverses of regular systems.

THEOREM 3.3. *Suppose that Σ is a regular linear system with the generating operators A, B, C, D and $0 \in \rho(A)$. Let Σ_- denote its reciprocal system with transfer function $\mathbf{G}_-$. If D is invertible, then the inverse system Σ^{inv} of Σ is regular. Moreover, Σ_-^{inv}, the inverse system of Σ_-, is the reciprocal system of Σ^{inv}.*

Proof. By Proposition 2.7, if D is invertible, then Σ^{inv} is a regular linear system with generating operators $A^{inv}, C^{inv} = -D^{-1}C_\Lambda, B^{inv} = BD^{-1}, D^{inv} = D^{-1}$, where

$$\begin{aligned} A^{inv}x &= (A - BD^{-1}C_\Lambda)x \quad \text{for all} \quad x \in \mathcal{D}(A^{inv}),\\ \mathcal{D}(A^{inv}) &= \{x \in \mathcal{D}(C_\Lambda)\,|(A - BD^{-1}C_\Lambda)x \in \mathcal{X}\}. \end{aligned}$$

By the same proposition, the inverse of the reciprocal system Σ_-^{inv} has the generating operators

$$\begin{aligned} A_-^{inv} = A^{-1} + A^{-1}BD_-^{-1}CA^{-1}, B_-^{inv} = A^{-1}BD_-^{-1},\\ C_-^{inv} = D_-^{-1}CA^{-1}, D_-^{inv} = D_-^{-1}, \end{aligned} \tag{3.3}$$

where $D_- = \mathfrak{G}(0)$. To show that the reciprocal system of Σ^{inv} is well defined, we first show that A^{inv} has the bounded inverse A_-^{inv}. Consider for all $x \in \mathcal{X}$

$$\begin{aligned} A^{inv}A_-^{inv}x &= (A - BD^{-1}C_\Lambda)(A^{-1} + A^{-1}BD_-^{-1}CA^{-1})x\\ &= (I + BD_-^{-1}CA^{-1} - BD^{-1}C_\Lambda A^{-1})x\\ &\quad -(BD^{-1}C_\Lambda A^{-1}BD_-^{-1}CA^{-1})x\\ &= x + BD^{-1}(D - D_- - C_\Lambda A^{-1}B)D_-^{-1}CA^{-1}x\\ &= x, \end{aligned}$$

where in the second line the terms are well defined in $\mathcal{X}_{-1}$. Next consider for $x \in \mathcal{D}(A^{inv}) \subset \mathcal{D}(C_\Lambda)$

$$\begin{aligned} A_-^{inv}A^{inv}x &= (A^{-1} + A^{-1}BD_-^{-1}CA^{-1})(A - BD^{-1}C)x\\ &= x - A^{-1}BD^{-1}C_\Lambda x + A^{-1}BD_-^{-1}C_\Lambda x\\ &\quad -A^{-1}BD_-^{-1}C_\Lambda A^{-1}BD^{-1}C_\Lambda x\\ &= x + A^{-1}BD_-^{-1}(-D_- + D - C_\Lambda A^{-1}B)D^{-1}C_\Lambda x\\ &= x. \end{aligned}$$

So the reciprocal system of Σ^{inv} is well defined and it has the generating operators $(A^{inv})^{-1}, (A^{inv})^{-1}B^{inv}, -C^{inv}(A^{inv})^{-1}, D^{inv} - C^{inv}(A^{inv})^{-1}B^{inv}$. We show that these are precisely the generating operators of Σ_-^{inv}. We consider the control operator

$$\begin{aligned} (A^{inv})^{-1}B^{inv} &= (A^{inv})^{-1}BD^{-1} = A_-^{inv}BD^{-1}\\ &= (A^{-1} + A^{-1}BD_-^{-1}C_\Lambda A^{-1})BD^{-1}\\ &= A^{-1}BD_-^{-1}(D_- + C_\Lambda A^{-1}B)D^{-1}\\ &= A^{-1}BD_-^{-1}. \end{aligned}$$

The observation operator is

$$\begin{aligned} C^{inv}(A^{inv})^{-1} &= -D^{-1}C_\Lambda(A^{inv})^{-1} = -D^{-1}C_\Lambda(A^{-1} + A^{-1}BD_-^{-1}CA^{-1})\\ &= -D^{-1}(D_- + C_\Lambda A^{-1}B)D_-^{-1}CA^{-1}\\ &= -D_-^{-1}CA^{-1}. \end{aligned}$$

Finally, the feedthrough term is

$$\begin{aligned} D^{inv} - C^{inv}(A^{inv})^{-1}B^{inv} &= D^{-1} + D^{-1}C_\Lambda[A^{-1} + A^{-1}BD_-^{-1}C_\Lambda A^{-1}]BD^{-1} \\ &= D^{-1}[D + C_\Lambda A^{-1}B + C_\Lambda A^{-1}BD_-^{-1}C_\Lambda A^{-1}B]D^{-1} \\ &= D^{-1}[D + C_\Lambda A^{-1}BD_-^{-1}(D_- + C_\Lambda A^{-1}B)]D^{-1} \\ &= D^{-1}[D + C_\Lambda A^{-1}BD_-^{-1}D]D^{-1} \\ &= D_-^{-1}. \qquad \square \end{aligned}$$

4. The linear quadratic optimal control problem. In this section, we consider the well-posed linear system $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$ on the Hilbert spaces $\mathcal{U}, \mathcal{X}, \mathcal{Y}$, and we denote its generating operators by A, B, C and its transfer function by $\mathbf{G}$. We suppose that Σ satisfies the following assumptions:

(a) $\mathbb{F} \in \mathcal{L}(\mathbf{L}_2(0,\infty;\mathcal{U}), \mathbf{L}_2(0,\infty;\mathcal{Y}))$.
(b) $\Psi \in \mathcal{L}(\mathcal{X}, \mathbf{L}_2(0,\infty;\mathcal{Y}))$.
(c) $\mathcal{U}$ is a separable Hilbert space.
(d) $0 \in \rho(A)$.

Assumption (b) implies that $\hat{\Psi}x \in \mathbf{H}_2(\mathcal{Y})$ for all $x \in \mathcal{X}$, and so we have $\mathbf{G} = \mathfrak{G}$ for $s \in \rho(A) \cap \mathbb{C}_0^+$ (see Lemma 2.3). In particular, with assumption (d) we obtain $\mathbf{G}(0) = \mathfrak{G}(0)$. Consequently, we do not have to distinguish between the transfer function and the characteristic function in what follows. Assumption (b) is equivalent to $\mathbf{G} \in \mathbf{H}_\infty(\mathcal{L}(\mathcal{U},\mathcal{Y}))$. Assumption (c) is made because we choose to quote results from Weiss and Weiss [22] and they use a frequency domain approach. However, it is possible to prove all the results without (c) using a time domain approach similar to that in Mikkola [9]. Let us consider the following cost functional associated with Σ:

$$J(x_0,u) = \int_0^\infty \left\langle \begin{bmatrix} Q & N^* \\ N & R \end{bmatrix} \begin{bmatrix} y(t) \\ u(t) \end{bmatrix}, \begin{bmatrix} y(t) \\ u(t) \end{bmatrix} \right\rangle_{\mathcal{Y}\times\mathcal{U}} dt, \tag{4.1}$$

where $R = R^* \in \mathcal{L}(\mathcal{U}), Q = Q^* \in \mathcal{L}(\mathcal{Y})$, and $N \in \mathcal{L}(\mathcal{Y},\mathcal{U})$. Note that no positivity assumptions are made on R or on Q. For each initial state $x_0 \in \mathcal{X}$, u and y are related as in Theorem 2.6. The *optimal control problem* is to find the input function $u^{opt} \in \mathbf{L}_2(0,\infty;\mathcal{U})$ that minimizes $J(x_0,u)$. We denote this the optimal control problem by (Σ, J). We substitute $y = \Psi x_0 + \mathbb{F}u$ into (4.1) to obtain

$$J(x_0,u) = \left\langle \begin{bmatrix} \Psi^*Q\Psi & \Psi^*(Q\mathbb{F} + N^*) \\ (\mathbb{F}^*Q + N)\Psi & \mathcal{R} \end{bmatrix} \begin{bmatrix} x_0 \\ u \end{bmatrix}, \begin{bmatrix} x_0 \\ u \end{bmatrix} \right\rangle, \tag{4.2}$$

where the scalar product is from $\mathcal{X} \times \mathbf{L}_2(0,\infty;\mathcal{U})$ and

$$\mathcal{R} = R + N\mathbb{F} + \mathbb{F}^*N^* + \mathbb{F}^*Q\mathbb{F}. \tag{4.3}$$

Under our assumptions (a), (b), we see that $J(x_0,u)$ is finite for all $x_0 \in \mathcal{X}$ and $u \in \mathbf{L}_2(0,\infty;\mathcal{U})$. $\mathcal{R}$ is a Toeplitz operator whose symbol is the *Popov function* associated with the above control problem.

DEFINITION 4.1. *The* Popov function $\Pi : i\mathbb{R} \to \mathcal{L}(\mathcal{U})$ *associated with the well-posed linear system* Σ *and the cost function* J *in* (4.1) *is defined for almost every* $\omega \in \mathbb{R}$ *by*

$$\Pi(i\omega) = R + N\mathbf{G}(i\omega) + \mathbf{G}(i\omega)^*N^* + \mathbf{G}(i\omega)^*Q\mathbf{G}(i\omega), \tag{4.4}$$

where $\mathbf{G}$ *is the transfer function of* Σ.

We remark that a sufficient condition for the Popov function Π to be well defined is that $\mathbb{F}$ be a bounded operator from $\mathbf{L}_2(0,\infty;\mathcal{U})$ to $\mathbf{L}_2(0,\infty;\mathcal{Y})$. For in this case, $\mathbf{G}\in\mathbf{H}_\infty(\mathcal{L}(\mathcal{U},\mathcal{Y}))$ and it has an extension to $s=i\omega$ in the sense that $\lim_{\sigma\to 0}\mathbf{G}(\sigma+i\omega)u$ exists for all $u\in\mathcal{U}$, for almost all $\omega\in\mathbb{R}$ (see Theorem 4.5 in Rosenblum and Rovnyak [10]). Moreover, $\Pi\in\mathbf{L}_\infty(\mathcal{L}(\mathcal{U}))$.

We summarize the main results from Weiss and Weiss [22] for the well-posed case (or see Staffans [11]).

THEOREM 4.2. *Let $\Sigma=(\mathbb{T},\Phi,\Psi,\mathbb{F})$ be a well-posed linear system satisfying assumptions* (a), (b), (c)*, and consider the associated cost function J from* (4.1)*. If the corresponding Popov function Π from* (4.4) *is coercive, i.e., there exists $\epsilon>0$ such that $\Pi(i\omega)\geq\epsilon I$ for almost all $\omega\in\mathbb{R}$, then it has a spectral factorization*

$$\Pi(i\omega)=\mathbf{\Xi}(i\omega)^*\mathbf{\Xi}(i\omega) \tag{4.5}$$

for almost all $\omega\in\mathbb{R}$. $\mathbf{\Xi}$ and its inverse are in $\mathbf{H}_\infty(\mathcal{L}(\mathcal{U},\mathcal{Y}))$ and it has the well-posed realization $\Sigma^\xi=(\mathbb{T},\Phi,\Psi^\xi,\mathbb{F}^\xi)$, where Ψ^ξ is an extended output map for $\mathbb{T}$ defined by

$$\Psi^\xi=(\mathbb{F}^\xi)^{-*}(\mathbb{F}^*Q+N)\Psi. \tag{4.6}$$

The optimal control is $u^{opt}=-\mathcal{R}^{-1}(\mathbb{F}^*Q+N)\Psi x_0$ and the optimal cost has the form

$$J(x_0,u^{opt})=\langle Xx_0,x_0\rangle, \tag{4.7}$$

where $X=X^*\in\mathcal{L}(\mathcal{X})$ is defined by

$$X=\Psi^*Q\Psi-\Psi^*(Q\mathbb{F}+N^*)\mathcal{R}^{-1}(\mathbb{F}^*Q+N)\Psi \tag{4.8}$$

$$=\Psi^*Q\Psi-(\Psi^\xi)^*\Psi^\xi. \tag{4.9}$$

Proof. All results are stated explicitly in Weiss and Weiss [22] with the exception of (4.9). This follows from the two equations (11.5) and (12.3) in [22], namely,

$$\mathcal{R}=(\mathbb{F}^\xi)^*\mathbb{F}^\xi \quad\text{and}\quad (\mathbb{F}^\xi)^*\Psi^\xi=(\mathbb{F}^*Q+N)\Psi. \qquad \square$$

In order to obtain Riccati equations resembling the usual ones (1.5) it is necessary to assume that Σ is weakly regular. The main results for this case from [22] are as follows.

THEOREM 4.3. *Let $\Sigma=(\mathbb{T},\Phi,\Psi,\mathbb{F})$ be a weakly regular linear system satisfying assumptions* (a), (b), (c) *and with feedthrough operator zero. If the Popov function is coercive and its spectral factor $\mathbf{\Xi}$ is regular with an invertible feedthrough operator D^ξ, then it has the generating operators A,B,C^ξ,D^ξ, where*

$$C^\xi=D^\xi((D^\xi)^*D^\xi)^{-1}(B^*_{\Lambda_w}X+NC), \tag{4.10}$$

where X maps $D(A)$ into $D(B^*_{\Lambda_w})$ and

$$\mathbf{\Xi}(s)=D^\xi+C^\xi_\Lambda(sI-A)^{-1}B. \tag{4.11}$$

Moreover, X satisfies the Riccati equation

$$A^*X+XA+C^*QC=(B^*_{\Lambda_w}X+NC)^*((D^\xi)^*D^\xi)^{-1}(B^*_{\Lambda_w}X+NC), \tag{4.12}$$

where all terms have values in $\mathcal{L}(\mathcal{X}_1,\mathcal{X}_{-1})$, and the optimal control is given by $u^{opt}(t)=-(D^{\xi^*}D^\xi)^{-1}(B^*_{\Lambda_w}X+NC)_\Lambda x^{opt}(t)$.

Note that even if Σ is regular, the spectral factor need not be. Moreover, even if the spectral factor is regular, $(D^{\xi})^* D^{\xi} = R$ need not hold as in the finite-dimensional theory; D_{ξ} is defined as the feedthrough operator of the spectral factor. While there are known sufficient conditions for these to hold (see the introduction), it is not as nice a theory as we had hoped for.

So far we have not used assumption (d). This is now invoked to show that for the general well-posed system the operator X defined by (4.8) is always the solution of a certain Riccati equation. Regularity is not needed. To achieve this we introduce the reciprocal system of Σ: $\Sigma_- = (\mathbb{T}_-, \Phi_-, \Psi_-, \mathbb{F}_-)$ with the generating operators $A^{-1}, A^{-1}B, -CA^{-1}, \mathbf{G}(0) = G_0$ and a cost functional J_- of the form (4.1) with $(\mathbb{T}, \Phi, \Psi, \mathbb{F})$ replaced by $(\mathbb{T}_-, \Phi_-, \Psi_-, \mathbb{F}_- - G_0)$ and R, N replaced by N_-, R_- given by

$$N_- = N + G_0^* Q; \quad R_- = R + NG_0 + G_0^* N^* + G_0^* Q G_0. \tag{4.13}$$

Note that this is an algebraic device to reduce a control problem for the reciprocal system Σ_-, which has nonzero feedthrough, to an equivalent control problem for a system with zero feedthrough. (This is the formulation of the control problem in Theorems 4.2 and 4.3.) We denote this optimal control problem by (Σ_-, J_-), and we call it the *reciprocal optimal control problem associated with the optimal control problem* (Σ, J). The following result justifies this notation.

THEOREM 4.4. *Let Σ be a well-posed linear system that satisfies assumptions* (a)–(d). *Then the minimum cost of the control problem (Σ, J) equals the minimum cost of the reciprocal control problem (Σ_-, J_-).*

Proof. The cost functionals can be expressed as

$$\begin{aligned} J(x,u) = {} & \langle \Psi^* Q \Psi x, x\rangle + \langle \mathcal{R} u, u\rangle_2 \\ & + \langle (\mathbb{F}^* Q + N)\Psi x, u\rangle_2 + \langle u, (\mathbb{F}^* Q + N)\Psi x\rangle_2, \end{aligned} \tag{4.14}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product in $\mathcal{X}$ (or subsequently $\mathcal{U}$ or $\mathcal{Y}$) and $\langle \cdot, \cdot \rangle_2$ denotes the scalar product in $\mathbf{L}_2(0, \infty; \mathcal{U})$ (or $\mathbf{L}_2(0, \infty; \mathcal{Y})$).

$$\begin{aligned} J_-(x,v) = {} & \langle \Psi_-^* Q \Psi_- x, x\rangle + \langle \mathcal{R}_- v, v\rangle_2 \\ & + \langle (\mathbb{F}_-^* Q + N)\Psi_- x, v\rangle_2 + \langle v, (\mathbb{F}_-^* Q + N)\Psi_- x\rangle_2, \end{aligned} \tag{4.15}$$

where

$$\mathcal{R}_- = R_- + N_- \mathbb{F}_- + \mathbb{F}_-^* N_-^* + \mathbb{F}_-^* Q \mathbb{F}_-, \tag{4.16}$$

R_-, N_- are defined by (4.13), and we have used

$$(\mathbb{F}_-^* - G_0)Q + N_- = \mathbb{F}_-^* Q + N.$$

The Popov function for the reciprocal control problem is given by

$$\begin{aligned} \Pi_-(i\omega) &= R_- + N_-(\mathbf{G}_-(i\omega) - G_0) + (\mathbf{G}_-(i\omega) - G_0)^* N_-^* \\ &\quad + (\mathbf{G}_-(i\omega) - G_0)^* Q (\mathbf{G}_-(i\omega) - G_0) \\ &= R + N\mathbf{G}_-(i\omega) + \mathbf{G}_-(i\omega)^* N^* + \mathbf{G}_-(i\omega)^* Q \mathbf{G}_-(i\omega). \end{aligned} \tag{4.17}$$

Thus, using Lemma 3.2, we see that

$$\Pi_-\left(\frac{1}{i\omega}\right) = \Pi(i\omega) \quad \text{for almost all } \omega \in \mathbb{R}. \tag{4.18}$$

Next we make use of a useful operator $\mathcal{H}$ on $\mathbf{H}_2(\mathcal{U})$ (or $\mathbf{H}_2(\mathcal{Y})$)

$$(\mathcal{H}f)(s) = \frac{1}{s} f\left(\frac{1}{s}\right) \quad \text{for} \quad f \in \mathbf{H}_2(\mathcal{U}). \tag{4.19}$$

It is readily verified that it is one-to-one, has a bounded inverse, and, in particular, has the properties

$$\langle \mathcal{H}f, g\rangle_{\mathbf{H}_2} = \langle f, \mathcal{H}g\rangle_{\mathbf{H}_2}, \ \langle \mathcal{H}f, \mathcal{H}g\rangle_{\mathbf{H}_2} = \langle f, g\rangle_{\mathbf{H}_2} \tag{4.20}$$

for f, g $\in \mathbf{H}_2(\mathcal{U})$, where $\langle \cdot, \cdot \rangle_{\mathbf{H}_2}$ is the scalar product in $\mathbf{H}_2(\mathcal{U})$ (or $\mathbf{H}_2(\mathcal{Y})$). We show that

$$\widehat{\Psi_- x} = \mathcal{H}(\widehat{\Psi x}), \tag{4.21}$$

where $\hat{u}$ denotes the Laplace transform of u. Now for $x \in D(A)$ we have

$$\begin{aligned} \text{right-hand side (RHS) of (4.21)} &= \mathcal{H}(C(sI - A)^{-1}x) \\ &= \frac{1}{s} C\left(\frac{1}{s} I - A\right)^{-1} x \\ &= -CA^{-1}(sI - A^{-1})^{-1}x \\ &= \text{left-hand side (LHS) of (4.21)}, \end{aligned}$$

and as in Lemma 2.3 this extends to $\mathcal{X}$. Next we prove

$$\Psi^* Q \Psi = \Psi_-^* Q \Psi_- \tag{4.22}$$

by considering the following expression for $x, y \in \mathcal{X}$:

$$\begin{aligned} \langle \Psi^* Q \Psi x, y\rangle &= \frac{1}{2\pi} \langle Q\widehat{\Psi x}, \widehat{\Psi y}\rangle_{\mathbf{H}_2} \\ &= \frac{1}{2\pi} \langle \mathcal{H} Q \widehat{\Psi x}, \mathcal{H}\widehat{\Psi y}\rangle_{\mathbf{H}_2} \\ &= \frac{1}{2\pi} \langle Q\widehat{\Psi_- x}, \widehat{\Psi_- y}\rangle_{\mathbf{H}_2} \quad \text{by (4.21)} \\ &= \langle \Psi_- y, Q \Psi_- x\rangle_2. \end{aligned}$$

This holds for all $x, y \in \mathcal{X}$, and so (4.22) holds. We now define the input v in (4.15) via $\hat{v} = \mathcal{H}\hat{u}$. From (4.19) $v \in \mathbf{L}_2(0, \infty; \mathcal{U})$ if and only if $u \in \mathbf{L}_2(0, \infty; \mathcal{U})$. With this definition we show that for all $u \in \mathbf{L}_2(0, \infty; \mathcal{U})$

$$\langle N\Psi x, u\rangle_2 = \langle N\Psi_- x, v\rangle_2. \tag{4.23}$$

$$\begin{aligned} \text{LHS of (4.23)} &= \frac{1}{2\pi} \langle \widehat{\Psi x}, N^* \hat{u}\rangle_{\mathbf{H}_2} \\ &= \frac{1}{2\pi} \langle \mathcal{H}\widehat{\Psi x}, N^* \mathcal{H}\hat{u}\rangle_{\mathbf{H}_2} \quad \text{from (4.20)} \\ &= \frac{1}{2\pi} \langle \widehat{\Psi_- x}, N^* \hat{v}\rangle_{\mathbf{H}_2} \quad \text{from (4.21)} \\ &= \text{RHS of (4.23)}. \end{aligned}$$

Next for all $u \in \mathbf{L}_2(0,\infty;\mathcal{U}), x \in \mathcal{X}$ we verify

$$\langle \mathbb{F}^* Q \Psi x, u\rangle_2 \;=\; \langle \mathbb{F}_-^* Q \Psi_- x, v\rangle_2. \tag{4.24}$$

$$\begin{aligned}
\text{LHS of (4.24)} &= \langle \Psi x, Q\mathbb{F}u\rangle_2 \\
&= \frac{1}{2\pi}\langle \widehat{\Psi x}, Q\widehat{\mathbb{F}u}\rangle_{\mathbf{H}_2} \\
&= \frac{1}{2\pi}\langle \mathcal{H}(\widehat{\Psi x}), Q\mathcal{H}(\mathbf{G}\hat{u})\rangle_{\mathbf{H}_2} \quad \text{by (4.20)} \\
&= \frac{1}{2\pi}\langle \widehat{\Psi_- x}, Q\mathbf{G}_-\hat{v}\rangle_{\mathbf{H}_2} \quad \text{by (4.21) and Lemma 3.2} \\
&= \langle \Psi_- x, Q\mathbb{F}_- v\rangle_2 \\
&= \text{RHS of (4.24)}.
\end{aligned}$$

Next for all $u \in \mathbf{L}_2(0,\infty;\mathcal{U})$ we show that

$$\langle \mathcal{R}u, u\rangle_2 \;=\; \langle \mathcal{R}_- v, v\rangle_2. \tag{4.25}$$

$$\begin{aligned}
\text{LHS of (4.25)} &= \frac{1}{2\pi}\langle \Pi\hat{u}, \hat{u}\rangle_{\mathbf{H}_2} \\
&= \frac{1}{2\pi}\langle \mathcal{H}(\Pi\hat{u}), \mathcal{H}(\hat{u})\rangle_{\mathbf{H}_2} \quad \text{by (4.20)} \\
&= \frac{1}{2\pi}\langle \Pi_-\hat{v}, \hat{v}\rangle_{\mathbf{H}_2} \quad \text{from (4.18)} \\
&= \text{RHS of (4.25)}.
\end{aligned}$$

Finally, combining (4.22), (4.21), (4.23), and (4.25), we conclude that

$$J(x,u) = J_-(x,v) \quad \text{for all } \; u \in \mathbf{L}_2(0,\infty;\mathcal{U}), x \in \mathcal{X},$$

which completes the proof. □

Now the reciprocal system certainly satisfies all the assumptions of Theorem 4.3, and this leads to our main result.

THEOREM 4.5. *Suppose that* Σ *is a well-posed linear system satisfying the conditions* (a)–(d). *If its Popov function* Π *associated with the optimal control problem* (Σ, J) *is coercive, then the reciprocal optimal control problem* (Σ_-, J_-) *has the minimum cost*

$$J_-(x_0, v^{opt}) = \langle x_0, X_- x_0\rangle, \tag{4.26}$$

where X_- *is given by*

$$X_- = \Psi_-^* Q \Psi_- - \Psi_-^*(Q\mathbb{F}_- + N_-^*)\mathcal{R}_-^{-1}(\mathbb{F}_-^* Q + N_-)\Psi_-,$$

where $\mathcal{R}_-$ *is given by* (4.16) *and* R_-, N_- *by* (4.13). X_- *is a solution of the Riccati equation*

$$A^{-*}X + XA^{-1} + A^{-*}C^*QCA^{-1} \tag{4.27}$$

$$= (B^*A^{-*}X - N_-CA^{-1})^* R_-^{-1}(B^*A^{-*}X - N_-CA^{-1}), \tag{4.28}$$

and the optimal control $v^{opt} \in \mathbf{L}_2(0,\infty;\mathcal{U})$ *is given by* $v^{opt}(t) = -R_-^{-1}(B^*A^{-*}X - N_-CA^{-1})x_-^{opt}(t)$.

The Popov function defined by (4.17) *has a spectral factor given by*

$$\mathbf{\Xi}_-(s) = D_-^\xi + C_-^\xi(sI - A^{-1})^{-1}A^{-1}B, \tag{4.29}$$

where D_-^ξ *is any invertible solution of* $R_- = (D_-^\xi)^*D_-^\xi$ *and* C_-^ξ *is infinite-time admissible and is defined by*

$$C_-^\xi = D_-^\xi R_-^{-1}(B^*A^{-*}X - N_-CA^{-1}). \tag{4.30}$$

The optimal cost for the control problem (Σ, J) *equals that of the reciprocal optimal control problem* (Σ_-, J_-), *and their optimal controls are related by* $\hat{u}^{opt} = \mathcal{H}\hat{v}^{opt}$.

Moreover, the realization $\Sigma_-^\xi = (\mathbb{T}_-, \Phi_-, \Psi_-^\xi, \mathbb{F}_-^\xi)$ *of* $\mathbf{\Xi}_-$ *is the reciprocal system of* $\Sigma^\xi = (\mathbb{T}, \Phi, \Psi^\xi, \mathbb{F}^\xi)$, *and the observation operator for* $\mathbf{\Xi}_-$ *is given by*

$$-C^\xi A^{-1} = C_-^\xi = D_-^\xi R_-^{-1}(B^*A^{-*}X - N_-CA^{-1}). \tag{4.31}$$

Proof. Note first from Lemma 3.2 that Σ_- satisfies conditions (a), (b), (c) if and only if Σ does. Now (4.18) shows that $\Pi_- \geq \epsilon I$ if and only if $\Pi \geq \epsilon I$. So Π_- has a spectral factorization and the spectral factor is necessarily regular by Lemma 2.5 and Theorem 4.2. In fact, it is regular in the uniform topology and the feedthrough operator D_-^ξ is invertible. We apply Theorem 4.3 to (Σ_-, J_-) to conclude that the reciprocal control problem has the solution as stated.

From Theorem 4.4 we know that the optimal cost is the same as for the original control problem (Σ, J) and the optimal control for the original system u^{opt} is defined in terms of v^{opt} via $\hat{u}^{opt} = \mathcal{H}\hat{v}^{opt}$ (see (4.19)). Since the spectral factorizations are unique up to left multiplication by a unitary $U \in \mathcal{L}(\mathcal{U})$, using (4.18), we can assume without loss of generality that

$$\mathbf{\Xi}(s) = \mathbf{\Xi}_-\left(\frac{1}{s}\right) \quad \text{for all} \quad s \in \mathbb{C}_0^+. \tag{4.32}$$

Theorems 4.3 and 4.2 show that $\Sigma^\xi = (\mathbb{T}, \Phi, \Psi^\xi, \mathbb{F}^\xi)$ and $\Sigma_-^\xi = (\mathbb{T}_-, \Phi_-, \Psi_-^\xi, \mathbb{F}_-^\xi)$. So to show that Σ_-^ξ is the reciprocal system of Σ^ξ, we need only to establish (4.31). It is clear from the proof of Theorem 4.4, particularly (4.21), that (4.31) holds if and only if for all $x \in \mathcal{X}$

$$\mathcal{H}\widehat{\Psi^\xi}x = \widehat{\Psi_-^\xi}x, \tag{4.33}$$

where $\mathcal{H}$ is defined by (4.19). We recall from (4.6) in Theorem 4.2 the formulas for the extended output maps of the spectral factor systems Σ^ξ and Σ_-^ξ, respectively:

$$\Psi^\xi = (\mathbb{F}^\xi)^{-*}(\mathbb{F}^*Q + N)\Psi, \tag{4.34}$$

$$\Psi_-^\xi = (\mathbb{F}_-^\xi)^{-*}(\mathbb{F}_-^*Q + N)\Psi_-. \tag{4.35}$$

Let $u \in \mathbf{L}_2(0,\infty;\mathcal{U})$, and define $v \in \mathbf{L}_2(0,\infty;\mathcal{U})$ by $\hat{v} = \mathcal{H}\hat{u}$. Now (4.34) holds if and only if for all $x \in \mathcal{X}$ and all $u \in \mathbf{L}_2(0,\infty;\mathcal{U})$ there holds

$$\langle \mathbb{F}^\xi u, \Psi^\xi x\rangle_2 = \langle \mathbb{F}u, Q\Psi x\rangle_2 + \langle u, N\Psi x\rangle_2.$$

As in the proof of Theorem 4.4, using Laplace transforms and the properties of the $\mathcal{H}$ operator, we obtain equivalent expressions for the left- and right-hand sides of the above expression:

$$\begin{aligned}
\text{LHS} &= \frac{1}{2\pi}\langle\widehat{\mathbb{F}^{\xi}u}, \widehat{\Psi^{\xi}x}\rangle_{\mathbf{H}_2} \\
&= \frac{1}{2\pi}\langle\Xi\hat{u}, \widehat{\Psi^{\xi}x}\rangle_{\mathbf{H}_2} \\
&= \frac{1}{2\pi}\langle\mathcal{H}(\boldsymbol{\Xi}\hat{u}), \mathcal{H}\widehat{\Psi^{\xi}x}\rangle_{\mathbf{H}_2} \quad \text{by (4.20)} \\
&= \frac{1}{2\pi}\langle\boldsymbol{\Xi}_{-}\hat{v}, \mathcal{H}\widehat{\Psi^{\xi}x}\rangle_{\mathbf{H}_2} \quad \text{by (4.19) and Lemma 3.2}
\end{aligned}$$

and

$$\begin{aligned}
\text{RHS} &= \langle\mathbb{F}u, Q\Psi x\rangle_2 + \langle u, N\Psi x\rangle_2 \\
&= \frac{1}{2\pi}\langle\widehat{\mathbb{F}u}, Q\widehat{\Psi x}\rangle_{\mathbf{H}_2} + \frac{1}{2\pi}\langle\hat{u}, N\widehat{\Psi x}\rangle_{\mathbf{H}_2} \\
&= \frac{1}{2\pi}\langle\mathbf{G}\hat{u}, Q\widehat{\Psi x}\rangle_{\mathbf{H}_2} + \frac{1}{2\pi}\langle\hat{u}, N\widehat{\Psi x}\rangle_{\mathbf{H}_2} \\
&= \frac{1}{2\pi}\langle\mathcal{H}\mathbf{G}\hat{u}, Q\mathcal{H}\widehat{\Psi x}\rangle_{\mathbf{H}_2} + \frac{1}{2\pi}\langle\mathcal{H}\hat{u}, N\mathcal{H}\widehat{\Psi x}\rangle_{\mathbf{H}_2} \quad \text{by (4.20)} \\
&= \frac{1}{2\pi}\langle\mathbf{G}_{-}\hat{v}, Q\widehat{\Psi_{-}x}\rangle_{\mathbf{H}_2} + \frac{1}{2\pi}\langle\hat{v}, N\widehat{\Psi_{-}x}\rangle_{\mathbf{H}_2} \quad \text{by (4.19) and Lemma 3.2} \\
&= \langle\mathbb{F}_{-}v, Q\Psi_{-}x\rangle_2 + \langle v, N\Psi_{-}x\rangle_2 \\
&= \langle v, (\mathbb{F}_{-}^{*}Q + N)\Psi_{-}x\rangle_2 \\
&= \langle\mathbb{F}_{-}^{\xi}v, \Psi_{-}^{\xi}x\rangle_2 \quad \text{by (4.35)} \\
&= \frac{1}{2\pi}\langle\boldsymbol{\Xi}_{-}\hat{v}, \widehat{\psi_{-}^{\xi}x}\rangle_{\mathbf{H}_2}.
\end{aligned}$$

Equating the last expressions of the left- and right-hand sides, we obtain

$$\langle\boldsymbol{\Xi}_{-}\hat{v}, \mathcal{H}(\widehat{\Psi^{\xi}x}) - \widehat{\Psi_{-}^{\xi}x}\rangle_{\mathbf{H}_2} = 0.$$

Since this holds for all $v \in \mathbf{L}_2(\mathcal{U})$ and all $x \in \mathcal{X}$ and $\boldsymbol{\Xi}_{-}$ is boundedly invertible over $\mathbf{H}_\infty(\mathcal{U})$, we have established (4.33). ∎

So under minimal conditions we have shown that the X in (4.8) is always a solution of the reciprocal Riccati equation (4.27) with bounded operators, and we have the following explicit formula for the spectral factor $\boldsymbol{\Xi}$ in the right half-plane given by

$$\boldsymbol{\Xi}(s) = D_{-}^{\xi} + C_{-}^{\xi}\left(\frac{1}{s}I - A^{-1}\right)^{-1}A^{-1}B, \tag{4.36}$$

without making any regularity assumptions. This represents a complete solution to the problem, but it is interesting to compare our solution with those in the existing literature. We recall from Theorem 4.2 that under the assumptions that $\boldsymbol{\Xi}$ is regular and its feedthrough operator is invertible, C^{ξ} satisfies (4.10) and X satisfies (4.12), whereas we have obtained a different formula (4.31) for C^{ξ} and a different Riccati equation (4.27) for X.

Next we give new explicit necessary and sufficient conditions for the spectral factor $\mathbf{\Xi}$ to be (weakly) regular.

LEMMA 4.6. *Suppose that Σ is a weakly regular linear system satisfying the conditions* (a)–(d) *and that its Popov function Π associated with the optimal control problem (Σ, J) is coercive.*

1. *The spectral factor $\mathbf{\Xi}$ is weakly regular if and only if the following limit exists in the weak topology for all $u \in \mathcal{U}$:*

$$Pu = \text{weak} \lim_{\mu \to 0} B^* A^{-*} X(\mu I - A^{-1})^{-1} A^{-1} Bu \tag{4.37}$$

$$= -\text{weak} \lim_{\lambda \to \infty} B^* A^{-*} X \lambda (\lambda I - A)^{-1} Bu. \tag{4.38}$$

In this case, the feedthrough operator of $\mathbf{\Xi}$ is given by

$$D^\xi = D^\xi_- R_-^{-1}(R_- + P + N_-(D - G(0))), \tag{4.39}$$

$$X(sI - A)^{-1}B : \mathcal{U} \to \mathcal{D}(B^*_{\Lambda_w}) \quad \text{for } s \in \rho(A),$$

$$P^* = -B^*_{\Lambda_w} X A^{-1} B, \tag{4.40}$$

and

$$X : \quad \mathcal{D}(A) \to \mathcal{D}(B^*_{\Lambda_w}). \tag{4.41}$$

2. *If Σ is regular, the limit P exists in the strong topology, and D^ξ is invertible, then Σ^ξ has the generating operators*

$$A, B, D^\xi((D^\xi)^* D^\xi)^{-1}(B^*_{\Lambda_w} X + NC), D^\xi. \tag{4.42}$$

In addition, Σ^{inv}_-, the inverse system of Σ^ξ_-, is the reciprocal system of Σ^{inv}, the inverse system of Σ^ξ; in particular, their semigroup generators are related by $A^{inv}_- = (A^{inv})^{-1}$, where

$$A^{inv} = A - B((D^\xi)^* D^\xi)^{-1}(B^*_{\Lambda_w} X + NC),$$
$$A^{inv}_- = A^{-1} - A^{-1} B R_-^{-1}(B^* A^{-*} X - N_- C A^{-1}).$$

Proof. 1. From (4.36) and (4.30) we have

$$\begin{aligned} &\mathbf{\Xi}(s) \\ &= D^\xi_- R_-^{-1} \left[R_- + (B^* A^{-*} X - N_- C A^{-1}) \left(\frac{1}{s} I - A^{-1} \right)^{-1} \right] A^{-1} B \\ &= D^\xi_- R_-^{-1} \left[R_- - (B^* A^{-*} X - N_- C A^{-1}) s (sI - A)^{-1} \right] B \\ &= D^\xi_- R_-^{-1} [R_- - B^* A^{-*} X s (sI - A)^{-1} B \\ &\quad + N_-(\mathbf{G}(s) - \mathbf{G}(0))], \end{aligned} \tag{4.43, 4.44}$$

where we have used (2.13) with $\beta = 0$. Since $\mathbf{G}$ is weakly regular, $\mathbf{\Xi}$ will be weakly regular if and only if (4.38) holds.

2. (4.39) follows by taking the weak limit in (4.36) as $s \to \infty$ and using (4.13).

3. To prove (4.40) consider for $u, v \in \mathcal{U}$

$$\begin{aligned} -\langle v, Pu \rangle &= \lim_{\lambda \to \infty} \langle v, B^* A^{-*} X \lambda (\lambda I - A)^{-1} Bu \rangle \\ &= \lim_{\lambda \to \infty} \langle B^* \lambda (\lambda I - A^*)^{-1} X A^{-1} Bv, u \rangle. \end{aligned}$$

This together with (2.19) shows that $XA^{-1}B : \mathcal{U} \to \mathcal{D}(B^*_{\Lambda_w})$ and $P^* = -B^*_{\Lambda_w} XA^{-1}B$. The resolvent identity completes the proof of (4.40).

4. To prove (4.41) we use the expression (4.9) for X and show that $(\Psi^\xi)^*\Psi^\xi$ and $\Psi^* Q \Psi$ map $\mathcal{D}(A)$ into $\mathcal{D}(B^*_{\Lambda_w})$. Now Σ^ξ is weakly regular, $\boldsymbol{\Xi} \in \mathbf{H}_\infty(\mathcal{L}(\mathcal{U}))$, and Ψ^ξ is an extended output map, so Theorem 11.1 in [22] applied to $\Psi^{new} = (\mathbb{F}^\xi)^*\Psi^\xi$ shows that $(\Psi^\xi)^*\Psi^\xi$ maps $\mathcal{D}(A)$ into $\mathcal{D}(B^*_{\Lambda_w})$. The proof for $\Psi^* Q \Psi$ is similar.

5. The statements about regularity follow as in 1. Since $\boldsymbol{\Xi}$ is regular and D^ξ is invertible, Theorem 4.3 now applies to obtain the formulas (4.42). An application of Theorem 3.3 completes the proof. □

We have the following corollary.

COROLLARY 4.7. *Under the assumptions of Theorem* 4.5 *the operators* A, B, $B^*A^{-*}XA, 0$ *are the generating operators of a well-posed linear system* Σ^1. *If* Σ *is (weakly) regular, then* Σ^1 *is (weakly) regular if and only if* Σ^ξ *is. In this case, the transfer function is* $\boldsymbol{\Xi}^1(s) = (B^*A^{-*}X)_{\Lambda_w}(sI - A)^{-1}B$. *If the degree of unboundedness of* B *is* $< \frac{1}{2}$, *then* Σ, Σ^ξ, *and* Σ^1 *are uniformly regular.*

Proof. Since CA^{-1} and C^ξ_- in (4.30) are infinite-time admissible observation operators for $T_-(\cdot)$, so is $B^*A^{-*}X$. So from Lemma 3.2 we see that $B^*A^{-*}XA$ is an infinite-time admissible observation operator for $T(\cdot)$. From (4.44) and (2.13) with $\beta = 0$ for Σ^1 we obtain

$$\boldsymbol{\Xi}(s) = D^\xi_- R_-^{-1}(R_- - (\boldsymbol{\Xi}^1(s) - \boldsymbol{\Xi}^1(0)) + N_-(\mathbf{G}(s) - \mathbf{G}(0))),$$

which shows that Σ^1 is well-posed, since Σ^ξ and Σ are well-posed. Moreover, if Σ is (weakly) regular, Σ^1 will be (weakly) regular if and only if Σ^ξ is. If the degree of unboundedness of B is $\le \frac{1}{2}$, the statement on uniform regularity follows from Lemma 2.5. □

We remark that the existence of the weak limit P is equivalent to the condition $XA^{-1}B : \mathcal{U} \to \mathcal{D}(B^*_{\Lambda_w})$. This was obtained earlier in [14] as well as formulas for $(D^\xi)^* D^\xi$. We derive similar ones in the next lemma.

LEMMA 4.8. *Suppose that* Σ *is a weakly regular linear system satisfying the conditions* (a)–(d) *and that its Popov function* Π *associated with the optimal control problem* (Σ, J) *is coercive.*

1. *If the limit* P *in* (4.38) *exists, then so does the limit*

$$V = \textit{weak} \lim_{\lambda \to \infty} B^*_{\Lambda_w} X(\lambda I - A)^{-1}B \tag{4.45}$$

and

$$(D^\xi)^* D^\xi = R + V. \tag{4.46}$$

2. *If* Σ *is regular, the limit* P *in* (4.38) *exists in the strong topology, and* D^ξ *is invertible, then the limit* V *exists in the strong topology and satisfies*

$$V = (B^*_{\Lambda_w} X)_\Lambda A^{-1}B - B^*_{\Lambda_w} XA^{-1}B.$$

Proof. 1. The implication (4.40) $\Longrightarrow$ (4.45) has been shown in Lemma 9.11.5 (e) of [9]. To prove (4.46) we denote

$$\begin{aligned} E(\lambda, \mu) &= \langle \mu(\mu I - A)^{-1}Bv, A^{-*}X\lambda(\lambda I - A)^{-1}Bu \rangle \\ &= \langle \mu(\mu I - A)^{-1}A^{-1}Bv, X\lambda(\lambda I - A)^{-1}Bu \rangle \end{aligned}$$

and calculate the following expressions:

$$\begin{aligned}\lim_{\lambda\to\infty}\lim_{\mu\to\infty} E(\lambda,\mu) &= \lim_{\lambda\to\infty}\langle A^{-1}Bv, X\lambda(\lambda I-A)^{-1}Bu\rangle\\ &= -\langle v, Pu\rangle;\end{aligned}$$

$$\begin{aligned}&\lim_{\mu\to\infty}\lim_{\lambda\to\infty} E(\lambda,\mu)\\ &= \lim_{\mu\to\infty}\lim_{\lambda\to\infty}\langle B^*\lambda(\lambda I-A^*)^{-1}X\mu(\mu I-A)^{-1}A^{-1}Bv,u\rangle\\ &= \lim_{\mu\to\infty}\langle B^*_{\Lambda_w}XA^{-1}\mu(\mu I-A)^{-1}Bv,u\rangle \text{ by (2.19) and (4.41)}\\ &= \langle (V-P^*)v,u\rangle \text{ by (4.45).}\end{aligned}$$

Similarly, if we denote

$$F(\lambda,\mu)=\langle \mu(\mu I-A)^{-1}Bv, XA^{-1}\lambda(\lambda I-A)^{-1}Bu\rangle,$$

using (4.41) and (4.40), we obtain

$$\begin{aligned}\lim_{\lambda\to\infty}\lim_{\mu\to\infty} F(\lambda,\mu) &= \lim_{\lambda\to\infty}\langle v, B^*_{\Lambda_w}XA^{-1}\lambda(\lambda I-A)^{-1}Bu\rangle \text{ by (4.41)}\\ &= \langle v,(V-P^*)u\rangle \quad \text{by (4.45),}\end{aligned}$$

$$\begin{aligned}\lim_{\mu\to\infty}\lim_{\lambda\to\infty} F(\lambda,\mu) &= \lim_{\mu\to\infty}\langle v, B^*\mu(\mu I-A^*)^{-1}XA^{-1}Bu\rangle\\ &= \langle v, B^*_{\Lambda_w}XA^{-1}Bu\rangle \text{ by (2.19) and (4.40)}\\ &= -\langle Pv,u\rangle \text{ by (4.40).}\end{aligned}$$

Further, denoting

$$Q(\lambda,\mu)=\langle \mu(\mu I-A)^{-1}Bv, A^{-*}C^*QCA^{-1}\lambda(\lambda I-A)^{-1}Bu\rangle,$$

using (2.19), $G_0=-C_\Lambda A^{-1}B$, and the regularity of Σ, we obtain

$$\lim_{\mu\to\infty}\lim_{\lambda\to\infty} Q(\lambda,\mu)=\lim_{\lambda\to\infty}\lim_{\mu\to\infty} Q(\lambda,\mu)=\langle G_0v, QG_0u\rangle.$$

Finally, denoting

$$R(\lambda,\mu)=\langle \mu(\mu I-A)^{-1}Bv, (C^\xi_-)^*C^\xi_-\lambda(\lambda I-A)^{-1}Bu\rangle,$$

using (4.31), (4.38), and the regularity of Σ, we obtain

$$\begin{aligned}\lim_{\mu\to\infty}\lim_{\lambda\to\infty} R(\lambda,\mu) &= \lim_{\lambda\to\infty}\lim_{\mu\to\infty} R(\lambda,\mu)\\ &= \langle (P-N_-G_0)v, R_-^{-1}(P-N_-G_0)u\rangle.\end{aligned}$$

Next we deduce a useful identity from (4.27) by applying the following operation for arbitrary $u,v\in\mathcal{U}$:

$$\langle \mu(\mu I-A)^{-1}Bv, (4.27)\ \lambda(\lambda I-A)^{-1}Bu\rangle,$$

which yields

$$E(\lambda,\mu)+F(\lambda,\mu)+Q(\lambda,\mu)=R(\lambda,\mu). \tag{4.47}$$

Using our calculations above for the limits as λ and $\mu \to \infty$ and denoting $M = P - N_- G_0$, we obtain

$$-P - P^* + V + G_0^* Q G_0 = M^* R_-^{-1} M. \tag{4.48}$$

Finally, using (4.39), we calculate

$$\begin{aligned}(D^\xi)^* D^\xi &= R_- + M^* + M + M^* R_-^{-1} M \\ &= R_- + P^* - G_0^* N_-^* + P - N_- G_0 - P - P^* + V \\ &\quad + G_0^* Q G_0 \qquad \text{from (4.48)} \\ &= R_- - G_0^* Q G_0 - G_0^* N^* - N G_0 + V \\ &= R + V \quad \text{by (4.13)}.\end{aligned}$$

2. From part 2 of Lemma 4.6, Σ^ξ is regular with generating operators given by (4.42). So $A, B, B_{\Lambda_w}^* X, 0$ are the generating operators of a regular linear system Σ_2 with transfer function $\mathbf{G}_2$. From (2.13) we have

$$\begin{aligned}\mathbf{G}_2(\lambda) - \mathbf{G}_2(0) &= \lambda B_{\Lambda_w}^* X (\lambda I - A)^{-1} A^{-1} B \\ &= B_{\Lambda_w}^* X A^{-1} B + B_{\Lambda_w}^* X (\lambda I - A)^{-1} B,\end{aligned}$$

where we have used (4.40). Now the claim in (2) follows, since $\mathbf{G}_2$ is regular, and $\mathbf{G}_2(0) = -(B_{\Lambda_w}^* X)_\Lambda A^{-1} B$.

Note that $V = 0$ if and only if

$$\lim_{\lambda \to \infty} \lim_{\mu \to \infty} F(\lambda, \mu) = \lim_{\mu \to \infty} \lim_{\lambda \to \infty} F(\lambda, \mu)$$

or a similar statement for $E(\lambda, \mu)$ holds. □

Of course it is difficult to check whether the limit P exists and whether D^ξ is invertible. D^ξ is always left invertible, and so if $\mathcal{U}$ is finite-dimensional, it will be invertible, and X satisfies the Riccati equation (4.12) (as already remarked in [22]). However, if the degree of unboundedness of B is $< \frac{1}{2}$, then from Lemma 2.5 we can conclude that $\mathbf{\Xi}$ will be uniformly regular and D^ξ will be invertible even for infinite-dimensional $\mathcal{U}$ (see [22, Proposition 12.3]).

COROLLARY 4.9. *Suppose that B and C are infinite-time admissible operators for $T(\cdot)$ and the degree of unboundedness of B is $< \frac{1}{2}$. Then $A, B, C, 0$ are the generating operators of a uniformly regular linear system Σ. Suppose that Σ satisfies the conditions* (a)–(c) *and its Popov function Π associated with the optimal control problem (Σ, J) is coercive. Then $\mathbf{\Xi}$ is uniformly regular, D^ξ is invertible, and X from* (4.7) *satisfies the Riccati equation* (4.12) *with $(D^\xi)^* D^\xi = R$.*

Acknowledgment. I would like to thank Kalle Mikkola for his useful comments on comparisons of my approach with the existing literature.

REFERENCES

[1] F. M. CALLIER AND P. GRABOWSKI, *On the Circle Criterion for Boundary Control Systems in Factor Form: Lyapunov Approach*, Report 2000/07, Department of Mathematics, University of Namur, Belgium, 2000.

[2] F. M. CALLIER AND P. GRABOWSKI, *On the Circle Criterion for Boundary Control Systems in Factor Form: Lyapunov Stability and Lur'e Equations*, Report 2002/05, Department of Mathematics, University of Namur, Belgium, 2002.

[3] R. F. Curtain, *Reciprocals of well-posed linear systems: A survey*, in Proceedings of the 15th International Symposium on Mathematical Theory of Networks and Systems, University of Notre Dame, South Bend, IN, 2002, CD-ROM; also available online at http://www.nd.edu/mtns.

[4] R. F. Curtain, *Regular linear systems and their reciprocals: Applications to Riccati equations*, Systems Control Lett., 49 (2003), pp. 81–89.

[5] P. Grabowski, *On the spectral-Lyapunov approach to parametric optimization of distributed parameter systems*, IMA J. Math. Control Inform., 7 (1990), pp. 317–338.

[6] S. Hansen and G. Weiss, *New results on the operator Carleson measure criterion*, IMA J. Math. Control Inform., 14 (1997), pp. 3–32.

[7] I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations: Continuous and Approximation Theories, Part* 1*, Abstract Parabolic systems*, Cambridge University Press, Cambridge, UK, 2000.

[8] M. S. Livsic, *Operators, Oscillations, Waves (Open Systems)*, Transl. Math. Monogr., 34, AMS, Providence, RI, 1973.

[9] K. Mikkola, *Infinite-Dimensional Linear Systems: Optimal Control and Riccati Equations*, Ph.D. thesis, Institute of Mathematics, Helsinki University of Technology, Helsinki, Finland, 2002, available online at http://www.math.hut.fi/reports.

[10] M. Rosenblum and J. Rovnyak, *Hardy Classes and Operator Theory*, Oxford University Press, New York, 1985.

[11] O. J. Staffans, *Quadratic optimal control of stable well-posed linear systems*, Trans. Amer. Math. Soc., 349 (1997), pp. 3679–3716.

[12] O. J. Staffans, *Coprime factorizations and well-posed linear systems*, SIAM J. Control Optim., 36 (1998), pp. 1268–1292.

[13] O. J. Staffans, *Feedback representations of critical controls for well-posed linear systems*, Internat. J. Robust Nonlinear Control, 8 (1998), pp. 1189–1217.

[14] O. J. Staffans, *Quadratic optimal control of well-posed linear systems*, SIAM J. Control Optim., 37 (1998), pp. 131–164.

[15] O. J. Staffans and G. Weiss, *Transfer functions of regular linear systems, part* III: *Inversions and duality*, J. Integral Equations and Operator Theory, to appear.

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

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

[18] G. Weiss, *The representation of regular linear systems on Hilbert spaces*, in Control and Estimation of Distributed Parameter Systems, W. Schappacher, F. Kappel, and K. Kunisch, eds., Birkhäuser Verlag, Basel, 1989, pp. 401–416.

[19] G. Weiss, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.

[20] G. Weiss, *Transfer functions of regular linear systems, part* 1: *Characterizations of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.

[21] G. Weiss and R. F. Curtain, *Exponential stabilization of vibrating systems by collocated feedback*, in Proceedings of the 7th Mediterranean Conference on Control and Automation, Haifa, Israel, 1999, CD-ROM, IEEE, Piscataway, NJ, 1999.

[22] M. Weiss and G. Weiss, *Optimal control of stable weakly regular linear systems*, Math. Control Signals Systems, 10 (1997), pp. 287–330.

[23] G. Weiss and H. J. Zwart, *An example in linear quadratic optimal control*, Systems Control Lett., 33 (1998), pp. 339–349.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1703–1726

A SEMIMARTINGALE BACKWARD EQUATION AND THE VARIANCE-OPTIMAL MARTINGALE MEASURE UNDER GENERAL INFORMATION FLOW*

MICHAEL MANIA† AND REVAZ TEVZADZE‡

Abstract. We consider a financial market model, where the dynamics of asset prices are given by an R^d-valued continuous semimartingale and the information flow is right-continuous. Using the dynamic programming approach we express the variance-optimal martingale measure in terms of the value process of a suitable optimization problem and show that this value process uniquely solves the corresponding semimartingale backward equation. We consider two extreme cases when this equation admits an explicit solution. In particular, we give necessary and sufficient conditions in order that the variance-optimal martingale measure coincides with the minimal martingale measure as well as with the martingale measure appearing in the second extreme case.

Key words. backward semimartingale equation, variance-optimal martingale measure, incomplete markets, contingent claim pricing

AMS subject classifications. 90A09, 60H30, 90C39

DOI. 10.1137/S036301290240628X

1. Introduction.

Let $X = (X_t, t \in [0,T])$ be an R^d-valued semimartingale defined on a filtered probability space $(\Omega, \mathcal{F}, F = (F_t, t \in [0,T]), P)$, and let η be an F_T-measurable random variable. The components of X_t are interpreted as discounted prices of d risky assets in a financial market and η is a contingent claim. Denote by $\mathcal{M}^e$ the set of martingale measures of X, i.e., probability measures Q that are equivalent to P such that X is a local martingale under Q. Let $(Z_t(Q), t \in [0,T])$ be the density process of Q with respect to P.

It is well known that the prices of contingent claims can usually be computed as expectations under a suitable martingale measure. The choice of the pricing (martingale) measure depends on the context, e.g., on the criterion relative to which the quality of the hedging strategy is measured. It was shown by Schweizer [23] that if the quadratic criterion to measure the hedging error is used, then the variance minimizing hedging price c of the claim η is equal to $E^*\eta$, where the expectation is calculated with respect to the variance-optimal martingale measure. The variance-optimal martingale measure also plays a crucial role in determining the optimal mean-variance hedging strategy (see Pham, Rheinländer, and Schweizer [19], Delbaen et al. [7], Gouriéroux, Laurent, and Pham [11], and Rheinländer [20]).

The variance-optimal martingale measure is defined by the property that its density with respect to P has minimal L^2 norm among all signed martingale measures (see Schweizer [23] for a precise definition of the last term).

In general, the variance-optimal martingale measure is not a positive measure. Therefore, in some situations the price of a nonnegative contingent claim $E^*\eta$ calcu-

*Received by the editors April 24, 2002; accepted for publication (in revised form) May 5, 2003; published electronically November 14, 2003. Research for this work was supported by grants INTAS 97-30204 and INTAS 99-00559.

http://www.siam.org/journals/sicon/42-5/40628.html

†A. Razmadze Mathematical Institute, Georgian Academy of Sciences, Tbilisi 0193, Georgia (mania@rmi.acnet.ge).

‡Institute of Cybernetics, Georgian Academy of Sciences, Tbilisi 0186, Georgia (reztev@yahoo.com).

lated by the variance-optimal martingale measure is negative, and a signed measure Q^* is not always attractive for the characterization of the price system. However, if

(i) the discounted asset price process X is continuous and

(ii) the set $\mathcal{M}_2^e = \{Q \in \mathcal{M}^e : E(Z_T^Q)^2 < \infty\}$ is not empty,

then, as was shown by Delbaen and Schachermayer [5], the variance-optimal martingale measure Q^* is positive and equivalent to P and may be used as a pricing measure. Therefore, assuming (i) and (ii), we may define the variance-optimal martingale measure as a solution of the optimization problem

$$EZ_T^2(Q) \to \inf_{Q \in \mathcal{M}_2^e}, \tag{1.1}$$

and we shall focus our attention on the study of the structure of the optimal martingale measure, using the dynamic programming approach. The dynamic programming method to determine the variance-optimal martingale measure was used by Laurent and Pham [14] for a multidimensional diffusion model with Brownian filtration. They characterize the variance-optimal martingale measure in terms of the value function associated to the problem (1.1), and they gave explicit expressions for this value function under special conditions on the model coefficients.

Our aim is to give a construction of the variance-optimal martingale measure when the dynamics of the discounted asset price process is governed by a continuous semimartingale and the information flow is right-continuous, rejecting the condition of continuity of the filtration imposed in our previous paper [17]. We obtain a description of the variance-optimal martingale measure in terms of a value process corresponding to the problem (1.1) and show that this value function uniquely solves the corresponding semimartingale Bellman equation. Such a type of backward equations was introduced by Chitashvili [3] (see also [4], [16]) as a stochastic version of the Bellman equation for an optimal control problem. We consider two extreme cases, including two specific cases studied by Pham, Rheinländer, and Schweizer [19], Laurent and Pham [14], and Biagini, Guasoni, and Pratelli [2], when this equation admits an explicit solution. In particular, we give necessary and sufficient conditions when the variance-optimal martingale measure coincides with the minimal martingale measure as well as with the martingale measure appearing in the second extreme case.

2. Basic assumptions, definitions, and some auxiliary facts. Let $(\Omega, \mathcal{F}, F = (F_t, t \in [0,T]), P)$ be a filtered probability space satisfying the usual conditions of right-continuity and completeness, where $T < \infty$ is a fixed time horizon and $\mathcal{F} = F_T$.

1. *The space BMO and exponential martingales.* We recall some definitions and known results on the relation of BMO and exponential martingales used in what follows.

DEFINITION 1. *A uniformly integrable martingale $M = (M_t, t \in [0,T])$ belongs to the class BMO_2 if there is a constant $C > 0$ such that*

$$E((M_T - M_{\tau-})^2 | F_\tau) \le C, \quad P\text{-a.s.}$$

for every stopping time τ. The smallest constant with this property (or $+\infty$ if it does not exist) is called the BMO_2 norm of M and is denoted by $||M||_{BMO_2}$.

Remark. If $M \in \mathcal{M}^2_{\text{loc}}$, then $M \in BMO_2$ if and only if M is of bounded jumps and for a constant C

$$E(\langle M \rangle_T - \langle M \rangle_\tau | F_\tau) \le C, \quad P\text{-a.s.}$$

for every stopping time τ (see Dellacherie and Meyer [6]).

Denote by $\mathcal{E}(X)$ the Doléans exponential of a semimartingale X, which is the unique solution of the linear SDE

$$Z_t = 1 + \int_0^t Z_{s-} dX_s. \tag{2.1}$$

Moreover, any solution of this equation coincides with $\mathcal{E}(X)$ on the set $\{(\omega, t) : \mathcal{E}_t(X) \neq 0\}$. Therefore, any strictly positive martingale Z with $Z_0 = 1$ is represented as a Doléans exponential $\mathcal{E}(X)$, with $X = \frac{1}{Z_-} \cdot Z$. We use the notation $\psi \cdot X$ for the stochastic integral with respect to the semimartingale X.

PROPOSITION 1 (Kazamaki [13]). *If M is a right-continuous BMO_2-martingale satisfying*

$$\Delta M \geq -1 + h$$

for some $h > 0$, then $\mathcal{E}(M)$ is a uniformly integrable martingale.

Let Z be a positive process that satisfies condition (J): There exists a constant C such that

$$\frac{1}{C} Z_- < Z < C Z_-.$$

DEFINITION 2. *A strictly positive adapted process Z satisfies the reverse Hölder inequality $R_p(P)$, where $1 < p < \infty$, if there exists a constant C such that*

$$E\left(\left(\frac{Z_T}{Z_\tau} \right)^p \middle/ F_\tau \right) \leq C, \quad P\text{-a.s.}$$

for every stopping time τ.

The following assertion relates BMO and the reverse Hölder condition.

PROPOSITION 2 (Doléans-Dade and Meyer [8]). *Let M be a local martingale and $\mathcal{E}(M)$ its Doléans exponential. The following assertions are equivalent:*

(i) *M belongs to the class BMO_2 and $\Delta M \geq -1 + h$ for some $h > 0$.*

(ii) *$\mathcal{E}(M)$ is a uniformly integrable martingale satisfying condition* (J) *and the reverse Hölder inequality $R_p(P)$ for some $p > 1$.*

2. *Martingale measures.* Let the discounted asset price process X admit a decomposition

$$X_t = X_0 + \Lambda_t + M_t, \quad \Lambda \in \mathcal{A}_{\mathrm{loc}}, M \in \mathcal{M}^2_{\mathrm{loc}}. \tag{2.2}$$

X satisfies the structure condition if there is a predictable R^d-valued process λ such that

$$X_t = X_0 + \int_0^t d\langle M \rangle_s \lambda_s + M_t, \quad P\text{-a.s. for all } t \in [0, T], \text{ and} \tag{2.3}$$

$$K_t = \langle \lambda \cdot M \rangle_t < \infty, \quad P\text{-a.s. for all } t \in [0, T].$$

The process K is called the mean-variance tradeoff process of X (see Schweizer [22]).

Denote by $\mathcal{M}^e$ the set of measures Q equivalent to P on F such that X is a local martingale under Q. Let Z_t^Q be the density process of Q with respect to the basic measure P which can be expressed as a Doléans exponential $Z_t^Q = \mathcal{E}_t(M^Q)$ of a local

martingale M^Q. Here and in what follows we identify the measure Q with Z^Q or with M^Q.

If the local martingale $\hat{Z}_t = \mathcal{E}_t(-\lambda \cdot M)$ is a true martingale, $d\hat{P}/dP = \hat{Z}_T$ defines an equivalent probability measure called the minimal martingale measure for X.

Throughout the paper we assume that the following conditions are satisfied.

(A) The discounted asset price process X is a continuous R^d-valued semimartingale.

(B) The minimal martingale measure exists and $E\mathcal{E}_T^2(-\lambda \cdot M) < \infty$.

We shall often use the following stronger condition.

(B*) The minimal martingale measure exists and satisfies the reverse Hölder condition $R_2(P)$.

Remark 1. (A) and (B) imply that the semimartingale X satisfies the structure condition; hence X admits the decomposition (2.3), where M is a continuous local martingale.

Remark 2. Since for any $Q \in \mathcal{M}_2^e$ the density process Z^Q is a strictly positive square integrable martingale, the process $M^Q = \frac{1}{Z_-^Q} \cdot Z^Q$ belongs to $\mathcal{M}_{\text{loc}}^2$; hence the square bracket $\langle M^Q \rangle$ exists for any $Q \in \mathcal{M}_2^e$.

Since X is assumed to be continuous, any element Q of $\mathcal{M}^e$ is given by the density Z_t^Q of the form

$$\mathcal{E}_t(-\lambda \cdot M + N), \tag{2.4}$$

where N is a local martingale strongly orthogonal to M.

Denote by $\mathcal{M}_2^e$ the subset of $\mathcal{M}^e$ with square integrable densities, i.e.,

$$\mathcal{M}_2^e = \left\{ Q \sim P : \frac{dQ}{dP} \in L^2(P), X \text{ is a } Q\text{-local martingale} \right\}.$$

Denote by $\mathcal{N}(X)$ the class of local martingales N strongly orthogonal to M such that the process $(\mathcal{E}_t(-\lambda \cdot M + N), t \in [0,T])$ is a martingale under P.

Let $\mathcal{N}_2(X)$ be the subclass of $\mathcal{N}(X)$ of local martingales N such that the process $(\mathcal{E}_t(-\lambda \cdot M + N), t \in [0,T])$ is a strictly positive square integrable martingale under P. Then

$$\mathcal{M}_2^e = \left\{ Q \sim P : \frac{dQ}{dP} = \mathcal{E}_T(-\lambda \cdot M + N), N \in \mathcal{N}_2(X) \right\}. \tag{2.5}$$

3. *The optimality principle.* Consider the value process V_t,

$$V_t = \operatorname*{ess\,inf}_{Q \in \mathcal{M}_2^e} E(\mathcal{E}_{tT}^2(M^Q)/F_t) = \operatorname*{ess\,inf}_{N \in \mathcal{N}_2(X)} E(\mathcal{E}_{tT}^2(-\lambda \cdot M + N)/F_t), \tag{2.6}$$

where

$$\mathcal{E}_{\tau T}(M^Q) = \frac{\mathcal{E}_T(M^Q)}{\mathcal{E}_\tau(M^Q)} = \mathcal{E}_T(M^Q - M^Q_{\cdot \wedge \tau})$$

for any $Q \in \mathcal{M}^e$ and any stopping time τ. We shall use also the notation

$$\langle M \rangle_{tT} = \langle M \rangle_T - \langle M \rangle_t$$

for the square bracket of the martingale M.

The following assertion is proved in a standard manner (see, e.g., [9],[10]).

PROPOSITION 3. (a) *There exists a right-continuous with left limits (RCLL) semimartingale still denoted by* V_t *such that for each* $t \in [0,T]$

$$V_t = \operatorname*{ess\,inf}_{Q \in \mathcal{M}_2^e} E(\mathcal{E}_{tT}^2(M^Q)/F_t).$$

V_t *is the largest RCLL process equal to* 1 *at time* T *such that* $V_t\mathcal{E}_t^2(M^Q)$ *is a submartingale for every* $Q \in \mathcal{M}_2^e$.

(b) *The following properties are equivalent:*

(i) Q^* *is variance-optimal, i.e.,* $V_0 = \inf_{Q \in \mathcal{M}_2^e} E\mathcal{E}_T^2(M^Q) = E\mathcal{E}_T^2(M^{Q^*})$.

(ii) Q^* *is variance-optimal for all conditional criteria; i.e., for all* $t \in [0,T]$

$$V_t = E(\mathcal{E}_{tT}^2(M^{Q^*})/F_t) \quad a.s.$$

(iii) $V_t\mathcal{E}_t^2(M^{Q^*})$ *is a P-martingale.*

We recall that any local martingale m admits a decomposition (see, e.g., Jacod [12] or Liptser and Shiryayev [15])

$$m = m^c + m^d = m^c + m^{dq} + m^{dp}, \tag{2.7}$$

where

- m^c is a continuous local martingale,
- m^{dq} is compensated sum of totally inaccessible jumps of m (m^{dq} is quasi-left-continuous),
- m^{dp} is compensated sum of predictable jumps of m, and
- $m^d = m^{dq} + m^{dp}$ is orthogonal to any continuous martingale and is called the purely discontinuous part of m.

Let Y be a special semimartingale with the decomposition

$$Y_t = Y_0 + B_t + L_t, \quad B \in \mathcal{A}_{\mathrm{loc}} \cap \mathcal{P}, L \in \mathcal{M}_{\mathrm{loc}}, \tag{2.8}$$

and let

$$L_t = \int_0^t \psi_s dM_s + \bar{L}_s, \quad \langle \bar{L}, M \rangle = 0, \tag{2.9}$$

be the Galtchouk–Kunita–Watanabe (GKW) decomposition of L with respect to the martingale M, and $\mathcal{P}$ is the class of predictable processes.

Equations (2.7)–(2.9) imply that any special semimartingale Y admits a decomposition

$$Y_t = Y_0 + B_t + \int_0^t \psi_s dM_s + \bar{L}_t^c + \bar{L}_t^d, \quad \langle \bar{L}^c, M \rangle = 0. \tag{2.10}$$

Note that, since M is continuous, we have that

$$L_t^c = \int_0^t \psi_s dM_s + \bar{L}_t^c, \quad \bar{L}_t^d = L_t^d, \quad \Delta\bar{L}_t = \Delta L_t. \tag{2.11}$$

Finally, we recall that the process X is said to belong to the class D if the random variables $X_\tau I_{(\tau \le T)}$ for all stopping times τ are uniformly integrable.

3. Backward semimartingale equation for the value process. Let us consider the optimization problem

$$E\mathcal{E}_T^2(M^Q) \to \inf_{Q\in\mathcal{M}_2^e}. \tag{3.1}$$

Conditions (A) and (B) imply that the value process of problem (3.1) defined by (2.6) is a special semimartingale with respect to the measure P, since the process $V_t\mathcal{E}_t^2(-\lambda\cdot M)$ is a P-submartingale and the process $\mathcal{E}_t^{-2}(\lambda\cdot M)$ is locally bounded. Let

$$V_t = m_t + A_t,\ \ m\in M_{\mathrm{loc}},\ \ A\in\mathcal{A}_{\mathrm{loc}}, \tag{3.2}$$

be the canonical decomposition of V, and let

$$m_t = \int_0^t \varphi_s dM_s + \bar{m}_s,\quad \langle\bar{m}, M\rangle = 0, \tag{3.3}$$

be the GKW decomposition of m with respect to M.

We say that the process B strongly dominates the process A and write $A\prec B$ if the difference $B-A\in\mathcal{A}^+_{\mathrm{loc}}$, i.e., is a locally integrable increasing process.

Let $(A^Q, Q\in\mathcal{Q})$ be a family of processes of finite variations, zero at time zero. Denote by $\operatorname{ess\,inf}_{Q\in\mathcal{Q}}(A^Q)$ the largest process of finite variation, zero at time zero, which is strongly dominated by the process $(A_t^Q, t\in[0,T])$ for every $Q\in\mathcal{Q}$; i.e., this is an "ess inf" of the family $(A^Q, Q\in\mathcal{Q})$ with respect to the strong order $\prec$.

Throughout the paper the symbols "ess inf" are used in the sense of the strong order.

Denote by $\tilde{A}$ the compensator (or the dual predictable projection) of A.

THEOREM 1. *Let conditions* (A) *and* (B) *be satisfied. Then, the value process V is a solution of the semimartingale backward equation*

$$Y_t = Y_0 - \operatorname*{ess\,inf}_{Q\in\mathcal{M}_2^e}\ \Big[\int_0^t Y_s d\langle M^{Q,c}\rangle_s + 2\langle M^{Q,c}, L^c\rangle_t$$

$$+\sum_{s\le t}(Y_s\widetilde{(\Delta M_s^Q)^2} + 2\Delta L_s\Delta M_s^Q)\Big] + L_t \tag{3.4}$$

with the boundary condition

$$Y_T = 1. \tag{3.5}$$

Proof. Let us first show that the process $[V, 2M^Q + [M^Q]]$ or, equivalently, the process $\sum_{s\le t}(2\Delta V_s\Delta M_s^Q + \Delta V_s(\Delta M_s^Q)^2)$ is P-locally integrable for any $Q\in\mathcal{M}_2^e$. By the Itô formula

$$\mathcal{E}_t^2(M^Q) = 1 + \int_0^t \mathcal{E}_{s-}^2(M^Q) d(2M_s^Q + [M^Q]_s), \tag{3.6}$$

and hence $2M_t^Q + [M^Q]_t = \int_0^t \mathcal{E}_{s-}^{-2}(M^Q)d\mathcal{E}_s^2(M^Q)$. Therefore, since $\mathcal{E}_{t-}^{-2}(M^Q)$ is locally bounded, it is sufficient to show that the process $[V, \mathcal{E}^2(M^Q)]_t$ is locally integrable. However, this is clearly true since V_t, as mentioned above, is a special semimartingale and $\mathcal{E}_t^2(M^Q)V_t$ is a submartingale and thus also a special semimartingale.

Thus the process

$$[V, 2M^Q + [M^Q]]_t = 2\langle m, M^{Q,c}\rangle_t + \sum_{s\le t}(2\Delta V_s \Delta M_s^Q + \Delta V_s(\Delta M_s^Q)^2) \tag{3.7}$$

is locally P-integrable for any $Q \in \mathcal{M}_2^e$; hence there exists the dual predictable projection of this process, and since $\sum_{s\le t}\Delta A_s \Delta M_s^Q$ is a local martingale, we have that

$$[V, 2\widetilde{M^Q + [M^Q]}]_t = 2\langle m, M^{Q,c}\rangle_t + \sum_{s\le t}(2\Delta m_s \Delta \widetilde{M_s^Q} + \Delta V_s(\Delta M_s^Q)^2). \tag{3.8}$$

Using (2.1), (3.6), the decomposition of V, and the Itô formula for $\mathcal{E}_t^2(M^Q)V_t$, we have

$$\mathcal{E}_t^2(M^Q)V_t = V_0 + \int_0^t \mathcal{E}_{s-}^2(M^Q)dV_s + \int_0^t V_{s-}\mathcal{E}_{s-}^2(M^Q)d(2M_s^Q + [M^Q]_s)$$

$$+ \int_0^t \mathcal{E}_{s-}^2(M^Q)d[V, 2M^Q + [M^Q]]_s = V_0 + \int_0^t \mathcal{E}_{s-}^2(M^Q)d(m_s + 2(V_- \cdot M^Q)_s)$$

$$+ \int_0^t \mathcal{E}_{s-}^2(M^Q)d\big(A_s + (V_- \cdot [M^Q])_s + 2[M^Q, V]_s] + [V, [M^Q]]_s\big). \tag{3.9}$$

Since the processes

$$[M^Q]_t - \langle M^{Q,c}\rangle_t - \sum_{s\le t}\widetilde{(\Delta M_s^Q)^2}$$

and

$$[V, 2M^Q + [M^Q]]_t - 2\langle m^c, M^{Q,c}\rangle_t - \sum_{s\le t}(\Delta V_s \widetilde{(\Delta M_s^Q)^2} + 2\Delta m_s \Delta M_s^Q)$$

are local martingales and $V_t\mathcal{E}_t^2(M^Q)$ is a submartingale, from (3.9) and from the uniqueness of the canonical decomposition of submartingales we obtain that

$$\int_0^t \mathcal{E}_{u-}^2(M^Q)d\Bigg[A_u + 2\langle m^c, M^{Q,c}\rangle_u + \int_0^t V_{s-}d\langle M^{Q,c}\rangle_s$$

$$+ \sum_{s\le t}(V_s \widetilde{(\Delta M_s^Q)^2} + 2\Delta m_s \Delta M_s^Q)\Bigg]$$

is an integrable increasing process. Therefore, since $\mathcal{E}_{t-}(M^Q)$ is strictly positive, we obtain that

$$A_t + 2\langle m^c, M^{Q,c}\rangle_t + \int_0^t V_{s-}d\langle M^{Q,c}\rangle_s \tag{3.10}$$

$$+ \sum_{s\le t}(V_s \widetilde{(\Delta M_s^Q)^2} + 2\Delta m_s \Delta M_s^Q) \in \mathcal{A}_{\mathrm{loc}}^+$$

for every $Q \in \mathcal{M}_2^e$. According to Delbaen and Schachermayer [5], there exists an optimal martingale measure Q^* which is equivalent to P. Therefore, by the optimality principle (Proposition 3) $V_t\mathcal{E}_t^2(M^{Q^*})$ is a P-martingale, and using the same arguments (as above) we have that

$$A_t + 2\langle m^c, M^{Q^*,c}\rangle_t + \int_0^t V_{s-} d\langle M^{Q^*,c}\rangle_s \tag{3.11}$$

$$+\sum_{s\le t}(V_s(\Delta M_s^{Q^*})\widetilde{^2} + 2\Delta m_s \Delta M_s^{Q^*}) = 0.$$

Hence relations (3.10) and (3.11) imply that

$$A_t = -\operatorname*{ess\,inf}_{Q\in\mathcal{M}_2^e}\Big[2\langle m^c, M^{Q,c}\rangle_t + \int_0^t V_{s-} d\langle M^{Q,c}\rangle_s \tag{3.12}$$

$$+\sum_{s\le t}(V_s(\Delta M_s^{Q})\widetilde{^2} + 2\Delta m_s \Delta M_s^{Q})\Big]$$

and V satisfies (3.4). □

COROLLARY 1. *Suppose that there exists $\hat{Q} \in \mathcal{M}_2^e$ with continuous density process $Z^{\hat{Q}}$. Then a.s. for all $t \in [0,T]$*

$$\Delta L_t \le \Delta Y_t$$

for any solution Y of (3.4)*, where L is the martingale part of Y.*

Proof. Since Y solves (3.4) we have for any $Q \in \mathcal{M}_2^e$ that

$$B_t + 2\langle L^c, M^{Q,c}\rangle_t + \int_0^t Y_{s-} d\langle M^{Q,c}\rangle_s \tag{3.13}$$

$$+\sum_{s\le t}(Y_s(\Delta M_s^{Q})\widetilde{^2} + 2\Delta L_s \Delta M_s^{Q}) \in \mathcal{A}_{\mathrm{loc}}^+.$$

Taking $Q = \hat{Q}$ (3.13) implies that $\Delta B_t \ge 0$ a.s. for all $t \in [0,T]$; hence $\Delta Y_t = \Delta B_t + \Delta L_t \ge \Delta L_t$. □

Remark. If there exists an equivalent local martingale measure $\tilde{Q}$ which satisfies the reverse Hölder inequality $R_2(P)$, then the variance-optimal martingale measure satisfies the reverse Hölder inequality $R_2(P)$ since for any stopping time τ

$$E(\mathcal{E}_{\tau,T}^2(M^{Q^*})/F_\tau) = \operatorname*{ess\,inf}_{Q\in\mathcal{M}_2^e} E(\mathcal{E}_{\tau,T}^2(M^{Q})/F_\tau) \le E(\mathcal{E}_{\tau,T}^2(M^{\tilde{Q}})/F_\tau) \le C.$$

LEMMA 1. *Assume that there is an equivalent local martingale measure $Q \in \mathcal{M}_2^e$ such that the associated local martingale M^Q belongs to BMO_2; then the martingale part L of any bounded solution Y of* (3.4) *belongs to the class BMO_2.*

Proof. Since $|Y_t| \le C$, we have that $|\Delta Y_t| \le 2C$ and $|\Delta L_t| \le 2C$ (see, e.g., Proposition 2.14 from Jacod [12]). Therefore, $[Y] \in \mathcal{A}_{\mathrm{loc}}^+$, $L \in \mathcal{M}_{\mathrm{loc}}^2$, and without loss

of generality we may assume that $[Y]$ is integrable and $L \in \mathcal{M}^2$; otherwise one can use the localization arguments.

By the Itô formula

$$Y_t^2 = Y_0^2 + 2\int_0^t Y_{s-}dY_s + [Y]_t, \tag{3.14}$$

and from the boundary condition $Y_T = 1$ we have

$$[Y]_T - [Y]_\tau + 2\int_\tau^T Y_{s-}d(B_s + L_s) = 1 - Y_\tau^2 \le 1 \tag{3.15}$$

for any stopping time τ.

Since Y satisfies (3.4), relation (3.13) implies that

$$B_t + \int_0^t \widetilde{Y_s d[M^Q]_s} + 2\langle M^Q, L\rangle_t \in \mathcal{A}^+_{\mathrm{loc}}. \tag{3.16}$$

On the other hand, $C[M^Q]_t - \int_0^t Y_s d[M^Q]_s \in \mathcal{A}^+_{\mathrm{loc}}$ by boundedness of Y, and hence

$$C\langle M^Q\rangle_t - \int_0^t \widetilde{Y_s d[M^Q]_s} \in \mathcal{A}^+_{\mathrm{loc}}. \tag{3.17}$$

Therefore,

$$B_t + C\langle M^Q\rangle_t + 2\langle M^Q, L\rangle_t \tag{3.18}$$

is also a locally integrable increasing process and (3.15) implies that

$$[Y]_T - [Y]_\tau + 2\int_\tau^T Y_{s-}dL_s - 2C\int_\tau^T Y_{s-}d\langle M^Q\rangle_s \tag{3.19}$$

$$- 4\int_\tau^T Y_{s-}d\langle M^Q, L\rangle_s \le 1.$$

Taking conditional expectations in (3.19), having inequality $Y_{t-} \le C$ in mind, we have

$$E([Y]_T - [Y]_\tau|F_\tau) \le 1 + 2C^2E(\langle M^Q\rangle_T - \langle M^Q\rangle_\tau/F_\tau)$$

$$+ 4CE\left(\int_\tau^T |d\langle M^Q, L\rangle_s|/F_\tau\right). \tag{3.20}$$

Since $E(\langle L\rangle_T - \langle L\rangle_\tau|F_\tau) \le E([Y]_T - [Y]_\tau|F_\tau)$, (3.20) and the Kunita–Watanabe inequality imply that

$$E(\langle L\rangle_T - \langle L\rangle_\tau/F_\tau) \le 1 + 2C^2E(\langle M^Q\rangle_T - \langle M^Q\rangle_\tau/F_\tau)$$

$$+ 4CE^{1/2}(\langle M^Q\rangle_T - \langle M^Q\rangle_\tau/F_\tau)E^{1/2}(\langle L\rangle_T - \langle L\rangle_\tau/F_\tau).$$

Since $M^Q \in BMO_2$, we obtain that

$$E(\langle L\rangle_T - \langle L\rangle_\tau / F_\tau) \le c_1 + c_2 E^{1/2}(\langle L\rangle_T - \langle L\rangle_\tau / F_\tau)$$

for some positive constants c_1 and c_2 which do not depend on τ. It follows from the latter inequality that $E(\langle L\rangle_T - \langle L\rangle_\tau / F_\tau)$ is bounded for every stopping time τ by one and the same constant. This implies $L \in BMO_2$, since L is of bounded jumps. □

COROLLARY 2. *Suppose that there exists a martingale measure Q, with continuous density process Z^Q, which satisfies the reverse Hölder inequality $R_2(P)$. Then the martingale part m of the value process belongs to the class BMO_2.*

Proof. The reverse Hölder condition implies that the value process is bounded (see the remark after Theorem 1) and that $M^Q \in BMO_2$ (Proposition 2). Thus it follows from Theorem 1 that the value process is a bounded solution of (3.4); hence $m \in BMO_2$ by Lemma 1. □

Let C be the predictable support of the set $]0,\infty[\cup\{\Delta m \ne 0\}$, and let $(\tau_n, n \ge 1)$ be a sequence of predictable stopping times such that $C = \cup[\tau_n]$.

Let us consider the martingale

$$M_t^{Q^*} = -\int_0^t \lambda_s dM_s + N_t^*, \tag{3.21}$$

where

$$N_t^* = -\int_0^t \frac{1}{V_{s-}} d\bar{m}_s^c - \int_0^t \frac{1}{V_s} d\bar{m}_s^{dq} \tag{3.22}$$

$$+ \sum_{\tau_n \le t} \frac{E(\frac{\Delta\bar{m}_{\tau_n}}{V_{\tau_n}}|F_{\tau_n-}) - \Delta\bar{m}_{\tau_n} E(\frac{1}{V_{\tau_n}}|F_{\tau_n-})}{V_{\tau_n} E(\frac{1}{V_{\tau_n}}|F_{\tau_n-})}.$$

Since $\bar{m}^{dq}$ is quasi-left-continuous and $V \ge 1$ (by Jensen's inequality), the stochastic integral $\frac{1}{V}\cdot\bar{m}^{dq}$ is well defined by the Chou–Lepingle lemma (see, e.g., Jacod [12]) as a unique purely discontinuous local martingale I such that

$$\Delta I_t = \frac{1}{V_t}\Delta\bar{m}_t^{dq}.$$

Using the equality $\Delta\bar{m} = \Delta m = \Delta V - \Delta A$ and that A_τ is $F_{\tau-}$-measurable, it is easy to see that the equality

$$\frac{E(\frac{\Delta\bar{m}_{\tau_n}}{V_{\tau_n}}|F_{\tau_n-}) - \Delta\bar{m}_{\tau_n} E(\frac{1}{V_{\tau_n}}|F_{\tau_n-})}{V_{\tau_n} E(\frac{1}{V_{\tau_n}}|F_{\tau_n})} = -1 + \frac{1}{V_{\tau_n} E(\frac{1}{V_{\tau_n}}|F_{\tau_n-})} \tag{3.23}$$

holds.

Later we shall show that the measure Q^* defined by $dQ^* = \mathcal{E}_T(M^{Q^*})dP$ is variance-optimal. Let us show that Q^* is an equivalent martingale measure of X.

LEMMA 2. *Let conditions* (A) *and* (B*) *be satisfied. Then Q^* defined by* (3.21)–(3.22) *is an equivalent martingale measure of X.*

Proof. Let us first show that

$$\Delta M_t^{Q^*} \ge -1 + \frac{1}{C}. \tag{3.24}$$

It is evident that the moments of jumps of martingales M^{Q^*} and m coincide. If a moment of jump σ of M^{Q^*} is totally inaccessible, then from (3.22) we have that

$$\Delta M_\sigma^{Q^*} = -\frac{1}{V_\sigma}\Delta \bar{m}_\sigma^{dq} = -\frac{1}{V_\sigma}\Delta m_\sigma \geq -\frac{1}{V_\sigma}\Delta V_\sigma$$

$$= -1 + \frac{V_{\sigma-}}{V_\sigma} \geq -1 + \frac{1}{C}. \tag{3.25}$$

Here we used Corollary 1 of Theorem 1 and the inequality $1 \leq V_t \leq C$ valid for all $t \in [0, T]$.

If τ is a predictable moment of jump of m, then equalities (3.22), (3.23) and inequality $1 \leq V_t \leq C$ imply that

$$\Delta M_\tau^{Q^*} = -1 + \frac{1}{V_\tau E(\frac{1}{V_\tau}/F_{\tau-})} \geq -1 + \frac{1}{C}. \tag{3.26}$$

Therefore, (3.24) holds.

Let us show now that $M^{Q^*} \in BMO_2$. For this it is sufficient to show that each martingale entering (3.22) is in BMO_2.

$\lambda \cdot M \in BMO_2$ by Condition (B*) and Proposition 2.

The martingales $\frac{1}{V} \cdot \bar{m}^c$ and $\frac{1}{V} \cdot \bar{m}^{dq}$ belong to BMO_2 since $V \geq 1$,

$$\langle \bar{m}^c \rangle_T - \langle \bar{m}^c \rangle_\tau + \langle \bar{m}^{dq} \rangle_T - \langle \bar{m}^{dq} \rangle_\tau \leq \langle m \rangle_T - \langle m \rangle_\tau,$$

and $m \in BMO_2$ by Corollary 2.

It is easy to see that the last term of (3.22) is a martingale from the class BMO_2, since it is of bounded jumps and (denoting this martingale by l)

$$E(\langle l \rangle_{\tau T} | F_\tau) = E\left[\sum_{\tau \leq \tau_n \leq T} E\left(\left(\frac{E(\frac{\Delta m_{\tau_n}}{V_{\tau_n}}|F_{\tau_n-}) - \Delta m_{\tau_n} E(\frac{1}{V_{\tau_n}}|F_{\tau_n-})}{V_{\tau_n} E(\frac{1}{V_{\tau_n}}|F_{\tau_n-})}\right)^2 |F_{\tau_n-}\right)|F_\tau\right]$$

$$\leq 2cE(\langle m \rangle_{\tau T}|F_\tau).$$

Thus $M^{Q^*} \in BMO_2$, which together with (3.24) implies that $\mathcal{E}_t(M^{Q^*})$ is a true martingale (Proposition 1). Besides it follows from (3.24) that $\mathcal{E}_T(M^{Q^*}) > 0$ so that $Q^* \in \mathcal{M}^e$. □

LEMMA 3. *Suppose that there exists $Q \in M^e$ satisfying the reverse Hölder condition $R_2(P)$. Then for any predictable stopping time τ*

$$\operatorname*{ess\,inf}_{N \in \mathcal{M}^2 : E(\Delta N_\tau | F_{\tau-}) = 0} E\left[\left(\sqrt{V_\tau}\Delta N_\tau + \frac{1}{\sqrt{V_\tau}}\Delta m_\tau\right)^2 |F_{\tau-}\right] = \frac{E^2(\frac{\Delta m_\tau}{V_\tau}|F_{\tau-})}{E(\frac{1}{V_\tau}|F_{\tau-})}, \tag{3.27}$$

and the infimum is attained for N^* with

$$\Delta N_\tau^* = \frac{E(\frac{\Delta m_\tau}{V_\tau}|F_{\tau-}) - \Delta m_\tau E(\frac{1}{V_\tau}|F_{\tau-})}{V_\tau E(\frac{1}{V_\tau}|F_{\tau-})}. \tag{3.28}$$

Proof. The $R_2(P)$ condition implies that V is bounded above and that $m \in BMO_2$. Therefore, conditional expectations in (3.27) and (3.28) exist. Using the

Cauchy–Schwarz inequality for any N with $E(\Delta N_\tau|F_{\tau-})=0$, we have

$$\begin{aligned}
&E\left(\frac{1}{V_\tau}|F_{\tau-}\right)E\left[\left(\sqrt{V_\tau}\Delta N_\tau+\frac{1}{\sqrt{V_\tau}}\Delta m_\tau\right)^2|F_{\tau-}\right]\\
&\quad\geq E^2\left[\frac{1}{\sqrt{V_\tau}}\left(\sqrt{V_\tau}\Delta N_\tau+\frac{1}{\sqrt{V_\tau}}\Delta m_\tau\right)|F_{\tau-}\right]\\
&\quad=E^2\left(\frac{\Delta m_\tau}{V_\tau}|F_{\tau-}\right).
\end{aligned}$$

On the other hand, it is easy to verify that

$$E\left[\left(\sqrt{V_\tau}\Delta N^*_\tau+\frac{1}{\sqrt{V_\tau}}\Delta m_\tau\right)^2|F_{\tau-}\right]=\frac{E^2(\frac{\Delta m_\tau}{V_\tau}|F_{\tau-})}{E(\frac{1}{V_\tau}|F_{\tau-})},$$

which proves the assertion of the lemma. □

LEMMA 4. *Let conditions* (A) *and* (B*) *be satisfied. Then*

$$\begin{aligned}
&\operatorname*{ess\,inf}_{N\in\mathcal{N}_2(X)}\left[\int_0^t V_{s-}d\langle N^c\rangle_s+2\langle\bar m^c,N\rangle_t+\sum_{s\leq t}(V_s\widetilde{(\Delta N_s)^2}+2\Delta\bar m_s\Delta N_s)\right]\\
&\quad=-\int_0^t\frac{1}{V_{s-}}d\langle\bar m^c\rangle_s-\sum_{s\leq t}\widetilde{\frac{1}{V_s}(\Delta\bar m_s)^2}+\sum_{\tau_n\leq t}\frac{E^2(\frac{\Delta\bar m_{\tau_n}}{V_{\tau_n}}|F_{\tau_n-})}{E(\frac{1}{V_{\tau_n}}|F_{\tau_n-})}.
\end{aligned}\tag{3.29}$$

Proof. Let us write (3.29) in the form

$$\operatorname*{ess\,inf}_{N\in\mathcal{N}_2(X)}A_t(N)=\bar A_t\tag{3.30}$$

with evident notation. Let us first show that

$$A_t(N)-\bar A_t\in\mathcal{A}^+_{\mathrm{loc}}.\tag{3.31}$$

We have

$$\begin{aligned}
A_t(N)-\bar A_t&=\left\langle\int\sqrt{V}_s dN^c_s+\int\frac{1}{\sqrt{V}_s}d\bar m^c_s\right\rangle_t\\
&\quad+\sum_{s\leq t}\left(\widetilde{\frac{1}{\sqrt{V}_s}\Delta\bar m_s}+\sqrt{V}_s\Delta N_s\right)^2-\sum_{\tau_n\leq t}\frac{E^2(\frac{\Delta\bar m_{\tau_n}}{V_{\tau_n}}|F_{\tau_n-})}{E(\frac{1}{V_{\tau_n}}|F_{\tau_n-})}\\
&=\left\langle\int\sqrt{V}_s dN^c_s+\int\frac{1}{\sqrt{V}_s}d\bar m^c_s\right\rangle_t+\sum_{\sigma_n\leq t}\left(\widetilde{\frac{1}{\sqrt{V}_{\sigma_n}}\Delta\bar m_{\sigma_n}}+\sqrt{V}_{\sigma_n}\Delta N_{\sigma_n}\right)^2\\
&\quad+\sum_{\tau_n\leq t}E\left[\left(\frac{1}{\sqrt{V_{\tau_n}}}\Delta\bar m_{\tau_n}+\sqrt{V}_{\tau_n}\Delta N_{\tau_n}\right)^2|F_{\tau_n-}\right]-\sum_{\tau_n\leq t}\frac{E^2(\frac{\Delta\bar m_{\tau_n}}{V_{\tau_n}}|F_{\tau_n-})}{E(\frac{1}{V_{\tau_n}}|F_{\tau_n-})}\\
&=A^1_t+A^2_t+A^3_t-A^4_t.
\end{aligned}\tag{3.32}$$

Here we used Proposition 1.49 of Jacod [12], and we recall that τ_n, σ_n are sequences of predictable and totally inaccessible stopping times, respectively, exhausting the jumps of m and that $\Delta m = \Delta \bar{m}$.

Lemma 3 implies that $A^3_t - A^4_t$ is an increasing process. It is evident that A^1 and A^2 also are increasing processes. Therefore, (3.31) holds. Let us show that

$$\bar{A}_t - \operatorname*{ess\,inf}_{N \in \mathcal{N}_2(X)} A_t(N) \in \mathcal{A}^+_{\mathrm{loc}}. \tag{3.33}$$

Define a sequence of stopping times $(s_n,\ \ n \geq 1)$ by

$$s_k = \inf\{t : \mathcal{E}_t(N^*) \geq k\},$$

where N^* is given by (3.22). Let $N^k_t = N^*_{t \wedge s_k}$.

Since N^* is of bounded jumps, it is evident that $\mathcal{E}_t(N^k) = \mathcal{E}_{t \wedge s_k}(N^*) \leq \text{const}$, $\mathcal{E}_t(-\lambda \cdot M + N^k)$ is a square integrable martingale with $\Delta N^k > -1$ (by Lemma 2), and N^k belongs to the class $N_2(X)$. Therefore, for every $k \geq 1$ we have that

$$A_t(N^k) - \operatorname*{ess\,inf}_{N \in \mathcal{N}_2(X)} A_t(N) \in \mathcal{A}^+_{\mathrm{loc}} \tag{3.34}$$

by definition of ess inf with respect to the strong order $\prec$.

It is easy to see that

$$A_t(N^k) = -\int_0^{t \wedge s_k} \frac{1}{V_{s-}} d\langle \bar{m}^c \rangle_s - \sum_{s \leq t \wedge s_k} \widetilde{\frac{1}{V_s}} (\Delta \bar{m}_s)^2 + \sum_{\tau_n \leq t \wedge s_k} \frac{E^2(\frac{\Delta \bar{m}_{\tau_n}}{V_{\tau_n}} | F_{\tau_n -})}{E(\frac{1}{V_{\tau_n}} | F_{\tau_n -})}$$

and $A_t(N^k) \to \bar{A}_t$ a.s. for every t as $k \to \infty$. Therefore, (3.33) follows from (3.34). It is evident that relations (3.31) and (3.33) imply the assertion of this lemma. □

Now we shall prove the main statement of the paper. Recall that any special semimartingale Y can be expressed in the form

$$Y_t = Y_0 + B_t + \int_0^t \psi_s dM_s + \bar{L}_t,$$

where $\bar{L}$ is a local martingale orthogonal to M.

THEOREM 2. *Let conditions* (A) *and* (B*) *be satisfied. Then the value process V is a unique solution of the semimartingale backward equation*

$$Y_t = Y_0 - \int_0^t Y_{s-} d\langle \lambda \cdot M \rangle_s + 2\langle \lambda \cdot M, \psi \cdot M \rangle_t \tag{3.35}$$

$$- \sum_{\tau_n \leq t} \frac{E^2(\frac{\Delta \bar{L}_{\tau_n}}{Y_{\tau_n}} | F_{\tau_n -})}{E(\frac{1}{Y_{\tau_n}} | F_{\tau_n -})} + \int_0^t \widetilde{\frac{1}{Y_s}} d[\bar{L}]_s + \int_0^t \psi_s dM_s + \bar{L}_t, \ \ Y_T = 1$$

in the class of semimartingales Y satisfying the two-sided inequality

$$c \leq Y_t \leq C \ \ \textit{for all} \ \ t \in [0,T] \ \ \textit{a.s.} \tag{3.36}$$

for some constants $0 < c \leq C$.

Moreover, the martingale measure Q^* is variance-optimal if and only if it is given by $dQ^* = \mathcal{E}_T(M^{Q^*})dP$, where

$$M_t^{Q^*} = -\int_0^t \lambda_s dM_s - \int_0^t \frac{1}{V_s} d(\bar{m}_s^c + \bar{m}_s^{dq}) + \sum_{\tau_n \le t} \left(-1 + \frac{1}{V_{\tau_n} E(\frac{1}{\Delta V_{\tau_n}} | F_{\tau_n -})} \right) \tag{3.37}$$

and $(\tau_n,\ n \ge 1)$ is a sequence of stopping times exhausting the predictable moments of jumps for V.

Proof. It follows from Lemma 4 that

$$-\operatorname*{ess\,inf}_{N \in \mathcal{N}_2^e} \left[\int_0^t V_{s-} d\langle M^{Q,c} \rangle_s + 2\langle M^{Q,c}, m \rangle_t + \sum_{s \le t} (V_s (\Delta M_s^Q)^{\widetilde{2}} + 2\Delta \bar{m}_s \Delta M_s^Q) \right]$$

$$= -\int_0^t V_{s-} d\langle \lambda \cdot M \rangle_s + 2\langle \lambda \cdot M, \varphi \cdot M \rangle_t + \int_0^t \widetilde{\frac{1}{V_s}} d[\bar{m}_s]$$

$$- \sum_{\tau_n \le t} \frac{E^2(\frac{\Delta \bar{m}_{\tau_n}}{V_{\tau_n}} | F_{\tau_n -})}{E(\frac{1}{V_{\tau_n}} | F_{\tau_n -})}. \tag{3.38}$$

Therefore, according to Theorem 1, the value process satisfies (3.35).

Let us show that the optimal martingale measure is of the form (3.37). Let the martingale measure Q^0 be optimal. By the optimality principle $V_t \mathcal{E}_t^2(M^{Q^0})$ is martingale. Since V solves (3.4), this implies that

$$\operatorname*{ess\,inf}_{Q \in \mathcal{M}_2^e} \left[\int_0^t V_{s-} d\langle M^{Q,c} \rangle_t + 2\langle M^{Q,c}, \bar{m}^c \rangle_t + \sum_{s \le t} (V_s (\Delta M_s^Q)^{\widetilde{2}} + \Delta m_s \Delta M_s^Q) \right]$$

$$= \int_0^t V_{s-} d\langle M^{Q^0,c} \rangle_t + 2\langle M^{Q^0,c}, \bar{m}^c \rangle_t$$

$$+ \sum_{s \le t} (V_s (\Delta M_s^{Q^0})^{\widetilde{2}} + 2\Delta m_s \Delta M_s^{Q^0}). \tag{3.39}$$

Since M^{Q^0} is represented in the form $-\lambda \cdot M + N^0$ for some $N^0 \in \mathcal{N}_2(X)$, it follows from (3.38) and (3.39) that

$$-\int_0^t \widetilde{\frac{1}{V_s}} d[\bar{m}]_s + \sum_{\tau_n \le t} \frac{E^2(\frac{\Delta \bar{m}_{\tau_n}}{V_{\tau_n}} | F_{\tau_n -})}{E(\frac{1}{V_{\tau_n}} | F_{\tau_n -})}$$

$$= \int_0^t V_{s-} d\langle N^{0,c} \rangle_s + 2\langle N^{0,c}, \bar{m}^c \rangle_t + \sum_{s \le t} (V_s (\Delta N_s^0)^{\widetilde{2}} + 2\Delta \bar{m}_s \Delta N_s^0),$$

and hence

$$\left\langle \sqrt{V}\cdot N^{0,c}+\frac{1}{\sqrt{V}}\cdot\bar{m}^{c}\right\rangle_t+\sum_{s\le t}\left(\sqrt{V_s}\Delta\widetilde{N_s^{0,dq}}+\frac{1}{\sqrt{V_s}}\Delta\bar{m}_s\right)^2$$
$$+\sum_{s\le t}\left(\sqrt{V_s}\Delta\widetilde{N_s^{0,dp}}+\frac{1}{\sqrt{V_s}}\Delta\bar{m}_s\right)^2$$
$$-\sum_{\tau_n\le t}\frac{E^2(\frac{\Delta\bar{m}_{\tau_n}}{V_{\tau_n}}|F_{\tau_n-})}{E(\frac{1}{V_{\tau_n}}|F_{\tau_n-})}=0.$$

This equality implies that the process N^0 may have the jumps at the moments of jumps of $\bar{m}$ only. Therefore,

$$\sum_{s\le t}\left(\sqrt{V_s}\Delta\widetilde{N_s^{0,dp}}+\frac{1}{\sqrt{V_s}}\Delta\bar{m}_s^{dp}\right)^2=\sum_{\tau_n\le t}E\left[\left(\Delta N^0_{\tau_n}+\frac{1}{V_{\tau_n}}\Delta\bar{m}_{\tau_n}\right)^2|F_{\tau_n-}\right],$$

and from Lemma 3 we obtain that

$$N^{0,c}=-\frac{1}{V}\cdot\bar{m}^{c},\quad N^{0,dq}=-\frac{1}{V}\cdot\bar{m}^{dq},$$

and

$$\Delta N^0_{\tau_n}=\frac{E(\frac{\Delta\bar{m}_{\tau_n}}{V_{\tau_n}}|F_{\tau_n-})-\Delta\bar{m}_{\tau_n}E(\frac{1}{V_{\tau_n}}|F_{\tau_n-})}{V_{\tau_n}E(\frac{1}{V_{\tau_n}}|F_{\tau_n-})}.$$

Thus the processes M^{Q^*} (defined by (3.37)) and M^{Q^0} are indistinguishable. Therefore, $Q^*\in\mathcal{M}_2^e$ and the variance-optimal martingale measure is unique and admits representation (3.37).

Uniqueness. Let Y be a solution of (3.35) satisfying (3.36). Since Y solves (3.35) using the Itô formula for $\mathcal{E}_t^2(M^Q)Y_t$, we obtain that for every $Q\in\mathcal{M}_2^e$ the process $Y_t\mathcal{E}_t^2(M^Q)$ is a local submartingale under P. Since Y is positive and bounded, $Y_t\mathcal{E}_t^2(M^Q)$ is a submartingale of class D, and it follows from the boundary condition that for every $Q\in\mathcal{M}_2^e$

$$Y_t\mathcal{E}_t^2(M^Q)\le E(Y_T\mathcal{E}_T^2(M^Q)|F_t)\le E(\mathcal{E}_T^2(M^Q)|F_t).$$

Hence $Y_t\le E(\mathcal{E}_{tT}^2(M^Q)|F_t)$ for all $Q\in\mathcal{M}_2^e$ and

$$Y_t\le Q\in M_2^e\,\text{ess inf}E(\mathcal{E}_{tT}^2(M^Q)|F_t)=V_t. \tag{3.40}$$

Let us prove the converse inequality.

Similarly to (3.38), one can show that

$$B_t=-\int_0^t Y_{s-}d\langle\lambda\cdot M\rangle_s+2\langle\lambda\cdot M,\psi\cdot M\rangle_t \tag{3.41}$$

$$-\sum_{\tau_n\le t}\frac{E^2(\frac{\Delta\bar{L}_{\tau_n}}{V_{\tau_n}}|F_{\tau_n-})}{E(\frac{1}{V_{\tau_n}}|F_{\tau_n-})}+\int_0^t\widetilde{\frac{1}{Y_s}d[\bar{L}]_s}.$$

Consider the martingale

$$M_t^{Q^0} = -\int_0^t \lambda_s dM_s - \int_0^t \frac{1}{V_s} d(\bar{L}_s^c + \bar{L}_s^{dq}) + \sum_{\tau_n \le t} \left(-1 + \frac{1}{Y_{\tau_n} E(\frac{1}{Y_{\tau_n}} | F_{\tau_n -})} \right), \tag{3.42}$$

which belongs to BMO_2 by Lemmas 1 and 2.

Using again the Itô formula, (3.41), and (3.42), one can show that the process $Y_t \mathcal{E}_t^2(M^{Q^0})$ is a local martingale. Since any positive local martingale is a supermartingale, we have that

$$Y_t \mathcal{E}_t^2(M^Q) \ge E(Y_T \mathcal{E}_T^2(M^{Q^0}) | F_t),$$

which gives that

$$E\mathcal{E}_T^2(M^{Q^0}) \le Y_0 < \infty. \tag{3.43}$$

Since $\Delta M^{Q^0} \ge -1 + \frac{c}{C}$ and $M^{Q^0} \in BMO_2$, the process $(\mathcal{E}_t(M^{Q^0}), t \in [0,T])$ is a uniformly integrable martingale and (3.43) implies that this martingale is square-integrable. Thus $Q^0 \in \mathcal{M}_2^e$, and $Y_t \mathcal{E}_t^2(M^{Q^0})$ is a martingale of the class D, since Y is bounded. Now the martingale property and the boundary condition imply that

$$Y_t = E(\mathcal{E}_{tT}^2(-\lambda \cdot M + N^0) | F_t).$$

Therefore, (3.40) and the latter equality imply that $Y_t = V_t$ for all $t \in [0,T]$; hence V is the unique solution of (3.35).

The same arguments show that the measure $dQ^* = \mathcal{E}_T(M^{Q^*}) dP$, where M^{Q^*} is given by (3.37), belongs to the class $\mathcal{M}_2^e$ and $V_t \mathcal{E}_t^2(M^{Q^*})$ is a martingale; hence Q^* is variance-optimal by the optimality principle. □

Remark 1. Since

$$\int_0^t \widetilde{\frac{1}{Y_s}} d[\bar{L}]_s = \sum_{\tau_n \le t} E\left(\frac{(\Delta \bar{L}_{\tau_n})^2}{Y_{\tau_n}} | F_{\tau_n -} \right) + \int_0^t \frac{1}{Y_{s-}} d\langle \bar{L}^{dq} \rangle_s + \int_0^t \frac{1}{Y_{s-}} d\langle \bar{L}^c \rangle_s,$$

one can write (3.35) in the following equivalent form:

$$\begin{aligned} Y_t = Y_0 - \int_0^t Y_{s-} d\langle \lambda \cdot M \rangle_s + 2\langle \lambda \cdot M, \psi \cdot M \rangle_t + \int_0^t \frac{1}{Y_{s-}} d\langle \bar{L}^c \rangle_s \\ + \int_0^t \frac{1}{Y_{s-}} d\langle \bar{L}^{dq} \rangle_s + \sum_{\tau_n \le t} \left[E\left(\frac{(\Delta \bar{L}_{\tau_n})^2}{Y_{\tau_n}} | F_{\tau_n -} \right) - \frac{E^2(\frac{\Delta \bar{L}_{\tau_n}}{Y_{\tau_n}} | F_{\tau_n -})}{E(\frac{1}{Y_{\tau_n}} | F_{\tau_n -})} \right] \\ + \int_0^t \psi_s dM_s + \bar{L}_t. \end{aligned} \tag{3.44}$$

Remark 2. Theorem 2 can also be formulated in the following form: The triple $(V, \varphi, \bar{m})$ is a unique solution of (3.35) (or (3.44)) in the class of processes $(Y, \psi, \bar{L})$ such that $0 < c \le Y \le C$, $\psi \cdot M \in BMO_2$, $\bar{L} \in BMO_2$.

COROLLARY 3. *The martingale measure Q^* is variance-optimal if and only if*

$$Z_T^* = c + \int_0^T h_s dX_s \tag{3.45}$$

for a constant c and an X-integrable process h such that the process $\int_0^t h_s dX_s$ is a Q-martingale for any $Q \in \mathcal{M}_2^e$. Moreover, $c = V_0$ and the integrand h can be expressed as

$$h_t = V_0 \mathcal{E}_{t-}\left(\left(\frac{\varphi}{V} - \lambda\right) \cdot X\right)\left(\frac{\varphi_t}{V_t} - \lambda_t\right), \tag{3.46}$$

where φ is defined by (3.3).

Proof. The first statement is proved in [5]. Let us show the validity of representation (3.46). Let Q^* be an optimal martingale measure. By Theorem 2, Q^* admits representation (3.37). It follows from the Itô formula for $V_t \mathcal{E}_t(M^{Q^*})$, using expressions (3.38) and (3.37), that

$$V_t \mathcal{E}_t(M^{Q^*}) = V_0 + \int_0^t V_{s-} \mathcal{E}_{s-}(M^{Q^*}) \left(\frac{\varphi_s}{V_s} - \lambda_s\right) dX_s. \tag{3.47}$$

Therefore, from (2.1)

$$\begin{aligned} V_t \mathcal{E}_t(M^{Q^*}) &= V_0 \mathcal{E}_t\left(\left(\frac{\varphi}{V} - \lambda\right) \cdot X\right) \\ &= V_0 \left(1 + \int_0^t \mathcal{E}_{s-}\left(\left(\frac{\varphi}{V} - \lambda\right) \cdot X\right)\left(\frac{\varphi_s}{V_s} - \lambda_s\right) dX_s\right), \end{aligned} \tag{3.48}$$

and using the boundary condition, we obtain representation (3.45) with h expressed by (3.46). Besides, it follows from (3.48) and from the optimality principle that $V_0 + \int_0^t h_s dX_s$ is a Q^*-martingale.

It follows from (3.45), (3.48) that

$$\int_0^t h_s dX_s \geq -V_0.$$

The latter inequality implies (see Ansel and Stricker [1]) that the process $\int_0^t h_s dX_s$ is a Q-local martingale; hence it is also a supermartingale for any $Q \in \mathcal{M}_2^e$.

On the other hand, since Q^* is optimal, by Lemma 1 of Schweizer [23] $E^Q \mathcal{E}_T(M^{Q^*}) =$ const for any $Q \in \mathcal{M}_2^e$ which implies that $E^Q \mathcal{E}_T(M^{Q^*}) = V_0$, and from (3.48) we have

$$E^Q \int_0^T h_s dX_s = E^Q \mathcal{E}_T(M^{Q^*}) - V_0 = 0.$$

Hence $\int_0^t h_s dX_s$ is a martingale for all $Q \in \mathcal{M}_2^e$. □

4. Some particular cases. In this section we consider the martingale analogues of two specific extreme cases, already studied by Pham, Rheinländer, and Schweizer [19], Laurent and Pham [14], and Biagini, Guasoni, and Pratelli [2], when the semimartingale backward equation (3.4) admits an explicit solution. Besides, we give necessary and sufficient conditions in order that the variance-optimal martingale measure coincides with the minimal martingale measure as well as with the martingale measure appearing in the second extreme case.

Let $X_t(2) = M_t + 2\int_0^t d < M >_s \lambda_s$. Condition (B) implies that the process $(\mathcal{E}_t(-2\lambda \cdot M), t \in [0,T])$ is a martingale, and one can define a probability measure

$Q(2)$ by $dQ(2) = \mathcal{E}_T(-2\lambda \cdot M)dP$. Note that the process $X(2)$ is a local martingale under $Q(2)$, by Girsanov's theorem.

PROPOSITION 4. *Let conditions* (A) *and* (B*) *be satisfied. Then the variance-optimal martingale measure coincides with the minimal martingale measure if and only if*

$$e^{\langle \lambda \cdot M \rangle_T} = c + \int_0^T h_s dX_s(2) \tag{4.1}$$

for a constant c *and an* $X(2)$*-integrable* h *such that the process* $(\int_0^t h_s dX_s(2),\ t \in [0,T])$ *is a martingale under* $Q(2)$.

Moreover, in this case (3.35) *admits an explicit solution*

$$V_t = E(\mathcal{E}^2_{tT}(-\lambda \cdot M)|F_t) = E^{Q(2)}(e^{\langle \lambda \cdot M \rangle_{tT}}|F_t)$$

with the martingale part

$$\int_0^t e^{-\langle \lambda \cdot M \rangle_s} h_s dM_s.$$

Proof. Let (4.1) be satisfied. Consider the process

$$Y_t = E(\mathcal{E}^2_{tT}(-\lambda \cdot M)|F_t) = E(\mathcal{E}_{tT}(-2\lambda \cdot M)e^{\langle \lambda \cdot M \rangle_{tT}}|F_t).$$

It is evident that $1 \le Y \le C$. Using (4.1), we have that $Y_t = e^{-\langle \lambda \cdot M \rangle_t} \bar{M}_t$, where $\bar{M}_t = c + \int_0^t h_s dX_s(2)$. By the Itô formula we obtain that

$$Y_t = Y_0 - \int_0^t e^{-\langle \lambda \cdot M \rangle_s} \bar{M}_s d\langle \lambda \cdot M \rangle_s + \int_0^t e^{-\langle \lambda \cdot M \rangle_s} h_s dX_s(2),$$

and Y satisfies the equation

$$Y_t = Y_0 - \int_0^t Y_s d\langle \lambda \cdot M \rangle_s + 2\int_0^t e^{-\langle \lambda \cdot M \rangle_s} d\langle h \cdot M, \lambda \cdot M \rangle_s + \int_0^t e^{-\langle \lambda \cdot M \rangle_s} h_s dM_s.$$

Thus Y is a bounded solution of (3.35) with $\psi_s = h_s e^{-<\lambda \cdot M>_s}$, $s \in [0,T]$, and $\bar{L} = 0$. Therefore, Theorem 2 implies that $Y = V$, which means that the minimal martingale measure Q^{min} is variance-optimal.

Assume that Q^{min} is variance-optimal. Then by Theorem 2 the process N^* defined by (3.22) is indistinguishable from 0. Since the martingales from (3.22) are orthogonal to each other, we have that $\bar{m}^c = 0$, $\bar{m}^{dq} = 0$, and for any predictable stopping time τ

$$\frac{E(\frac{\Delta \bar{m}_\tau}{V_\tau}|F_{\tau-}) - \Delta \bar{m}_\tau E(\frac{1}{V_\tau}|F_{\tau-})}{V_\tau E(\frac{1}{V_\tau}|F_{\tau-})} = 0.$$

The latter equality implies that $\Delta \bar{m}_\tau$ is $F_{\tau-}$-measurable, and hence $\Delta \bar{m}_\tau = E(\Delta \bar{m}_\tau|F_{\tau-}) = 0$ for any predictable moment τ, and we obtain that $\bar{m} = 0$. Therefore, the value process satisfies

$$V_t = V_0 - \int_0^t V_s d\langle \lambda \cdot M \rangle_s + \int_0^t \varphi_s dX_s(2).$$

Solving this equation with respect to V and using the boundary condition $V_T = 1$, we have that

$$e^{\langle \lambda \cdot M \rangle_T} = V_0 \mathcal{E}_T\left(\frac{\varphi}{V} \cdot X(2)\right) = V_0 + V_0 \int_0^T \mathcal{E}_s\left(\frac{\varphi}{V} \cdot X(2)\right) \frac{\varphi_s}{V_s} dX_s(2).$$

Since $\lambda \cdot M \in BMO_2$ and $V \geq 1$, it follows from Lemma 1 and Proposition 7 of Doléans-Dade and Meyer [8] that $\frac{\varphi}{V} \cdot X(2)$ belongs to the class $BMO_2(Q(2))$ and hence the process $\int_0^t \mathcal{E}_s(\frac{\varphi}{V} \cdot X(2))\frac{\varphi_s}{V_s} dX_s(2)$ is a $Q(2)$-martingale. □

PROPOSITION 5. *Let conditions* (A) *and* (B*) *be satisfied. Then the variance-optimal martingale measure is of the form*

$$\mathcal{E}_T(M^{Q^*}) = c\mathcal{E}_T(-\lambda \cdot X) \tag{4.2}$$

if and only if

$$e^{-\langle \lambda \cdot M \rangle_T} = \hat{c} + \hat{m}_T \tag{4.3}$$

for a constant $\hat{c}$ and a martingale $\hat{m}$ orthogonal to M.

Under (4.3), (3.35) *admits an explicit solution*

$$V_t = \frac{1}{E(\exp\{- < \lambda \cdot M >_{tT}\}/F_t)}.$$

Proof. Assume that the variance-optimal martingale measure is of the form (4.2). By Theorem 2 the optimal density is of the form $\mathcal{E}_T(-\lambda \cdot M + N^*)$, where N^* admits representation (3.22). Therefore, (4.2) implies that

$$\mathcal{E}_T(-\lambda \cdot M + N^*) = c\mathcal{E}_T(-\lambda \cdot X)$$

and

$$\mathcal{E}_T(N^*) = c\frac{\mathcal{E}_T(-\lambda \cdot X)}{\mathcal{E}_T(-\lambda \cdot M)} = ce^{-\langle \lambda \cdot M \rangle_T}.$$

Thus

$$e^{-\langle \lambda \cdot M \rangle_T} = \frac{1}{c} + \frac{1}{c}\int_0^T \mathcal{E}_{s-}(N^*) dN_s^*,$$

and, according to Lemma 2, $N^* \in BMO_2$, $\Delta N^* \geq -1 + h$, and hence the process $\hat{m}_t = \frac{1}{c}\int_0^t \mathcal{E}_{s-}(N^*) dN_s^*$ is a true martingale.

Conversely, let (4.3) be satisfied. By Theorem 2 the optimal density is of the form

$$Z_T^{Q^*} = \mathcal{E}_T(-\lambda \cdot M + N^*), \tag{4.4}$$

where N^* is defined by (3.37). On the other hand, Corollary 3 implies (see (3.46) and (3.48)) that the optimal density can be expressed in the alternate form

$$Z_T^{Q^*} = c\mathcal{E}_T((\bar{\varphi} - \lambda) \cdot X), \tag{4.5}$$

where $\bar{\varphi} = \frac{\varphi}{V}$ and φ is defined by (3.3).

Using equality

$$\mathcal{E}_T((\bar{\varphi} - \lambda) \cdot X) = \mathcal{E}_T(-\lambda \cdot M)\mathcal{E}_T(\bar{\varphi} \cdot X(2)) \exp\{-\langle \lambda \cdot M \rangle_T\},$$

from (4.3) and (4.5) we have

$$Z_T^{Q^*} = c\mathcal{E}_T(-\lambda \cdot M)\mathcal{E}_T(\bar{\varphi} \cdot X(2))(\hat{c} + \hat{m}_T). \tag{4.6}$$

Therefore, from (4.4) and (4.6) we obtain the equality

$$\mathcal{E}_T(N^*) = c\mathcal{E}_T(\bar{\varphi}\cdot X(2))(\hat{c}+\hat{m}_T). \tag{4.7}$$

Lemma 1, inequality (3.24), and the two-sided inequality $1 \le V \le C$ imply that N^* and $\bar{\varphi}\cdot M$ belong to BMO_2. Therefore, by Proposition 7 of Doléans-Dade and Meyer [8], $\bar{\varphi}\cdot X(2) \in BMO_2(Q(2))$, and since N^* is orthogonal to M, it is also in $BMO_2(Q(2))$. Hence the corresponding Doléans exponentials are uniformly integrable martingales by Proposition 1. From condition (4.3) we have that

$$\hat{c}+\hat{m}_t = E(\exp\{- < \lambda\cdot M >_T\}|F_t);$$

hence $\hat{m}$ is a bounded $Q(2)$-martingale, and since it is orthogonal to M the process $(\mathcal{E}_t(\bar{\varphi}\cdot X(2))(\hat{c}+\hat{m}_t), t\in[0,T])$ is a uniformly integrable martingale. Therefore, taking conditional expectations in (4.7) with respect to the measure $Q(2)$, we obtain that this equality holds for any $t\in[0,T]$, i.e.,

$$\mathcal{E}_t(N^*) = c\mathcal{E}_t(\bar{\varphi}\cdot X(2))(\hat{c}+\hat{m}_t). \tag{4.8}$$

Let $N_t^0 = \int_0^t \frac{1}{\hat{c}+\hat{m}_{s-}} d\hat{m}_s$. Then $1+\frac{1}{\hat{c}}\hat{m}_t = \mathcal{E}_t(N^0)$, and, using the Yor formula from (4.8), we obtain that

$$\mathcal{E}_t(N^*) = c\hat{c}\mathcal{E}_t(\bar{\varphi}\cdot X(2)+N^0) \tag{4.9}$$

for each $t\in[0,T]$. However, this implies that $c=\hat{c}^{-1}$ and

$$N_t^* = \int_0^t \bar{\varphi}_s dM_s + 2\langle\bar{\varphi}\cdot M, \lambda\cdot M\rangle_t + N_t^0,$$

and since the martingales N^* and N^0 are orthogonal to M and $\langle\bar{\varphi}\cdot M,\lambda\cdot M\rangle$ is of bounded variation, we obtain that the stochastic integral $\bar{\varphi}\cdot X(2)$ is indistinguishable from zero. Hence $\bar{\varphi}=0$ a.e. with respect to the Doléans measure of $\langle M\rangle$, which implies that $\bar{\varphi}\cdot X(2)$ is also the zero process. Therefore, from (4.5) (or (3.46)) we obtain that the density of the variance-optimal martingale measure admits the form (4.2).

Finally, the last equality implies that $N^*=N^0$, and one can write the optimal density in the following alternate form:

$$\mathcal{E}_t(M^{Q^*}) = \mathcal{E}_t\left(-\lambda\cdot M + \int_0^{\cdot}\frac{1}{\hat{c}+\hat{m}_{s-}}d\hat{m}_s\right) = \mathcal{E}_t(-\lambda\cdot M)\left(1+\frac{\hat{m}_t}{\hat{c}}\right).$$

Therefore, the value process is equal to

$$V_t = \frac{E(\mathcal{E}_{tT}^2(-\lambda\cdot M)(c+\hat{m}_T)^2/F_t)}{(c+\hat{m}_t)^2}.$$

Using (4.3) and the fact that $\hat{m}$ is a martingale under $Q(2)$, we obtain that

$$\frac{E(\mathcal{E}_{tT}^2(-\lambda\cdot M)(c+\hat{m}_T)^2/F_t)}{(c+\hat{m}_t)^2} = \frac{e^{-\langle\lambda\cdot M\rangle_t}E^{Q(2)}(c+\hat{m}_T/F_t)}{(c+\hat{m}_t)^2} = \frac{e^{-\langle\lambda\cdot M\rangle_t}}{(c+\hat{m}_t)}$$

$$= \frac{1}{E(\exp\{- < \lambda\cdot M >_{tT}\}/F_t)},$$

and the last term satisfies equation (3.35), since it coincides with the value process. □

Similar results were obtained in Mania, Santacroce, and Tevzadze [18] and Santacroce [21] for the minimal entropy and p-optimal martingale measures, respectively, under condition of continuity of the filtration.

Example. Let W be a standard Brownian motion defined on a complete probability space (Ω, F, F, P) with filtration. Let π be a standard Poisson process with intensity α, independent from W, defined on the same space. Assume that $F = (F_t, t \in [0,T])$ coincides with the P-augmented filtration generated by W and π.

Assume that there are two assets, a stock and a bond, traded on the market. For simplicity the bond price is supposed to be 1 at all times and the stock price dynamics is given by

$$dX_t = X_t(\mu_t dt + \sigma_t dW_t), \quad t \in [0,T]. \tag{4.10}$$

The market coefficients, the process μ of stock appreciation rate and the volatility σ, are progressively measurable with respect to F. We also require that for any $t \in [0,T]$ the volatility is nonsingular almost surely.

Straightforward calculations yield that in this case $\lambda = \mu X^{-1}\sigma^{-2}$,

$$\int_0^t \lambda_s dM_s = \int_0^t \theta_s dW_s, \qquad \langle \lambda \cdot M \rangle_t = \int_0^t \theta_s^2 ds$$

is the mean variance tradeoff, and $\theta = \sigma^{-1}\mu$ is the market price of risk.

By the Itô representation theorem, any locally square integrable martingale L adapted to the filtration F admits an integral representation

$$L_t = \int_0^t \psi_s dW_s + \int_0^t \bar{\psi}_s d(\pi_s - \alpha s) \tag{4.11}$$

for some F-predictable processes $(\psi, \bar{\psi})$ such that $\int_0^T \psi_t^2 dt + \int_0^T \bar{\psi}_t^2 dt < \infty$ a.s.

As before, we denote by $\mathcal{M}^e$ the set of equivalent martingale measures of X. From (2.4) and (4.11) the density of any martingale measure is expressed as

$$Z_t^\nu = \mathcal{E}_t\left(-\int_0^\cdot \theta_s dW_s + \int_0^\cdot \nu_s d(\pi_s - \alpha s)\right), \quad t \in [0,T], \tag{4.12}$$

for some F-predictable process ν such that $\int_0^T \nu_t^2 dt < \infty$. Let

$$\mathcal{K}_2(\sigma) = \{\nu : EZ_T^\nu = 1, E(Z_T^\nu)^2 < \infty\}.$$

Then the subclass $\mathcal{M}_2^e$ of equivalent martingale measures is given by

$$\mathcal{M}_2^e = \{P^\nu : dP^\nu/dP = Z_T^\nu, \nu \in \mathcal{K}_2(\sigma)\} \tag{4.13}$$

and condition (B) is equivalent to $\nu = 0 \in \mathcal{K}_2(\sigma)$.

For simplicity we assume the following.

(C) The mean-variance tradeoff is bounded; i.e., $\int_0^T ||\theta_s||^2 ds \le C$ a.s. for some $C > 0$.

Note that condition (C) implies that the minimal martingale measure exists and satisfies the reverse Hölder $R_2(P)$ inequality, since for any stopping time τ

$$E(\mathcal{E}^2_{\tau T}(-\lambda\cdot M)/F_\tau) = E(\mathcal{E}_{\tau T}(-2\lambda\cdot M)e^{\langle\lambda\cdot M\rangle_{\tau T}}|F_\tau) \le e^C. \tag{4.14}$$

Therefore, conditions (A) and (B*) are satisfied and it follows from Theorem 2 that the martingale measure Q^* is variance-optimal if and only if it is given by

$$M_t^{Q^*} = -\int_0^t \theta_s dW_s - \int_0^t \frac{1}{V_s}\bar{\varphi}_s d(\pi_s - \alpha s), \tag{4.15}$$

where the triple $(V,\varphi,\bar{\varphi})$ uniquely solves the backward stochastic differential equation (BSDE)

$$V_t = V_0 - \int_0^t \left(V_s\theta_s^2 - 2\theta_s\varphi_s - \frac{\alpha}{V_s}\bar{\varphi}_s^2\right) ds \tag{4.16}$$

$$+ \int_0^t \varphi_s dW_s + \int_0^t \bar{\varphi}_s d(\pi_s - \alpha s), \quad V_T = 1,$$

in the class of semimartingales satisfying (3.36).

Now we consider the abovementioned extreme cases for this example.

Case 1 (an "almost complete" model). If the market price of risk is adapted to the filtration F^W generated by the Brownian motion W, i.e., $\theta = \theta(t,W), t\in[0,T])$, then it follows from Example 1 of Pham, Rheinländer, and Schweizer [19] that the variance-optimal martingale measure coincides with the minimal martingale measure.

Let $X(2)$ be the process satisfying

$$dX_t(2) = X_t(2)(2\mu_t dt + \sigma_t dW_t), \quad t\in[0,T], \tag{4.17}$$

and denote by $Q(2)$ the measure defined by $dQ(2) = \mathcal{E}_T(-2\theta\cdot W)dP$. Under $Q(2)$ the process $X(2)$ satisfies the SDE

$$dX_t(2) = X_t(2)\sigma_t dW_t(2), \quad t\in[0,T],$$

where

$$W_t(2) = 2\int_0^t \theta(s,W)ds + W_t$$

is the Brownian motion with respect to the measure $Q(2)$ by Girsanov's theorem.

If (as in Example 1 of Pham, Rheinländer, and Schweizer [19]) the market price of risk θ is F^W-adapted, then by the representation property of W (and thus of $W(2)$ due to the Bayes rule) the random variable $e^{\int_0^T\theta_s^2 ds}$ admits an integral representation

$$e^{\int_0^T\theta_s^2 ds} = c + \int_0^T \gamma_s dW_s(2) \tag{4.18}$$

for some F^W-predictable γ with $\int_0^T\gamma_t^2 dt < \infty$. Note that if (4.18) is satisfied by some F-predictable γ, then the market price of risk θ need not be F^W-adapted, but condition (4.1) is satisfied (with $h = \gamma/\sigma$) and according to Proposition 4

the variance-optimal martingale measure coincides with the minimal martingale measure.

Case 2. Assume that the market price of risk is adapted to the filtration F^π generated by the Poisson process π, i.e.,

$$\theta = (\theta(t,\pi), t \in [0,T]).$$

Since θ is F^π-adapted, by the integral representation theorem there exists an F^π-adapted process g such that

$$\exp\left\{-\frac{1}{2}\int_0^T \theta_s^2 ds\right\} = c + \int_0^T g_s d(\pi_s - \alpha s).$$

Therefore, condition (4.3) is satisfied, and Proposition 5 implies that the variance-optimal martingale measure is of the form (4.2). Moreover, condition (4.3) is satisfied if g is F-predictable, and such a representation is also necessary for the variance-optimal martingale measure to admit the form (4.2).

Acknowledgment. We would like to thank the referees for valuable remarks and suggestions which led to many improvements in the paper.

REFERENCES

[1] J. P. Ansel and C. Stricker, *Couverture de actifs contingents et prix maximum*, Ann. Inst. H. Poincaré Probab. Statist., 30 (1994), pp. 303–315.
[2] F. Biagini, P. Guasoni, and M. Pratelli, *Mean variance hedging for stochastic volatility models*, Math. Finance, 10 (2000), pp. 109–123.
[3] R. J. Chitashvili, *Martingale ideology in the theory of controlled stochastic processes*, in Probability Theory and Mathematical Statistics, Lecture Notes in Math. 1021, Springer-Verlag, Berlin, 1983, pp. 73–92.
[4] R. J. Chitashvili and M. Mania, *Optimal locally absolutely continuous change of measure: Finite set of decisions* II, Stochastics, 21 (1987), pp. 187–229.
[5] F. Delbaen and W. Schachermayer, *Variance-optimal martingale measure for continuous processes*, Bernoulli, 2 (1996), pp. 81–105.
[6] C. Dellacherie and P. A. Meyer, *Probabilités et potentiel* II, Hermann, Paris, 1980.
[7] F. Delbaen, P. Monat, W. Schachermayer, W. Schweizer, and C. Stricker, *Weighted norm inequalities and hedging in incomplete markets*, Finance Stoch., 1 (1997), pp. 181–227.
[8] C. Doléans-Dade and P. A. Meyer, *Inegalites de normes avec poinds*, in Seminaire de Probabilites XIII, Lecture Notes in Math. 721, Springer-Verlag, Berlin, 1979, pp. 313–331.
[9] N. El Karoui and M.-C. Quenez, *Dynamic programming and pricing of contingent claims in an incomplete market*, SIAM J. Control Optim., 33 (1995), pp. 29–66.
[10] R. J. Elliott, *Stochastic Calculus and Applications*, Springer-Verlag, Berlin, 1982.
[11] C. Gouriéroux, J. P. Laurent, and H. Pham, *Mean-variance hedging and numeraire*, Math. Finance, 8 (1998), pp. 179–200.
[12] J. Jacod, *Calcul Stochastique et Problèmes des Martingales*, Lecture Notes in Math. 714, Springer-Verlag, Berlin, 1979.
[13] N. Kazamaki, *A sufficient condition for the uniform integrability of exponential martingales*, Math. Rep. Toyama Univ., 2 (1979), pp. 1–11.
[14] J. P. Laurent and H. Pham, *Dynamic programming and mean-variance hedging*, Finance Stoch., 3 (1999), pp. 83–110.
[15] R. Sh. Liptser and A. N. Shiryayev, *Martingale Theory*, Nauka, Moscow, 1986.
[16] M. Mania, *A general problem of an optimal equivalent change of measure and contingent claim pricing in an incomplete market*, Stochastic Process. Appl., 90 (2000), pp. 19–42.
[17] M. Mania and R. Tevzadze, *A semimartingale Bellman equation and the variance-optimal martingale measure*, Georgian Math. J., 7 (2000), pp. 765–792.
[18] M. Mania, M. Santacroce, and R. Tevzadze, *A semimartingale backward equation and the Bellman equation related to the minimal entropy martingale measure*, Finance Stoch., 7 (2003), pp. 385–402.

[19] H. Pham, T. Rheinländer, and M. Schweizer, *Mean-variance hedging for continuous processes: New proofs and examples*, Finance Stoch., 2 (1998), pp. 173–198.

[20] T. Rheinländer, *Optimal Martingale Measures and Their Applications in Mathematical Finance*, Zur Erlangung des Akademischen Grades eines Doktors der Naturwissenschaften, Technische Universität Berlin, Berlin, Germany, 1999.

[21] M. Santacroce, *Semimartingale Backward Equation for the p-Optimal Martingale Measure: Some Extreme Cases*, Tech. report, Dip. di Statistica, Probabilita' e Statistiche Applicate, Univ. degli Studi di Roma "La Sapienza," Rome, Italy, 2001.

[22] M. Schweizer, *Approximating random variables by stochastic integrals*, Ann. Probab., 22 (1994), pp. 1536–1575.

[23] M. Schweizer, *Approximation pricing and the variance-optimal martingale measure*, Ann. Probab., 24 (1996), pp. 206–236.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1727–1744

LIPSCHITZ CONTINUITY OF OPTIMAL CONTROLS FOR STATE CONSTRAINED PROBLEMS*

GRANT N. GALBRAITH† AND RICHARD B. VINTER‡

Abstract. This paper provides new conditions under which optimal controls are Lipschitz continuous for dynamic optimization problems with functional inequality constraints, a control constraint expressed in terms of a general closed convex set and a coercive cost function. It is shown that the linear independence condition on active state constraints, present in the earlier literature, can be replaced by a less restrictive, positive linear independence condition that requires linear independence merely with respect to nonnegative weighting parameters. Smoothness conditions on the data, imposed in earlier work, are also relaxed. The new conditions for Lipschitz continuity of optimal controls are obtained by a detailed analysis of the implications of first order optimality conditions in the form of a nonsmooth maximum principle.

Key words. optimal control, Lipschitz controls, normal necessary conditions

AMS subject classifications. Primary, 49N60, 49J30; Secondary, 49J52

DOI. 10.1137/S0363012902404711

1. Introduction.

Consider the following optimal control problem with functional inequality state constraints, distinguishing features of which are that the controlled differential equation (the dynamic constraint) is linear in the control variable and that an integral term is included in the cost function:

$$(\mathcal{P})\quad \begin{cases} \text{Minimize } l(x(S),x(T)) + \int_S^T L(t,x(t),u(t))\,dt \\ \text{over } x \in W^{1,1}([S,T],\mathbb{R}^n) \text{ and measurable } u:[S,T]\to\mathbb{R}^m \\ \text{satisfying} \\ \dot{x}(t) = f(t,x(t)) + G(t,x(t))u(t) \quad \text{for a.e. } t\in[S,T], \\ h_j(t,x(t)) \le 0 \quad \text{for all } t\in[S,T],\ j=1,\dots,r, \\ u(t)\in U \quad \text{for a.e. } t\in[S,T], \\ (x(S),x(T))\in C, \end{cases} \tag{1.1}$$

with data an interval $[S,T]$, functions $L:[S,T]\times\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$, $f:[S,T]\times\mathbb{R}\to\mathbb{R}^n$, $G:[S,T]\times\mathbb{R}^n\to\mathbb{R}^{n\times m}$, $h_j:\mathbb{R}\times\mathbb{R}^n\to\mathbb{R}$ for $j=1,\dots,r$, and closed sets $U\subset\mathbb{R}^m$ and $C\subset\mathbb{R}^n\times\mathbb{R}^n$.

(Here, $W^{1,1}([S,T],\mathbb{R}^n)$, abbreviated as $W^{1,1}$, is the space of absolutely continuous $\mathbb{R}^n$-valued functions on $[S,T]$.)

A control function is a measurable function $u:[S,T]\to\mathbb{R}^m$ such that $u(t)\in U$ for a.e. $t\in[S,T]$. A process (x,u) comprises a control function u and a $W^{1,1}$ function x satisfying the constraints of $(\mathcal{P})$. We say the process $(\bar{x},\bar{u})$ is a minimizer if it achieves the minimum. In this case, $\bar{u}$ and $\bar{x}$ are referred to as an optimal control and an optimal state trajectory (corresponding to $\bar{u}$), respectively.

This paper focuses on conditions on the data for the above control problem that guarantee Lipschitz continuity of the optimal control $\bar{u}$. The issue of minimizer regularity is important for several reasons. One is its relevance to computations; prior

*Received by the editors March 27, 2002; accepted for publication (in revised form) March 8, 2003; published electronically November 14, 2003.
http://www.siam.org/journals/sicon/42-5/40471.html
†Mathematics Department, LSE, Houghton Street, London WC2A 2AE, UK (G.Galbraith@lse.ac.uk).
‡Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, Exhibition Road, London, UK, SW7 2AZ (r.vinter@ic.ac.uk).

knowledge of minimizer regularity influences the choice of discretization procedures since, typically, higher order schemes can achieve improved rates of convergence only when minimizers are sufficiently regular [6]. It also affects the selection of sample period in digital implementation of control strategies. It is further relevant to physical modeling, where a variational formulation of the underlying dynamics must be matched to observed phenomena, including regularity [1].

A key advance in the quest for conditions ensuring Lipschitz continuity of optimal controls in the presence of both state and control functional inequality constraints was provided in Hager's 1979 paper [7]. Here, Lipschitz continuity was established under hypotheses that, in the special case when no control constraints are imposed, include the following:

(i) The data is of class C^2, the cost integrand is jointly convex in both the (x, u) variables and uniformly coercive in the u variable, and the dynamics are affine with respect to the (x, u) variables.

(ii) There exists a process (x, u) such that u is continuous and x and u lie in the interiors of, respectively, the state and control constraint sets for each time ("interiority"), and $C = C_0 \times \mathbb{R}^n$ for some $C_0 \subset \mathbb{R}^n$(no right endpoint constraint).

(iii) There exists $\gamma > 0$ such that, for each $t \in [S, T]$,

$$\left| G^T(t, \bar{x}(t)) \sum_j \alpha_j \nabla_x h_j(t, \bar{x}(t)) \right| \geq \gamma \left(\sum_j |\alpha_j|^2 \right)^{\frac{1}{2}},$$

where $\nabla_x h_j$ is interpreted as a column vector and the summations are taken over values of the index j for which the state constraint is active ("linear independence of active state constraints").

Lipschitz continuity of optimal controls under these hypotheses was established in [7] by consideration of the implications of the maximum principle for optimal control problems with an affine state equation, a convex cost, and convex functional inequality constraints.

Malanowski [8] refined Hager's analysis to establish Lipschitz continuity of optimal controls under less restrictive conditions that allow dynamics nonlinear with respect to the state variable and a cost integrand which is possibly nonconvex with respect to the state variable. Alternative proofs and additional regularity properties of optimal controls under certain circumstances ("piecewise analyticity") were later proved by Dontchev and Hager [3] and Dontchev and Kolmanovsky [4].

This paper establishes Lipschitz continuity of optimal controls under hypotheses that are less restrictive that those invoked hitherto in a number of respects. The most significant improvement is that the linear independence hypothesis (iii) of Hager, present in different forms in [3], [8], is replaced by a less demanding *positive* linear independence hypothesis on the state constraints (hypothesis (H4) below). We also allow a general convex constraint on the control variable ("$u(t) \in U$ for some closed convex set U") in place of the collection of functional inequality constraints in previous work, and we relax differentiability hypotheses on the data in a number of respects.

The positive linear independence hypothesis that we employ has previously arisen in connection with conditions for normality of multiplier sets in nonlinear programming; specifically it provides a dual formulation of the Mangasarian–Fromowitz constraint qualification (see [9]). However, consideration of positive linear independence, in the context of optimal control regularity analysis, appears to be new.

The conditions for Lipschitz continuity of optimal controls are obtained with the help of a more detailed analysis of the implications of the maximum principle (now in a nonsmooth manifestation) than has previously been undertaken. A key step (Lemma 4.6) is to consider the properties of trajectory subarcs with the property that all state constraints active at some intermediate time are active also at the endtimes; the significance of such subarcs for regularity investigations was earlier emphasized by Hager [7, Thm. 2.1]. The analysis greatly simplifies if the control constraints are absent, the cost is quadratic in the u variable, and there is only one state constraint. (See [12, Ch. 11].)

We highlight also the role of the Legendre–Fenchel transform in our analysis. This provides an important explicit representation of the optimal control (see (4.1) below) in terms of the costate variables and state constraint multipliers.

Research efforts following Hager's 1979 paper were directed, in part, toward assembling a set of hypotheses ensuring regularity of minimizers, uniqueness of multipliers, and smooth dependence on parameters and constructing a framework for numerical solution techniques involving "dual" concepts. The present paper is of narrower focus, concentrating exclusively on conditions for regularity of optimal controls. If this alone is our goal, then the linear independence hypotheses of the earlier literature can be relaxed to positive linear independence. Note, however, that, under this positive linear independence hypothesis, the state constraint multipliers may fail to be unique.

Finally, we give some notation. $|\,.\,|$ denotes the Euclidean norm. The closed unit ball in Euclidean space is written B. $C^{\oplus}(S,T)$ denotes the space of nonnegative Borel measures on the Borel subsets of $[S,T]$. For a given subset $A \subset \mathbb{R}^k$, Ψ_A denotes the indicator function

$$\Psi_A(y) = \begin{cases} 0 & \text{if} \quad y \in A, \\ +\infty & \text{otherwise.} \end{cases}$$

We make use of two standard constructs from nonsmooth analysis (see, for example, [11] for full details), the normal cone and the subgradient, defined as follows.

DEFINITION 1.1. *Take a closed set $C \subset \mathbb{R}^n$ and a point $\bar{x} \in C$. We say that $y \in \mathbb{R}^n$ is a normal to C at $\bar{x}$ if there exist $y_i \to y$ and $x_i \to \bar{x}$ (in C) such that for all i,*

$$\langle y_i, x - x_i \rangle \le o(|x - x_i|)$$

for all $x \in C$. The normal cone to C at $\bar{x}$, written $N_C(\bar{x})$, is the set of all normals to C at $\bar{x}$. (It is also referred to as the limiting normal cone*.)*

Given a lower semicontinuous (lsc) function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, we denote by $\partial f(\bar{x})$ the subgradient of f at $\bar{x}$ (also known as the limiting subgradient*), defined as*

$$\partial f(\bar{x}) := \{y : (y, -1) \in N_{\operatorname{epi} f}(\bar{x}, f(\bar{x}))\},$$

in which epi f *denotes the set* $\{(x,\alpha) \in \mathbb{R}^n \times \mathbb{R} : \alpha \ge f(x)\}$.

2. The maximum principle and normality. Denote by $\mathcal{H} : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ the unmaximized Hamiltonian

$$\mathcal{H}(t,x,p,u,\lambda) = \langle p, f(t,x) + G(t,x)u \rangle - \lambda L(t,x,u). \tag{2.1}$$

Let $(\bar{x}, \bar{u})$ be a minimizing process. Under mild hypotheses and, in particular, under hypotheses (H1)–(H3) of section 3, necessary conditions of optimality, known as the

(state constrained) maximum principle [12], provide the following information about $(\bar{x}, \bar{u})$.

There exist "multipliers" $p \in W^{1,1}([S,T]; \mathbb{R}^n)$, $\mu_j \in C^{\oplus}(S,T)$ for $j = 1, \ldots, r$, and $\lambda \geq 0$ such that, writing

$$q(t) \;=\; p(t) + \sum_{j=1}^{r} \int_{[S,t)} \nabla_x h_j(s, \bar{x}(s)) \mu_j(ds), \tag{2.2}$$

we have

$$(p, \mu, \lambda) \;\neq\; (0,0,0), \tag{2.3}$$

$$-\dot{p}(t) \;\in\; \operatorname{con} \partial_x \mathcal{H}(\, t,\, \bar{x}(t),\, q(t^-),\, \bar{u}(t),\, \lambda\,) \text{ a.e. } t \in [S,T], \tag{2.4}$$

$$\begin{aligned} \mathcal{H}(\, t,\, \bar{x}(t),\, q(t^-),\, \bar{u}(t),\, \lambda\,) \;&=\; \max_{u \in U} \mathcal{H}(\, t,\, \bar{x}(t),\, q(t^-),\, u,\, \lambda\,) \\ &\text{a.e. } t \in [S,T], \end{aligned} \tag{2.5}$$

$$\operatorname{supp}\{\mu_j\} \;\subset\; \{t \;:\; h_j(t, \bar{x}(t)) = 0\} \quad \textit{for } j = 1, \ldots, r, \tag{2.6}$$

$$\begin{aligned} &\left(p(S), - \left[p(T) + \sum_{j=1}^{r} \int_{[S,T]} \nabla_x h_j(t, \bar{x}(t)) \mu_j(dt) \right] \right) \\ &\qquad \in \lambda \partial l(\bar{x}(S), \bar{x}(T)) + N_C(\bar{x}(S), \bar{x}(T)). \end{aligned} \tag{2.7}$$

A process for which these conditions are satisfied is said to be an *extremal*.

The methodology behind the ensuing analysis is to deduce regularity properties of optimal controls from the conditions of the maximum principle. It is inevitable then that some kind of hypothesis on the data for problem $(\mathcal{P})$ is imposed, ensuring that the maximum principle supplies useful information about the minimizer $(\bar{x}, \bar{u})$. This hypothesis is *normality*. If it is possible to satisfy the conditions of the maximum principle with a set of multipliers $(p, \mu_1, \ldots, \mu_r, \lambda)$ in which $\lambda = 0$, the maximum principle makes no reference to the cost function and degenerates into a relationship between the constraints. "Normality" means that this kind of degeneracy is excluded.

DEFINITION 2.1. *A process $(\bar{x}, \bar{u})$ is said to be a normal extremal if there exist $p \in W^{1,1}([S,T]; \mathbb{R}^n)$ and $\mu_j \in C^{\oplus}(S,T)$, $j = 1, \ldots, r$, such that the relationships (2.2)–(2.7) are satisfied with $\lambda = 1$.*

We address the question of when minimizers are normal extremals in section 5.

3. Conditions for Lipschitz continuity of normal extremals. We shall invoke the following hypotheses; reference is made here to the process $(\bar{x}, \bar{u})$ of interest. In the hypotheses, $\Omega \subset [S,T] \times \mathbb{R}^n$ is some "tube" about $\bar{x}$. That is,

$$\Omega \;=\; \{(t,x) \in [S,T] \times \mathbb{R}^n \;:\; |x - \bar{x}(t)| \leq \bar{\varepsilon}\}$$

(for some given $\bar{\varepsilon} > 0$). We denote by $\mathcal{J}(t, \bar{x})$ the collection of active state constraints at time t; that is,

$$\mathcal{J}(t, \bar{x}) \;=\; \{j \;:\; h_j(t, \bar{x}(t)) = 0\}.$$

(H1) *G, l, and f are locally Lipschitz continuous functions.*

(H2) *For $j = 1, \dots, r$, h_j is of class $\mathcal{C}^{1+}$ on Ω; i.e., h_j is continuously differentiable with locally Lipschitz continuous gradient.*

(H3) *U is a closed convex set. For each $(t, x) \in \Omega$, $L(t, x, .)$ is finite-valued, convex, and continuously differentiable. $L(t, x, .)$ is uniformly coercive in the sense that there exists a monotone function $\theta : [0, \infty) \to \mathbb{R}$, such that $\theta(s)/s \to \infty$ as $s \to \infty$ and*

$$L(t, x, v) \ > \theta(|v|) \quad \textit{for all } (t, x) \in \Omega \textit{ and } v \in U.$$

Both L and $\nabla_u L$ are locally Lipschitz continuous. $L(t, x, .)$ is strictly convex in the following uniform sense: for each $M > 0$ there is a constant $k_M > 0$, such that, for any $(t, x) \in \Omega$ and $u_1, u_2 \in M\mathbb{B}$, we have

$$\langle y_2 - y_1, u_2 - u_1 \rangle \geq k_M |u_2 - u_1|^2, \tag{3.1}$$

where $y_2 = \nabla_u L(t, x, u_2)$ and $y_1 = \nabla_u L(t, x, u_1)$.

(H4) *For every $t \in [S, T]$ and every set of nonnegative numbers $\{\alpha_j\}_{j \in \mathcal{J}(t, \bar{x})}$, not all zero, we have*

$$\sum_{j \in \mathcal{J}(t, \bar{x})} \alpha_j G^T(t, \bar{x}(t)) \nabla_x h_j(t, \bar{x}(t)) \notin \textit{span}\, N_U(\bar{u}(t)).$$

(For a subset $D \subset \mathbb{R}^k$, span D denotes the intersection of all linear subspaces of $\mathbb{R}^k$ that contain D.)

The stage is now set for statement of conditions for Lipschitz continuity of optimal controls.

THEOREM 3.1. *Let $(\bar{x}, \bar{u})$ be a normal extremal. Assume* (H1)–(H4). *Then $\bar{u}$ is Lipschitz continuous.*

Comments.

(a) Of course interest focuses primarily on cases when optimal processes are normal extremals, for then Theorem 3.1 gives conditions for Lipschitz continuity of optimal controls. We discuss conditions for normality in section 5. Note, however, that as far as applications to Hamiltonian mechanics are concerned, normal extremals (and related issues of regularity) are of direct interest, since the action principle interprets motions as normal extremals, which may fail to be minimizers of the action functional.

(b) The key difference between the hypotheses of Theorem 3.1 and those formerly invoked for regularity of optimal controls concerns the "nondegeneracy" of the state constraints. The linear independence hypothesis of [7] (condition (iii) of section 1) has been replaced by the positive linear independence hypothesis (H4). (H4) is a less restrictive hypothesis in which nonzero linear combinations of active state constraint function gradients are required to be nonzero *only for linear combinations with nonnegative weights.* A simple case when (iii) is always violated, but (H4) is possibly satisfied, is when there are two state constraint functions such that, at some time t' when they are both active, we have $\nabla_x h_1(t', \bar{x}(t')) = \alpha \nabla_x h_2(t', \bar{x}(t'))$ for some $\alpha > 0$. Another case is when the number of active state constraints exceeds the dimension of

the state space; here the gradients of the state constraint functions cannot be linearly independent, but they will be positively linear independent if the gradients are, in some sense, "unidirectional."

(c) Suppose that the cost integrand L can be decomposed as

$$L(t,x,u) = L_1(t,x) + L_2(t,x,u)\,.$$

Then the analysis of this paper, almost without change, allows us to deduce Lipschitz continuity of optimal controls when L_2 satisfies (H3) and L_1 satisfies the following condition: $L_1(t,x)$ is locally bounded, measurable in t for each x, and locally Lipschitz continuous in x uniformly in t. We draw attention to this refinement, since the optimal control problems with quadratic cost integrand

$$L(t,x,u) = x^T Q(t)x + u^T R(t)u$$

are of widespread interest. Our analysis establishes Lipschitz continuity of optimal controls for such problems when $Q(.)$ is merely measurable and essentially bounded. ($R(.)$ is required to be Lipschitz continuous and such that $R(t)$ is positive definite for all t.)

4. Proof of Theorem 3.1. Define the extended-real-valued function $L_0 : [S,T]\times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}\cup\{+\infty\}$

$$L_0(t,x,u) \;=\; L(t,x,u) + \Psi_U(u),$$

in which Ψ_U is the indicator function of the set U. Note that, since

$$\max_{u\in U}\mathcal{H}(t,\bar{x}(t),q(t^-),1) \;=\; \langle q(t^-), f(t,\bar{x}(t))\rangle + \max_{u\in\mathbb{R}^m}\{\langle q(t^-), G(t,\bar{x}(t))u\rangle - L_0(t,\bar{x}(t),u)\}\,,$$

we have from the "maximization of the Hamilitonian" condition (2.5) that

$$\begin{aligned}&\langle G^T(t,\bar{x}(t))q(t^-),\bar{u}(t)\rangle - L(t,\bar{x}(t),\bar{u}(t))\\ &\quad= \max_{u\in\mathbb{R}^m}\{\langle G^T(t,\bar{x}(t))q(t^-),u\rangle - L_0(t,\bar{x}(t),u)\} \quad \text{a.e. } t\in[S,T].\end{aligned}$$

By the rules governing subdifferentials of convex functions, this last condition implies that

$$\bar{u}(t) \;=\; \partial_y L_0^*(t,\bar{x}(t),G^T(t,\bar{x}(t))q(t^-)) \quad \text{a.e. } t\in[S,T]. \tag{4.1}$$

Here, $L_0^*(t,x,.):\mathbb{R}^m\to\mathbb{R}\cup\{+\infty\}$ is the Fenchel dual function of $L_0(t,x,.)$ for each (t,x).

The representation (4.1) of the optimal control in terms of the Fenchel dual function L_0^* has a crucial role in the following analysis. We pause to investigate some of its properties.

LEMMA 4.1.

(i) *For each $(t,x,y)\in\Omega\times\mathbb{R}^m$, $\partial_y L_0^*(t,x,y)$ is single-valued and continuously differentiable. (Write it henceforth $\nabla_y L_0^*(t,x,y)$.)*

(ii) *$(t,x,y)\to\nabla_y L_0^*(t,x,y)$ is locally Lipschitz continuous.*

Proof. Take any $(t,x)\in\Omega$ and $y\in\mathbb{R}^m$. The nonemptiness of $\partial_y L_0^*(t,x,y)$ follows from the representation of the subdifferential

$$\partial_y L_0^*(t,x,y) \;=\; \left\{u \;:\; u\cdot y - L_0(t,x,u) = \max_{v\in\mathbb{R}^m}\{v\cdot y - L_0(t,x,v)\}\right\}$$

and the coercivity of L (see hypothesis (H3)), which ensures existence of a maximizing v. Take any compact neighborhood $\mathcal{N}$ of (t,x) and a number $N>0$. Representation (4.1) and hypothesis (H3) also ensure the existence of $M>0$ such that

$$u' \in \partial_y L_0^*(t',x',y') \quad \text{and} \quad (t'x') \in \Omega,\ y' \in y + NB$$

imply

$$|u'| \ \leq \ M\,. \tag{4.2}$$

Let k_1 be a Lipschitz constant for L on $\mathcal{N} \times MB$. Choose arbitrary $(t',x') \in \mathcal{N}$ and $y' \in y + NB$. Choose also

$$u \in \partial_y L_0^*(t,x,y) \quad \text{and} \quad u' \in \partial_y L_0^*(t',x',y')\,.$$

By a fundamental property of "convex" subdifferentials,

$$y \in \partial_u L_0(t,x,u) \quad \text{and} \quad y' \in \partial_u L_0(t,x',u')\,.$$

However, since $\nabla_u L(t,x,.)$ is continuously differentiable,

$$\partial_u L_0(t,x,u) \ = \ \nabla_u L(t,x,u) + N_U(u)\,.$$

It follows that

$$y = \nabla_u L(t,x,u) + e, \quad y' = \nabla_u L(t',x',u') + e'$$

for some $e \in N_U(u)$ and $e' \in N_U(u')$. In consequence of the local Lipschitz continuity of $\nabla_u L$ and (4.2), there exists $k_1>0$, independent of the choice of t',x',y',u', such that

$$|\nabla_u L(t,x,u') - \nabla_u L(t',x',u')| \ \leq \ k_1 |(t',x') - (t,x)|\,.$$

Let $\tilde{y} = \nabla_u L(t,x,u') + e'$. Then

$$|y' - \tilde{y}| \ \leq \ k_1 |(t',x') - (t,x)|\,. \tag{4.3}$$

We have

$$\begin{aligned} |\tilde{y} - y||u' - u| \ \geq \ \langle \tilde{y} - y, u' - u\rangle &= \langle \nabla_u L(t,x,u') - \nabla_u L(t,x,u), u' - u\rangle \\ &\quad + \langle e', u' - u\rangle \ + \ \langle e, u - u'\rangle\,. \end{aligned}$$

However, there exists $k_2 > 0$ (independent of our choice of $((t',x'),y')$ in $\mathcal{N} \times (y + NB)$) such that

$$\langle \nabla_u L(t,x,u') - \nabla_u L(t,x,u), u' - u\rangle \ \geq \ k_2 |u' - u|^2\,.$$

Also, by the definition of the "convex" normal cone

$$\langle e', u' - u\rangle \geq 0 \quad \text{and} \quad \langle e, u - u'\rangle \geq 0\,.$$

It follows that

$$|u' - u| \ \leq \ k_2^{-1} |\tilde{y} - y|\,.$$

By (4.3) and the triangle inequality

$$\begin{aligned}|u'-u| &\le k_2^{-1}(|y'-y|+|y'-\tilde{y}|)\\ &\le k_2^{-1}(|y'-y|+k_1|(t',x')-(t,x)|)\\ &\le \max\{1,k_1\}k_2^{-1}\sqrt{2}|(t',x',y')-(t,x,y)|\,.\end{aligned}$$

This inequality implies that $\partial_y L_0^*(t,x,y)$ is single-valued. Since a convex function with a single-valued subdifferential is continuously differentiable, $\partial_y L_0^*(t,x,.)$ is continuously differentiable. The above inequality also implies that $(t,x,y)\to\partial_y L_0^*(t,x,y)$ is locally Lipschitz continuous.

LEMMA 4.2. *There exist $\bar{k}>0$ and $\varepsilon>0$ such that, for any $t\in[S,T]$, $y\in\mathbb{R}^m$, and $\{\alpha_j\}_{j=1}^r$ such that*

(4.4)
$\alpha_j\ge 0$ *for each j and $\alpha_j=0$ if $h_j(t',\bar{x}(t'))<0$ for all $t'\in(t-\varepsilon,t+\varepsilon)\cap[S,T]$,*

we have

$$\langle v,\nabla_y L_0^*(t,\bar{x}(t),y+v)-\nabla_y L_0^*(t,\bar{x}(t),y)\rangle \;\ge\; \bar{k}\left|\sum_j\alpha_j\right|^2,$$

where

$$v \;=\; G^T(t,\bar{x})\sum_j\alpha_j\nabla_x h_j(t,\bar{x}(t))\,.$$

Proof. To simplify notation, we shall write $\nabla_y L_0^*(t,x,y+v)$ as $\nabla_y L_0^*(y+v)$ and suppress the argument (t,x) in expressions involving $G(t,x)$, etc.

Take any $\{\alpha_j\}$ such that $\alpha_j\ge 0$ for each j. Write

$$u'=\nabla_y L_0^*(y+v)\quad\text{and}\quad u=\nabla_y L_0^*(y),$$

where v is as in the lemma. Then

$$y+v\in\partial_u L_0(u')\quad\text{and}\quad y\in\partial_u L_0(u)\,.$$

Since L is continuously differentiable (with respect to the control variable),

$$y+v=\nabla_u L(u')+e'\quad\text{and}\quad y=\nabla_u L(u)+e \tag{4.5}$$

for some $e'\in N_U(u')$ and $e\in N_U(u)$.

By the strong convexity hypothesis (H3), there exists $k_1>0$, independent of our choice of t, y and $\{\alpha_j\}$, such that

$$\langle\nabla_u L(u')-\nabla_u L(u),u'-u\rangle \;\ge\; k_1|\nabla_u L(u')-\nabla_u L(u)|^2\,.$$

From (4.5)

$$\langle v,u'-u\rangle+\langle e',u-u'\rangle+\langle e,u'-v\rangle \;\ge\; k_1\left|\sum_j\alpha_j G^T\nabla_x h_j-e'+e\right|^2.$$

By properties of the (convex) normal cone

$$\langle e', (u - u')\rangle \le 0 \quad \text{and} \quad \langle e, (u' - u)\rangle \le 0\,.$$

Further, it can be deduced from the constraint qualification (H4) that there exist k_2 and $\varepsilon > 0$, independent of our choice of (t, x, y), e', e and $\{\alpha_j\}$ satisfying (4.4), such that

$$\left|\sum_j \alpha_j G^T \nabla_x h_j + e' - e\right| \;\ge\; k_2 \left|\sum_j \alpha_j\right|\,.$$

Assembling these inequalities, we conclude that

$$\langle v, u' - u\rangle \;\ge\; \bar{k}\left|\sum_j \alpha_j\right|^2\,,$$

where $\bar{k} = k_1 k_2^2$. This is what the lemma asserts.

The following lemma, stated without proof, is a direct consequence of Lemma 4.1, the representation of $\bar{u}$ given by (4.1), the fact that $q(.)$ is a function of bounded variation, and the maximum principle conditions.

LEMMA 4.3. *We can choose $\bar{u}$ (from the equivalence class of a.e. equal functions) to have left and right limits at all points in (S, T) and one sided limits at the endpoints. (This version of) $\bar{u}$ is a bounded function. The functions $\bar{x}$ and p are Lipschitz continuous.*

Next we establish that μ has no atoms at interior points, and we list some related properties.

LEMMA 4.4. *μ has no atoms in (S, T). Consequently, $q(.)$ is continuous on (S, T) and has one sided limits at its endpoints. $\bar{u}$ is continuous on $[S, T]$ (strictly speaking, has a continuous version). For each $t \in (S, T)$ and $j \in \mathcal{J}(t, \bar{x})$, we have*

$$\nabla_t h_j(t, \bar{x}(t)) + \langle \nabla_x h_j(t, \bar{x}(t)), f(t, \bar{x}(t)) + G(t, \bar{x}(t))\bar{u}(t)\rangle \;=\; 0\,. \tag{4.6}$$

Proof. Take any $t \in (S, T)$. Choose $j \in \{1, 2, \ldots, r\}$. If $j \notin \mathcal{J}(t, \bar{x})$, then $\mu_j(\{t\}) = 0$, by the complementary slackness condition (2.6). If, on the other hand, $j \in \mathcal{J}(t, \bar{x}(t))$, then $h_j(t, \bar{x}(t)) = 0$. It follows that

$$\delta^{-1}(h_j(t + \delta, \bar{x}(t + \delta)) - h_j(t, \bar{x}(t))) \;\le\; 0$$

and

$$\delta^{-1}(h_j(t, \bar{x}(t)) - h_j(t - \delta, \bar{x}(t - \delta))) \;\ge\; 0$$

for δ sufficiently small. Passing to the limit as $\delta \downarrow 0$ and recalling that $\bar{u}$ has left and right limits, we obtain

$$\nabla_t h_j + \langle \nabla_x h_j, f + G\bar{u}(t^+)\rangle \;\le\; 0 \tag{4.7}$$

and

$$\nabla_t h_j + \langle \nabla_x h_j, f + G\bar{u}(t^-)\rangle \;\ge\; 0\,. \tag{4.8}$$

(Here, h_j, f, etc. are evaluated at $(t, \bar{x}(t))$.)

We deduce from these inequalities that

$$\langle \nabla_x h_j, G(\bar{u}(t^+) - \bar{u}(t^-)) \rangle \leq 0 .$$

Noting (4.1) and appropriately weighting and summing this inequality over all j's in $\mathcal{J}(t, \bar{x})$ give

$$\left\langle \sum_{j \in \mathcal{J}(t,\bar{x})} \mu_j(\{t\}) G^T \nabla_x h_j, \nabla_y L_0^*(G^T q(t^+)) - \nabla_y L_0^*(G^T q(t^-)) \right\rangle \leq 0 .$$

By Lemma 4.2, however, there exists $k_1 > 0$ such that

$$\left\langle \sum_{j \in \mathcal{J}(t,\bar{x})} \mu_j(\{t\}) G^T \nabla_x h_j, \nabla_y L_0^*(G^T q(t^+)) - \nabla_y L_0^*(G^T q(t^-)) \right\rangle \geq k_1 \left| \sum_j \mu_j(\{t\}) \right|^2 .$$

It follows that $\sum_j \mu_j(\{t\}) = 0$. We have shown that μ has no atoms in (S, T).

We conclude from the definition of $q(.)$ that $q(.)$ is continuous on (S, T) and has one sided limits at the endpoints. The same is true then of $\bar{u}$, in view of Lemma 4.1. By redefining $\bar{u}$ to take at its endpoint values one sided limits, we can arrange that $\bar{u}$ is continuous. Finally, we observe that (4.6) follows from (4.7) and (4.8).

In view of the preceding lemma, we can unambiguously write $\int_{[s,t]} \mu_j(d\sigma)$ as $\int_s^t \mu_j(d\sigma)$ for any $[s, t] \subset [S, T]$.

The next objective is to find a constant K such that, for any interval $[s, t] \subset [S, T]$, we have $\int_s^t \mu_j(d\sigma) \leq K|t - s|$.

The following lemma establishes such a bound in the special case when $[s, t]$ has the following property: all state constraints that are active at *some* point in the interior of $[s, t]$ are also active at *both* endpoints. To investigate this special case, it is helpful to introduce some additional notation:

$$\mathcal{A}_{[s,t]} := \{j \in \{1, \ldots, r\} : h_j(\tau, \bar{x}(\tau)) = 0 \text{ for some } \tau \in (s, t)\} .$$

(The right side will be recognized as "the set of indices corresponding to state constraints that are active at some point in (s, t).")

LEMMA 4.5. *There exists $K > 0$ such that, given any interval $[s, t] \subset (S, T)$ with the property*

$$h_j(s, \bar{x}(s)) = h_j(t, \bar{x}(t)) = 0 \quad \textit{for all } j \in \mathcal{A}_{[s,t]} ,$$

we have

$$\sum_j \int_s^t \mu_j(d\sigma) \leq K|t - s| \quad \textit{for all } j \in \mathcal{A}_{[s,t]}.$$

Proof. Suppose the assertions of the lemma are false. Then there exist $K_i \uparrow \infty$ and intervals $[s_i, t_i] \subset (S, T)$, $i = 1, 2, \ldots$, such that

$$h_j(s_i, \bar{x}(s_i)) = h_j(t_i, \bar{x}(t_i)) = 0 \quad \text{for all } j \in \mathcal{A}_{[s_i,t_i]} \tag{4.9}$$

and

$$\sum_j \int_{s_i}^{t_i} \mu_j(d\sigma) \;\geq\; K_i|t_i - s_i| \quad \text{for all } i\,. \tag{4.10}$$

By extracting a subsequence, we can arrange that there is a fixed index set $\mathcal{A}$ such that $\mathcal{A} = \mathcal{A}_{[s_i,t_i]}$ for all i. Notice that (4.10) implies that $|t_i - s_i| \to 0$ as $i \to \infty$. Since

$$\sum_j \int_{s_i}^{t_i} \mu_j(d\sigma) \;=\; \sum_j \int_{s_i}^{\frac{s_i+t_i}{2}} \mu_j(d\sigma) + \sum_j \int_{\frac{s_i+t_i}{2}}^{t_i} \mu_j(d\sigma)\,,$$

we can arrange, by extraction of a further subsequence, that either of the two cases (a) or (b) occur:

(a) $\sum_j \int_{s_i}^{\frac{s_i+t_i}{2}} \mu_j(d\sigma) \geq \frac{K_i}{2}|t_i - s_i|$ for all i;

(b) $\sum_j \int_{\frac{s_i+t_i}{2}}^{t_i} \mu_j(d\sigma) \geq \frac{K_i}{2}|t_i - s_i|$ for all i.

Let us first assume that (a) is true. For each i define

$$P_i \;=\; \int_{s_i}^{t_i} \sum_j \left(h_j(t_i, \bar{x}(t_i)) - h_j(s, \bar{x}(s))\right) \mu_j(ds)\,.$$

Note, however, that, for each $j \in \mathcal{A}$,

$$\text{supp}\{\mu_j\} \;\subset\; \{s \;:\; h_j(s, \bar{x}(s)) = 0\} \quad \text{and} \quad h_j(t_i, \bar{x}(t_i)) = 0\,.$$

It follows that

$$P_i = 0\,.$$

Writing

$$h_j(t_i, \bar{x}(t_i)) - h_j(s, \bar{x}(s)) \;=\; \int_s^{t_i} \left(\nabla_t h_j + \langle \nabla_x h_j, (f + G(\bar{u}))\rangle\right) dt$$

and carrying out an integration by parts give

$$P_i \;=\; \int_{s_i}^{t_i} \sum_j \int_{s_i}^{t} \mu_j(d\sigma) \left(\nabla_t h_j + \langle \nabla_x h_j, (f + G(\bar{u}(t)))\rangle\right) dt\,.$$

(Here and below, $\nabla_t h_j$, $\nabla_x h_j$, f, and G are evaluated at $(t, \bar{x}(t))$.) However,

$$P_i \;=\; a_i + b_i,$$

where

$$a_i = \int_{s_i}^{t_i} \sum_j \int_{s_i}^{t} \mu_j(ds) \left[\nabla_t h_j + \langle \nabla_x h_j, (f + Gu_i(t))\rangle\right] dt\,,$$

$$b_i = \int_{s_i}^{t_i} \sum_j \int_{s_i}^{t} \mu_j(ds) \langle \nabla_x h_j, G(\bar{u}(t) - u_i(t))\rangle dt\,.$$

In these formulae,

$$u_i \;=\; \nabla_y L_0^* \left(t, \bar{x}(t), G^T \left[p(t) + \int_{[S,t]} \sum_j \nabla_x h_j(s, \bar{x}(s)) \mu_j(ds) \right.\right.$$
$$\left.\left. - \sum_j \nabla_x h_j(t, \bar{x}(t)) \int_{s_i}^t \mu_j(ds) \right] \right) dt\,.$$

Taking note of (4.6) and the local Lipschitz continuity of f, $\nabla_t h_j$, and $\nabla_x h_j$, we deduce from Lemmas 4.1 and 4.3 that there exists $k_1 > 0$ (independent of i) such that, for each j and $t \in [s_i, t_i]$,

$$|\nabla_t h + \langle \nabla_x h_j, (f + G u_i(t)) \rangle| \;\le\; k_1 \left(1 + \sum_j \int_{s_i}^{t_i} \mu_j(ds) \right) (t - s_i).$$

We conclude that

$$a_i \;\ge\; -\frac{k_1}{2} \left(\sum_j \int_{s_i}^{t_i} \mu_j(ds) \right) \left(1 + \sum_j \int_{s_i}^{t_i} \mu_j(ds) \right) (t_i - s_i)^2\,.$$

On the other hand, Lemma 4.2 tells us that there exists $k_2 > 0$, independent of i, such that

$$b_i \ge k_2 \int_{s_i}^{t_i} \left| \sum_j \int_{s_i}^t \mu_j(ds) \right|^2 dt$$
$$\ge k_2 \left| \sum_j \int_{s_i}^{\frac{s_i + t_i}{2}} \mu_j(ds) \right|^2 \left(\frac{t_i - s_i}{2} \right).$$

Bearing in mind that

$$\sum_j \int_{s_i}^{\frac{s_i + t_i}{2}} \mu_j(ds) \;\ge\; \frac{1}{2} \sum_j \int_{s_i}^{t_i} \mu_j(ds) \;\ge\; \frac{1}{2} K_i |t_i - s_i|\,,$$

we deduce that

$$a_i + b_i \;\ge\; -\frac{k_1}{2} K_i |t_i - s_i|^3 (1 + K_i |t_i - s_i|) \;+\; \frac{k_2}{8} K_i^2 (t_i - s_i)^3\,.$$

However, the expression on the right is positive for i sufficiently large. This is not possible since $P_i = a_i + b_i = 0$ for all i. The case (b) is treated in analogous fashion by considering the equation $Q_i = 0$, where

$$Q_i \;=\; \sum_j \int_{s_i}^{t_i} \sum_j \left(h_j(s, \bar{x}(s)) - h_j(s_i, \bar{x}(s_i)) \right) \mu_j(ds)\,.$$

Now define

$$\mathcal{N}_{[s,t]} \;:=\; \text{cardinality}(\mathcal{A}_{[s,t]})\,.$$

For $\bar{r} \in \{0, \dots, r\}$ denote by $(H_{\bar{r}})$ the following condition.

$(H_{\bar{r}})$ There exists $K_{\bar{r}} \geq 0$ with the following property: given any subinterval $[s,t] \subset [S,T]$ such that $\mathcal{N}_{[s,t]} \leq \bar{r}$, we have

$$\sum_j \int_{[s,t]} \mu_j(d\sigma) \ \leq \ K_{\bar{r}}|t-s|.$$

LEMMA 4.6. *Condition $(H_{\bar{r}})$ is satisfied for $\bar{r} = r$.*

Proof. $(H_{\bar{r}})$ is satisfied with $\bar{r} = 0$ since, in this case, $\mu_j = 0$ for all $j \in \{1, \dots, r\}$. Fix $\bar{r} \in \{0, \dots, r-1\}$, and assume that

$$(H_{\bar{r}}) \text{ is true.}$$

We shall show that $(H_{\bar{r}+1})$ is true. The assertions of the lemma then follow by induction.

Take any $[s,t] \subset [S,T]$ such that $\mathcal{N}_{[s,t]} \leq \bar{r}+1$.We must find $K_{\bar{r}+1}$ (independent of $[s,t]$) such that

$$\sum_j \int_{[s,t]} \mu_j(d\sigma) \ \leq \ K_{\bar{r}+1}|t-s| \,. \tag{4.11}$$

We can assume that $\mathcal{N}_{[s,t]} = \bar{r}+1$, for otherwise (4.11) is true with $K_{\bar{r}+1} = K_{\bar{r}}$. Our next goal is to find a point $\bar{s} \in [s,t]$ such that

$$\sum_j \int_{[s,\bar{s}]} \mu_j(d\sigma) \ \leq \ K_{\bar{r}}|\bar{s}-s| \tag{4.12}$$

and either of the following two conditions holds:

(a) $\bar{s} = t$, or

(b) $h_j(\bar{s}, \bar{x}(\bar{s})) = 0$ for all $j \in \mathcal{A}_{[s,t]}$.

If $h_j(s, \bar{x}(s)) = 0$ for all $j \in \mathcal{A}_{[s,t]}$, we can set $\bar{s} = s$, and (4.12) and condition (b) are satisfied. So we can assume that

$$h_j(\bar{s}, \bar{x}(\bar{s})) < 0 \text{ for some } j \in \mathcal{A}_{[s,t]}. \tag{4.13}$$

We now construct an increasing sequence $\{s_i\} \subset (s,t]$ that terminates after N steps, in which case we set $\bar{s} = s_N$, or which is an infinite sequence, in which case we set $\bar{s} = \lim_{i\to\infty} s_i$. In either case, $\bar{s}$ will have the desired properties, as we now confirm.

Define

$$s_1 \ = \ \sup_{\sigma \in (s,t)} \{\sigma \,:\, \mathcal{N}_{[s,\sigma]} \leq \bar{r}\} \,.$$

By condition (4.13), $s_1 > s$. We have

$$\sum_j \int_{[s,s_1]} \mu_j(d\sigma) \ \leq \ K_{\bar{r}}|s_1 - s| \,.$$

If $s_1 = t$, set $\bar{s} = s_1$. Then condition (a) is satisfied, and so is (4.12), by the induction hypothesis. If $s_1 < t$ and

$$h_j(s_1, \bar{x}(s_1)) \ = \ 0 \quad \text{for all } j \in \mathcal{A}_{[s,t]},$$

also set $\bar{s} = s_1$. In this case condition (b) and (4.12) are satisfied. Otherwise, $s_1 < t$ and

$$h_{\bar{j}}(s_1, \bar{x}(s_1)) \ < \ 0 \quad \text{for some } \bar{j} \in \mathcal{A}_{[s,t]}\,.$$

In this case define $s_2 (> s_1)$ to be

$$s_2 \ = \ \sup_{\sigma \in (s_1, t)} \{\sigma \ : \ \mathcal{N}_{[s_1,\sigma]} \leq \bar{r}\}. \tag{4.14}$$

By the induction hypothesis $\sum_j \int_{[s_1,s_2]} \mu_j(ds) \leq \ K_{\bar{r}}|s_2 - s_1|$, from which we conclude that

$$\sum_j \int_{[s,s_2]} \mu_j(d\sigma) \ \leq \sum_j \int_{[s,s_1]} \mu_j(d\sigma) + \sum_j \int_{[s_1,s_2]} \mu_j(ds) \leq \ K_{\bar{r}}|s_2 - s|\,.$$

Observe also that, if $s_2 < t$,

$$\max_{\sigma \in [s_1,s_2]} h_j(\sigma, \bar{x}(\sigma)) \ = \ 0 \quad \text{for all } j \in \mathcal{A}_{[s,t]},$$

for otherwise s_2 cannot provide the supremum in (4.14). If $s_2 = t$, set $\bar{s} = s_2$. In this case (4.12) and condition (a) are satisfied. If $s_2 < t$ and $h_j(s_2, \bar{x}(s_2)) = 0$ for all $j \in \mathcal{A}_{[s,t]}$, set $\bar{s} = s_2$; (4.12) and condition (b) are satisfied.

If neither condition (a) nor (b) are satisfied (when $\bar{s} = s_2$), construct $s_3 \in (s_2, t]$, and so on.

This procedure either provides an element $\bar{s} \in (s, t]$ satisfying (4.12) and either condition (a) or (b) in a finite number of steps or generates an infinite increasing sequence $\{s_i\}$ in $(s, t]$. In the latter case

$$\sum_j \int_{[s,s_i]} \mu_j(d\sigma) \leq \ K_{\bar{r}}|s_i - s| \quad \text{for all } i$$

and

$$\max_{s_i \leq \sigma \leq s_{i+1}} h_j(\sigma, \bar{x}(\sigma)) \ = \ 0 \quad \text{for all } j \in \mathcal{A}_{[s,t]}\,.$$

Let $\bar{s} = \lim_{i \to \infty} s_i$. We have $\bar{s} \in (s, t]$. Furthermore, the preceding relationships ensure that

$$\sum_j \int_{[s,\bar{s}]} \mu_j(d\sigma) \leq \ K_{\bar{r}}|\bar{s} - s| \tag{4.15}$$

and

$$h_j(\bar{s}, \bar{x}(\bar{s})) \ = \ 0 \quad \text{for all } j \in \mathcal{A}_{[s,t]}\,.$$

Similarly, working from the right endpoint of $[s, t]$, we can find $\bar{t} \in [s, t]$ such that

$$\sum_j \int_{[\bar{t},t]} \mu_j(d\sigma) \leq \ K_{\bar{r}}|t - \bar{t}| \tag{4.16}$$

and either

(a′) $\bar{t} = s$ or

(b′) $h_j(\bar{t}, \bar{x}(\bar{t})) = 0$ for all $j \in \mathcal{A}_{[s,t]}$.

If either (a) or (a′) are true, then (4.11) is true with $K_{\bar{r}+1} = K_{\bar{r}}$. If, on the other hand, $\bar{s} < t$ and $s < \bar{t}$, then $\bar{s} \leq \bar{t}$ and

$$h_j(\bar{s}, \bar{x}(\bar{s})) = h_j(\bar{t}, \bar{x}(\bar{t})) \;=\; 0 \quad \text{for all } j \in \mathcal{A}_{[s,t]}.$$

If follows from Lemma 4.5 that, for some $K > 0$ (that does not depend on $[s,t]$),

$$\sum_j \int_{[\bar{s},\bar{t}]} \mu_j(d\sigma) \;\leq\; K|\bar{t} - \bar{s}|\,.$$

However, then, by (4.15) and (4.16),

$$\sum_j \int_{[s,t]} \mu_j(d\sigma) \;=\; \sum_j \int_{[s,\bar{s}]\cup[\bar{s},\bar{t}]\cup[\bar{s},t]} \mu_j(d\sigma) \;\leq\; \tilde{K}|t - s|\,,$$

where $\tilde{K} = \max\{K, K_r\}$. Since $\tilde{K}$ does not depend on $[s,t]$, the lemma is proved.

Completion of the proof of Theorem 3.1 is now straightforward. Since

$$\text{cardinality}(\mathcal{A}_{[S,T]}) \;\leq\; r,$$

we deduce from Lemma 4.6 that there exists $K_r > 0$ such that, for every $[s,t] \subset [S,T]$,

$$\sum_j \int_{[s,t]} \mu_j(d\sigma) \;\leq\; K_r|t - s|\,.$$

Since $p(.)$ is Lipschitz continuous,

$$q(t) \;\left(:= p(t) + \int_{[S,t)} \sum_j \nabla_x h_j(s, \bar{x}(s))\mu_j(ds) \right)$$

is also Lipschitz continuous on (S,T).

It merely remains to conclude from Lemma 4.1 that the version of $\bar{u}$ chosen to coincide with the function $t \to \partial_y L_0^*(t, \bar{x}(t), G^T(t, \bar{x}(t))q(t))$ on the interior of $[S,T]$ and to assume the function's one sided limits at the endpoints, is Lipschitz continuous.

5. Conditions for normality. Theorem 3.1 provides conditions for Lipschitz continuity of optimal controls in circumstances when minimizers are normal extremals. It is of interest then to know when minimizers can be interpreted as normal extremals.

In this section we give two "constraint qualifications" (i.e., conditions on the data relating to the dynamic, pathwise, and endpoint constraints of problem $(\mathcal{P})$) that, when added to (H1)–(H4), ensure normality.

The first involves the (Clarke) generalized Jacobian: take $y \in \mathbb{R}^k$ and a function $\psi : \mathbb{R}^n \to \mathbb{R}^k$ which is Lipschitz continuous on a neighborhood of y. The generalized Jacobian $\text{co}\,\partial\psi(y)$ of ψ at y is defined to be

$$\text{co}\,\partial\psi(y) \;:=\; \text{co}\{\xi \in \mathbb{R}^{k\times n} \;:\; \exists y_i \;\to y\,,\, \xi_i \to \xi,\; \text{and } \xi_i = \nabla\psi(y_i) \text{ for all } i\}$$

(In this definition, $\nabla\psi(y_i)$ refers to the Fréchet derivative. There is no ambiguity of notation, since, in the case $k = 1$, $\text{co}\,\partial\psi(y)$ coincides with the convex hull of the

subdifferential $\partial\psi(y)$ defined in section 1 (see [2]).) $\operatorname{co}\partial_x\psi(y)$ denotes the (partial) generalized Jacobian of a function of several variables including x, with respect to the x variable alone.

Define

$$C_1 := \{x \in \mathbb{R}^n : (\bar{x}(S), x) \in C\}. \tag{5.1}$$

$(\mathrm{CQ})_1$ *For every measurable function* $A(.) : [S,T] \to \mathbb{R}^{n\times n}$ *such that*

$$A(t) \in co\,\partial_x(f(t,x) + G(t,x)\bar{u}(t))|_{x=\bar{x}(t)} \quad a.e.\ t \in [S,T],$$

there exists a measurable function $u : [S,T] \to \mathbb{R}^m$ *such that*

$$\nabla_t h_j(t,\bar{x}(t)) + \langle \nabla_x h_j(t,\bar{x}(t)), y_u(t)\rangle < 0 \quad \text{for all } j \in \mathcal{J}(t,\bar{x}) \text{ and } t \in [S,T] \tag{5.2}$$

and

$$\langle v, y_u(T)\rangle < 0 \quad \text{for all nonzero vectors } v \in N_{C_1}(\bar{x}(T)),$$

where y_u *is the solution to*

$$\dot{y}_u(t) = A(t)y_u(t) + G(t,\bar{x}(t))(u(t) - \bar{u}(t)) \quad a.e.\ t \in [S,T], \tag{5.3}$$

$$y_u(S) = 0. \tag{5.4}$$

$(\mathrm{CQ})_1$ can be regarded as a generalization of the Slater-type "interiority" hypothesis invoked by Hager [7] to allow for nonlinear nonsmooth dynamics.

$(\mathrm{CQ})_1$ falls somewhat short of providing directly verifiable hypotheses for normality, since it involves assessing controllability properties of a time-varying linear system, in the presence of pathwise control and state constraints.

The next constraint qualification places merely pointwise restrictions on the data and is accordingly of a more directly verifiable nature. The application of this alternative constraint qualification is restricted, however, to optimal control problems for which the right endpoint of state trajectories are free. On the other hand, it covers problems for which C takes the form

$$C = \{x_0\} \times \mathbb{R}^n$$

("fixed left endpoint and free right endpoint") and for which

$$h_j(S, x_0) = 0$$

("fixed initial state located in the state constraint set boundary"). This is a case of interest in which $(\mathrm{CQ})_1$ is never satisfied. (Equation (5.2) is violated at $t = 0$.)

$(\mathrm{CQ})_2$ *The endpoint constraint set takes the form* $C = C_0 \times \mathbb{R}^n$ *for some* $C_0 \subset \mathbb{R}^n$. *The functions* h_k, $i = 1,\dots,r$ *are twice continuously differentiable. There exist constants* $\varepsilon, \delta, \gamma > 0$ *and a continuous function* $\nu : [S,T] \times \mathbb{R}^n \to U$ *such that if* $(t,\xi) \in [S,T] \times \mathbb{R}^n$ *is any point satisfying*

$$|\xi - \bar{x}(t)| \le \varepsilon \quad \text{and} \quad h_j(t,\bar{x}(t)) \ge -\delta \quad \text{for some } j,$$

then

$$\nabla_t h_j(t,\xi) + \langle \nabla_x h_j(t,\xi), f(t,\xi) + G(t,\xi)\nu(t,\xi)\rangle < -\gamma.$$

The following proposition provides conditions for optimal controls to be Lipschitz continuous, in which the "$(\bar{x}, \bar{u})$ is a normal extremal" hypothesis of Theorem 3.1 is replaced by either of the above constraint qualifications.

PROPOSITION 5.1. *Let $(\bar{x}, \bar{u})$ be a minimizer for $(\mathcal{P})$. Assume Hypotheses* (H1)–(H4) *are satisfied. Assume also that either* $(\mathrm{CQ})_1$ *or* $(\mathrm{CQ})_2$ *is satisfied. Then $\bar{u}$ is Lipschitz continuous.*

Proof. Take $(\bar{x}, \bar{u})$ to be a minimizer for $(\mathcal{P})$. Then $(\bar{x}, \bar{u})$ is a minimizer also for the related optimal control problem (we label it $(\mathcal{Q})$), in which the endpoint constraint set C is replaced by $\{\bar{x}(S)\} \times C_1$, where C_1 was defined by (5.1).

If $(\mathrm{CQ})_2$ is satisfied (together with (H1)–(H4)), then it is known (see [10]) that the minimizer $(\bar{x}, \bar{u})$ is a normal extremal. The fact that $\bar{u}$ is Lipschitz continuous (after, if required, adjustment on a null-set) now follows from Theorem 3.1, applied to $(\bar{x}, \bar{u})$, regarded as a minimizer for $(\mathcal{Q})$.

Suppose next that condition $(\mathrm{CQ})_1$ (together with (H1)–(H4)) is satisfied. Then the conditions are met under which $(\bar{x}, \bar{u})$, regarded as a minimizer for $(\mathcal{Q})$, satisfies the maximum principle (see, e.g., [12, Chapter 6]). We conclude that there exist $p \in W^{1,1}([S,T]; \mathbb{R}^n)$, $\mu_j \in C^{\oplus}(S,T)$ for $j = 1, \ldots, r$, and $\lambda \geq 0$ such that, writing

$$q(t) = p(t) + \sum_{j=1}^{r} \int_{[S,t)} \nabla_x h_j(s, \bar{x}(s)) \mu_j(ds), \tag{5.5}$$

conditions (2.3), (2.5), and (2.6) are satisfied, and also

$$-\dot{p}(t) = A^T(t) p(t), \tag{5.6}$$

$$-\left(p(T) + \int_{[S,T]} \nabla_x h_j(t, \bar{x}(t)) \mu_j(dt) \right) = \lambda \eta + v \tag{5.7}$$

for some

$$\eta \in \partial_{x_1} l(\bar{x}(S), x_1)|_{x_1 = x(\bar{T})} \quad \text{and} \quad v \in N_{C_1}(\bar{x}(T)) \,.$$

We now show that these conditions can be satisfied only if $\lambda > 0$. Since multipliers can be scaled by an arbitrary positive constant, this will imply that the maximum principle (for $(\bar{x}, \bar{u})$, regarded as a solution to $(\mathcal{Q})$) is satisfied with $\lambda = 1$, i.e., $(\bar{x}, \bar{u})$ is a normal extremal for $(\mathcal{Q})$. The Lipschitz continuity of $\bar{u}$ will then follow from Theorem 3.1.

Suppose, to the contrary, that $\lambda = 0$. Then

$$((\mu_1, \ldots, \mu_r), p) \neq ((0, \ldots, 0), 0) \,.$$

Define

$$a(u(.)) \;:=\; \int_S^T \langle q(t), G(t, \bar{x}(t))(u(t) - \bar{u}(t)) \rangle dt \,.$$

Here, $u(.)$ is the function whose existence is hypothesized in $(\mathrm{CQ})_1$. From (2.5) we deduce that

$$a(u(.)) \;\geq\; 0 \,.$$

However, an analysis along the lines of that in [5] permits us to deduce from (2.5), (2.6), and (5.7) that

(5.8)
$$(0 \leq)\ a(u(.)) \ = \ \int_{[S,T]} \sum_{j=1}^{r} (\nabla_t h_j(t, \bar{x}(t)) + \langle \nabla_x h_j(t, \bar{x}(t)), y_u(t) \rangle) \mu_j(dt) + \langle v, y_u(T) \rangle \, .$$

We know that, since it is assumed that $\lambda = 0$,

$$((\mu_1, \ldots, \mu_r), p) \neq 0 \, .$$

However, if $\mu_j \neq 0$ for some j, the right side of (5.8) is strictly negative, which is a contradiction. So suppose $\mu_j = 0$ for all j. In this case $p \neq 0$. We must have $v \neq 0$ (since otherwise by (5.6), $p \equiv 0$). However, then, once again, we arrive at the contradiction that the right side of (5.8) is strictly negative.

Acknowledgments. Helpful comments by the reviewers on an earlier version of this paper are gratefully acknowledged.

REFERENCES

[1] J. Ball, *Constitutive inequalities and existence theorems in nonlinear elastostatics*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium Vol. I, Pitman, London, 1977, pp. 187–241.
[2] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
[3] A. L. Dontchev and W. W. Hager, *A new approach to Lipschitz continuity in state constrained optimal control*, Systems Control Lett., 35 (1998), pp. 137–143.
[4] A. L. Dontchev and I. Kolmanovsky, *On regularity of optimal control*, in Recent Developments in Optimization, Lecture Notes in Econom. and Math. Systems 429, R. Durier and C. Michelot, eds., Springer-Verlag, Berlin, 1995, pp. 125–135.
[5] M. M. A. Ferreira and R. B. Vinter, *When is the maximum principle for state constrained problems nondegenerate?*, J. Math. Anal. Appl., 187 (1994), pp. 438–467.
[6] W. W. Hager, *Rates of convergence for discrete approximations to unconstrained control problems*, SIAM J. Numer. Anal., 13 (1976), pp. 449–472.
[7] W. W. Hager, *Lipschitz continuity for constrained processes*, SIAM J. Control Optim., 17 (1979), pp. 321–338.
[8] K. M. Malanowski, *On regularity of solutions to optimal control problems for systems with control appearing linearly*, Arch. Automat. Telemech., 23 (1978), pp. 227–242.
[9] O. L. Mangasarian, *Nonlinear Programming*, Classics Appl. Math. 10, SIAM, Philadelphia, 1994.
[10] F. Rampazzo and R. B. Vinter, *A theorem on existence of neighbouring trajectories satisfying a state constraint, with applications to optimal control*, IMA J. Math. Control Inform., 16 (1999), pp. 335–351.
[11] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[12] R. B. Vinter, *Optimal Control*, Birkhäuser Boston, Boston, MA, 2000.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1745–1775

NONTANGENCY-BASED LYAPUNOV TESTS FOR CONVERGENCE AND STABILITY IN SYSTEMS HAVING A CONTINUUM OF EQUILIBRIA*

SANJAY P. BHAT† AND DENNIS S. BERNSTEIN‡

Abstract. This paper focuses on the stability analysis of systems having a continuum of equilibria. Two notions that are of particular relevance to such systems are convergence and semistability. Convergence is the property whereby every solution converges to a limit point that may depend on the initial condition. Semistability is the additional requirement that all solutions converge to limit points that are Lyapunov stable. We give new Lyapunov-function-based results for convergence and semistability of nonlinear systems. These results do not make assumptions of sign definiteness on the Lyapunov function. Instead, our results use a novel condition based on nontangency between the vector field and invariant or negatively invariant subsets of the level or sublevel sets of the Lyapunov function or its derivative and represent extensions of previously known stability results involving semidefinite Lyapunov functions. To illustrate our results we deduce convergence and semistability of the kinetics of the Michaelis–Menten chemical reaction and the closed-loop dynamics of a scalar system under a universal adaptive stabilizing feedback controller.

Key words. nontangency, Lyapunov stability, semistability, convergence, prolongations

AMS subject classifications. 34C99, 34D20, 37B25, 93D99

DOI. 10.1137/S0363012902407119

1. Introduction.

This paper focuses on the stability analysis of systems that have a continuum of equilibria. Examples of such systems include mechanical systems having rigid-body modes and isospectral matrix dynamical systems [15]. Such systems also arise in chemical kinetics, compartmental modeling, and adaptive control. Since every neighborhood of a nonisolated equilibrium contains another equilibrium, a nonisolated equilibrium cannot be asymptotically stable. Thus asymptotic stability is not the appropriate notion of stability for systems having a continuum of equilibria. However, given a system that has a continuum of equilibria, it is still natural to ask if the trajectories converge to limit points and if the limit points are Lyapunov stable. These questions lead us to consider the properties of *convergence* and *semistability*. For linear systems, semistability was originally defined in [9] and applied to matrix second-order systems in [3]. In the present paper, we extend the notion of semistability to nonlinear systems. Preliminary versions of some of the results of this paper appeared in [5, 6].

Convergence is the notion that every trajectory of the system converges to a limit point. The limit point, which is necessarily an equilibrium, depends in general on the initial conditions. In a convergent system, the limit points of trajectories may or may not be Lyapunov stable. Semistability is the additional requirement that trajectories converge to limit points that are Lyapunov stable. More precisely, an equilibrium is semistable if it is Lyapunov stable, and every trajectory starting in a neighborhood

*Received by the editors May 3, 2002; accepted for publication (in revised form) April 23, 2003; published electronically, November 14, 2003. A preliminary version of this paper appeared in the Proceedings of the American Control Conference, Arlington, VA, 2001.

http://www.siam.org/journals/sicon/42-5/40711.html

†Department of Aerospace Engineering, Indian Institute of Technology, Powai, Mumbai 400076, India (bhat@aero.iitb.ac.in).

‡Department of Aerospace Engineering, The University of Michigan, Ann Arbor, MI 48109-2140 (dsbaero@engin.umich.edu). This author was supported in part by the Air Force Office of Scientific Research under grant F49620-98-1-0037.

of the equilibrium converges to a (possibly different) Lyapunov stable equilibrium. It can be seen that, for an equilibrium, asymptotic stability implies semistability, while semistability implies Lyapunov stability.

The relationship between Lyapunov stability, semistability, and asymptotic stability can be understood by considering the motion of a particle translating along a fixed direction. Such a particle, when moving under the action of a linear elastic spring, possesses a unique equilibrium, which is Lyapunov stable. In the additional presence of viscous damping, all motions of the particle converge to the unique equilibrium state, which is thus asymptotically stable. On the other hand, a particle moving under the action of viscous damping in the absence of a position-dependent restoring force can remain at rest in any position and thus exhibits a continuum of equilibria, each of which is Lyapunov stable. All motions of such a particle converge to rest, and the equilibrium that the particle converges to is determined by the initial position and velocity of the particle. The motion of the particle is thus convergent, while every equilibrium of the dynamics is semistable.

Besides the damped motion of a particle, there are several applications which involve systems having a continuum of equilibria, and in which semistability is the appropriate notion of stability. For example, we can consider the stability of the lateral dynamics of an aircraft in level trimmed flight. For disturbances affecting the angle between the longitudinal axis and the velocity vector, the vertical tail is designed to influence yaw so as to cause the sideslip angle to converge to zero. However, the heading angle will not generally converge to its predisturbance value. The offset in the final heading angle is an indication of the existence of a continuum of semistable equilibria.

Another application of semistability involves the kinetics of chemical reactions. While periodic or chaotic behavior can occur in chemical reactions [27], it is of interest to determine conditions under which the concentrations of the reacting species converge. In this case, the limiting concentrations are not completely determined by the dynamics but depend upon the initial concentrations as well. The stability of chemical kinetics with respect to a stoichiometric subspace is considered in [11, 12, 28], while [4] applies Lyapunov theory to study the semistability of mass action chemical kinetics.

Chemical reactions are a special case of a more general class of systems known as compartmental systems, which involve mass or energy balance [18]. Compartmental systems arise in biomedical, environmental, economic, power, and thermodynamic applications. Since compartmental systems possess a continuum of equilibria, semistability is the appropriate notion of stability.

In control applications, it is often desirable to design the control system so that the closed-loop system, in the absence of exogenous inputs (commands and disturbances), has an equilibrium that is asymptotically stable. For such designs, semistability is not needed. However, adaptive controllers [17, 22, 23, 24] involve feedback gains that evolve in response to the plant trajectories; that is, the limiting values of the gains depend on the initial condition of the plant states. An adaptive closed-loop system is thus not asymptotically stable, yet convergence and Lyapunov stability of the plant/gain equilibria, that is, semistability, is desirable.

In all of the applications above, it is of interest to determine the convergence and semistability properties of the system. Accordingly, we wish to obtain Lyapunov tests for convergence and semistability.

It is obvious that if a system is convergent, then all of its trajectories converge

to the set of equilibria. However, as the following example shows, the converse is not true.

Example 1.1. Consider the system $\dot{y}(t) = f(y(t))$, where $f : \mathbb{R}^2 \to \mathbb{R}^2$ is the continuous vector field given by

$$\begin{aligned} f(x) = {} & \operatorname{sign}(x_1^2 + x_2^2 - 1)|x_1^2 + x_2^2 - 1|^{\alpha} f_r(x) \\ & + \operatorname{sign}(x_1^2 + x_2^2 - 1)|x_1^2 + x_2^2 - 1|^{\beta} f_\theta(x), \end{aligned} \tag{1}$$

with $\alpha, \beta \geq 1$ and the vector fields f_r and f_θ given by

$$f_r(x) = \begin{bmatrix} -x_1 \\ -x_2 \end{bmatrix}, \ f_\theta(x) = \begin{bmatrix} x_2 \\ -x_1 \end{bmatrix}. \tag{2}$$

The vector fields f_r and f_θ point in the radial and circumferential directions, respectively, and thus the parameters α and β determine the rates at which solutions move in these directions, respectively. This can be seen more clearly by rewriting (1) in terms of polar coordinates $r = \sqrt{x_1^2 + x_2^2}$ and $\theta = \tan^{-1}(x_2/x_1)$ as

$$\dot{r} = -r\operatorname{sign}(r^2 - 1)|r^2 - 1|^{\alpha}, \tag{3}$$

$$\dot{\theta} = -\operatorname{sign}(r^2 - 1)|r^2 - 1|^{\beta}. \tag{4}$$

It can be seen from (3) and (4) that the set of equilibria $f^{-1}(0)$ consists of the origin $x = 0$ and the unit circle $\mathrm{S}^1 = \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}$. All solutions of the system starting from nonzero initial conditions $y(0)$ that are not on the unit circle approach the unit circle. Solutions starting outside the unit circle spiral in clockwise toward the unit circle, while solutions starting inside the unit circle spiral out counterclockwise. Consequently, all solutions are bounded, and, for every choice of α and β, all solutions converge to the set of equilibria. However, if $\alpha \geq \beta + 1$, then the system is not convergent. This can be seen by using (3) and (4) to obtain

$$\frac{dr}{d\theta} = r|r^2 - 1|^{\alpha - \beta}. \tag{5}$$

For initial values $r(0) > 1$, solutions of (5) converge to the equilibrium value $r = 1$ for decreasing θ, while, for initial values $r(0) < 1$, solutions converge to $r = 1$ for increasing θ. For $\alpha \geq \beta + 1$, the right-hand side of (5) is locally Lipschitz in r. Consequently, (5) has a unique maximally defined solution given by $r \equiv 1$ for the initial condition $r(0) = 1$ for increasing as well as decreasing θ. Since convergence to the equilibrium value $r = 1$ on a finite θ-interval for increasing (decreasing) θ implies nonuniqueness of solutions for decreasing (increasing) θ for the initial condition $r(0) = 1$, it follows that the solution $r(\cdot)$ of (5) can approach $r = 1$ only as $\theta \to \infty$ if $r(0) < 1$ and as $\theta \to -\infty$ if $r(0) > 1$. This implies that, for $\alpha - \beta \geq 1$, the solutions of (3)–(4) that converge to the unit circle spiral around an infinite number of times, thus ruling out convergence. In subsequent sections, we will use the results of this paper to show convergence and semistability in the case $\alpha \leq \beta$. Figure 1 shows the phase portrait of the system for $\alpha = 2$ and $\beta = 1$, while Figure 2 shows the phase portrait of the system for $\alpha = \beta = 1$.

In the case of Figure 1, where $\alpha = 2$ and $\beta = 1$, it follows from (5) that $\frac{dr}{d\theta} \to 0$ as $r \to 1$ so that all nontrivial trajectories approach the unit circle tangentially. As shown above, the system depicted in Figure 1 is not convergent, and, moreover, every equilibrium on the unit circle is unstable. In contrast, all trajectories in Figure 2,

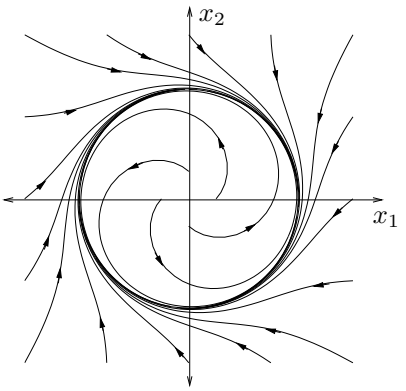


FIG. 1. *Phase portrait of* (3)–(4) *for* $\alpha = 2$, $\beta = 1$.

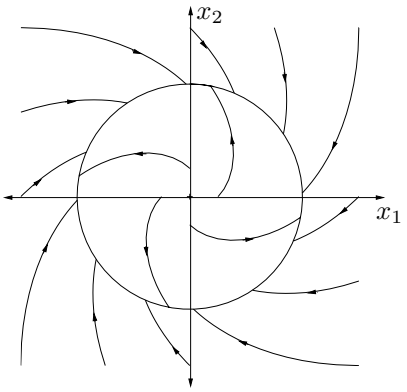


FIG. 2. *Phase portrait of* (3)–(4) *for* $\alpha = \beta = 1$.

where $\alpha = \beta = 1$, approach the unit circle nontangentially, the system is convergent, and all equilibria on the unit circle are semistable. Thus, Figures 1 and 2 suggest that nontangency of trajectories to the set of equilibria is a sufficient condition under which convergence to the set of equilibria implies convergence and semistability.

Intuitively, a vector field is nontangent to a set at a point if the vector field at the point is not contained in the tangent space to the set at that point. We shall apply this intuitive idea to the situation depicted in Figure 2, where the vector field is the vector field describing the dynamics and the set is the set of equilibria of the system. However, this intuitive notion presents two chief difficulties when the set under consideration is the set of singular points of the vector field, that is, the set of equilibria of the system. First, the vector field at an equilibrium is zero, and hence it is always contained in the tangent space to the set of equilibria. To capture the notion of nontangency in such a case, we introduce the *direction cone* of a vector field in section 4. The second difficulty is that, unlike as in Figures 1 and 2, the set of equilibria may not be sufficiently regular to possess a tangent space at the equilibrium point under consideration and may have corners or self-intersections. For example, consider a dynamical system that evolves on the nonnegative orthant and has the boundary of the orthant as its set of equilibria. In this case, the set of equilibria has a corner at the origin. We overcome this difficulty by considering the tangent

cone [2, 26], which extends the notion of a tangent space to a nonsmooth setting. In section 4, we formalize our intuitive notion of nontangency by defining nontangency of a vector field to a set at a point to be the condition that the tangent cone to the set at the point and the direction cone of the vector field at that point have no nonzero vector in common. Section 4 contains examples illustrating direction cones as well as nontangency. We also present a result that is useful in computing the direction cone of a vector field in applications.

We apply our notion of nontangency in section 5, where we show that the solution starting from a point converges if and only if the vector field is nontangent to the positive limit set of the point at some positive limit point. While this result cannot be applied directly in practice, we use it along with well-known properties of positive limit sets to show that, if the vector field is nontangent to the largest invariant subset of the zero-level set of the derivative of a Lyapunov function that is nonincreasing along the solutions, then every bounded solution converges to a limit.

Since the application of the convergence results of section 5 depend on verifying the boundedness of trajectories, we consider the property of boundedness in section 3. We extend well-known results for boundedness involving proper (that is, radially unbounded in the case where the state-space is $\mathbb{R}^n$) Lyapunov functions [21, 30] by introducing the notion of a *weakly proper* function. A function is weakly proper if the connected components of its sublevel sets are compact. We show that the existence of a weakly proper function that is nonincreasing along the trajectories implies that the trajectories are bounded. The usefulness of this result is illustrated in the examples given in section 3.

In section 6, we apply nontangency to Lyapunov stability. Here, prolongations [7, 8] play a role analogous to that played by positive limit sets in section 5. More specifically, we introduce the *restricted prolongation* of a point and show that an equilibrium point is Lyapunov stable if and only if the vector field is nontangent at the equilibrium to its restricted prolongation. The restricted prolongation of a point is a subset of its positive prolongation as defined in [7, 8]. While positive prolongations have been widely used in stability analysis [7], restricted prolongations have invariance properties that are needed for the results that we present. These properties, which are established in section 6, represent one of the key contributions of this paper.

In section 7, we use the results of sections 5 and 6 to obtain novel Lyapunov tests for Lyapunov stability, semistability, and asymptotic stability of nonlinear systems. These results do not make any assumptions about the sign definiteness of the Lyapunov function. Instead, they require only that the Lyapunov function derivative be nonpositive and the equilibrium be a local minimizer of the Lyapunov function on the set of points at which the Lyapunov function derivative is negative. For Lyapunov stability, the weaker assumptions on the Lyapunov function are supplemented by assuming nontangency of the vector field to invariant or negatively invariant subsets of the level set of the Lyapunov function containing the equilibrium and, for semistability, to invariant or negatively invariant subsets of the zero-level set of the Lyapunov function derivative. These results either extend or complement known results for Lyapunov stability and asymptotic stability involving semidefinite Lyapunov functions and Lyapunov function derivatives as given in [1, 16, 20, 21].

As mentioned above, chemical kinetics comprise one of the application areas for semistability theory. Since the kinetic equation for a system of chemical reactions governs concentrations of the reacting species, all solutions of physical interest take values in the nonnegative orthant. For such systems, which evolve on possibly closed

positively invariant subsets of $\mathbb{R}^n$, it is natural to consider relative stability, that is, stability with respect to perturbed initial conditions that belong to the positively invariant subset. Therefore, with applications to nonnegative dynamics in mind, we consider relative stability of dynamical systems that evolve on (not necessarily open) subsets of $\mathbb{R}^n$. Relative stability has been considered previously in [7, 16].

We illustrate the main results by applying them to examples from chemical kinetics and adaptive control. More specifically, we use the nontangency-based Lyapunov results to deduce convergence and semistability of the kinetics of the Michaelis–Menten chemical reaction [11] and the closed-loop dynamics of a scalar system under a universal adaptive stabilizing controller given in [17, 22].

2. Preliminaries. Let $\mathcal{G} \subseteq \mathbb{R}^n$, and let $\|\cdot\|$ denote a norm on $\mathbb{R}^n$. A subset $\mathcal{U}$ of $\mathcal{G}$ is *relatively open* in $\mathcal{G}$ if $\mathcal{U}$ is open in the subspace topology induced on $\mathcal{G}$ by the norm $\|\cdot\|$. Given $\mathcal{K} \subseteq \mathcal{G}$, we let int $\mathcal{K}$ and bd $\mathcal{K}$ denote the interior and boundary, respectively, of $\mathcal{K}$ in the subspace topology on $\mathcal{G}$. Thus int $\mathcal{K}$ is the largest subset of $\mathcal{K}$ that is relatively open in $\mathcal{G}$, while bd $\mathcal{K} = (\overline{\mathcal{K}} \cap \mathcal{G}) \backslash \text{int } \mathcal{K}$, where $\overline{\mathcal{K}}$ denotes the closure of $\mathcal{K}$ in $\mathbb{R}^n$. A set $\mathcal{U} \subseteq \mathcal{G}$ is *relatively bounded* in $\mathcal{G}$ if $\overline{\mathcal{U}}$ is compact and contained in $\mathcal{G}$. A point $x \in \mathbb{R}^n$ is a *subsequential limit* of a sequence $\{x_i\}$ in $\mathbb{R}^n$ if there exists a subsequence of $\{x_i\}$ that converges to x in the norm $\|\cdot\|$. A sequence $\{x_i\}$ in $\mathcal{G}$ is relatively bounded in $\mathcal{G}$ if it is relatively bounded when viewed as a set. Every sequence that is relatively bounded in $\mathcal{G}$ has at least one subsequential limit, and every subsequential limit of the sequence is contained in $\mathcal{G}$. When there is no possibility of confusion, we will use "relatively open (bounded)" instead of "relatively open (bounded) in $\mathcal{G}$." Also, in the case where $\mathcal{G} = \mathbb{R}^n$, we will use "open" and "bounded" instead of "relatively open" and "relatively bounded," respectively.

We recall that a set $\mathcal{K} \subseteq \mathcal{G}$ is *connected* if and only if every pair of relatively open sets $\mathcal{U}_i \subseteq \mathcal{G}$, $i = 1, 2$, satisfying $\mathcal{K} \subseteq \mathcal{U}_1 \cup \mathcal{U}_2$ and $\mathcal{U}_i \cap \mathcal{K} \neq \varnothing$, $i = 1, 2$, has a nonempty intersection. Also, a connected component of the set $\mathcal{K} \subseteq \mathcal{G}$ is a connected subset of $\mathcal{K}$ that is not properly contained in any connected subset of $\mathcal{K}$.

Consider the system of differential equations

$$\dot{y}(t) = f(y(t)), \tag{6}$$

where $f : \mathcal{D} \to \mathbb{R}^n$ is continuous on the open set $\mathcal{D} \subseteq \mathbb{R}^n$. We assume that, for every initial condition $y(0) \in \mathcal{D}$ and every $a > 0$, the differential equation (6) possesses a unique C^1 solution $y : [0, a) \to \mathcal{D}$ on the interval $[0, a)$. Letting $\psi(\cdot, x)$ denote the solution of (6) that exists on $[0, \infty)$ and satisfies the initial condition $y(0) = x$, the above assumptions imply that the map $\psi : [0, \infty) \times \mathcal{D} \to \mathcal{D}$ is continuous [14, Thm. V.2.1], satisfies $\psi(0, x) = x$, and possesses the semigroup property, that is, $\psi(t, \psi(h, x)) = \psi(t + h, x)$ for all $t, h \geq 0$ and $x \in \mathcal{D}$. Given $t \geq 0$, it will often be convenient to denote the map $\psi(t, \cdot) : \mathcal{D} \to \mathcal{D}$ by ψ_t. The *orbit* $\mathcal{O}_x$ of a point $x \in \mathcal{D}$ is the set $\{\psi(t, x) : t \geq 0\}$.

A set $\mathcal{U} \subseteq \mathbb{R}^n$ is *positively invariant* if $\psi_t(\mathcal{U}) \subseteq \mathcal{U}$ for all $t \geq 0$. The set $\mathcal{U}$ is *negatively invariant* if, for every $z \in \mathcal{U}$ and every $t \geq 0$, there exists $x \in \mathcal{U}$ such that $\psi(t, x) = z$ and $\psi(\tau, x) \in \mathcal{U}$ for all $\tau \in [0, t]$. Hence, if $\mathcal{U}$ is negatively invariant, then $\mathcal{U} \subseteq \psi_t(\mathcal{U})$ for all $t \geq 0$, although the converse is not generally true. Finally, the set $\mathcal{U}$ is *invariant* if $\psi_t(\mathcal{U}) = \mathcal{U}$ for all $t \geq 0$. Note that a set is invariant if and only if it is positively as well as negatively invariant. Also, it is easy to show that each connected component of a positively invariant (respectively, negatively invariant, invariant) set is positively invariant (respectively, negatively invariant, invariant).

In the rest of the paper, $\mathcal{G} \subseteq \mathcal{D}$ will denote a positively invariant set so that $\mathcal{O}_x \subseteq \mathcal{G}$ for all $x \in \mathcal{G}$. Except for local compactness, which is invoked in section 4 and subsequent sections, we require no additional hypotheses on $\mathcal{G}$.

An *equilibrium point* of (6) is a point $x \in \mathcal{D}$ satisfying $f(x) = 0$ or, equivalently, $\psi(t, x) = x$ for all $t \geq 0$. We let $\mathcal{E} = f^{-1}(0) \cap \mathcal{G}$, the set of all equilibrium points of (6) in $\mathcal{G}$. An *isolated* equilibrium is an isolated point of $\mathcal{E}$. An equilibrium point $x \in \mathcal{E}$ is *Lyapunov stable* relative to $\mathcal{G}$ if, for every relatively open neighborhood $\mathcal{U}_\varepsilon \subseteq \mathcal{G}$ of x, there exists a relatively open neighborhood $\mathcal{U}_\delta \subseteq \mathcal{G}$ of x such that $\psi_t(\mathcal{U}_\delta) \subseteq \mathcal{U}_\varepsilon$ for all $t \geq 0$. If $x \in \mathcal{E}$ is Lyapunov stable relative to $\mathcal{G}$, then every relatively open neighborhood of x contains a positively invariant, relatively open neighborhood of x [8, section V.1].

The system (6) is *convergent* relative to $\mathcal{G}$ if, for every $x \in \mathcal{G}$, $\lim_{t\to\infty} \psi(t, x)$ exists and is contained in $\mathcal{G}$. It follows from the continuity of ψ and the semigroup property that, if $x \in \mathcal{G}$ is such that $\lim_{t\to\infty} \psi(t, x)$ exists and is contained in $\mathcal{G}$, then, for every $h > 0$, $\psi_h(\lim_{t\to\infty} \psi(t, x)) = \lim_{t\to\infty} \psi(t + h, x) = \lim_{t\to\infty} \psi(t, x)$ so that $\lim_{t\to\infty} \psi(t, x) \in \mathcal{E}$.

An equilibrium point $x \in \mathcal{G}$ is *semistable* relative to $\mathcal{G}$ if there exists a relatively open neighborhood $\mathcal{U} \subseteq \mathcal{G}$ of x such that, for every $z \in \mathcal{U}$, $\lim_{t\to\infty} \psi(t, z)$ exists, is contained in $\mathcal{G}$, and is Lyapunov stable relative to $\mathcal{G}$. Note that if the equilibrium $x \in \mathcal{G}$ is semistable relative to $\mathcal{G}$, then every equilibrium in some relatively open neighborhood of x is Lyapunov stable relative to $\mathcal{G}$. In particular, every equilibrium that is semistable relative to $\mathcal{G}$ is also Lyapunov stable relative to $\mathcal{G}$.

An equilibrium point $x \in \mathcal{G}$ is *asymptotically stable* relative to $\mathcal{G}$ if x is Lyapunov stable relative to $\mathcal{G}$ and there exists a relatively open neighborhood $\mathcal{U} \subseteq \mathcal{G}$ of x such that, for every $z \in \mathcal{U}$, $\lim_{t\to\infty} \psi(t, z) = x$. It is easy to see that an equilibrium is asymptotically stable relative to $\mathcal{G}$ if and only if it is an isolated equilibrium and is semistable relative to $\mathcal{G}$.

Given a function $V : \mathcal{G} \to \mathbb{R}$, a point $x \in \mathcal{G}$ is a *local minimizer* of V relative to $\mathcal{K} \subseteq \mathcal{G}$ if there exists a relatively open neighborhood $\mathcal{U} \subseteq \mathcal{G}$ of x such that $V(x) \leq V(z)$ for all $z \in \mathcal{U} \cap \mathcal{K}$. The point $x \in \mathcal{G}$ is a *global minimizer* of V relative to $\mathcal{G}$ if $V(x) \leq V(z)$ for all $z \in \mathcal{G}$. Local and global maximizers of V are defined similarly.

Given a continuous function $V : \mathcal{G} \to \mathbb{R}$, we define

$$\dot{V}(x) = \lim_{h\to 0^+} \frac{1}{h}[V(\psi(h, x)) - V(x)] \tag{7}$$

for every $x \in \mathbb{R}^n$ such that the limit in (7) exists. It is easy to see that if $x \in \mathcal{E}$ and $V : \mathcal{G} \to \mathbb{R}$, then $\dot{V}(x)$ is defined and equals zero.

Some of the results that we present involve an equilibrium point that is also a local or global maximizer of $\dot{V}$ for some function V. Since $\dot{V}$ is zero at equilibrium points, an equilibrium x is a local maximizer of $\dot{V}$ relative to $\mathcal{G}$ if and only if $\dot{V}$ assumes nonpositive values in some relatively open neighborhood of x. Similarly, an equilibrium x is a global maximizer of $\dot{V}$ relative to $\mathcal{G}$ if and only if $\dot{V}$ assumes nonpositive values in $\mathcal{G}$. We will find it convenient to state the familiar requirements of local and global negative semidefiniteness on $\dot{V}$ in terms of the equilibrium being a local or global maximizer, respectively, of $\dot{V}$.

3. Boundedness of orbits. A function $U : \mathcal{G} \to \mathbb{R}$ is *proper* if $U^{-1}(\mathcal{I})$ is a compact subset of $\mathcal{G}$ for all compact subsets $\mathcal{I}$ of $\mathbb{R}$, and *weakly proper* if, for every $c \in \mathbb{R}$, every connected component of the set $\{x \in \mathcal{G} : U(x) \leq c\} = U^{-1}((-\infty, c])$ is compact. If U is proper and bounded below on $\mathcal{G}$, then U is weakly proper.

We present a Lyapunov test for boundedness of orbits that will be useful in some of the examples we present in this paper. The test involves weakly proper functions.

PROPOSITION 3.1. *Suppose there exists a weakly proper, continuous function $U : \mathcal{G} \to \mathbb{R}$ such that $\dot{U}$ is defined on $\mathcal{G}$ and such that $\dot{U}(x) \leq 0$ for all $x \in \mathcal{G}$. Then, for every $x \in \mathcal{G}$, $\mathcal{O}_x$ is relatively bounded in $\mathcal{G}$.*

Proof. Consider $x \in \mathcal{G}$, and let $c = U(x)$. The assumptions on $\dot{U}$ imply that the function $t \mapsto U(\psi(t,x))$ is nonincreasing. Hence $U(\psi(t,x)) \leq c$ for all $t \geq 0$ so that $\psi(t,x) \in \mathcal{K} \stackrel{\text{def}}{=} \{z \in \mathcal{G} : U(z) \leq c\}$ for all $t \geq 0$. Thus $\mathcal{O}_x \subseteq \mathcal{K}$. Since $\mathcal{O}_x$ is connected, it follows that $\mathcal{O}_x$ is contained in a connected component $\mathcal{M}$ of $\mathcal{K}$. By weak properness, $\mathcal{M}$ is compact, and thus $\overline{\mathcal{O}}_x$ is contained in $\mathcal{M}$. Hence $\mathcal{O}_x$ is relatively bounded in $\mathcal{G}$. ∎

Proposition 3.1 is an extension of the following well-known sufficient condition for boundedness of orbits. See, for instance, [21, Thm. 4], [30, Thm. 8.7].

COROLLARY 3.1. *Suppose there exists a proper, continuous function $U : \mathcal{G} \to \mathbb{R}$ such that $\dot{U}$ is defined on $\mathcal{G}$ and such that $U(x) \geq 0$ and $\dot{U}(x) \leq 0$ for all $x \in \mathcal{G}$. Then, for every $x \in \mathcal{G}$, $\mathcal{O}_x$ is relatively bounded in $\mathcal{G}$.*

Proof. The result follows from Proposition 3.1 by noting that the function U is weakly proper. ∎

Example 3.1. Consider the uncertain linear system

$$\dot{y}(t) = -ay(t) + bu(t), \tag{8}$$

where $a, b \in \mathbb{R}$ and b is nonzero but otherwise unknown. A universal adaptive stabilizing controller for the system (8) is given by [17, 22, 24]

$$u(t) = -k^2(t) \cos k(t) y(t), \tag{9}$$

where the adaptive gain parameter k satisfies the update law

$$\dot{k}(t) = y^2(t). \tag{10}$$

The corresponding closed-loop system on $\mathcal{G} = \mathbb{R}^2$ is described by (10) and

$$\dot{y}(t) = -[a + bk^2(t) \cos(k(t))] y(t). \tag{11}$$

We will use Proposition 3.1 to show that all orbits of the closed-loop system (10)–(11) are bounded.

Consider the function $U : \mathbb{R}^2 \to \mathbb{R}$ given by $U(x) = \frac{1}{2}y^2 + g(k)$, where $x = (y,k)$ and $g : \mathbb{R} \to \mathbb{R}$ is given by $g(k) = ak + b(k^2 - 2) \sin k + 2bk \cos k$. It is easy to verify that, for the closed-loop system, $\dot{U}(x) = 0$ for every $x \in \mathbb{R}^2$. We claim that the function U is weakly proper.

Let $c \in \mathbb{R}$, let $\mathcal{K} = \{x \in \mathbb{R}^2 : U(x) \leq c\}$, and consider $x_0 = (y_0, k_0) \in \mathcal{K}$. Let $\mathcal{M}$ be the connected component of $\mathcal{K}$ containing x_0. By continuity of U, $\mathcal{K}$ is closed. Since $\mathcal{M}$ is a connected component of a closed set, $\mathcal{M}$ is closed.

First suppose $b > 0$, and let $p : \mathbb{R} \to \mathbb{R}$ be given by $p(k) = bk^2 + ak - 2b$. There exists an integer $m > 0$ such that, for $k_1 = -(2m + \frac{3}{2})\pi$ and $k_2 = (2m + \frac{1}{2})\pi$, it follows that $p(k_i) > c$, $i = 1, 2$, and $k_0 \in [k_1, k_2]$. A simple computation shows that $p(k_i) = g(k_i)$, $i = 1, 2$, so that, for every $y \in \mathbb{R}$, the points $x_i = (y, k_i)$, $i = 1, 2$, satisfy $U(x_i) = \frac{1}{2}y^2 + p(k_i) > c$. Letting ρ denote the projection $(y,k) \mapsto k$, it follows that $\rho(\mathcal{M})$ is a connected set that contains $k_0 \in [k_1, k_2]$ but does not contain k_1 and k_2. Hence it follows that $\rho(\mathcal{M}) \subseteq [k_1, k_2]$; that is, $k \in [k_1, k_2]$ for every

$x = (y,k) \in \mathcal{M}$. Denote $l = \min_{k\in[k_1,k_2]} g(k)$. Then, for every $x = (y,k) \in \mathcal{M}$, $\frac{1}{2}y^2 \le c - g(k) \le c - l$. Thus, for every $x = (y,k) \in \mathcal{M}$, $y \in [y_1, y_2]$, where $y_1 = -\sqrt{2(c-l)}$ and $y_2 = \sqrt{2(c-l)}$. Thus the closed set $\mathcal{M}$ is contained in the compact set $[y_1, y_2] \times [k_1, k_2]$ and hence compact. It follows that U is weakly proper in the case where $b > 0$.

Now suppose $b < 0$, and let $p : \mathbb{R} \to \mathbb{R}$ be given by $p(k) = -bk^2 + ak + 2b$. There exists an integer $m > 0$ such that, for $k_1 = -(2m + \frac{1}{2})\pi$ and $k_2 = (2m + \frac{3}{2})\pi$, it follows that $p(k_i) > c$, $i = 1, 2$, and $k_0 \in [k_1, k_2]$. Letting $l = \min_{k\in[k_1,k_2]} g(k)$, $y_1 = -\sqrt{2(c-l)}$, and $y_2 = \sqrt{2(c-l)}$, it can be shown by repeating the arguments given above that the closed set $\mathcal{M}$ is contained in the compact set $[y_1, y_2] \times [k_1, k_2]$ and hence compact. It follows that U is weakly proper in the case where $b < 0$.

Since the function U is weakly proper and $\dot{U} \equiv 0$, it follows from Proposition 3.1 that every orbit of the closed-loop system (10)–(11) is bounded.

4. Direction cones and nontangency. Given a set $\mathcal{K} \subseteq \mathbb{R}^n$, we let co $\mathcal{K}$ denote the union of the convex hulls of the connected components of $\mathcal{K}$ and let coco $\mathcal{K}$ denote the cone generated by co $\mathcal{K}$. Given $\mathcal{K} \subseteq \mathbb{R}^n$ and $x \in \mathbb{R}^n$, we denote $\operatorname{dist}(x, \mathcal{K}) = \inf_{y\in\mathcal{K}} \|x - y\|$. Finally, we let $\mathrm{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$ denote the unit sphere in $\mathbb{R}^n$.

Given $x \in \mathcal{G}$, the *direction cone* $\mathcal{F}_x$ of f at x relative to $\mathcal{G}$ is the intersection of all sets of the form $\overline{\text{coco}\,(f(\mathcal{U})\backslash\{0\})}$, where $\mathcal{U} \subseteq \mathcal{G}$ is a relatively open neighborhood of x. It is easy to see that if $x \in \mathcal{G}\backslash\text{int}\,\mathcal{E}$, then $\mathcal{F}_x$ is a closed cone containing $f(x)$, while if $x \in \text{int}\,\mathcal{E}$, then $\mathcal{F}_x = \varnothing$.

Let $\mathcal{K} \subseteq \mathbb{R}^n$ and $x \in \mathcal{K}$. A vector $v \in \mathbb{R}^n$ is *tangent* to $\mathcal{K}$ at $x \in \mathcal{K}$ if there exist a sequence $\{x_i\}$ in $\mathcal{K}$ converging to x and a sequence $\{h_i\}$ of positive real numbers converging to 0 such that $\lim_{i\to\infty} \frac{1}{h_i}(x_i - x) = v$. The *tangent cone* to $\mathcal{K}$ at x is the closed cone $T_x\mathcal{K}$ of all vectors tangent to $\mathcal{K}$ at x [2, p. 121], [26, Prop. 6.2]. It is easy to see that $0 \in T_x\mathcal{K}$. Moreover, if x is an isolated point of $\mathcal{K}$, then $T_x\mathcal{K} = \{0\}$. Also, if $\mathcal{K} \subseteq \mathcal{M}$, then, for every $x \in \mathcal{K}$, $T_x\mathcal{K} \subseteq T_x\mathcal{M}$. Finally, if $x \in \mathcal{K}$ and $\mathcal{U} \subseteq \mathbb{R}^n$ is an open neighborhood of x such that $\mathcal{K} \cap \mathcal{U}$ is a differentiable submanifold of $\mathbb{R}^n$, then $T_x\mathcal{K}$ is the usual tangent space to $\mathcal{K}$ at x.

Remark 4.1. Tangent cones are called *contingent cones* in [2]. We have followed the terminology used in [26].

The vector field f is *nontangent* to the set $\mathcal{K} \subseteq \mathcal{G}$ at the point $x \in \mathcal{K}$ relative to $\mathcal{G}$ if $T_x\mathcal{K} \cap \mathcal{F}_x \subseteq \{0\}$.

Remark 4.2. The notion of nontangency introduced here is different from the well-known notion of *transversality* [13]. Transversality between a vector field and a set is possible only at a point in the set where the vector field is not zero and the set is locally a differentiable submanifold of codimension one. On the other hand, nontangency is possible even if the vector field is zero and the set is not a differentiable submanifold of codimension one.

In the rest of the paper, we assume that $\mathcal{G}$ is *locally compact*, that is, every point in $\mathcal{G}$ is contained in a relatively open and relatively bounded set $\mathcal{U} \subseteq \mathcal{G}$. In particular, if $\mathcal{G}$ is either open or closed, then $\mathcal{G}$ is locally compact. Local compactness implies that every relatively open neighborhood of a point $x \in \mathcal{G}$ contains a relatively open neighborhood of x that is also relatively bounded in $\mathcal{G}$ [10, Thm. XI.6.2].

The following proposition is a key result of this paper. The result shows that if the vector field f is nontangent to the set $\mathcal{B}$ of all subsequential limits of sequences of points taken from a sequence of segments of orbits of (6), then the set $\mathcal{B}$ contains exactly one point. This result will be applied to positive limit sets in section 5

and restricted prolongations in section 6 to obtain nontangency-based conditions for convergence and stability.

PROPOSITION 4.1. *Let $x \in \mathcal{G}$, and let $\{x_i\}$ be a sequence in $\mathcal{G}$ converging to x. Let $\mathcal{I}_i \subseteq [0,\infty)$, $i = 1, 2, \ldots$, be intervals containing 0, and let $\mathcal{B} \subseteq \mathcal{G}$ be the set of all subsequential limits contained in $\mathcal{G}$ of sequences of the form $\{\psi(\tau_i, x_i)\}$, where $\tau_i \in \mathcal{I}_i$ for each i. Then $\mathcal{B} = \{x\}$ if and only if f is nontangent to $\mathcal{B}$ at x relative to $\mathcal{G}$.*

Proof. First, we note that $x \in \mathcal{B}$ since $x = \lim_{i\to\infty} \psi(0, x_i)$. Necessity now follows by noting that if $\mathcal{B} = \{x\}$, then $T_x\mathcal{B} = \{0\}$, and hence $T_x\mathcal{B} \cap \mathcal{F}_x \subseteq \{0\}$.

To prove sufficiency, suppose $z_0 \in \mathcal{B}$, $z_0 \neq x$. If the sequence $\{f(x_i)\}$ is eventually zero, then every sequence of the form $\{\psi(\tau_i, x_i)\}$ converges to x, and, consequently, $\mathcal{B}$ is a singleton. Hence we may assume without loss of generality that $f(x_i) \neq 0$ for every i. Let $\{\mathcal{U}_k\}$ be a nested sequence of neighborhoods of x that are relatively bounded and relatively open in $\mathcal{G}$, contained in $\mathcal{U}$, and such that $\overline{\mathcal{U}_{k+1}} \subset \mathcal{U}_k$ and $x_k \in \mathcal{U}_k$ for every $k = 1, 2, \ldots$, $\cap_k \mathcal{U}_k = \{x\}$, and $z_0 \notin \mathcal{U}_1$. Since $z_0 \in \mathcal{B}$, there exists a sequence $\{\tau_i\}$ such that $\tau_i \in \mathcal{I}_i$ for every i, and $\lim_{i\to\infty} \psi(\tau_i, x_i) = z_0 \notin \mathcal{U}_1$. By continuity of ψ, for every k, there exists a sequence $\{h_j^k\}_{j=k}^{\infty}$ in $[0,\infty)$ such that, for every $j \geq k$, $h_j^k \in \mathcal{I}_j$, $h_j^k \leq \tau_j$, $\psi(\tau, x_j) \in \mathcal{U}_k$ for every $\tau \in [0, h_j^k)$, and $\psi(h_j^k, x_j) \in \operatorname{bd} \mathcal{U}_k$. For each k, let $z_k \in \operatorname{bd} \mathcal{U}_k$ be a subsequential limit of the relatively bounded sequence $\{\psi(h_j^k, x_j)\}_{j=k}^{\infty}$. Then, for every k, it follows that $z_k \in \mathcal{B}$, $z_k \neq x$, and $\lim_{k\to\infty} z_k = x$. Now consider a subsequential limit v of the bounded sequence $\{\|z_k - x\|^{-1}(z_k - x)\}$. Clearly, $v \in T_x\mathcal{B}$. Also, $\|v\| = 1$ so that $v \neq 0$. We claim that $v \in \mathcal{F}_x$.

Let $\mathcal{V} \subseteq \mathcal{G}$ be a relatively open neighborhood of x, and consider $\varepsilon > 0$. By construction, there exists k such that $\big\|v - \|z_k - x\|^{-1}(z_k - x)\big\| < \varepsilon/3$. Moreover, since $\cap_i \mathcal{U}_i = \{x\}$, we can assume that $\mathcal{U}_k \subseteq \mathcal{V}$. Since z_k belongs to the boundary of a relatively open neighborhood of x, $\delta \stackrel{\text{def}}{=} \|z_k - x\| > 0$. Since $z_k = \lim_{i\to\infty} \psi(h_i^k, x_i)$ and $x = \lim_{i\to\infty} x_i$, there exists i such that $x_i \in \mathcal{V}$, $\|x - x_i\| < \varepsilon\delta/3$, and $\|z_k - \psi(h_i^k, x_i)\| < \varepsilon\delta/3$. Let $\mathcal{A}$ be the connected component of $f(\mathcal{V})\backslash\{0\}$ containing $f(x_i) \neq 0$. Since f is continuous, it follows that $f(\psi(\tau, x_i)) \in \overline{\mathcal{A}}$ for all $\tau \in [0, h_i^k]$. Therefore, $w \stackrel{\text{def}}{=} \psi(h_i^k, x_i) - x_i = \int_0^{h_i^k} f(\psi(\tau, x_i))d\tau$ is contained in the convex cone generated by $\overline{\mathcal{A}}$ [29, Thm. I.6.13]. Since $\mathcal{A}$ is connected, coco $\mathcal{A}$ is simply the convex cone generated by $\mathcal{A}$. Since $\mathcal{A}$ and $\overline{\mathcal{A}}$ generate the same closed convex cone, we have coco $\overline{\mathcal{A}} \subseteq \overline{\text{coco } \overline{\mathcal{A}}} = \overline{\text{coco } \mathcal{A}} \subseteq \overline{\text{coco } (f(\mathcal{V})\backslash\{0\})}$. Thus $w \in \overline{\text{coco } (f(\mathcal{V})\backslash\{0\})}$. Now,

$$\begin{aligned}\big\|v - \delta^{-1}w\big\| &= \big\|v - \delta^{-1}(z_k - x) - \delta^{-1}(\psi(h_i^k, x_i) - z_k) - \delta^{-1}(x - x_i)\big\| \\ &\leq \big\|v - \|z_k - x\|^{-1}(z_k - x)\big\| + \delta^{-1}\|\psi(h_i^k, x_i) - z_k\| + \delta^{-1}\|x - x_i\| \\ &< \varepsilon.\end{aligned}$$

We conclude that, for every $\varepsilon > 0$, there exists $w \in \overline{\text{coco } (f(\mathcal{V})\backslash\{0\})}$ and $\delta > 0$ such that $\|v - \delta^{-1}w\| < \varepsilon$. It follows that $v \in \overline{\text{coco } (f(\mathcal{V})\backslash\{0\})}$. Since the neighborhood $\mathcal{V}$ was arbitrary, it follows that $v \in \mathcal{F}_x$. Thus, if $\mathcal{B} \neq \{x\}$, then there exists $v \in \mathbb{R}^n$ such that $v \neq 0$ and $v \in T_x\mathcal{B} \cap \mathcal{F}_x$; that is, f is not nontangent to $\mathcal{B}$ at x relative to $\mathcal{G}$. Sufficiency now follows. ∎

Since any application of Proposition 4.1 will involve finding the direction cone, we next give a result that provides a convenient means of determining the direction cone in applications. For this purpose, it will be useful to introduce the limiting direction set of a vector field at a point.

Let $x \in \mathcal{G}\backslash\operatorname{int} \mathcal{E}$. A vector $v \in \mathrm{S}^{n-1}$ is a *limiting direction* of f at x relative to $\mathcal{G}$ if there exists a sequence $\{x_i\}$ in $\mathcal{G}\backslash\mathcal{E}$ such that $\lim_{i\to\infty} x_i = x$ and

$\lim_{i\to\infty} \frac{1}{\|f(x_i)\|} f(x_i) = v$. The *limiting direction set* $\mathcal{L}_x$ of f at x relative to $\mathcal{G}$ is the set of all limiting directions of f at x relative to $\mathcal{G}$. Clearly, $\mathcal{L}_x$ is nonempty, compact, and contained in S^{n-1}. Moreover, for every $\varepsilon > 0$, there exists a relatively open neighborhood $\mathcal{U}_\varepsilon \subseteq \mathcal{G}$ of x such that, for every $z \in \mathcal{U}_\varepsilon \backslash \mathcal{E}$, $\operatorname{dist}(\frac{1}{\|f(z)\|} f(z), \mathcal{L}_x) < \varepsilon$.

Consider $x \in \mathcal{G} \backslash \operatorname{int} \mathcal{E}$, and let $\mathcal{U} \subseteq \mathcal{G}$ be a relatively open neighborhood of x. For every sequence $\{x_i\}$ in $\mathcal{G}\backslash\mathcal{E}$ converging to $x \in \mathcal{G}$, the sequence $\{\frac{1}{\|f(x_i)\|} f(x_i)\}$ is eventually in the cone generated by $f(\mathcal{U})\backslash\{0\}$ and hence in coco $(f(\mathcal{U})\backslash\{0\})$ so that every subsequential limit of the sequence $\{\frac{1}{\|f(x_i)\|} f(x_i)\}$ is contained in coco $(f(\mathcal{U})\backslash\{0\})$. Since $\mathcal{U}$ was chosen arbitrarily, it follows that $\mathcal{L}_x \subseteq \mathcal{F}_x$. The following result shows that if no connected component of the limiting direction set contains the origin in its convex hull, then the direction cone is contained in the union of the convex cones generated by the connected components of the limiting direction set. This result provides a convenient means for determining the direction cone in applications.

PROPOSITION 4.2. *Let $x \in \mathcal{G}\backslash\operatorname{int} \mathcal{E}$, and suppose $0 \notin \operatorname{co} \mathcal{L}_x$. Then $\mathcal{F}_x \subseteq \operatorname{coco} \mathcal{L}_x$.*

Proof. See the appendix. ∎

Remark 4.3. A special case where Proposition 4.2 applies is the case where $\mathcal{L}_x \subseteq \mathrm{S}^{n-1}$ is finite. Suppose $x \in \mathcal{G}$ is such that $\mathcal{L}_x$ is finite. Then $\operatorname{co} \mathcal{L}_x = \mathcal{L}_x \subseteq \mathrm{S}^{n-1}$, and thus $0 \notin \operatorname{co} \mathcal{L}_x$. Moreover, $\operatorname{coco} \mathcal{L}_x$ is the union of rays generated by the points of $\mathcal{L}_x$. Proposition 4.2 implies that $\mathcal{F}_x \subseteq \operatorname{coco} \mathcal{L}_x$. However, since $\mathcal{L}_x \subset \mathcal{F}_x$ and since $\mathcal{F}_x$ is a cone, the rays generated by points of $\mathcal{L}_x$ are contained in $\mathcal{F}_x$, that is, $\operatorname{coco} \mathcal{L}_x \subseteq \mathcal{F}_x$. Thus, in the case where the limiting direction set is finite, the direction cone is the union of rays generated by points of the limiting direction set.

The following example shows that, in general, $\mathcal{F}_x \not\subseteq \operatorname{coco} \mathcal{L}_x$.

Example 4.1. Consider the system (6), where $f : \mathbb{R}^3 \to \mathbb{R}^3$ is given by

$$f(x) = -(x_1^2 + x_2^2)^6 \begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix} + (x_1^2 + x_2^2) \begin{bmatrix} x_2 \\ -x_1 \\ 0 \end{bmatrix} + (x_1^2 + x_2^2)^2 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{12}$$

Letting $\mathcal{G} = \mathbb{R}^3$, the set of equilibria is $\mathcal{E} = \{x \in \mathcal{G} : x_1 = x_2 = 0\}$.

Let $a \in \mathbb{R}$, and consider $x = [0 \ 0 \ a]^{\mathrm{T}} \in \mathcal{E}$. To compute $\mathcal{L}_x$, it will be convenient to introduce the function $r : \mathcal{G} \to \mathbb{R}$ given by $r(z) = \sqrt{z_1^2 + z_2^2}$ and the function $\theta : \mathcal{G}\backslash\mathcal{E} \to [0, 2\pi)$ such that, for every $z \in \mathcal{G}\backslash\mathcal{E}$, $z_1 = r(z)\cos(\theta(z))$ and $z_2 = r(z)\sin(\theta(z))$.

For every $z \in \mathcal{G}$, $\|f(z)\| = (r(z))^3 \sqrt{1 + (r(z))^2 + (r(z))^{20}}$, while, for every $z \in \mathcal{G}\backslash\mathcal{E}$,

$$\frac{1}{\|f(z)\|} f(z) = \frac{1}{\sqrt{1 + (r(z))^2 + (r(z))^{20}}} \begin{bmatrix} -(r(z))^{10}\cos(\theta(z)) + \sin(\theta(z)) \\ -(r(z))^{10}\sin(\theta(z)) - \cos(\theta(z)) \\ r(z) \end{bmatrix}, \tag{13}$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^3$. Consider a sequence $\{x_i\}$ in $\mathcal{G}\backslash\mathcal{E}$ converging to x. Since $\lim_{i\to\infty} r(x_i) = 0$, it is easy to see from (13) that every subsequential limit of the sequence $\{\|f(x_i)\|^{-1} f(x_i)\}$ is of the form $[\sin\alpha \ -\cos\alpha \ 0]^{\mathrm{T}}$, where $\alpha \in \mathbb{R}$. On the other hand, for every $\alpha \in \mathbb{R}$, the sequence $\{x_i\}$ given by $x_i = \left[\frac{1}{i}\cos\alpha \ \frac{1}{i}\sin\alpha \ a\right]^{\mathrm{T}}$ converges to x while $\lim_{i\to\infty} \|f(x_i)\|^{-1} f(x_i) = [\sin\alpha \ -\cos\alpha \ 0]^{\mathrm{T}}$. Hence we conclude that the limiting direction set of f at x is a circle and is given by $\mathcal{L}_x = \{[\sin\alpha \ -\cos\alpha \ 0]^{\mathrm{T}} : \alpha \in \mathbb{R}\}$. $\mathcal{L}_x$ is connected, and hence it easily follows that $\operatorname{co} \mathcal{L}_x = \{w \in \mathbb{R}^3 : w_1^2 + w_2^2 \le 0, w_3 = 0\}$ and $\operatorname{coco} \mathcal{L}_x = \{w \in \mathbb{R}^3 : w_3 = 0\}$. Note that $0 \in \operatorname{co} \mathcal{L}_x$.

We claim that $\mathcal{F}_x \not\subseteq \operatorname{coco} \mathcal{L}_x$. To see this, consider the vector $w = [0\ 0\ 1]^{\mathrm{T}}$. Clearly, $w \notin \operatorname{coco} \mathcal{L}_x$. We claim that $w \in \mathcal{F}_x$.

Let $\mathcal{U}$ be an open neighborhood of x. Choose $\varepsilon > 0$ such that $\{z \in \mathcal{G} : \|z - x\| \leq \varepsilon\} \subseteq \mathcal{U}$, and consider $z_1 = [\varepsilon\ 0\ a]^{\mathrm{T}}$ and $z_2 = [-\varepsilon\ 0\ a]^{\mathrm{T}}$. Then $z_i \in \mathcal{U}$, $i = 1, 2$. It is easy to verify that $w = \frac{1}{2}\varepsilon^{-4}[f(z_1) + f(z_2)]$. Since $\mathcal{U}\backslash\mathcal{E}$ is connected and f is continuous, $f(\mathcal{U})\backslash\{0\} = f(\mathcal{U}\backslash\mathcal{E})$ is also connected. Therefore, $f(z_i)$, $i = 1, 2$, are contained in the same connected component of $f(\mathcal{U})\backslash\{0\}$. Hence $w \in \operatorname{coco}(f(\mathcal{U})\backslash\{0\})$. Since $\mathcal{U}$ was chosen to be arbitrary, it follows that $w \in \mathcal{F}_x$.

The following example illustrates direction cones, nontangency, and the use of Proposition 4.2.

Example 4.2. Consider the system described in Example 1.1 with $\mathcal{G} = \mathbb{R}^2$, and let $x \in \mathrm{S}^1$. As discussed in Example 1.1, x is an equilibrium point for the system. It is easy to show that

$$\begin{aligned} \mathcal{L}_x &= \{\pm f_\theta(x)\}, && \alpha > \beta, \\ &= \left\{\pm \tfrac{1}{\sqrt{2}}(f_\theta(x) + f_r(x))\right\}, && \alpha = \beta, \\ &= \{\pm f_r(x)\}, && \alpha < \beta, \end{aligned}$$

where f_r and f_θ are given in (2). Thus $\mathcal{L}_x$ is finite, and hence, by Remark 4.3,

$$\begin{aligned} \mathcal{F}_x &= \{k f_\theta(x) : k \in \mathbb{R}\}, && \alpha > \beta, \\ &= \{k(f_\theta(x) + f_r(x)) : k \in \mathbb{R}\}, && \alpha = \beta, \\ &= \{k f_r(x) : k \in \mathbb{R}\}, && \alpha < \beta. \end{aligned} \tag{14}$$

The unit circle S^1 is a differentiable submanifold of $\mathbb{R}^2$. Hence, for every $x \in \mathrm{S}^1$, $T_x\mathrm{S}^1$ is the tangent line to S^1 at x. Since the vector field f_θ points in the circumferential direction at every point, it follows that, for every $x \in \mathrm{S}^1$, $T_x\mathrm{S}^1 = \operatorname{span}\{f_\theta(x)\}$. It now follows from (14) that

$$\begin{aligned} T_x\mathrm{S}^1 \cap \mathcal{F}_x &= T_x\mathrm{S}^1, && \alpha > \beta, \\ &= \{0\}, && \alpha \leq \beta. \end{aligned} \tag{15}$$

Thus, for every $x \in \mathrm{S}^1$, f is nontangent to S^1 at x relative to $\mathcal{G}$ if and only if $\alpha \leq \beta$. It can be observed that Figures 1 and 2 reflect this fact.

5. Positive limit sets, convergence, and nontangency. In this section, we present nontangency-based Lyapunov results for convergence. These results use three key ideas. The first of these, given in Proposition 5.1, is that a solution of (6) converges to a limit if and only if its positive limit set is a singleton set. The second key idea, presented as Proposition 5.2, is to use Proposition 4.1 to show that the positive limit set of a solution of (6) is a singleton set if and only if the vector field f is nontangent to the positive limit set at some point. Since it is not generally possible to find the positive limit set of a solution in practice, it is difficult to check nontangency of the vector field f to the positive limit set in applications. Since nontangency to any outer estimate of the positive limit set implies nontangency to the positive limit set itself, the third key idea is to check nontangency of the vector field f to an outer estimate of the positive limit set that is easier to find in practice. Proposition 5.3 gives outer estimates of the positive limit sets in terms of invariant subsets of the level and sublevel sets of a Lyapunov function and its derivative. Theorem 5.1, the main result of this section, combines the ideas of Propositions 5.1, 5.2, and 5.3 to give

a sufficient condition for convergence that involves a nontangency condition between the vector field f and invariant subsets of the level sets of the derivative of a Lyapunov function.

Given $x \in \mathcal{G}$, the *positive limit set* of x relative to $\mathcal{G}$ is the set $\mathcal{O}_x^\infty$ of points $z \in \mathcal{G}$ such that there exists a divergent sequence $\{t_i\}$ in $[0, \infty)$ satisfying $\lim_{i\to\infty} \psi(t_i, x) = z$. The first part of the following result on positive limit sets is well known in the case $\mathcal{G} = \mathbb{R}^n$. See, for instance, [7, Thm. 5.5, 5.9], [8, p. 24], [19, p. 114], and [21]. The second part depends on the local compactness of $\mathcal{G}$.

PROPOSITION 5.1. *Let $x \in \mathcal{G}$. If $\mathcal{O}_x$ is relatively bounded in $\mathcal{G}$, then $\mathcal{O}_x^\infty$ is nonempty, compact, invariant, and connected, and, in addition, $\psi(t, x) \to \mathcal{O}_x^\infty$ as $t \to \infty$; that is, for every relatively open subset $\mathcal{U} \subseteq \mathcal{G}$ that contains $\mathcal{O}_x^\infty$, there exists $T > 0$ such that $\psi(t, x) \in \mathcal{U}$ for all $t > T$. Moreover, $\lim_{t\to\infty} \psi(t, x)$ exists and is contained in $\mathcal{G}$ if and only if $\mathcal{O}_x^\infty$ contains exactly one point.*

Proof. The first part of the result is well known for $\mathcal{G} = \mathbb{R}^n$. The proof is similar in the case where $\mathcal{G} \neq \mathbb{R}^n$ and is left to the reader. In the second part, the necessity is straightforward. To prove sufficiency, suppose that $\mathcal{O}_x^\infty = \{z\}$. Let $\mathcal{U}_\varepsilon \subseteq \mathcal{G}$ be a relatively open neighborhood of z. Since $\mathcal{G}$ is locally compact, there exists a neighborhood $\mathcal{U} \subseteq \mathcal{U}_\varepsilon$ of x that is relatively open and relatively bounded in $\mathcal{G}$. Since $z \in \mathcal{O}_x^\infty$, there exists a divergent sequence $\{t_i\}$ in $[0, \infty)$ such that $\psi(t_i, x) \in \mathcal{U}$ for all i. We claim that there exists $T > 0$ such that $\psi(t, x) \in \mathcal{U}_\varepsilon$ for all $t > T$. If not, then by the continuity of ψ, for every i, there exists $\tau_i > t_i$ such that $\psi(\tau_i, x) \in \operatorname{bd} \mathcal{U}$. In this case, the sequence $\{\psi(\tau_i, x)\}$ is relatively bounded in $\mathcal{G}$ and hence has a subsequential limit $w \in \mathcal{G}$. By construction, $w \in \operatorname{bd} \mathcal{U}$ and hence $w \neq z$. However, by definition, $w \in \mathcal{O}_x^\infty = \{z\}$, which is a contradiction. Hence we conclude that there exists $T > 0$ such that $\psi(t, x) \in \mathcal{U}_\varepsilon$ for all $t > T$. Since $\mathcal{U}_\varepsilon$ was chosen arbitrarily, it follows that $\lim_{t\to\infty} \psi(t, x) = z \in \mathcal{G}$. ∎

The following application of Proposition 4.1 gives a nontangency-based necessary and sufficient condition for a solution of (6) to converge to a limit.

PROPOSITION 5.2. *Let $x \in \mathcal{G}$, and suppose that $\mathcal{O}_x^\infty$ is nonempty. Then $\lim_{t\to\infty} \psi(t, x)$ exists and is contained in $\mathcal{G}$ if and only if there exists $z \in \mathcal{O}_x^\infty$ such that f is nontangent to $\mathcal{O}_x^\infty$ at z relative to $\mathcal{G}$.*

Proof. Consider $z \in \mathcal{O}_x^\infty$. There exists a divergent sequence $\{t_i\}$ in $[0, \infty)$ such that $\lim_{i\to\infty} \psi(t_i, x) = z$. For every i, denote $x_i = \psi(t_i, x)$ so that $\lim_{i\to\infty} x_i = z$. $\mathcal{O}_x^\infty$ is the set of subsequential limits of sequences of the form $\{\psi(h_i, x_i)\}$, where $h_i \in [0, \infty)$ for every i. Letting $\mathcal{I}_i = [0, \infty)$ for every i and $\mathcal{B} = \mathcal{O}_x^\infty$, it follows from Proposition 4.1 that $\mathcal{O}_x^\infty = \{z\}$ if and only if f is nontangent to $\mathcal{O}_x^\infty$ at z. The result now follows from the second part of Proposition 5.1. ∎

Example 5.1. Consider the system described in Example 1.1, and assume that $\alpha \geq \beta + 1$. Let $\mathcal{G} = \mathbb{R}^2$, and consider $x \in \mathcal{G} \backslash \mathrm{S}^1$. As discussed in Example 1.1, $\lim_{t\to\infty} \psi(t, x)$ does not exist. Indeed, $\mathcal{O}_x^\infty = \mathrm{S}^1$. As discussed in Example 4.2, f is not nontangent to $\mathcal{O}_x^\infty = \mathrm{S}^1$ at x, thus illustrating Proposition 5.2.

It should be noted that in order to apply Proposition 5.2 to a solution of (6), the positive limit set of the solution needs to be known. Since it is not generally possible to find the positive limit set of a solution, Proposition 5.2 is not directly useful in applications. However, Lyapunov functions can sometimes be used to obtain sets that contain the positive limit set of a solution. The following proposition gives two such containment results for positive limit sets. The result is a straightforward extension of [8, Thm. VIII.6.1, c)] and [21, Thm. 1]. Hence the proof is left to the reader. We note only that the proof uses the invariance properties of positive limit

sets given in Proposition 5.1.

PROPOSITION 5.3. *Suppose $V : \mathcal{G} \to \mathbb{R}$ is a continuous function such that $\dot{V}$ is defined on $\mathcal{G}$, and $\dot{V}(x) \leq 0$ for all $x \in \mathcal{G}$. Let $x \in \mathcal{G}$ be such that $\mathcal{O}_x$ is relatively bounded in $\mathcal{G}$. Let $\mathcal{P}$ and $\mathcal{N}$ denote the largest invariant subsets of the sets $\{z \in \mathcal{G} : V(z) \leq V(x)\}$ and $\dot{V}^{-1}(0)$, respectively. Then $\mathcal{O}_x^\infty \subseteq \mathcal{P} \cap \mathcal{N}$. In addition, if $\dot{V} \equiv 0$, then $\mathcal{O}_x^\infty$ is contained in the largest invariant subset of $V^{-1}(V(x))$.*

In the following example, we illustrate Propositions 5.2 and 5.3 by applying them to the chemical kinetics of the Michaelis–Menten chemical reaction.

Example 5.2. In the Michaelis–Menten chemical reaction, a substrate S is converted into a product P through an intermediate complex C in the presence of an enzyme E. The reaction is depicted as

$$\mathrm{S} + \mathrm{E} \underset{k_2}{\overset{k_1}{\rightleftharpoons}} \mathrm{C} \xrightarrow{k_3} \mathrm{P} + \mathrm{E},$$

where $k_i > 0$, $i = 1, 2, 3$, are chemical rate constants. In this example, we use Propositions 5.2 and 5.3 to show that the concentrations of species S, P, C, and E in this chemical reaction converge to equilibrium values.

Letting $y_1(t), y_2(t), y_3(t)$, and $y_4(t)$ denote the instantaneous nonnegative concentrations of the species S, C, E, and P, respectively, the law of mass action kinetics yields [11]

$$\dot{y}(t) = y_2(t)v_1 + y_1(t)y_3(t)v_2, \tag{16}$$

where $v_1 = \begin{bmatrix} k_2 & -(k_2+k_3) & k_2+k_3 & k_3 \end{bmatrix}^{\mathrm{T}}$ and $v_2 = \begin{bmatrix} -k_1 & k_1 & -k_1 & 0 \end{bmatrix}^{\mathrm{T}}$. Equation (16) is of the form (6), where $f : \mathbb{R}^4 \to \mathbb{R}^4$ is given by $f(x) = x_2v_1 + x_1x_3v_2$.

The nonnegative orthant $\mathcal{G} = \{x \in \mathbb{R}^4 : x_i \geq 0, i = 1, \ldots, 4\}$ is positively invariant under the dynamics (16) [4, 11]. Since the vectors v_1 and v_2 are linearly independent, it is easy to see that the set of equilibrium concentrations in $\mathcal{G}$ is $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$, where $\mathcal{E}_1 = \{x \in \mathcal{G} : x_1 = 0, x_2 = 0, x_3 > 0\}$ and $\mathcal{E}_2 = \{x \in \mathcal{G} : x_1 \geq 0, x_2 = 0, x_3 = 0\}$.

We claim that f is nontangent to $\mathcal{E}_1$ at every point in $\mathcal{E}_1$ relative to $\mathcal{G}$. Indeed, the direction cone $\mathcal{F}_x$ of f at every point $x \in \mathcal{G}$ is contained in the span of v_1 and v_2, while every vector that is tangent to $\mathcal{E}_1$ at some point $x \in \mathcal{E}_1$ is contained in the span of the vectors $v_3 = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}^{\mathrm{T}}$ and $v_4 = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}^{\mathrm{T}}$. Since $\{v_i : i = 1, 2, 3, 4\}$ is a set of linearly independent vectors, it follows that $\mathrm{span}\{v_1, v_2\} \cap \mathrm{span}\{v_3, v_4\} = \{0\}$. Thus $T_x\mathcal{E}_1 \cap \mathcal{F}_x \subseteq \{0\}$ for every $x \in \mathcal{E}_1$.

It is easy to verify that the function $U : \mathcal{G} \to \mathbb{R}$ given by $U(x) = x_1 + 2x_2 + x_3 + x_4$ is proper and satisfies $\dot{U} \equiv 0$. It follows from Corollary 3.1 that every orbit in $\mathcal{G}$ is relatively bounded in $\mathcal{G}$. Hence, by Proposition 5.1, $\mathcal{O}_x^\infty$ is nonempty for every $x \in \mathcal{G}$.

Now consider the function $V : \mathcal{G} \to \mathbb{R}$ defined by $V(x) = \frac{1}{2}x_2^2 + x_1(x_2 + x_3) - \frac{1}{2}x_3^2$. Then $\dot{V}(x) = -k_3x_2(x_2 + x_3) \leq 0$ for all $x \in \mathcal{G}$. Thus $\dot{V}^{-1}(0) = \{x \in \mathcal{G} : x_2 = 0\}$. If a solution y of the differential equation (16) satisfies $y_2 \equiv 0$, then $\dot{y}_2 \equiv 0$ and hence $y_1y_3 \equiv 0$. It therefore follows that the largest invariant subset of $\dot{V}^{-1}(0)$ is the set $\mathcal{E}$ of equilibrium concentrations. Proposition 5.3 now implies that $\mathcal{O}_x^\infty \subseteq \mathcal{E}$ for every $x \in \mathcal{G}$.

Next, consider the function $W : \mathcal{G} \to \mathbb{R}$ defined by $W(x) = x_2 + x_3$. It is easy to verify that $\dot{W} \equiv 0$. Hence, by Proposition 5.3, for every $x \in \mathcal{G}$, $\mathcal{O}_x^\infty$ is contained in the level set $W^{-1}(W(x))$ of W containing x. Thus it follows that, for every $x \in \mathcal{G}$, $\mathcal{O}_x^\infty \subseteq W^{-1}(W(x)) \cap \mathcal{E}$. Since $W^{-1}(0) = \mathcal{E}_2$, it follows that $W^{-1}(W(x)) \cap \mathcal{E} \subseteq \mathcal{E}_1$ for every $x \in \mathcal{G}$ satisfying $W(x) > 0$. Hence we conclude that, for every $x \in \mathcal{G} \backslash \mathcal{E}_2$, $\mathcal{O}_x^\infty \subseteq \mathcal{E}_1$.

Now consider $x \in \mathcal{E}_2$. Then x is an equilibrium, and concentrations starting from the initial value x clearly converge to the equilibrium value x. Next, consider $x \in \mathcal{G}\backslash\mathcal{E}_2$. Then $\mathcal{O}_x^\infty \subseteq \mathcal{E}_1$ and, for every $z \in \mathcal{O}_x^\infty$, $T_z\mathcal{O}_x^\infty \cap \mathcal{F}_z \subseteq T_z\mathcal{E}_1 \cap \mathcal{F}_z \subseteq \{0\}$. Hence Proposition 5.2 implies that concentrations starting from the initial values x converge to equilibrium values. We conclude that the system described by (16) is convergent relative to $\mathcal{G}$; that is, the concentrations of all species in the Michaelis–Menten reaction converge to equilibrium values.

The following result gives a sufficient condition for a trajectory of (6) to converge to a limit. Unlike Proposition 5.2, the following result is not based on nontangency.

PROPOSITION 5.4. *Let $x \in \mathcal{G}$. If $\mathcal{O}_x^\infty$ contains an equilibrium z that is Lyapunov stable relative to $\mathcal{G}$, then $z = \lim_{t\to\infty} \psi(t,x)$; that is, $\mathcal{O}_x^\infty = \{z\}$.*

Proof. Suppose $z \in \mathcal{O}_x^\infty$ is Lyapunov stable relative to $\mathcal{G}$. Let $\mathcal{U}_\varepsilon \subseteq \mathcal{G}$ be a relatively open neighborhood of z. By Lyapunov stability, there exists a relatively open neighborhood $\mathcal{U}_\delta \subset \mathcal{G}$ of z such that $\psi_t(\mathcal{U}_\delta) \subseteq \mathcal{U}_\varepsilon$ for every $t \geq 0$. Since $z \in \mathcal{O}_x^\infty$, there exists $h \geq 0$ such that $\psi(h,x) \in \mathcal{U}_\delta$. Therefore, $\psi(t+h,x) = \psi_t(\psi(h,x)) \in \psi_t(\mathcal{U}_\delta) \subseteq \mathcal{U}_\varepsilon$ for every $t > 0$. Since $\mathcal{U}_\varepsilon \subseteq \mathcal{G}$ was chosen arbitrarily, it follows that $z = \lim_{t\to\infty} \psi(t,x)$. It immediately follows that $\lim_{i\to\infty} \psi(t_i,x) = z$ for every divergent sequence $\{t_i\}$ and thus $\mathcal{O}_x^\infty = \{z\}$. ∎

Our next result applies Propositions 5.2, 5.3, and 5.4 to obtain a sufficient condition for convergence. The result uses Proposition 5.3 to obtain a set containing all positive limit sets and then uses Propositions 5.2 and 5.4 to show convergence.

THEOREM 5.1. *Suppose $\mathcal{O}_x$ is relatively bounded in $\mathcal{G}$ for all $x \in \mathcal{G}$, and assume that there exists a continuous function $V : \mathcal{G} \to \mathbb{R}$ such that $\dot{V}$ is defined on $\mathcal{G}$ and satisfies $\dot{V}(x) \leq 0$ for all $x \in \mathcal{G}$. Let $\mathcal{S} \subseteq \mathcal{E}$ denote the set of equilibria that are Lyapunov stable relative to $\mathcal{G}$, and let $\mathcal{N}_0$ denote the largest invariant subset of $\dot{V}^{-1}(0)$. For every $k = 0,1,2,\ldots$, let $\mathcal{M}_k \subseteq \mathcal{N}_k$ denote the set of points in $\mathcal{N}_k$ where f is not nontangent to $\mathcal{N}_k$ relative to $\mathcal{G}$, and let $\mathcal{N}_{k+1} \subseteq \mathcal{M}_k$ denote the largest invariant subset of $\mathcal{M}_k$. If $\mathcal{M}_k \subseteq \mathcal{S}$ for some k, then the system* (6) *is convergent relative to $\mathcal{G}$.*

Proof. Consider $x \in \mathcal{G}$. Since $\mathcal{O}_x$ is relatively bounded in $\mathcal{G}$, Proposition 5.1 implies that $\mathcal{O}_x^\infty$ is nonempty and invariant. To prove the result, we first show that if $\mathcal{O}_x^\infty$ contains more than one element, then $\mathcal{O}_x^\infty \subseteq \mathcal{M}_k\backslash\mathcal{S}$ for every k.

Suppose $\mathcal{O}_x^\infty$ contains more than one element. By Proposition 5.3, it follows that $\mathcal{O}_x^\infty \subseteq \mathcal{N}_0$. Now assume that $\mathcal{O}_x^\infty \subseteq \mathcal{N}_k$ for some $k = 0,1,\ldots$. Since $\mathcal{O}_x^\infty$ contains more than one element, it follows from Proposition 5.2 that, for every $z \in \mathcal{O}_x^\infty$, f is not nontangent to $\mathcal{O}_x^\infty$ at z relative to $\mathcal{G}$, that is, $\{0\} \not\supseteq T_z\mathcal{O}_x^\infty \cap \mathcal{F}_z \subseteq T_z\mathcal{N}_k \cap \mathcal{F}_z$. Since $\mathcal{M}_k = \{z \in \mathcal{N}_k : T_z\mathcal{N}_k \cap \mathcal{F}_z \not\subseteq \{0\}\}$, it follows that $\mathcal{O}_x^\infty \subseteq \mathcal{M}_k$. Since $\mathcal{O}_x^\infty$ is invariant, $\mathcal{O}_x^\infty \subseteq \mathcal{N}_{k+1}$. It follows by induction that $\mathcal{O}_x^\infty \subseteq \mathcal{N}_{k+1} \subseteq \mathcal{M}_k$ for every k. Also, it follows from Proposition 5.4 that $\mathcal{O}_x^\infty \cap \mathcal{S} = \varnothing$. Thus $\mathcal{O}_x^\infty \subseteq \mathcal{M}_k\backslash\mathcal{S}$ for every k.

It follows from the above arguments that if $\mathcal{M}_k\backslash\mathcal{S} = \varnothing$ for some k, then $\mathcal{O}_x^\infty$ contains only one element, and hence, by Proposition 5.1, $\lim_{t\to\infty} \psi(t,x)$ exists and is contained in $\mathcal{G}$. The result now follows. ∎

Example 5.3. In this example, we use Theorem 5.1 to show that the closed-loop adaptive system given by (10) and (11) in Example 3.1 is convergent.

It was shown in Example 3.1 that every orbit of the system (10)–(11) is bounded. Consider the function $V : \mathbb{R}^2 \to \mathbb{R}$ given by $V(x) = e^{-k}$, $x = (y,k) \in \mathbb{R}^2$. For the closed-loop system (10)–(11), it follows that $\dot{V}(x) = -e^{-k}y^2 \leq 0$ for all $x = (y,k) \in \mathbb{R}^2$. Thus $\dot{V}^{-1}(0) = \{(y,k) \in \mathbb{R}^2 : y = 0\} = \mathcal{E}$, the set of equilibria, and the largest invariant subset of $\dot{V}^{-1}(0)$ is $\mathcal{N}_0 = \dot{V}^{-1}(0) = \mathcal{E}$.

To investigate nontangency, let $f : \mathbb{R}^2 \to \mathbb{R}^2$ denote the right-hand side of (10)–(11), and let $\mathcal{Z} = \{(0,k) : a + bk^2 \cos k = 0\}$. We note that $\mathcal{Z} \subseteq \mathcal{E}$. Since $b \neq 0$, the set $\mathcal{Z}$ is nonempty, and every point of $\mathcal{Z}$ is isolated. For every $x = (y,k) \in \mathcal{N}_0 \backslash \mathcal{Z}$ and every sequence $\{x_i\} = \{(y_i, k_i)\}$ in $\mathbb{R}^2 \backslash \mathcal{E}$ converging to x and satisfying $\operatorname{sign}(y_i) = \operatorname{sign}(y_j)$, $i, j = 1, 2, \ldots$, we have

$$\begin{aligned}\lim_{i\to\infty} \frac{1}{\|f(x_i)\|} f(x_i) &= \lim_{i\to\infty} \frac{1}{\sqrt{y_i^2 + (a + bk_i^2 \cos k_i)^2}} \begin{bmatrix} -\operatorname{sign}(y_i)(a + bk_i^2 \cos k_i) \\ |y_i| \end{bmatrix} \\ &= \begin{bmatrix} -\operatorname{sign}(y_1)\operatorname{sign}(a + bk^2 \cos k) \\ 0 \end{bmatrix} \\ &\in \{[\pm 1 \ 0]^{\mathrm{T}}\}.\end{aligned}$$

Thus, for every $x \in \mathcal{N}_0 \backslash \mathcal{Z}$, $\mathcal{L}_x$ is finite. By Remark 4.3, for every $x \in \mathcal{N}_0 \backslash \mathcal{Z}$, $\mathcal{F}_x = \operatorname{coco} \mathcal{L}_x = \{(c, 0) : c \in \mathbb{R}\}$. On the other hand, for every $x \in \mathcal{N}_0$, $T_x \mathcal{N}_0 = \{(0, c) : c \in \mathbb{R}\}$. Hence f is nontangent to $\mathcal{N}_0$ at every point $x \in \mathcal{N}_0 \backslash \mathcal{Z}$.

In order to apply Theorem 5.1, we note that, in the notation of Theorem 5.1, $\mathcal{N}_1 \subseteq \mathcal{M}_0 \subseteq \mathcal{Z}$. If $\mathcal{N}_1$ is empty, then $\mathcal{M}_1 \subseteq \mathcal{N}_1$ is empty. If $\mathcal{N}_1$ is nonempty, $\mathcal{Z}$ consists only of isolated points and hence $T_x \mathcal{N}_1 = \{0\}$ for every $x \in \mathcal{N}_1$. Consequently, f is nontangent to $\mathcal{N}_1$ at every point in $\mathcal{N}_1$ and hence $\mathcal{M}_1$ is empty. In either case, $\mathcal{M}_1 \subseteq \mathcal{S}$ vacuously. It now follows from Theorem 5.1 that the system (10)–(11) is convergent relative to $\mathbb{R}^2$.

The following two results follow easily from Theorem 5.1.

COROLLARY 5.1. *Suppose $\mathcal{O}_x$ is relatively bounded in $\mathcal{G}$ for all $x \in \mathcal{G}$, and assume that there exists a continuous function $V : \mathcal{G} \to \mathbb{R}$ such that $\dot{V}$ is defined on $\mathcal{G}$ and satisfies $\dot{V}(x) \leq 0$ for all $x \in \mathcal{G}$. Let $\mathcal{N}$ be the largest invariant subset of $\dot{V}^{-1}(0)$. If f is nontangent to $\mathcal{N}$ at every $z \in \mathcal{N}$ relative to $\mathcal{G}$, then the system* (6) *is convergent relative to $\mathcal{G}$.*

Proof. Suppose f is nontangent to $\mathcal{N}$ at every $z \in \mathcal{N}$ relative to $\mathcal{G}$. In the notation of Theorem 5.1, $\mathcal{N}_0 = \mathcal{N}$, while $\mathcal{M}_0 = \varnothing$. Thus $\mathcal{M}_0 \subseteq \mathcal{S}$ vacuously, and the result follows from Theorem 5.1. ∎

COROLLARY 5.2. *Suppose $\mathcal{O}_x$ is relatively bounded in $\mathcal{G}$ for all $x \in \mathcal{G}$, and assume that there exists a continuous function $V : \mathcal{G} \to \mathbb{R}$ such that $\dot{V}$ is defined on $\mathcal{G}$ and satisfies $\dot{V}(x) \leq 0$ for all $x \in \mathcal{G}$. Let $\mathcal{N}$ be the largest invariant subset of $\dot{V}^{-1}(0)$. If every point in $\mathcal{N}$ is Lyapunov stable relative to $\mathcal{G}$, then the system* (6) *is convergent relative to $\mathcal{G}$.*

Proof. Suppose every point in $\mathcal{N}$ is Lyapunov stable relative to $\mathcal{G}$. Then, in the notation of Theorem 5.1, $\mathcal{M}_0 \subseteq \mathcal{N}_0 = \mathcal{N} = \mathcal{S}$. The result now follows from Theorem 5.1. ∎

In the following example, we illustrate Corollary 5.1 by applying it to the system considered in Example 1.1.

Example 5.4. Consider the system described in Example 1.1. Suppose $\alpha \leq \beta$, and let $\mathcal{G} = \mathbb{R}^2$. Consider the Lyapunov function $V : \mathcal{G} \to \mathbb{R}$ given by $V(x) = \frac{1}{4}(x_1^2 + x_2^2 - 1)^2$. It is easy to compute $\dot{V}(x) = -(x_1^2 + x_2^2)|x_1^2 + x_2^2 - 1|^{1+\alpha}$ so that $\dot{V}(x) \leq 0$ for all $x \in \mathcal{G}$ and $\dot{V}^{-1}(0) = \mathrm{S}^1 \cup \{0\}$. Since every point in $\dot{V}^{-1}(0)$ is an equilibrium, the largest invariant subset of $\dot{V}^{-1}(0)$ is $\mathcal{N} = \dot{V}^{-1}(0)$. We know from Example 4.2 that f is nontangent to $\mathcal{N}$ at every point in S^1 relative to $\mathcal{G}$. Since $\{0\}$ is an isolated point of $\mathcal{N}$, it follows that f is nontangent to $\mathcal{N}$ at 0 relative to $\mathcal{G}$. Thus the hypotheses of Corollary 5.1 are satisfied, and we conclude that the system considered in Example 1.1 is convergent relative to $\mathcal{G} = \mathbb{R}^2$ in the case where $\alpha \leq \beta$.

6. Restricted prolongations, Lyapunov stability, and nontangency. In this section, we develop three key ideas that will be needed for the nontangency-based Lyapunov tests for Lyapunov stability, semistability, and asymptotic stability that we present in the next section. The first of these ideas, given in Proposition 6.2, is that an equilibrium is Lyapunov stable if and only if its restricted prolongation, as defined below, is a singleton set. The second key idea, given in Proposition 6.3, is to use Proposition 4.1 to show that the restricted prolongation of an equilibrium of (6) is a singleton set if and only if the vector field f is nontangent to the restricted prolongation at the equilibrium. Since it is not generally possible to find the restricted prolongation of an equilibrium in practice, nontangency of the vector field f to the restricted prolongation is difficult to verify in applications. Since nontangency to any outer estimate of the restricted prolongation implies nontangency to the restricted prolongation itself, the third key idea is to determine outer estimates of the restricted prolongation that are easier to find in practice. Proposition 6.4 gives outer estimates of restricted prolongations in terms of connected components of invariant and negatively invariant subsets of level and sublevel sets of a Lyapunov function and its derivative. This result depends on the invariance properties of restricted prolongations given in Proposition 6.1.

Given a point $x \in \mathcal{G}$ and a relatively open and relatively bounded neighborhood $\mathcal{U} \subseteq \mathcal{G}$ of x, the *restricted prolongation* of x with respect to $\mathcal{U}$ is the set $\mathcal{R}_x^{\mathcal{U}} \subseteq \overline{\mathcal{U}}$ of all subsequential limits of sequences of the form $\{\psi(t_i, x_i)\}$, where $\{t_i\}$ is a sequence in $[0, \infty)$ and $\{x_i\}$ is a sequence in $\mathcal{U}$ converging to x such that the set $\psi([0, t_i] \times \{x_i\})$ is contained in $\overline{\mathcal{U}}$ for every i. It is easy to see that, for every $x \in \mathcal{G}$ and every relatively bounded and open neighborhood $\mathcal{U} \subseteq \mathcal{G}$ of x, $\mathcal{R}_x^{\mathcal{U}}$ contains x and is thus nonempty. The *restricted prolongation* of x relative to $\mathcal{G}$ is the union $\mathcal{R}_x$ of all sets of the form $\mathcal{R}_x^{\mathcal{U}}$ where $\mathcal{U} \subseteq \mathcal{G}$ is a relatively open and relatively bounded neighborhood of x. It can be shown that $\mathcal{R}_x$ is the set of all subsequential limits of sequences of the form $\{\psi(t_i, x_i)\}$, where $\{t_i\}$ is a sequence in $[0, \infty)$ and $\{x_i\}$ is a sequence in $\mathcal{G}$ converging to x such that the set $\cup_i \psi([0, t_i] \times \{x_i\})$ is relatively bounded in $\mathcal{G}$. The restricted prolongation is a subset of the positive prolongation as defined in [7, 8].

The following result gives invariance properties of restricted prolongations.

PROPOSITION 6.1. *Suppose $x \in \mathcal{G}$, and let $\mathcal{U} \subseteq \mathcal{G}$ be a relatively open and relatively bounded neighborhood of x. Then $\mathcal{R}_x^{\mathcal{U}}$ and $\mathcal{R}_x$ are connected. Moreover, if x is an equilibrium, then $\mathcal{R}_x^{\mathcal{U}}$ is negatively invariant and $\mathcal{R}_x$ is invariant.*

Proof. To prove the first part of the proposition, suppose $\mathcal{R}_x^{\mathcal{U}}$ is not connected. Then there exist open and disjoint subsets $\mathcal{V}$ and $\mathcal{W}$ of $\mathbb{R}^n$ such that $\mathcal{R}_x^{\mathcal{U}} \subseteq \mathcal{V} \cup \mathcal{W}$, $\mathcal{R}_x^{\mathcal{U}} \cap \mathcal{V} \neq \varnothing$, and $\mathcal{R}_x^{\mathcal{U}} \cap \mathcal{W} \neq \varnothing$. Since $x \in \mathcal{R}_x^{\mathcal{U}}$, we may assume without loss of generality that $x \in \mathcal{V}$.

Consider $z \in \mathcal{R}_x^{\mathcal{U}} \cap \mathcal{W}$. There exist a sequence $\{x_i\}$ in $\mathcal{G}$ converging to x and a sequence $\{t_i\}$ in $[0, \infty)$ such that $\lim_{i \to \infty} \psi(t_i, x_i) = z$ and $\psi([0, t_i] \times \{x_i\}) \subseteq \overline{\mathcal{U}}$ for every i. Since $\lim_{i \to \infty} \psi(0, x_i) = x$ and $\lim_{i \to \infty} \psi(t_i, x_i) = z$, there exists $k > 0$ such that, for every $i > k$, the connected set $\psi([0, t_i] \times \{x_i\})$ intersects the disjoint, relatively open sets $\mathcal{V}$ and $\mathcal{W}$. We conclude that, for every $i > k$, there exists $\tau_i \in [0, t_i]$ such that $\psi(\tau_i, x_i) \in (\mathcal{V} \cup \mathcal{W})^{\mathrm{c}} = \mathcal{V}^{\mathrm{c}} \cap \mathcal{W}^{\mathrm{c}}$, where $\mathcal{A}^{\mathrm{c}}$ denotes the complement $\mathbb{R}^n \backslash \mathcal{A}$ of the set $\mathcal{A} \subseteq \mathbb{R}^n$. The sequence $\{\psi(\tau_i, x_i)\}$ is relatively bounded in $\mathcal{G}$ and contained in the closed set $\mathcal{V}^{\mathrm{c}} \cap \mathcal{W}^{\mathrm{c}}$. Therefore, the sequence $\{\psi(\tau_i, x_i)\}$ has a subsequential limit in $\mathcal{V}^{\mathrm{c}} \cap \mathcal{W}^{\mathrm{c}}$. However, by definition, every subsequential limit of the sequence $\{\psi(\tau_i, x_i)\}$ is contained in $\mathcal{R}_x^{\mathcal{U}} \subset \mathcal{V} \cup \mathcal{W}$. This contradiction proves that $\mathcal{R}_x^{\mathcal{U}}$ is connected.

It now follows that $\mathcal{R}_x$ is a union of connected sets of the form $\mathcal{R}_x^{\mathcal{U}}$, all of which contain x, and is thus connected [10, Thm. V.1.5].

To prove the second part of the proposition, suppose $x \in \mathcal{G}$ is an equilibrium, and consider $z \in \mathcal{R}_x^{\mathcal{U}}$. There exist a sequence $\{t_i\}$ in $[0, \infty)$ and a sequence $\{x_i\}$ in $\mathcal{G}$ converging to x such that $z = \lim_{i\to\infty} \psi(t_i, x_i)$ and, for every i, $\psi(h, x_i) \in \overline{\mathcal{U}}$ for every $h \in [0, t_i]$.

Now, let $t \geq 0$. If $z = x$, then $\psi(\tau, x) = x = z$ for every $\tau \in [0, t]$. Hence suppose $z \neq x$. If the sequence $\{t_i\}$ has a bounded subsequence, say, $\{t_{i_k}\}$, then we may assume that $\lim_{k\to\infty} t_{i_k} = T$ so that $z = \lim_{k\to\infty} \psi(t_{i_k}, x_{i_k}) = \psi(T, x) = x$. However, this contradicts our assumption that $z \neq x$. Hence we conclude that the sequence $\{t_i\}$ diverges. Therefore, there exists $N > 0$ such that $t_i > t$ for all $i > N$. The sequence $\{\psi(t_{i+N} - t, x_{i+N})\}_{i=1}^{\infty}$ is contained in $\overline{\mathcal{U}}$ and hence relatively bounded in $\mathcal{G}$. Let $y \in \mathcal{G}$ be a subsequential limit point of this sequence. Clearly, $y \in \mathcal{R}_x^{\mathcal{U}}$. Also, by continuity and the semigroup property, $\psi(t, y) = z$, and, for every $\tau \in [0, t]$, $\psi(\tau, y)$ is a subsequential limit of the bounded sequence $\{\psi(t_{i+N} + \tau - t, x_{i+N})\}_{i=1}^{\infty}$ and hence contained in $\mathcal{R}_x^{\mathcal{U}}$. This proves the negative invariance of $\mathcal{R}_x^{\mathcal{U}}$.

From the negative invariance of $\mathcal{R}_x^{\mathcal{U}}$, it follows that $\mathcal{R}_x$ is the union of negatively invariant sets and hence negatively invariant. To prove positive invariance of $\mathcal{R}_x$, let $\tau \geq 0$ and $z \in \mathcal{R}_x$. There exist a sequence $\{t_i\}$ in $[0, \infty)$ and a sequence $\{x_i\}$ in $\mathcal{G}$ converging to x and a compact set $\mathcal{M} \subseteq \mathcal{G}$ such that $z = \lim_{i\to\infty} \psi(t_i, x_i)$ and, for every i and every $h \in [0, t_i]$, $\psi(h, x_i) \in \mathcal{M}$. Now, for every i and every $h \in [0, t_i + \tau]$, $\psi(h, x_i)$ is contained in the compact subset $\psi([0, \tau] \times \mathcal{M})$ of $\mathcal{G}$. Hence a subsequence of the sequence $\{\psi(t_i + \tau, x_i)\}$ converges in $\mathcal{G}$. The limit of this subsequence is $\psi(\tau, z)$ by continuity and the semigroup property and is contained in $\mathcal{R}_x$ by definition. Thus $\psi(\tau, z) \in \mathcal{R}_x$ for every $z \in \mathcal{R}_x$ and $\tau \geq 0$. This proves the positive invariance and hence the invariance of $\mathcal{R}_x$. □

The utility of prolongations in stability analysis stems from the well-known fact that an equilibrium point is Lyapunov stable if and only if the positive prolongation of the equilibrium consists only of the equilibrium point. See, for instance, [7, Prop. 7.3] and [8, Thm. V.1.12]. We prove the same result for restricted prolongations in Proposition 6.2 below. Since, as mentioned above, the restricted prolongation of a point is a subset of the positive prolongation of the point, Proposition 6.2 is a sharper version of the results [7, Prop. 7.3] and [8, Thm. V.1.12]. However, our reason for considering restricted prolongations instead of positive prolongations in this paper is that restricted prolongations are invariant and connected as proved in Proposition 6.1 above. Positive prolongations, on the other hand, may neither be connected nor invariant under our assumptions on the system (6). Since the invariance and connectedness of restricted prolongations play a crucial role in our main results, we introduce restricted prolongations instead of using positive prolongations.

PROPOSITION 6.2. *Suppose $x \in \mathcal{G}$, and let $\mathcal{U} \subseteq \mathcal{G}$ be a relatively open and relatively bounded neighborhood of x. Then the following statements are equivalent.*

(i) *The point x is a Lyapunov stable equilibrium relative to $\mathcal{G}$.*
(ii) $\mathcal{R}_x^{\mathcal{U}} = \{x\}$.
(iii) $\mathcal{R}_x = \{x\}$.

Proof. If x is a Lyapunov stable equilibrium relative to $\mathcal{G}$, then the positive prolongation of x, which contains $\mathcal{R}_x$, is $\{x\}$ [7, Prop. 7.3], [8, Th. V.1.12], and hence $\mathcal{R}_x = \{x\}$. Thus (i) implies (iii).

Since $x \in \mathcal{R}_x^{\mathcal{U}}$ and $\mathcal{R}_x^{\mathcal{U}} \subseteq \mathcal{R}_x$, (iii) implies (ii).

To show that (ii) implies (i), suppose x is not a Lyapunov stable equilibrium

relative to $\mathcal{G}$. Then there exist a neighborhood $\mathcal{V} \subseteq \mathcal{U}$ of x that is relatively bounded and relatively open in $\mathcal{G}$, a sequence $\{x_i\}$ in $\mathcal{V}$ converging to x, and a sequence $\{t_i\}$ in $[0, \infty)$ such that $\psi(t_i, x_i) \in \text{bd } \mathcal{V}$ for every i. Without loss of generality, we may assume that the sequence $\{t_i\}$ is chosen such that, for every i, $\psi(h, x_i) \in \mathcal{V}$ for all $h \in [0, t_i)$. Now, every subsequential limit of the relatively bounded sequence $\{\psi(t_i, x_i)\}$ is distinct from x by construction and is contained in $\mathcal{R}_x^{\mathcal{U}}$ by definition. Thus the negation of (i) implies the negation of (ii). Hence it follows that (ii) implies (i). ∎

The following result characterizes Lyapunov stable equilibrium points in terms of nontangency.

PROPOSITION 6.3. *Let $x \in \mathcal{G}$, and let $\mathcal{U} \subseteq \mathcal{G}$ be a relatively open and relatively bounded neighborhood of x. Then the following statements are equivalent.*

(i) *The point x is a Lyapunov stable equilibrium relative to $\mathcal{G}$.*
(ii) *The vector field f is nontangent to $\mathcal{R}_x^{\mathcal{U}}$ at x relative to $\mathcal{G}$.*
(iii) *The vector field f is nontangent to $\mathcal{R}_x$ at x relative to $\mathcal{G}$.*

Proof. If x is a Lyapunov stable equilibrium relative to $\mathcal{G}$, then it follows from Proposition 6.2 that $\mathcal{R}_x = \{x\}$ and hence $T_x\mathcal{R}_x = \{0\}$. Thus (i) implies (iii).

Since $\mathcal{R}_x^{\mathcal{U}} \subseteq \mathcal{R}_x$, it follows that $T_x\mathcal{R}_x^{\mathcal{U}} \subseteq T_x\mathcal{R}_x$ and hence (iii) implies (ii).

Now, suppose (ii) holds so that $T_x\mathcal{R}_x^{\mathcal{U}} \cap \mathcal{F}_x \subseteq \{0\}$. Let $z \in \mathcal{R}_x^{\mathcal{U}}$. There exist a sequence $\{x_i\}$ converging to x and a sequence $\{t_i\}$ in $[0, \infty)$ such that $\cup_i \psi([0, t_i] \times \{x_i\}) \subset \overline{\mathcal{U}}$ and $\lim_{i\to\infty} \psi(t_i, x_i) = z$.

First, suppose that the sequence $\{t_i\}$ converges to 0. Then by continuity, $z = \lim_{i\to\infty} \psi(t_i, x_i) = \psi(0, x) = x$. Next, suppose the sequence $\{t_i\}$ does not converge to 0. Then there exists a subsequence $\{t_{i_k}\}$ of the sequence $\{t_i\}$ such that $\inf_k t_{i_k} > 0$. Let $\mathcal{I}_k = [0, t_{i_k}]$ for each k, and let $\mathcal{B} \subseteq \overline{\mathcal{U}}$ denote the set of all subsequential limits of sequences of the form $\{\psi(\tau_k, x_{i_k})\}_{k=1}^{\infty}$, where $\tau_k \in \mathcal{I}_k$ for every k. By construction, $z \in \mathcal{B}$ and $\mathcal{B} \subseteq \mathcal{R}_x^{\mathcal{U}}$. Therefore, $T_x\mathcal{B} \cap \mathcal{F}_x \subseteq T_x\mathcal{R}_x^{\mathcal{U}} \cap \mathcal{F}_x \subseteq \{0\}$; that is, f is nontangent to $\mathcal{B}$ at x relative to $\mathcal{G}$. It now follows from Proposition 4.1 that $\mathcal{B} = \{x\}$. Hence $z = x$. Since $z \in \mathcal{R}_x^{\mathcal{U}}$ was arbitrary, it follows that $\mathcal{R}_x^{\mathcal{U}} = \{x\}$. Proposition 6.2 now implies that (i) holds. ∎

Example 6.1. Consider the system described in Example 1.1, and assume that $\alpha \geq \beta + 1$. Let $\mathcal{G} = \mathbb{R}^2$. As discussed in Example 1.1, every point in S^1 is an unstable equilibrium. From Figure 1, we observe that, for every $z \in \mathrm{S}^1$, $\mathcal{R}_z = \mathrm{S}^1$. Since S^1 consists only of equilibrium points, it follows that $\mathcal{R}_z$ is invariant for every $z \in \mathrm{S}^1$. This illustrates Proposition 6.1. As seen in Example 4.2, the vector field f is not nontangent to S^1 at any $z \in \mathrm{S}^1$, thus illustrating the equivalence of (i) and (iii) in Proposition 6.3.

As mentioned earlier, it is not generally possible to find the restricted prolongation of an equilibrium, and hence Proposition 6.3 cannot be directly used in applications. However, the following proposition shows that Lyapunov functions can be used to obtain sets that contain the restricted prolongation of an equilibrium. The proof uses the properties of invariance and connectedness of restricted prolongations given in Proposition 6.1.

PROPOSITION 6.4. *Suppose $V : \mathcal{G} \to \mathbb{R}$ is a continuous function such that $\dot{V}$ is defined on $\mathcal{G}$. Let $x \in \mathcal{G}$ be an equilibrium, and let $\mathcal{W} \subseteq \mathcal{G}$ be a relatively open neighborhood of x. Let $\mathcal{M}_x$ and $\mathcal{N}_x$ denote the largest connected subsets of $V^{-1}(V(x)) \cap \mathcal{W}$ and $\overline{\dot{V}^{-1}(0)} \cap \mathcal{W}$, respectively, that contain x and are negatively invariant. Let $\mathcal{P}_x$ denote the largest connected subset of $\{z \in \mathcal{G} : V(z) \leq V(x)\}$ that contains x and is invariant. Then the following statements hold.*

(i) *If x is a global maximizer of $\dot{V}$ relative to $\mathcal{G}$, then $\mathcal{R}_x \subseteq \mathcal{P}_x$.*

(ii) *If x is a local maximizer of $\dot{V}$ relative to $\mathcal{G}$ and a local minimizer of V relative to the set $\mathcal{K} \stackrel{\text{def}}{=} \mathcal{G} \backslash \overline{\dot{V}^{-1}(0)}$, then there exists a relatively open and relatively bounded neighborhood $\mathcal{V} \subseteq \mathcal{W}$ of x such that $\mathcal{R}_x^{\mathcal{V}} \subseteq \mathcal{M}_x$.*

(iii) $\mathcal{M}_x \subseteq \mathcal{N}_x$.

Proof. (i) Suppose x is a global maximizer of $\dot{V}$ relative to $\mathcal{G}$ so that $\dot{V}(z) \leq \dot{V}(x) = 0$ for every $z \in \mathcal{G}$. Consider $z \in \mathcal{R}_x$. Let $\{x_i\}$ be a sequence in $\mathcal{G}$ converging to x and $\{t_i\}$ a sequence in $[0, \infty)$ such that $\lim_{i \to \infty} \psi(t_i, x_i) = z$. Since V is continuous and decreasing along the solutions of (6), it follows that $V(z) = \lim_{i \to \infty} V(\psi(t_i, x_i)) \leq \lim_{i \to \infty} V(x_i) = V(x)$. Thus $\mathcal{R}_x$ is contained in the set $\{z \in \mathcal{G} : V(z) \leq V(x)\}$. By Proposition 6.1, $\mathcal{R}_x$ is invariant and connected. Also, $x \in \mathcal{R}_x$. Hence it follows that $\mathcal{R}_x \subseteq \mathcal{P}_x$.

(ii) Let x be a local maximizer of $\dot{V}$ relative to $\mathcal{G}$ and a local minimizer of V relative to the set $\mathcal{K}$. Let $\mathcal{U} \subseteq \mathcal{W}$ be a relatively open neighborhood of x such that $\dot{V}(z) \leq \dot{V}(x) = 0$ for every $z \in \mathcal{U}$ and $V(x) \leq V(w)$ for every $w \in \mathcal{U} \cap \mathcal{K}$. Consider $w \in \mathcal{U} \cap \overline{\mathcal{K}}$. There exists a sequence $\{w_i\}$ in $\mathcal{U} \cap \mathcal{K}$ that converges to w. Then $V(w_i) \geq V(x)$ for every i. It now follows by continuity of V that $V(w) = \lim_{i \to \infty} V(w_i) \geq V(x)$. Next let $\mathcal{V} \subseteq \mathcal{G}$ be a relatively open and relatively bounded neighborhood of x such that $\overline{\mathcal{V}} \subseteq \mathcal{U}$. Then, for every $w \in \overline{\mathcal{V}} \cap \overline{\mathcal{K}}$, $V(w) \geq V(x)$.

Now consider $z \in \mathcal{R}_x^{\mathcal{V}}$. Let $\{x_i\}$ be a sequence in $\mathcal{V}$ converging to x and $\{t_i\}$ a sequence in $[0, \infty)$ such that the sequence $\{\psi(t_i, x_i)\}$ converges to z and, for every i, $\psi(\tau, x_i) \in \overline{\mathcal{V}} \subset \mathcal{U}$ for every $\tau \in [0, t_i]$. Then, $V(z) = \lim_{i \to \infty} V(\psi(t_i, x_i)) \leq \lim_{i \to \infty} V(x_i) = V(x)$. Thus $V(z) \leq V(x)$. We claim that $V(z) = V(x)$. To see this, first suppose $z \in \mathcal{R}_x^{\mathcal{V}} \cap \overline{\mathcal{K}} \subseteq \overline{\mathcal{V}} \cap \overline{\mathcal{K}}$. Then, since $V(w) \geq V(x)$ for every $w \in \overline{\mathcal{V}} \cap \overline{\mathcal{K}}$, it follows that $V(z) = V(x)$. Next assume that $z \in \mathcal{R}_x^{\mathcal{V}} \cap (\mathcal{G} \backslash \overline{\mathcal{K}}) = \mathcal{R}_x^{\mathcal{V}} \cap \operatorname{int} \dot{V}^{-1}(0)$. Since $\mathcal{G} \backslash \overline{\mathcal{K}}$ is relatively open in $\mathcal{G}$ and $z \in \mathcal{G} \backslash \overline{\mathcal{K}}$, there exists $m > 0$ such that $\psi(t_i, x_i) \in \mathcal{G} \backslash \overline{\mathcal{K}}$ for every $i > m$. For each $i > m$, let τ_i denote the smallest number in $[0, t_i]$ such that $\psi(t, x_i) \in \mathcal{G} \backslash \overline{\mathcal{K}}$ for all $t \in (\tau_i, t_i]$. The sequence $\{\tau_i\}_{i=m}^{\infty}$ either has a positive subsequence or is eventually zero.

First consider the case where $\{\tau_i\}_{i=m}^{\infty}$ has a positive subsequence, say, $\{\tau_{i_k}\}_{k=1}^{\infty}$. By the continuity of ψ, it follows that $\psi(\tau_{i_k}, x_{i_k}) \in \overline{\mathcal{K}}$ for every k. Since V is nonincreasing along the solutions of (6) in $\mathcal{U}$, $V(\psi(\tau_{i_k}, x_{i_k})) \leq V(x_{i_k})$ for all k, and hence $\lim_{k \to \infty} V(\psi(\tau_{i_k}, x_{i_k})) \leq \lim_{k \to \infty} V(x_{i_k}) = V(x)$. On the other hand, x is a global minimizer of V on the set $\overline{\mathcal{V}} \cap \overline{\mathcal{K}}$ so that $V(\psi(\tau_{i_k}, x_{i_k})) \geq V(x)$ for all k. Thus $\lim_{k \to \infty} V(\psi(\tau_{i_k}, x_{i_k})) = V(x)$ for all k. Since $\psi(t, x_i) \in \mathcal{G} \backslash \overline{\mathcal{K}} \subseteq \dot{V}^{-1}(0)$ for all $t \in (\tau_i, t_i]$ and every i, it follows that $V(\psi(t_{i_k}, x_{i_k})) = V(\psi(\tau_{i_k}, x_{i_k}))$ for all k. Hence $V(z) = \lim_{k \to \infty} V(\psi(t_{i_k}, x_{i_k})) = \lim_{k \to \infty} V(\psi(\tau_{i_k}, x_{i_k})) = V(x)$.

Finally, consider the case where the sequence $\{\tau_i\}_{i=m}^{\infty}$ is eventually zero. In this case, there exists $K > 0$ such that, for every $i > K$ and every $t \in [0, t_i]$, it follows that $V(\psi(t, x_i)) = V(x_i)$. Hence $V(z) = \lim_{i \to \infty} V(\psi(t_i, x_i)) = \lim_{i \to \infty} V(x_i) = V(x)$.

We have thus shown that if $z \in \mathcal{R}_x^{\mathcal{V}}$, then $V(z) = V(x)$. Therefore, $\mathcal{R}_x^{\mathcal{V}} \subseteq V^{-1}(V(x)) \cap \overline{\mathcal{V}} \subset V^{-1}(V(x)) \cap \mathcal{W}$. By Proposition 6.1, $\mathcal{R}_x^{\mathcal{V}}$ is negatively invariant and connected, and $x \in \mathcal{R}_x^{\mathcal{V}}$. Hence $\mathcal{R}_x^{\mathcal{V}} \subseteq \mathcal{M}_x$.

(iii) Consider $z \in \mathcal{M}_x$, and let $t > 0$. By negative invariance, there exists $w \in \mathcal{M}_x$ such that $\psi(t, w) = z$ and $\psi(\tau, w) \in \mathcal{M}_x \subseteq V^{-1}(V(x))$ for all $\tau \in [0, t]$. Hence $V(\psi(\tau, w)) = V(x)$ for every $\tau \in [0, t]$, and thus $\dot{V}(\psi(\tau, w)) = 0$ for every $\tau \in [0, t)$. Let $\{t_i\}$ be a sequence in $[0, t)$ converging to t. Then, by the continuity of ψ, $\{\psi(t_i, w)\}$ is a sequence in $\dot{V}^{-1}(0)$ that converges to z. It follows that $z \in \overline{\dot{V}^{-1}(0)}$. Thus $\mathcal{M}_x \subseteq \overline{\dot{V}^{-1}(0)}$. Since $\mathcal{M}_x$ is negatively invariant, connected, contains x, and is contained in $\mathcal{W}$, the result follows. ∎

It is interesting to note the parallels between the ideas used in this section and those used in the previous section. The propositions given in section 5 allow us to conclude convergence of a solution of (6) under the assumption of nontangency of the vector field f to an outer estimate of the positive limit set of that solution. The results of this section allow us to conclude the Lyapunov stability of an equilibrium of (6) under the assumption of nontangency of the vector field f to an outer estimate of the restricted prolongation of the equilibrium. The outer estimates of the positive limit set in section 5 and the restricted prolongation in this section are in terms of invariant and negatively invariant subsets of level and sublevel sets of a Lyapunov function and its derivative. Moreover, the results of sections 5 and 6 depend on the invariance properties of the positive limit set and the restricted prolongation, respectively.

7. Stability theorems. In this section, we use the results of the previous two sections to derive Lyapunov results for Lyapunov stability, semistability, and asymptotic stability. The main results are Theorems 7.1 and 7.2. These results do not make any assumptions about the sign definiteness of the Lyapunov function. Instead, these results require only that the Lyapunov function derivative be nonpositive and the equilibrium be a local minimizer of the Lyapunov function on the set of points where the Lyapunov function derivative is negative. The weaker assumptions on the Lyapunov function are supplemented by assuming nontangency of the vector field to the level set of the Lyapunov function containing the equilibrium or to the closure of the zero-level set of the Lyapunov function derivative. In both Theorems 7.1 and 7.2, Propositions 6.4 and 5.3 are used to "trap" the restricted prolongation and the positive limit set, respectively, in the level sets of the Lyapunov function and its derivative. Propositions 6.3 and 5.2 are then used to deduce stability and convergence from nontangency.

THEOREM 7.1. *Suppose $V : \mathcal{G} \to \mathbb{R}$ is a continuous function such that $\dot{V}$ is defined on $\mathcal{G}$. Let $x \in \mathcal{E}$ be a local maximizer of $\dot{V}$ relative to $\mathcal{G}$ and a local minimizer of V relative to the set $\mathcal{K} \stackrel{\text{def}}{=} \mathcal{G} \backslash \overline{\dot{V}^{-1}(0)}$. Let $\mathcal{W} \subseteq \mathcal{G}$ be a relatively open neighborhood of x. For every $z \in \mathcal{E} \cap \mathcal{W}$, let $\mathcal{M}_z$ denote the largest connected subset of $V^{-1}(V(z)) \cap \mathcal{W}$ that is negatively invariant and contains z. Let $\mathcal{N}$ denote the largest negatively invariant subset of $\overline{\dot{V}^{-1}(0)} \cap \mathcal{W}$ and, for every $z \in \mathcal{N}$, let $\mathcal{N}_z$ denote the connected component of $\mathcal{N}$ containing z.*

Then the following statements hold.

(i) *If f is nontangent to $\mathcal{M}_x$ at x relative to $\mathcal{G}$, then x is Lyapunov stable relative to $\mathcal{G}$.*

(ii) *If f is nontangent to $\mathcal{N}_x$ at x relative to $\mathcal{G}$, then x is Lyapunov stable relative to $\mathcal{G}$.*

(iii) *If there exists a relatively open neighborhood $\mathcal{U} \subseteq \mathcal{W}$ of x such that every equilibrium in $\mathcal{U}$ is a local minimizer of V relative to $\mathcal{K}$ and, for every $z \in \mathcal{N} \cap \mathcal{U}$, f is nontangent to $\mathcal{N}_z$ at z relative to $\mathcal{G}$, then x is semistable relative to $\mathcal{G}$.*

(iv) *If x is an isolated equilibrium, and there exists a relatively open neighborhood $\mathcal{U} \subseteq \mathcal{W}$ of x such that, for every $z \in \mathcal{N} \cap \mathcal{U}$, f is nontangent to $\mathcal{N}_z$ at z relative to $\mathcal{G}$, then x is asymptotically stable relative to $\mathcal{G}$.*

Proof. (i) Suppose $T_x\mathcal{M}_x \cap \mathcal{F}_x \subseteq \{0\}$. By (ii) of Proposition 6.4, there exists a relatively open and bounded neighborhood $\mathcal{V} \subseteq \mathcal{W}$ of x such that $\mathcal{R}_x^{\mathcal{V}} \subseteq \mathcal{M}_x$. Therefore, $T_x\mathcal{R}_x^{\mathcal{V}} \cap \mathcal{F}_x \subseteq T_x\mathcal{M}_x \cap \mathcal{F}_x \subseteq \{0\}$. Hence, by Proposition 6.3, x is a Lyapunov stable equilibrium relative to $\mathcal{G}$.

(ii) Note that since x is an equilibrium, $x \in \dot{V}^{-1}(0)$. Suppose $T_x\mathcal{N}_x \cap \mathcal{F}_x \subseteq \{0\}$.

By (iii) of Proposition 6.4, $\mathcal{M}_x \subseteq \mathcal{N}_x$. Hence $T_x\mathcal{M}_x \cap \mathcal{F}_x \subseteq T_x\mathcal{N}_x \cap \mathcal{F}_x \subseteq \{0\}$. Therefore, by (i), x is a Lyapunov stable equilibrium relative to $\mathcal{G}$.

(iii) Let $\mathcal{U} \subseteq \mathcal{W}$ be a relatively open neighborhood of x such that every equilibrium in $\mathcal{U}$ is a local minimizer of V relative to $\mathcal{K}$ and $T_z\mathcal{N}_z \cap \mathcal{F}_z \subseteq \{0\}$ for every $z \in \mathcal{N} \cap \mathcal{U}$. Since x is a local maximizer of $\dot{V}$ relative to $\mathcal{G}$, we may assume without loss of generality that $\dot{V}(z) \leq \dot{V}(x) = 0$ for every $z \in \mathcal{U}$. It follows that every equilibrium in $\mathcal{U}$ is a local maximizer of $\dot{V}$ relative to $\mathcal{G}$. Since every equilibrium in $\mathcal{U}$ is contained in $\mathcal{N}$, it follows from (ii) that every equilibrium in $\mathcal{U}$ is Lyapunov stable relative to $\mathcal{G}$. In particular, x is Lyapunov stable relative to $\mathcal{G}$. By Lyapunov stability of x and local compactness of $\mathcal{G}$, there exists a positively invariant neighborhood $\mathcal{V} \subset \mathcal{U}$ of x that is relatively open and relatively bounded in $\mathcal{G}$, and such that $\overline{\mathcal{V}} \subset \mathcal{U}$. For every $z \in \mathcal{V}$, $\mathcal{O}_z \subseteq \mathcal{V}$ is relatively bounded in $\mathcal{G}$. Therefore, by Propositions 5.1 and 5.3, $\mathcal{O}_z^\infty \subseteq \overline{\mathcal{V}} \subset \mathcal{W}$ is nonempty and contained in $\overline{\dot{V}^{-1}(0)}$. The invariance of $\mathcal{O}_z^\infty$ implies that $\mathcal{O}_z^\infty$ is contained in the largest invariant subset of $\overline{\dot{V}^{-1}(0)} \cap \mathcal{W}$. Since every invariant set is also negatively invariant, it follows that $\mathcal{O}_z^\infty \subseteq \mathcal{N}$ for every $z \in \mathcal{V}$. Consider $z \in \mathcal{V}$ and $w \in \mathcal{O}_z^\infty$. Since $\mathcal{O}_z^\infty$ is connected and contained in $\mathcal{N}$, it follows that $\mathcal{O}_z^\infty \subseteq \mathcal{N}_w$. Therefore, $T_w\mathcal{O}_z^\infty \cap \mathcal{F}_w \subseteq T_w\mathcal{N}_w \cap \mathcal{F}_w \subseteq \{0\}$. Now Proposition 5.2 implies that $\lim_{t\to\infty} \psi(t, z)$ exists. Since $z \in \mathcal{V}$ was chosen arbitrarily, it follows that every trajectory in $\mathcal{V}$ converges to a limit. The positive invariance of $\mathcal{V}$ implies that the limit of every trajectory in $\mathcal{V}$ is contained in $\overline{\mathcal{V}}$. Since every equilibrium in $\overline{\mathcal{V}} \subset \mathcal{U}$ is Lyapunov stable relative to $\mathcal{G}$, it follows that x is semistable relative to $\mathcal{G}$.

(iv) Suppose x is an isolated equilibrium, and let $\mathcal{U} \subseteq \mathcal{W}$ be a relatively open neighborhood of x such that $T_z\mathcal{N}_z \cap \mathcal{F}_z \subseteq \{0\}$ for every $z \in \mathcal{N} \cap \mathcal{U}$. Without loss of generality, we may assume that x is the only equilibrium in $\mathcal{U}$. Then, it follows from (iii) that x is semistable relative to $\mathcal{G}$. Asymptotic stability now follows by noting that every isolated equilibrium that is semistable relative to $\mathcal{G}$ is asymptotically stable relative to $\mathcal{G}$. ∎

Remark 7.1. Note that Theorem 7.1 does not require $\dot{V}$ to be continuous. However, in the case where $\dot{V}$ is continuous, we have $\overline{\dot{V}^{-1}(0)} = \dot{V}^{-1}(0)$, and the set $\mathcal{K}$ in Theorem 7.1 is the set of points where the Lyapunov derivative is negative. Thus, in the case where $\dot{V}$ is continuous, Theorem 7.1 requires the equilibrium to be a local minimizer of V only relative to the set of points where the Lyapunov function is strictly decreasing along the solutions of (6).

Example 7.1. In this example, we apply Theorem 7.1 to the mass action kinetics (16) of the Michaelis–Menten chemical reaction introduced in Example 5.2.

Recall that the set of equilibrium concentrations of the reaction in the nonnegative orthant $\mathcal{G}$ is $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$, where $\mathcal{E}_1 = \{x \in \mathcal{G} : x_1 = 0, x_2 = 0, x_3 > 0\}$ and $\mathcal{E}_2 = \{x \in \mathcal{G} : x_1 \geq 0, x_2 = 0, x_3 = 0\}$. Also recall from Example 5.2 that the right-hand side f of the differential equation (16) is nontangent to $\mathcal{E}_1$ at every point in $\mathcal{E}_1$ relative to $\mathcal{G}$.

We claim that every equilibrium in $\mathcal{E}_1$ is semistable relative to $\mathcal{G}$. To show this, let $\alpha \in (1, 1 + k_3/k_2)$, and consider the function $V : \mathcal{G} \to \mathbb{R}$ defined by $V(x) = \alpha x_1 + x_2$. Then, $V(x) \geq 0$ for every $x \in \mathcal{G}$ and $V^{-1}(0) = \overline{\mathcal{E}_1}$. Thus every point in $\mathcal{E}_1$ is a local minimizer of V relative to $\mathcal{G}$. Since $\dot{V} : \mathcal{G} \to \mathbb{R}$ is given by $\dot{V}(x) = [\alpha k_2 - (k_2 + k_3)]x_2 + k_1(1 - \alpha)x_1x_3$, it follows that $\dot{V}(x) \leq 0$ for every $x \in \mathcal{G}$ and $\dot{V}^{-1}(0) = \overline{\dot{V}^{-1}(0)} = \mathcal{E}$ so that the largest negatively invariant subset of $\overline{\dot{V}^{-1}(0)}$ is $\mathcal{E}$.

Now consider an equilibrium $x \in \mathcal{E}_1$. There exists a relatively open neighborhood $\mathcal{U} \subseteq \mathcal{G}$ of x such that $\mathcal{U} \cap \mathcal{E} = \mathcal{U} \cap \mathcal{E}_1$. It now follows that every equilibrium $z \in \mathcal{U} \cap \mathcal{E}$

is a local maximizer of $\dot{V}$ and a local minimizer of V relative to $\mathcal{G}$ and, as shown in Example 5.2, that f is nontangent to $\mathcal{E}$ at every $z \in \mathcal{U} \cap \mathcal{E}$ relative to $\mathcal{G}$. Hence, applying (iii) of Theorem 7.1 with $\mathcal{W} = \mathcal{G}$, it follows that every equilibrium in $\mathcal{E}_1$ is semistable relative to $\mathcal{G}$.

The following corollary of Theorem 7.1 is an extension of Theorems 1 and 2 from [16].

COROLLARY 7.1. *Suppose $V : \mathcal{G} \to \mathbb{R}$ is a continuous function such that $\dot{V}$ is defined on $\mathcal{G}$. Let $x \in \mathcal{E}$ be a local maximizer of $\dot{V}$ relative to $\mathcal{G}$ and a local minimizer of V relative to the set $\mathcal{K} \stackrel{\text{def}}{=} \mathcal{G} \backslash \overline{\dot{V}^{-1}(0)}$. Let $\mathcal{W} \subseteq \mathcal{G}$ be a relatively open neighborhood of x. For every $z \in \mathcal{W}$, let $\mathcal{M}_z$ denote the largest connected subset of $V^{-1}(V(z)) \cap \mathcal{W}$ that is negatively invariant and contains z. Let $\mathcal{N}$ denote the largest negatively invariant subset of $\overline{\dot{V}^{-1}(0)} \cap \mathcal{W}$. Then the following statements hold.*

(i) *If $\mathcal{M}_x = \{x\}$, then x is Lyapunov stable relative to $\mathcal{G}$.*

(ii) *If x is an isolated point of $\mathcal{N}$, then x is asymptotically stable relative to $\mathcal{G}$.*

Proof. (i) Suppose $\mathcal{M}_x = \{x\}$. Then Lyapunov stability of x follows from (i) of Theorem 7.1 by noting that $T_x\mathcal{M}_x = \{0\}$.

(ii) Suppose $\{x\}$ is an isolated point of $\mathcal{N}$. Since every equilibrium in $\mathcal{W}$ is contained in $\mathcal{N}$, it follows that x is an isolated equilibrium. Also, the connected component $\mathcal{N}_x$ of $\mathcal{N}$ containing x is $\{x\}$, and hence $T_x\mathcal{N}_x = \{0\}$. The statement (ii) now follows by applying Theorem 7.1, (iv). ∎

Remark 7.2. Theorems 1 and 2 of [16] follow from (i) and (ii) of Corollary 7.1, respectively. However, while Corollary 7.1 requires only that the equilibrium be a local minimizer of the Lyapunov function relative to the set of points at which the Lyapunov function is strictly decreasing, the results of [16] require the equilibrium to be a local minimizer of the Lyapunov function relative to $\mathcal{G}$. Thus Corollary 7.1 is an extension of the main results of [16]. It is shown in [16] that the main result of [1] follows from Theorem 1 in [16]. Since Corollary 7.1 is an extension of the main results of [16], the arguments presented in [16] can be used to show that the main result of [1] follows from Corollary 7.1.

Remark 7.3. It is interesting to note that neither Corollary 7.1 nor the results of [16] can be applied in the case of Example 7.1, because the level sets of V and $\dot{V}$ containing the equilibrium point of interest contain a continuum of equilibria. Consequently, the equilibrium point of interest is not an isolated point of the largest negatively invariant subset of the level sets of V or $\dot{V}$, and the results mentioned above do not apply.

The following corollary of Theorem 7.1 does not require finding negatively invariant subsets of the level sets of the Lyapunov function and its derivative.

COROLLARY 7.2. *Suppose $V : \mathcal{G} \to \mathbb{R}$ is a continuous function such that $\dot{V}$ is defined on $\mathcal{G}$. Let $x \in \dot{V}^{-1}(0)$ be a local maximizer of $\dot{V}$ relative to $\mathcal{G}$ and a local minimizer of V relative to the set $\mathcal{K} \stackrel{\text{def}}{=} \mathcal{G} \backslash \overline{\dot{V}^{-1}(0)}$. Then the following statements hold.*

(i) *If f is nontangent to $V^{-1}(V(x))$ at x relative to $\mathcal{G}$, then x is a Lyapunov stable equilibrium relative to $\mathcal{G}$.*

(ii) *If f is nontangent to $\overline{\dot{V}^{-1}(0)}$ at x relative to $\mathcal{G}$, then x is a Lyapunov stable equilibrium relative to $\mathcal{G}$.*

(iii) *If there exists a relatively open neighborhood $\mathcal{U} \subseteq \mathcal{G}$ of x such that every equilibrium in $\mathcal{U}$ is a local minimizer of V relative to $\mathcal{K}$ and f is nontangent*

to $\overline{\dot{V}^{-1}(0)}$ *at every point in* $\mathcal{U} \cap \overline{\dot{V}^{-1}(0)}$ *relative to* $\mathcal{G}$*, then* x *is a semistable equilibrium relative to* $\mathcal{G}$*.*

(iv) *If* x *is an isolated point of the set* $\dot{V}^{-1}(0)$*, then* x *is an asymptotically stable equilibrium relative to* $\mathcal{G}$*.*

Proof. Let $\mathcal{V} \subseteq \mathcal{G}$ be a relatively open neighborhood of x such that $\dot{V}(z) \leq \dot{V}(x) = 0$ for every $z \in \mathcal{V}$ and $V(x) \leq V(z)$ for every $z \in \mathcal{K} \cap \mathcal{V}$. There exists $\tau > 0$ such that $\psi(t,x) \in \mathcal{V}$ for all $t \in [0,\tau)$. Therefore, $V(\psi(t,x)) \geq V(x)$ for every $t \in [0,\tau)$. However, since $\dot{V}(\psi(t,x)) \leq 0$ for every $t \in [0,\tau)$, it follows that $V(\psi(t,x)) \leq V(x)$ for every $t \in [0,\tau)$. We conclude that $V(\psi(t,x)) = V(x)$ and $\dot{V}(\psi(t,x)) = 0$ for every $t \in [0,\tau)$. In other words, $\psi(t,x) \in V^{-1}(V(x)) \cap \dot{V}^{-1}(0)$ for every $t \in [0,\tau)$. In particular, $x \in \dot{V}^{-1}(0)$. Also, if $\{t_i\}$ is a sequence in $[0,\tau)$ that converges to zero, then $\{\psi(t_i,x)\}$ is a sequence in $V^{-1}(V(x)) \cap \dot{V}^{-1}(0)$ converging to x such that $\lim_{i\to\infty} \frac{1}{t_i}[\psi(t_i,x) - x] = f(x)$. Thus $f(x) \in T_x\dot{V}^{-1}(0) \cap T_xV^{-1}(V(x))$.

(i) Suppose $T_xV^{-1}(V(x)) \cap \mathcal{F}_x \subseteq \{0\}$. If $\mathcal{F}_x = \varnothing$, then $x \in \operatorname{int} \mathcal{E}$ and $f(x) = 0$. On the other hand, if $\mathcal{F}_x \neq \varnothing$, then $f(x) \in T_xV^{-1}(V(x)) \cap \mathcal{F}_x \subseteq \{0\}$ so that $f(x) = 0$. In either case, it follows that x is an equilibrium. Lyapunov stability of x now follows from (i) of Theorem 7.1 by noting that if $\mathcal{M}_x$ is the largest connected, negatively invariant subset of $V^{-1}(V(x))$ containing x, then $T_x\mathcal{M}_x \cap \mathcal{F}_x \subseteq T_xV^{-1}(V(x)) \cap \mathcal{F}_x \subseteq \{0\}$.

(ii) Suppose $T_x\overline{\dot{V}^{-1}(0)} \cap \mathcal{F}_x \subseteq \{0\}$. Arguing as in the proof of (i), it can be shown that x is an equilibrium. Lyapunov stability of x now follows from (ii) of Theorem 7.1 by noting that if $\mathcal{N}_x$ is the largest connected, negatively invariant subset of $\overline{\dot{V}^{-1}(0)}$ containing x, then $T_x\mathcal{N}_x \cap \mathcal{F}_x \subseteq T_x\overline{\dot{V}^{-1}(0)} \cap \mathcal{F}_x \subseteq \{0\}$.

(iii) Suppose $\mathcal{U} \subseteq \mathcal{G}$ is a relatively open neighborhood of x such that every equilibrium in $\mathcal{U}$ is a local minimizer of V relative to the set $\mathcal{K}$ and $T_z\overline{\dot{V}^{-1}(0)} \cap \mathcal{F}_z \subseteq \{0\}$ for every $z \in \mathcal{U} \cap \overline{\dot{V}^{-1}(0)}$. Arguing as in the proof of (i), it can be shown that x is an equilibrium. The result now follows from (iii) of Theorem 7.1 by noting that, if $\mathcal{N}$ is the largest negatively invariant subset of $\overline{\dot{V}^{-1}(0)}$, then, for every $z \in \mathcal{N} \cap \mathcal{U} \subseteq \overline{\dot{V}^{-1}(0)} \cap \mathcal{U}$, it follows that $T_z\mathcal{N} \cap \mathcal{F}_z \subseteq T_z\overline{\dot{V}^{-1}(0)} \cap \mathcal{F}_z \subseteq \{0\}$.

(iv) Suppose x is an isolated point of $\dot{V}^{-1}(0)$. Then x is also an isolated point of $\overline{\dot{V}^{-1}(0)}$. Since $\psi(t,x) \in \dot{V}^{-1}(0)$ for all $t \in [0,\tau)$, it follows from the continuity of ψ that $\psi(t,x) = x$ for all $t \in [0,\tau)$. In other words, x is an equilibrium. Since every equilibrium is contained in $\dot{V}^{-1}(0)$, it follows that x is an isolated equilibrium. Since x is an isolated point of $\overline{\dot{V}^{-1}(0)}$, it follows that $T_x\overline{\dot{V}^{-1}(0)} = \{0\}$. Hence (iii) implies that x is semistable relative to $\mathcal{G}$. Since an isolated semistable equilibrium is asymptotically stable, (iv) follows. ∎

Example 7.2. Consider the system described in Example 1.1. Suppose $\alpha \leq \beta$, and let $\mathcal{G} = \mathbb{R}^2$. Consider the Lyapunov function $V : \mathcal{G} \to \mathbb{R}$ given by $V(x) = \frac{1}{4}(x_1^2 + x_2^2 - 1)^2$ introduced in Example 5.4, and recall that $\dot{V}(x) = -(x_1^2+x_2^2)|x_1^2+x_2^2-1|^{1+\alpha}$ so that $\dot{V}(x) \leq 0$ for all $x \in \mathcal{G}$ and $\overline{\dot{V}^{-1}(0)} = \mathrm{S}^1 \cup \{0\}$. The set $\mathcal{K} = \mathcal{G}\backslash\overline{\dot{V}^{-1}(0)}$ in Corollary 7.2 is given in this case by $\mathcal{K} = \mathcal{G}\backslash(\mathrm{S}^1 \cup \{0\})$. Hence every point $x \in \mathrm{S}^1$ is a local minimizer of V relative to $\mathcal{K}$. We know from Example 4.2 that f is nontangent to S^1 and hence to $\dot{V}^{-1}(0)$ at every $z \in \mathrm{S}^1$ relative to $\mathcal{G}$. Hence by (iii) of Corollary 7.2, it follows that every $x \in \mathrm{S}^1$ is semistable relative to $\mathcal{G} = \mathbb{R}^2$.

The following theorem gives stability conditions in terms of invariant sets rather than negatively invariant sets as in Theorem 7.1. Also, the first part of the theorem does not require the equilibrium to be a local minimizer of the Lyapunov function.

THEOREM 7.2. *Suppose $V : \mathcal{G} \to \mathbb{R}$ is a continuous function such that $\dot{V}$ is defined on $\mathcal{G}$. For every $z \in \mathcal{G}$, let $\mathcal{P}_z$ denote the largest connected subset of $\{w \in \mathcal{G} : V(w) \leq V(z)\}$ that is invariant and contains z. Let $x \in \mathcal{E}$ be a global maximizer of $\dot{V}$ relative to $\mathcal{G}$. Let $\varepsilon > 0$, and let $\mathcal{Q}$ be the largest invariant subset of the set $\{w \in \mathcal{G} : V(w) < V(x) + \varepsilon\}$. For every $z \in \mathcal{Q}$, let $\mathcal{Q}_z$ denote the connected component of $\mathcal{Q}$ containing z. Then the following statements hold.*

(i) *If f is nontangent to $\mathcal{P}_x$ at x relative to $\mathcal{G}$, then x is Lyapunov stable relative to $\mathcal{G}$.*

(ii) *If there exists a relatively open neighborhood $\mathcal{U} \subseteq \mathcal{G}$ of x such that, for every $z \in \mathcal{U} \cap \mathcal{Q}$, f is nontangent to $\mathcal{Q}_z$ at z relative to $\mathcal{G}$, then x is semistable relative to $\mathcal{G}$.*

(iii) *If x is an isolated point of $\mathcal{Q}$, then x is asymptotically stable relative to $\mathcal{G}$.*

If, in addition, x is a local minimizer of V relative to the set $\mathcal{K} \stackrel{\text{def}}{=} \mathcal{G} \backslash \overline{\dot{V}^{-1}(0)}$, then the following statements hold.

(iv) *If f is nontangent to $\mathcal{P}_x \cap V^{-1}(V(x))$ at x relative to $\mathcal{G}$, then x is Lyapunov stable relative to $\mathcal{G}$.*

(v) *If f is nontangent to $\mathcal{P}_x \cap \overline{\dot{V}^{-1}(0)}$ at x relative to $\mathcal{G}$, then x is Lyapunov stable relative to $\mathcal{G}$.*

(vi) *If there exists a relatively open neighborhood $\mathcal{U} \subseteq \mathcal{G}$ of x such that every equilibrium in $\mathcal{U}$ is a local minimizer of V relative to the set $\mathcal{K}$ and, for every $z \in \mathcal{U} \cap \mathcal{Q}$, f is nontangent to $\mathcal{Q}_z \cap \overline{\dot{V}^{-1}(0)}$ at z relative to $\mathcal{G}$, then x is semistable relative to $\mathcal{G}$.*

Proof. Note that by (i) of Proposition 6.4, $\mathcal{R}_x \subseteq \mathcal{P}_x$.

(i) Suppose $T_x\mathcal{P}_x \cap \mathcal{F}_x \subseteq \{0\}$. We have $T_x\mathcal{R}_x \cap \mathcal{F}_x \subseteq T_x\mathcal{P}_x \cap \mathcal{F}_x \subseteq \{0\}$. Hence it follows from Proposition 6.3 that x is Lyapunov stable relative to $\mathcal{G}$.

(ii) Let $\mathcal{U} \subseteq \mathcal{G}$ be a relatively open neighborhood of x such that $T_z\mathcal{Q}_z \cap \mathcal{F}_z \subseteq \{0\}$ for every $z \in \mathcal{U} \cap \mathcal{Q}$. Since V is continuous, we can assume without loss of generality that $V(z) < V(x) + \varepsilon$ for every $z \in \mathcal{U}$. Then, for every $z \in \mathcal{U}$, $\{w \in \mathcal{G} : V(w) \leq V(z)\} \subseteq \{w \in \mathcal{G} : V(w) < V(x) + \varepsilon\}$. Consequently, $\mathcal{P}_z \subseteq \mathcal{Q}_z$ for every $z \in \mathcal{U}$. In particular, if $z \in \mathcal{U}$ is an equilibrium, then $z \in \mathcal{P}_z \subseteq \mathcal{Q}_z$ so that $T_z\mathcal{P}_z \cap \mathcal{F}_z \subseteq T_z\mathcal{Q}_z \cap \mathcal{F}_z \subseteq \{0\}$. It therefore follows from (i) that every equilibrium in $\mathcal{U}$ is Lyapunov stable relative to $\mathcal{G}$. In particular, x is Lyapunov stable relative to $\mathcal{G}$. By Lyapunov stability of x and local compactness of $\mathcal{G}$, there exists a positively invariant neighborhood $\mathcal{V} \subset \mathcal{U}$ of x that is relatively open and bounded in $\mathcal{G}$ and such that $\overline{\mathcal{V}} \subset \mathcal{U}$. Consider $z \in \mathcal{V}$. Then $\mathcal{O}_z \subseteq \mathcal{V}$ is relatively bounded in $\mathcal{G}$, and hence by Propositions 5.1 and 5.3, $\mathcal{O}_z^\infty \subseteq \overline{\mathcal{V}} \subset \mathcal{U}$ is nonempty, connected, and contained in the largest invariant subset of $\{w \in \mathcal{G} : V(w) \leq V(z)\}$ and hence in $\mathcal{Q}$. Next, let $w \in \mathcal{O}_z^\infty$. Since $\mathcal{O}_z^\infty \subseteq \mathcal{Q}$ is connected, it follows that $\mathcal{O}_z^\infty \subseteq \mathcal{Q}_w$. Now, since $w \in \overline{\mathcal{V}} \cap \mathcal{Q} \subseteq \mathcal{U} \cap \mathcal{Q}$, we have $T_w\mathcal{O}_z^\infty \cap \mathcal{F}_w \subseteq T_w\mathcal{Q}_w \cap \mathcal{F}_w \subseteq \{0\}$. Proposition 5.2 now implies that $\lim_{t\to\infty} \psi(t, z)$ exists and is contained in $\overline{\mathcal{V}} \subseteq \mathcal{G}$. Since $z \in \mathcal{V}$ was chosen arbitrarily, it follows that every trajectory in $\mathcal{V}$ converges to a limit. The positive invariance of $\mathcal{V}$ implies that the limit of every trajectory in $\mathcal{V}$ is contained in $\overline{\mathcal{V}}$. Since every equilibrium in $\overline{\mathcal{V}} \subset \mathcal{U}$ is Lyapunov stable relative to $\mathcal{G}$, it follows that x is semistable relative to $\mathcal{G}$.

(iii) Suppose x is an isolated point of $\mathcal{Q}$, and let $\mathcal{U} \subseteq \mathcal{G}$ be a relatively open neighborhood of x such that $\mathcal{U} \cap \mathcal{Q} = \{x\}$. Then $T_x\mathcal{Q} = \{0\}$, and hence every point $z \in \mathcal{U} \cap \mathcal{Q}$ satisfies $T_z\mathcal{Q} \cap \mathcal{F}_z \subseteq \{0\}$. By (ii), x is semistable relative to $\mathcal{G}$. Next, consider the relatively open neighborhood $\mathcal{V} = \{z \in \mathcal{G} : V(z) < V(x) + \varepsilon\}$ of x. Since every equilibrium in $\mathcal{V}$ is contained in $\mathcal{Q}$, we have $(\mathcal{U} \cap \mathcal{V}) \cap \mathcal{E} = \mathcal{U} \cap (\mathcal{V} \cap \mathcal{E}) \subseteq \mathcal{U} \cap \mathcal{Q} = \{x\}$.

Since $\mathcal{U} \cap \mathcal{V}$ is a relatively open neighborhood of x, it follows that x is an isolated equilibrium. The result now follows by noting that an isolated equilibrium that is semistable relative to $\mathcal{G}$ is also asymptotically stable relative to $\mathcal{G}$.

Now, suppose x is a local minimizer of V relative to the set $\mathcal{K} = \mathcal{G} \backslash \overline{\dot{V}^{-1}(0)}$. By (ii) and (iii) of Proposition 6.4, there exists a neighborhood $\mathcal{V}$ of x that is open and bounded relative to $\mathcal{G}$, and such that $\mathcal{R}_x^{\mathcal{V}} \subseteq V^{-1}(V(x)) \cap \overline{\dot{V}^{-1}(0)}$.

(iv) Suppose $T_x(\mathcal{P}_x \cap V^{-1}(V(x))) \cap \mathcal{F}_x \subseteq \{0\}$. Then $T_x \mathcal{R}_x^{\mathcal{V}} \cap \mathcal{F}_x \subseteq T_x(\mathcal{R}_x \cap V^{-1}(V(x))) \cap \mathcal{F}_x \subseteq T_x(\mathcal{P}_x \cap V^{-1}(V(x))) \cap \mathcal{F}_x \subseteq \{0\}$. It now follows from Proposition 6.3 that x is Lyapunov stable relative to $\mathcal{G}$.

(v) Suppose $T_x(\mathcal{P}_x \cap \overline{\dot{V}^{-1}(0)}) \cap \mathcal{F}_x \subseteq \{0\}$. Then $T_x \mathcal{R}_x^{\mathcal{V}} \cap \mathcal{F}_x \subseteq T_x(\mathcal{R}_x \cap \overline{\dot{V}^{-1}(0)}) \cap \mathcal{F}_x \subseteq T_x(\mathcal{P}_x \cap \overline{\dot{V}^{-1}(0)}) \cap \mathcal{F}_x \subseteq \{0\}$. It now follows from Proposition 6.3 that x is Lyapunov stable relative to $\mathcal{G}$.

(vi) Let $\mathcal{U} \subseteq \mathcal{G}$ be a relatively open neighborhood of x such that every equilibrium in $\mathcal{U}$ is a local minimizer of V relative to the set $\mathcal{K}$, and every point z in $\mathcal{U} \cap \mathcal{Q}$ satisfies $T_z(\mathcal{Q}_z \cap \overline{\dot{V}^{-1}(0)}) \cap \mathcal{F}_z \subseteq \{0\}$. Without loss of generality, we assume that $V(z) < V(x) + \varepsilon$ for every $z \in \mathcal{U}$. Then, for every $z \in \mathcal{U}$, $\mathcal{P}_z \subseteq \mathcal{Q}_z$ and hence $T_z(\mathcal{P}_z \cap \overline{\dot{V}^{-1}(0)}) \cap \mathcal{F}_z \subseteq T_z(\mathcal{Q}_z \cap \overline{\dot{V}^{-1}(0)}) \cap \mathcal{F}_z \subseteq \{0\}$. It therefore follows from (v) that every equilibrium in $\mathcal{U}$ is Lyapunov stable relative to $\mathcal{G}$. In particular, x is Lyapunov stable relative to $\mathcal{G}$. By Lyapunov stability of x and local compactness of $\mathcal{G}$, there exists a positively invariant, relatively open, and relatively bounded neighborhood $\mathcal{V} \subset \mathcal{G}$ of x such that $\overline{\mathcal{V}} \subset \mathcal{U}$. Consider $z \in \mathcal{V}$. Then $\mathcal{O}_z \subseteq \mathcal{V}$ is relatively bounded in $\mathcal{G}$, and hence by Propositions 5.1 and 5.3, $\mathcal{O}_z^\infty \subseteq \overline{\mathcal{V}}$ is nonempty, connected, and contained in the largest invariant subset of $\{w \in \mathcal{G} : V(w) \leq V(z)\} \subset \{w \in \mathcal{G} : V(w) < V(x) + \epsilon\}$ and hence in $\mathcal{Q}$. Next, let $w \in \mathcal{O}_z^\infty$. Since $\mathcal{O}_z^\infty \subseteq \mathcal{Q}$ is connected, it follows that $\mathcal{O}_z^\infty \subseteq \mathcal{Q}_w$. By Proposition 5.3, $\mathcal{O}_z^\infty$ is also contained in $\dot{V}^{-1}(0)$. Now, since $w \in \overline{\mathcal{V}} \cap \mathcal{Q} \subseteq \mathcal{U} \cap \mathcal{Q}$, we have $T_w \mathcal{O}_z^\infty \cap \mathcal{F}_w \subseteq T_w(\mathcal{Q}_w \cap \dot{V}^{-1}(0)) \cap \mathcal{F}_w \subseteq T_w(\mathcal{Q}_w \cap \overline{\dot{V}^{-1}(0)}) \cap \mathcal{F}_w \subseteq \{0\}$. Proposition 5.2 now implies that $\lim_{t \to \infty} \psi(t, z)$ exists and is contained in $\mathcal{G}$. Since $z \in \mathcal{V}$ was chosen to be arbitrary, it follows that every trajectory in $\mathcal{V}$ converges to an equilibrium in $\overline{\mathcal{V}} \subset \mathcal{U}$. Semistability of x relative to $\mathcal{G}$ now follows by noting that every equilibrium in $\overline{\mathcal{V}} \subset \mathcal{U}$ is Lyapunov stable relative to $\mathcal{G}$. □

Example 7.3. In this example, we use Theorem 7.2 to show that every equilibrium $(0, k)$ of the adaptive closed-loop system (10)–(11) satisfying $g'(k) = a + bk^2 \cos k > 0$ is semistable, where g' is the derivative of the function $g : k \mapsto ak + b(k^2 - 2) \sin k + 2bk \cos k$ introduced in Example 3.1.

Suppose $k_0 \in \mathbb{R}$ is such that $g'(k_0) > 0$. Since g' is continuous, there exist $\delta > 0$ and $r > 0$ such that $g'(k) > r$ for all $k \in [k_0 - \delta, k_0 + \delta]$. In particular, we note that g is increasing on the interval $[k_0 - \delta, k_0 + \delta]$.

We claim that the equilibrium $(0, k_0)$ of the system (10)–(11) is semistable. To show this, consider the Lyapunov function $V(y, k) = e^{-k}$ introduced in Example 5.3, and let $\varepsilon = V(0, k_0)(e^\delta - 1)$. It is easy to see that the set $\{(y, k) \in \mathbb{R}^2 : V(y, k) \leq V(0, k_0) + \varepsilon\} = \{(y, k) \in \mathbb{R}^2 : k \geq k_0 - \delta\}$. Let $\mathcal{Q}$ denote the largest invariant subset of $\{(y, k) \in \mathbb{R}^2 : k \geq k_0 - \delta\}$. Also, recall from Example 5.3 that $\dot{V}(y, k) = -e^{-k} y^2$ and $\dot{V}^{-1}(0) = \{(0, k) : k \in \mathbb{R}\} = \mathcal{E}$.

Let $U : \mathbb{R}^2 \to \mathbb{R}$ be the function $U(y, k) = \frac{1}{2} y^2 + g(k)$, and recall from Example 3.1 that U is weakly proper and $\dot{U} \equiv 0$. Choose $k_1 \in (k_0, k_0 + \delta)$, and let $c = U(0, k_1) = g(k_1)$. Note that since g is increasing on $[k_0, k_0 + \delta]$, $U(0, k_0) = g(k_0) < g(k_1) = c$. Let $\mathcal{U}$ be the connected component of the set $\{(y, k) \in \mathbb{R}^2 : U(y, k) < c\}$ containing $(0, k_0)$. The set $\mathcal{U}$ is open by continuity and bounded by weak properness of U.

Moreover, letting ρ denote the projection $(y,k) \mapsto k$, the set $\rho(\mathcal{U})$ is a connected set that contains k_0 and does not contain $k_1 > k_0$. Hence we conclude that, for every $(y,k) \in \mathcal{U}$, $k < k_1 < k_0 + \delta$. It follows that every $(y,k) \in \mathcal{U} \cap \mathcal{Q}$ satisfies $k \in [k_0 - \delta, k_0 + \delta]$. Also, since $\mathcal{U}$ is bounded, it follows that there exists $Y \geq 0$ such that every $(y,k) \in \mathcal{U}$ satisfies $|y| \leq Y$.

We claim that $\mathcal{U} \cap \mathcal{Q}$ is contained in $\dot{V}^{-1}(0)$. To see this, consider $(y,k) \in \mathcal{U} \cap \mathcal{Q}$, and let $\tau > 0$. By the invariance of $\mathcal{Q}$, there exists $(y_2,k_2) \in \mathcal{Q}$ such that $(\phi_1(\tau), \phi_2(\tau)) = (y,k)$, where $\phi : \mathbb{R} \to \mathbb{R}^2$ denotes the solution of (10)–(11) satisfying $\phi(0) = (y_2,k_2)$. Since $\dot{U} \equiv 0$, $U(\phi(t)) = U(y,k) < c$ for every $t \in [0,\tau]$. Since, for every $t \in [0,\tau]$, both (y,k) and $\phi(t)$ are contained in the connected subset $\phi([0,\tau])$ of the set $\{(y,k) : U(y,k) < c\}$, it follows that $\phi(t) \in \mathcal{U}$ for every $t \in \mathcal{U}$. Thus $\phi(t) \in \mathcal{U} \cap \mathcal{Q}$ for every $t \in [0,\tau]$. Consequently, $\phi_2(t) \in [k_0 - \delta, k_0 + \delta]$ for all $t \in [0,\tau]$. Next we compute $\frac{d}{dt}(\phi_1(t))^2 = -2g'(\phi_2(t))(\phi_1(t))^2$ so that $y^2 = (\phi_2(\tau))^2 = y_2^2 \exp[-2\int_0^\tau g'(\phi_2(t))dt] \leq Y^2 \exp(-2r\tau)$. Since $\tau > 0$ was chosen arbitrarily, it follows that $y = 0$. Thus $\mathcal{U} \cap \mathcal{Q} \subseteq \dot{V}^{-1}(0)$.

It was shown in Example 5.3 that, for every $(y,k) \in \dot{V}^{-1}(0)$ such that $g'(k) \neq 0$, the vector field f defined by the right-hand side of (10)–(11) is nontangent to $\dot{V}^{-1}(0)$. Hence, for every point $(0,k) \in \mathcal{U} \cap \mathcal{Q}$, $T_z\mathcal{Q} \cap \mathcal{F}_z = T_z(\mathcal{U} \cap \mathcal{Q}) \cap \mathcal{F}_z \subseteq T_z(\dot{V}^{-1}(0)) \cap \mathcal{F}_z \subseteq \{0\}$. Thus f is nontangent to $\mathcal{Q}$ at every point in $\mathcal{U} \cap \mathcal{Q}$. By (ii) of Theorem 7.2, we conclude that the equilibrium $(0,k_0)$ is semistable.

Remark 7.4. It is worthwhile to note that Theorem 7.1 cannot be applied in Example 7.3 using the Lyapunov function V, because none of the equilibria of the closed-loop system (10)–(11) are local minimizers of the function V. In fact, the Lyapunov function V chosen in the example does not have any local minimizers.

Remark 7.5. The convergence of the closed-loop system (10)–(11) is also proved in [17, p. 46]. However, unlike the proof given in [17], our proof of convergence given in Example 5.3 is based on Lyapunov analysis. Moreover, in Example 7.3 we go beyond convergence and identify a class of semistable equilibria for the closed-loop system.

Remark 7.6. Theorems 7.1 and 7.2 involve hypotheses on the Lyapunov function and its derivative that are weaker than sign definiteness or sign semidefiniteness. For instance, in Theorem 7.1 and (iv)–(vi) of Theorem 7.2, the Lyapunov function is required to have a local nonstrict minimum at the equilibrium point only with respect to the set of points where the Lyapunov function derivative is negative, while in (i)–(iii) of Theorem 7.2, the Lyapunov function is not required to have even a local minimum at the equilibrium point of interest. Consequently, Theorems 7.1 and 7.2 can be used to deduce stability of all equilibria in a continuum by considering a single Lyapunov function for all the equilibria. This makes Theorems 7.1 and 7.2 particularly suited for applications to systems having a continuum of equilibria. Examples 7.1 and 7.3 illustrate these observations.

Our final result is the following corollary of Theorem 7.2.

COROLLARY 7.3. *Suppose $V : \mathcal{G} \to \mathbb{R}$ is a continuous function such that $\dot{V}$ is defined on $\mathcal{G}$. Let $x \in \mathcal{E}$ be a global minimizer of V relative to $\mathcal{G}$ and a global maximizer of $\dot{V}$ relative to $\mathcal{G}$. Let $\mathcal{P}$ denote the largest subset of $V^{-1}(V(x))$ that is invariant and connected and contains x. Let $\mathcal{N}$ denote the largest invariant subset of $\overline{\dot{V}^{-1}(0)}$. Then the following statements hold.*

(i) *If $\mathcal{P} = \{x\}$, then x is Lyapunov stable relative to $\mathcal{G}$.*

(ii) *If x is an isolated point of $\mathcal{N}$, then x is asymptotically stable relative to $\mathcal{G}$.*

Proof. (i) Since x is a global minimizer of V relative to $\mathcal{G}$, it follows that $\mathcal{P}$ is also the largest connected subset of the set $\{z \in \mathcal{G} : V(z) \leq V(x)\}$ that is invariant and contains x. If $\mathcal{P} = \{x\}$, then $T_x\mathcal{P} \cap \mathcal{F}_x \subseteq \{0\}$, and hence the result follows from (i) of Theorem 7.2.

(ii) Since $\mathcal{P}$ is an invariant subset of a level set of V, it follows that $\mathcal{P} \subseteq \dot{V}^{-1}(0)$. Thus $\mathcal{P}$ is a connected invariant subset of $\overline{\dot{V}^{-1}(0)}$ containing x, and hence $\mathcal{P} \subseteq \mathcal{N}$. Now suppose x is an isolated point of $\mathcal{N}$. Then $\mathcal{P} = \{x\}$, and hence, by (i), x is Lyapunov stable relative to $\mathcal{G}$. By Lyapunov stability of x and local compactness of $\mathcal{G}$, there exists a relatively open, relatively bounded, and positively invariant neighborhood $\mathcal{U} \subseteq \mathcal{G}$ of x such that $\overline{\mathcal{U}} \cap \mathcal{N} = \{x\}$. For every $z \in \mathcal{U}$, $\mathcal{O}_z \subseteq \mathcal{U}$ is relatively bounded. By Proposition 5.3, for every $z \in \mathcal{U}$, $\mathcal{O}_z^\infty \subseteq \overline{\mathcal{O}_z} \subseteq \overline{\mathcal{U}}$ is contained in $\mathcal{N}$. Since $\overline{\mathcal{U}} \cap \mathcal{N} = \{x\}$, it follows that $\mathcal{O}_z^\infty = \{x\}$ for every $z \in \mathcal{U}$. Asymptotic stability of x now follows. ∎

Remark 7.7. It is interesting to compare (ii) of Corollary 7.3 with the Krasovskii–LaSalle theorem for asymptotic stability [20, Thm. 14.1]. In the result from [20], the Lyapunov function and its derivative are assumed to be locally positive definite and locally negative semidefinite, respectively, at the equilibrium. In other words, the equilibrium is assumed to be a strict local minimizer for the Lyapunov function and a local nonstrict maximizer for the Lyapunov function derivative. On the other hand, Corollary 7.3 assumes the equilibrium to be a global nonstrict minimizer and a global nonstrict maximizer for the Lyapunov function and its derivative, respectively. In other words, Corollary 7.3 assumes the Lyapunov function to be only positive semidefinite globally instead of positive definite locally at the equilibrium. Thus, both (ii) of Corollary 7.3 and Theorem 14.1 of [20] assert asymptotic stability of the equilibrium under two alternative and complementary sets of assumptions.

Remark 7.8. It is also interesting to compare Corollary 7.3 with Corollary 7.1. While Corollary 7.3 requires the equilibrium to be a global minimizer and a global maximizer of V and $\dot{V}$, respectively, Corollary 7.1 requires only that the equilibrium be a local minimizer and a local maximizer, respectively, of V and $\dot{V}$. Thus the hypotheses of Corollary 7.1 on the Lyapunov function and its derivative are weaker than those of Corollary 7.3. However, Corollary 7.1 requires the equilibrium point to be an isolated point of the largest negatively invariant subsets of level sets of V and $\dot{V}$, while Corollary 7.3 requires only that the equilibrium be an isolated point of the largest invariant subsets of level sets of V and $\dot{V}$. Since invariant sets are also negatively invariant, it is clear that the hypotheses of Corollary 7.3 relating to invariant subsets are weaker than the corresponding hypotheses in Corollary 7.1. Thus both corollaries assert Lyapunov and asymptotic stability under alternative and complementary sets of conditions.

8. Conclusions. This paper introduces convergence and semistability as two notions of importance in the study of systems having a continuum of equilibria. The main contribution of this paper has been the introduction of the notion of nontangency and its application in the Lyapunov analysis of systems having a continuum of equilibria. Positive limit sets and restricted prolongations play a key role in the application of nontangency to Lyapunov analysis of convergence and stability, respectively. We introduce restricted prolongations in the paper, establish their invariance properties, and give inclusion results for restricted prolongations in terms of invariant and negatively invariant subsets of the level and sublevel sets of a Lyapunov function and its derivative. Using nontangency, we obtain Lyapunov results for convergence, Lyapunov stability, semistability, and asymptotic stability. The results on Lyapunov

stability and asymptotic stability involve hypotheses on the Lyapunov function and its derivative that are weaker than sign definiteness or semidefiniteness. This makes our results particularly suited for applications to systems having a continuum of equilibria. The weaker hypotheses on the Lyapunov function and its derivative are supplemented by assuming nontangency of the vector field to appropriate subsets of the level and sublevel sets of the Lyapunov function and its derivative. We illustrate the main results by applying them to an example from chemical kinetics and an example from adaptive control.

Appendix. Proof of Proposition 4.2. We present here the proof of Proposition 4.2. First, we recall some definitions related to set convergence [2, 26]. Consider a sequence $\{\mathcal{W}_k\}$ of subsets of $\mathbb{R}^n$. The limit superior of the sequence, denoted $\limsup_{k\to\infty} \mathcal{W}_k$, is the set of all subsequential limits of sequences $\{w_k\}$ in $\mathbb{R}^n$ such that $w_k \in \mathcal{W}_k$ for every k. The limit inferior of the sequence, denoted by $\liminf_{k\to\infty} \mathcal{W}_k$, is the set of limits of convergent sequences $\{w_k\}$ in $\mathbb{R}^n$ such that $w_k \in \mathcal{W}_k$ for every k. The sequence $\{\mathcal{W}_k\}$ *converges* to the set $\mathcal{W} \subseteq \mathbb{R}^n$ if $\mathcal{W} = \liminf_{k\to\infty} \mathcal{W}_k = \limsup_{k\to\infty} \mathcal{W}_k$.

The proof of Proposition 4.2 requires two basic results on set convergence. The first result is that if $\{\mathcal{W}_k\}$ is a sequence of subsets of a bounded subset of $\mathbb{R}^n$, then the sequence $\{\mathcal{W}_k\}$ has a subsequence that converges to a nonempty set. This result follows from Theorem 1.1.7 of [2] and Theorem 4.18 of [26]. The second result is given by the following lemma.

LEMMA A.1. *Suppose $\{\mathcal{W}_k\}$ is a sequence of connected subsets of a bounded set $\mathcal{B} \subset \mathbb{R}^n$ that converges to a set $\mathcal{W} \subseteq \mathbb{R}^n$. Then $\mathcal{W}$ is connected.*

Proof. Suppose $\mathcal{W}$ is not connected. Then there exist two disjoint open sets $\mathcal{U} \subseteq \mathbb{R}^n$ and $\mathcal{V} \subseteq \mathbb{R}^n$ such that $\mathcal{W} \subseteq \mathcal{U} \cup \mathcal{V}$ and $\mathcal{W} \cap \mathcal{U}$ and $\mathcal{W} \cap \mathcal{V}$ are nonempty. Since $\mathcal{W} = \liminf_{k\to\infty} \mathcal{W}_k$, there exist convergent sequences $\{u_k\}$ and $\{v_k\}$ in $\mathbb{R}^n$ such that $\lim_{k\to\infty} u_k \in \mathcal{W} \cap \mathcal{U}$, $\lim_{k\to\infty} v_k \in \mathcal{W} \cap \mathcal{V}$ and, for every k, $u_k, v_k \in \mathcal{W}_k$. Therefore, there exists $K > 0$ such that, for every $k > K$, $\mathcal{W}_k \cap \mathcal{U}$ and $\mathcal{W}_k \cap \mathcal{V}$ are nonempty. Since each $\mathcal{W}_k$ is connected, it follows that, for every $k > K$, there exists $w_k \in \mathcal{W}_k$ such that $w_k \notin \mathcal{U} \cup \mathcal{V}$. The sequence $\{w_k\}_{k=K}^{\infty}$ is contained in $\mathcal{B}$ and hence bounded. Let w be a subsequential limit of the sequence $\{w_k\}_{k=K}^{\infty}$. Since the sequence $\{w_k\}_{k=K}^{\infty}$ is contained in the closed set $\mathbb{R}^n \backslash (\mathcal{U} \cup \mathcal{V})$, it follows that $w \in \mathbb{R}^n \backslash (\mathcal{U} \cup \mathcal{V})$. On the other hand, $w \in \limsup_{k\to\infty} \mathcal{W}_k = \mathcal{W} \subseteq \mathcal{U} \cup \mathcal{V}$, which leads to a contradiction. Hence we conclude that $\mathcal{W}$ is connected. ∎

Proof of Proposition 4.2. Suppose $0 \notin \operatorname{co} \mathcal{L}_x$. For each $k = 1, 2, \ldots$, let $\mathcal{U}_k = \{w \in \mathbb{R}^n : \operatorname{dist}(w, \mathcal{L}_x) < 1/k\}$. For every k, $\mathcal{U}_k$ is an open set containing $\mathcal{L}_x$, $\overline{\mathcal{U}_k}$ is compact, and $\overline{\mathcal{U}_{k+1}} \subset \mathcal{U}_k$. Moreover, $\cap_k \overline{\mathcal{U}_k} = \mathcal{L}_x$.

For every k, there exists a relatively open neighborhood $\mathcal{V}_k \subseteq \mathcal{G}$ of x such that, for every $z \in \mathcal{V}_k \backslash \mathcal{E}$, $\|f(z)\|^{-1} f(z) \in \mathcal{U}_k$. It is easy to show that, for every k, every connected component of $f(\mathcal{V}_k)\backslash\{0\}$ is contained in the cone generated by a connected component of $\mathcal{U}_k$. Hence it follows that, for every k, $\overline{\operatorname{coco}(f(\mathcal{V}_k)\backslash\{0\})} \subseteq \overline{\operatorname{coco} \mathcal{U}_k}$. Consequently, $\mathcal{F}_x \subseteq \cap_{k=1}^{\infty} \overline{\operatorname{coco}\,(f(\mathcal{V}_k)\backslash\{0\})} \subseteq \cap_{k=1}^{\infty} \overline{\operatorname{coco} \mathcal{U}_k}$.

We claim that $\cap_{k=1}^{\infty} \overline{\operatorname{coco} \mathcal{U}_k} \subseteq \operatorname{co} \mathcal{L}_x$. To prove this, choose $v \in \cap_{k=1}^{\infty} \overline{\operatorname{coco} \mathcal{U}_k}$. There exist a sequence $\{\alpha_k\}$ in $(0, \infty)$ and a sequence $\{v_k\}$ in $\mathbb{R}^n$ such that, for every k, $v_k \in \operatorname{co} \mathcal{U}_k$ and

$$\|v - \alpha_k v_k\| < \frac{1}{k}. \tag{17}$$

For each k, let $\mathcal{W}_k$ be a connected component of $\mathcal{U}_k$ such that v_k is contained in the convex hull of $\mathcal{W}_k$. Each $\mathcal{W}_k$ is a subset of the bounded set $\mathcal{U}_1$. Hence there

exists an increasing sequence $\{k_j\}_{j=1}^{\infty}$ of integers such that the subsequence $\{\mathcal{W}_{k_j}\}_{j=1}^{\infty}$ converges. Let $\mathcal{W} = \lim_{j\to\infty} \mathcal{W}_{k_j}$. Then, by Lemma A.1, $\mathcal{W}$ is connected.

Next, consider $w \in \mathcal{W}$. There exists a sequence $\{w_j\}$ such that $w_j \in \mathcal{W}_{k_j} \subseteq \mathcal{U}_{k_j}$ for every j, and $\lim_{j\to\infty} w_j = w$. Since $\{\mathcal{U}_k\}$ is a decreasing sequence of sets, for every k, the sequence $\{w_j\}$ is eventually contained in $\mathcal{U}_k$. Hence $w \in \overline{\mathcal{U}_k}$ for every k; that is, $w \in \cap_k \overline{\mathcal{U}_k} = \mathcal{L}_x$. Since $w \in \mathcal{W}$ was arbitrary, it follows that $\mathcal{W} \subseteq \mathcal{L}_x$. Hence the connected set $\mathcal{W}$ is contained in a connected component of $\mathcal{L}_x$.

By Caratheodory's theorem [25, Thm. 17.1], for every j, there exist vectors $w_j^i \in \mathcal{W}_{k_j}$, $i = 1, \ldots, n$, and scalars $\lambda_j^i \in [0, 1]$, $i = 1, \ldots, n$, such that $\lambda_j^1 + \cdots + \lambda_j^n = 1$ and

$$v_{k_j} = \lambda_j^1 w_j^1 + \cdots + \lambda_j^n w_j^n. \tag{18}$$

For each $i = 1, \ldots, n$, let $\lambda^i \in [0, 1]$ and w^i be subsequential limits of the bounded sequences $\{\lambda_j^i\}_{j=1}^{\infty}$ and $\{w_j^i\}_{j=1}^{\infty}$, respectively. Then, for every i, $w^i \in \mathcal{W}$, while $\lambda^1 + \cdots + \lambda^n = 1$. Let $w = \lambda^1 w^1 + \cdots + \lambda^n w^n$. Then $w \in \operatorname{co} \mathcal{L}_x$, and hence $w \neq 0$.

By (18), there exists an increasing sequence $\{i_m\}_{m=1}^{\infty}$ of integers such that, for every m,

$$\|v_{i_m} - w\| < \frac{1}{m}. \tag{19}$$

For every m, (17) implies that $\alpha_{i_m}\|v_{i_m}\| \leq \|v\| + i_m^{-1}$, while (19) implies that $\|v_{i_m}\| \geq \|w\| - m^{-1}$. It therefore follows that, for every $m > \|w\|^{-1}$, $\alpha_{i_m} \leq (\|v\| + i_m^{-1})(\|w\| - m^{-1})^{-1}$. There exists $M > 0$ such that, for every $m > M$, $i_m^{-1} < \|v\|$ and $m^{-1} < \|w\|/2$. Then, for every $m > M$, $\alpha_{i_m} \leq 4\|v\|/\|w\|$. It follows that the subsequence $\{\alpha_{i_m}\}_{m=1}^{\infty}$ is bounded. Hence, by choosing subsequences appropriately, we may assume that there exists $\alpha \in [0, \infty)$ such that, for every $m > M$, $|\alpha_{i_m} - \alpha| < m^{-1}$. Then, for every $m > M$, $\|v - \alpha w\| \leq \|v - \alpha_{i_m} v_{i_m}\| + \alpha_{i_m}\|v_{i_m} - w\| + |\alpha_{i_m} - \alpha|\|w\| \leq i_m^{-1} + m^{-1}(\alpha_{i_m} + \|w\|)$. Since the subsequence $\{\alpha_{i_m}\}_{m=1}^{\infty}$ is bounded and the sequence $\{i_m\}$ is divergent, m can be chosen to make the right-hand side of the previous inequality arbitrarily small. It follows that $v = \alpha w$ and thus $v \in \operatorname{coco} \mathcal{L}_x$. The result now follows. ∎

REFERENCES

[1] D. Aeyels and R. Sepulchre, *Stability for dynamical systems with first integrals: A topological criterion*, Systems Control Lett., 19 (1992), pp. 461–465.

[2] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Boston, 1990.

[3] D. S. Bernstein and S. P. Bhat, *Lyapunov stability, semistability, and asymptotic stability of matrix second-order systems*, ASME Trans. J. Vibr. Acoustics, 117 (1995), pp. 145–153.

[4] D. S. Bernstein and S. P. Bhat, *Nonnegativity, reducibility and semistability of mass action kinetics*, in Proceedings of the IEEE Conference on Decision and Control (Phoenix, AZ), IEEE Control Systems Society, Piscataway, NJ, 1999, pp. 2206–2211.

[5] S. P. Bhat and D. S. Bernstein, *Lyapunov analysis of semistability*, in Proceedings of the American Control Conference, San Diego, CA, 1999, pp. 1608–1612.

[6] S. P. Bhat and D. S. Bernstein, *Nontangency-based Lyapunov tests for stability and convergence*, in Proceedings of the American Control Conference, Arlington, VA, 2001, pp. 4840–4845.

[7] N. P. Bhatia and O. Hajek, *Local Semi-Dynamical Systems*, Springer-Verlag, Berlin, 1969.

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

[9] S. L. Campbell and N. J. Rose, *Singular perturbation of autonomous linear systems*, SIAM J. Math. Anal., 10 (1979), pp. 542–551.

[10] J. Dugundji, *Topology*, Wm. C. Brown, Dubuque, Iowa, 1989.
[11] P. Erdi and J. Toth, *Mathematical Models of Chemical Reactions: Theory and Applications of Deterministic and Stochastic Models*, Princeton University Press, Princeton, NJ, 1988.
[12] M. Feinberg, *The existence and uniqueness of steady states for a class of chemical reaction networks*, Arch. Ration. Mech. Anal., 132 (1995), pp. 311–370.
[13] V. Guillemin and A. Pollack, *Differential Topology*, Prentice–Hall, Englewood Cliffs, NJ, 1974.
[14] P. Hartman, *Ordinary Differential Equations*, 2nd ed., Birkhäuser Boston, Cambridge, MA, 1982.
[15] U. Helmcke and J. B. Moore, *Optimization and Dynamical Systems*, Springer-Verlag, London, 1994.
[16] A. Iggidr, B. Kalitine, and R. Outbib, *Semidefinite Lyapunov functions stability and stabilization*, Math. Control Signals Systems, 9 (1996), pp. 95–106.
[17] A. Ilchmann, *Non-Indentifier-Based High Gain Adaptive Control*, Springer-Verlag, London, 1993.
[18] J. A. Jacquez and C. P. Simon, *Qualitative theory of compartmental systems*, SIAM Rev., 35 (1993), pp. 43–79.
[19] H. K. Khalil, *Nonlinear Systems*, 2nd ed., Prentice–Hall, Upper Saddle River, NJ, 1996.
[20] N. N. Krasovskii, *Problems of the Theory of Stability of Motion*, Stanford University Press, Stanford, CA, 1963.
[21] J. P. LaSalle, *Some extensions of Liapunov's second method*, IRE Trans., CT-7 (1960), pp. 520–527.
[22] D. R. Mudgett and A. S. Morse, *Adaptive stabilization of linear systems with unknown high-frequency gain*, IEEE Trans. Automat. Control, 30 (1985), pp. 549–554.
[23] K. S. Narendra and A. M. Annaswamy, *Stable Adaptive Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1989.
[24] R. D. Nussbaum, *Some remarks on a conjecture in parameter adaptive control*, Systems Control Lett., 3 (1983), pp. 243–246.
[25] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[26] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[27] S. K. Scott, *Chemical Chaos*, Oxford University Press, Oxford, UK, 1991.
[28] E. D. Sontag, *Structure and stability of certain chemical networks and applications to the kinetic proofreading model of T-cell receptor signal transduction*, IEEE Trans. Automat. Control, 46 (2001), pp. 1028–1047.
[29] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[30] T. Yoshizawa, *Stability Theory and the Existence of Periodic Solutions and Almost Periodic Solutions*, Springer-Verlag, New York, 1975.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1776–1794

LINEAR CONTROL SYSTEMS ON UNBOUNDED TIME INTERVALS AND INVARIANT MEASURES OF ORNSTEIN–UHLENBECK PROCESSES IN HILBERT SPACES*

MARCO FUHRMAN† AND ANNA MARIA PAGANONI†

Abstract. We consider linear control systems in a Hilbert space over an unbounded time interval of the form

$$y'_\alpha(t) = (A - \alpha I)y_\alpha(t) + Bu(t), \qquad t \in (-\infty, T],$$

with bounded control operator B, under appropriate stability assumptions on the operator A. We study how the space of states reachable at time T depends on the parameter $\alpha \geq 0$. We apply the results to study the dependence on α of the Cameron–Martin spaces of the invariant measures of the Ornstein–Uhlenbeck processes X_α defined by the equation driven by the Wiener process W:

$$dX_\alpha(t) = (A - \alpha I)X_\alpha(t)\, dt + B\, dW(t), \qquad t \geq 0.$$

Key words. infinite dimensional control systems, reachability, Ornstein–Uhlenbeck process

AMS subject classifications. Primary, 93C25, 37L55; Secondary, 93B03, 60H99

DOI. 10.1137/S0363012902414652

1. Introduction.

The aim of this paper is to study some controllability properties of a linear control system in a Hilbert space H over an unbounded time interval and to apply these results to study the behavior of the Cameron–Martin space of the invariant measure for a class of Ornstein–Uhlenbeck stochastic processes in H under perturbation.

Let us consider a linear control system in H of the form

$$y'(t) = Ay(t) + Bu(t), \qquad t \in (-\infty, T], \tag{1.1}$$

for t varying in a fixed unbounded time interval $(-\infty, T]$, where A is the infinitesimal generator of a strongly continuous semigroup of operators, B is a bounded operator from another Hilbert space U to H, and u is a control, which will always be assumed to belong to the space $L^2(-\infty, T; U)$ of square summable functions from $(-\infty, T]$ to U. The value of T is irrelevant in most of what follows, and we could even replace T by zero; nevertheless, we will keep the present, slightly more general notation. For the moment we assume for simplicity that A is exponentially stable, although this assumption will be relaxed in the following sections. Then the solution of (1.1) can be defined in a standard way for every control u. The initial condition at $-\infty$ is assumed to be zero. One defines in an obvious way the space of states reachable at time T over the interval $(-\infty, T]$, which is denoted by $\mathcal{K}_\infty$. By straightforward extensions of the results on a finite time interval, this space can be characterized as the image of the square root of the operator Q_∞ defined by

$$Q_\infty h = \int_0^\infty e^{tA} B B^* e^{tA^*} h\, dt, \qquad h \in H.$$

Thus $\mathcal{K}_\infty = im\, Q_\infty^{1/2}$. Clearly, this space does not depend on T.

*Received by the editors September 18, 2002; accepted for publication (in revised form) May 6, 2003; published electronically November 14, 2003.

http://www.siam.org/journals/sicon/42-5/41465.html

†Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy (marco.fuhrman@polimi.it, annamaria.paganoni@mate.polimi.it).

One of the main results of this paper is a precise description of the behavior of the space of reachable states for the class of perturbed control systems

$$y'_\alpha(t) = (A - \alpha I)y_\alpha(t) + Bu(t), \qquad t \in (-\infty, T], \tag{1.2}$$

when the parameter α ranges over $[0, \infty)$.

We note that the system (1.2) can be reduced to system (1.1) by the transformation $y(t) = e^{\alpha t}y_\alpha(t)$. This transformation is sometimes called the "exponential shift." For control systems on a finite time interval this argument shows that the space of reachable states does not depend on α. We note, however, that the exponential shift maps the control u to the function $t \mapsto e^{\alpha t}u(t)$, which does not necessarily belong to the space $L^2(-\infty, T; U)$ of square summable controls, due to the fact that the time interval $(-\infty, T]$ is unbounded. Keeping the space of controls fixed, the problem arises whether the space of reachable sets may change with α. One can also study a related problem, namely, whether a state which is reachable on $(-\infty, T]$ can also be reached over a finite time interval $[T - r, T]$ for some $r > 0$.

In Theorem 4.1 we give an answer to both questions. Somewhat imprecisely, the situation is as follows: denoting by $\mathcal{K}^\alpha_\infty$ the space of reachable states for the perturbed systems (1.2), it is clear that

$$0 \le \alpha \le \beta \qquad \text{implies} \qquad \mathcal{K}^\alpha_\infty \supset \mathcal{K}^\beta_\infty,$$

possibly with equality. It can happen that $\mathcal{K}^\alpha_\infty$ are strictly decreasing for all $\alpha \ge 0$. Or it can happen that $\mathcal{K}^\alpha_\infty$ are strictly decreasing for α less or equal to some critical value α_0, and then they remain unchanged for $\alpha > \alpha_0$; the case $\alpha_0 = 0$ may occur. Another case can also happen, namely, that $\mathcal{K}^\alpha_\infty$ do not depend on α. Finally, the states in $\mathcal{K}^\alpha_\infty$ can be reached in a finite time interval precisely when a small change of α does not change the reachable set. See Theorem 4.1 for precise statements. We also relate the critical value α_0 to the norms of some appropriately defined operators; see Corollary 4.7.

One may ask whether more general perturbations of the original system (1.1) could affect the space $\mathcal{K}_\infty$ of reachable states over $(-\infty, T]$. The results described above show the (perhaps surprising) fact that even the spaces $\mathcal{K}^\alpha_\infty$ are very "sensitive" to perturbations; an arbitrarily small change in the value of α may suffice to change the space $\mathcal{K}^\alpha_\infty$. Therefore, one expects that more general perturbations of the original system (1.1) will have the same effect, unless very stringent conditions are imposed. That is why we do not address this problem here but rather postpone it to future study.

We devote section 5 to giving examples where the various possibilities described above occur and also to studying some important classes of control systems. For example, we show that the spaces $\mathcal{K}^\alpha_\infty$ do not depend on α if the system is finite-dimensional or exactly null controllable.

In connection with all the properties and problems discussed so far, instead of the set of reachable states one may consider the set of approximately reachable states. In strong contrast with the previous results, the space of approximately reachable states for system (1.2) turns out to be independent of α; see section 4.3.

We believe that the previous results have an intrinsic interest, since they address basic structural properties of linear control systems, but our interest in this question arose from some probabilistic motivation, which we now shortly describe.

Let us consider the following stochastic evolution equation of Itô type:

$$\begin{cases} dX(t) = AX(t)\,dt + B\,dW(t), & t \geq 0, \\ X(0) = X_0, \end{cases}$$

where W is a (cylindrical) Wiener process in a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t))$ and X_0 is a gaussian $\mathcal{F}_0$-measurable random variable with values in H. Assuming the finite trace condition

$$\text{Trace } Q_\infty < \infty,$$

one can show that the equation uniquely defines a gaussian process X and that the centered gaussian measure μ with covariance operator Q_∞ is invariant for the process X. For these facts we refer to [8] and to the discussion in section 6. The process X is called the (nonsymmetric) Ornstein–Uhlenbeck process and has been intensively studied in recent years; see [3, 4, 5, 6, 13, 14].

Associated to the centered gaussian measure μ is the so-called Cameron–Martin space; see, e.g., [1]. It is the subspace of H consisting of all vectors h such that the image measure of μ under the mapping $x \to x + h$, $x \in H$, is absolutely continuous with respect to μ. The corresponding Radon–Nikodym density is then called the logarithmic derivative of μ along h. The Cameron–Martin space plays a basic role in the construction of the Sobolev classes of functions over the measure space (H, μ) (see [1]), in the definition and properties of the associated Dirichlet forms (see [2], [17]), and in the subsequent constructions needed for the analysis of the stochastic process X (see the references on the Ornstein–Uhlenbeck process cited above). It turns out that the Cameron–Martin space of μ coincides with $im\, Q_\infty^{1/2}$ and hence with the space of reachable states $\mathcal{K}_\infty$ for the control system (1.1). So our results can be applied to study how the Cameron–Martin space behaves under perturbation of the Ornstein–Uhlenbeck process. Thus, for $\alpha \geq 0$, we consider the perturbed processes X_α solution of

$$\begin{cases} dX_\alpha(t) = (A - \alpha I)X_\alpha(t)\,dt + B\,dW(t), & t \geq 0, \\ X_\alpha(0) = X_0, \end{cases}$$

and we consider the corresponding gaussian invariant measures μ_α. Since the Cameron–Martin space of μ_α coincides with $\mathcal{K}_\infty^\alpha$, our results give information on the dependence of the Cameron–Martin space of μ_α on α.

The plan of the paper is as follows: Section 2 is devoted to some preliminary technical facts. Section 3 contains the standing assumptions and some preliminary results on control systems on unbounded time intervals. Section 4 contains the main results, Theorem 4.1 and Corollary 4.7, while section 5 is devoted to examples. Finally, in section 6 it is shown how to apply results to the Ornstein–Uhlenbeck process.

2. Some notation and technical tools. Let H and U be separable Hilbert spaces, over the real or complex field, with norm and scalar product denoted by $|\cdot|$, $\langle\cdot\rangle$. We use $\|\cdot\|$ to denote the operator norm. Later we will consider H as the state space of a control system and U as the space of control parameters.

We start by recalling a few facts on Hilbert space valued integrals.

Let I be an interval in the real line and $f : I \to H$ a Borel measurable function. It is well known that f is integrable (in the sense of Bochner) if and only if $\int_I |f(t)|\,dt < \infty$. To allow more generality, in the following we will use the concept of weak integrals,

and so we are going to recall the definition and some of its elementary properties; for more on this subject the reader may consult, e.g., [9]. We say that f is weakly integrable if $\int_I |\langle f(t), h\rangle|\, dt < \infty$ for every $h \in H$ and there exists a (necessarily unique) element of H, denoted $\int_I f(t)\, dt$, such that

$$\left\langle \int_I f(t)\ dt, h \right\rangle = \int_I \langle f(t), h\rangle\ dt, \qquad h \in H.$$

If f is Bochner integrable, then $\int_I f(t)\, dt$ coincides with the Bochner integral. The definition of weak integrability, as stated above, is slightly redundant, as shown by the following well-known lemma that will be useful in what follows; see, for example, [10, 9].

LEMMA 2.1. *If $f : I \to H$ is Borel measurable and $\int_I |\langle f(t), h\rangle|\, dt < \infty$ for every $h \in H$, then f is weakly integrable.*

In the following sections we will systematically use the following lemma, whose proof can be found in [7] or [21].

LEMMA 2.2. *Let H, U, Z be Hilbert spaces, and let $A : U \to H$ and $B : Z \to H$ be bounded linear operators. Then the inclusion between the images*

$$im\ A \subset im\ B$$

holds if and only if there exists $c > 0$ such that

$$|A^* h| \le c|B^* h|, \qquad h \in H.$$

In this case, denoting by B^{-1} the pseudoinverse of B, we have $\|B^{-1}A\| \le c$.

In particular, if we define $Q = BB^*$, then $im\ B = im\ Q^{1/2}$ and, denoting $Q^{-1/2}$ the pseudoinverse of $Q^{1/2}$, we have $\|B^{-1}Q^{1/2}\| = \|Q^{-1/2}B\| = 1$.

3. Assumptions and preliminaries. Throughout the paper the following assumptions are assumed to hold.

HYPOTHESIS 3.1.

(i) *H and U are separable Hilbert spaces.*

(ii) *The operator A is the generator of a strongly continuous semigroup $\{e^{tA},\ t \ge 0\}$ of bounded linear operators in H. B is a bounded linear operator from U to H.*

(iii) *We have*

$$\int_0^\infty |B^* e^{tA^*} x|^2 dt < \infty, \qquad x \in H. \tag{3.1}$$

We remark that the validity of (3.1) does not imply that the operator A is exponentially stable. Condition (3.1) has received a fair amount of attention in some of the recent literature. At least in the case when $U = \mathbb{R}$ (and therefore B^* maps H to $\mathbb{R}$), in several cases condition (3.1) is equivalent to a bound of the form $|B^*(sI - A^*)^{-1}|_{H^*} \le c/\sqrt{\mathrm{Re}\ s}$ for complex s with $\mathrm{Re}\ s > 0$; see, for instance, [15, 20, 16].

Let us consider a control system in H on a finite time interval. We shortly recall some usual definitions and properties, mainly to fix notation. For further details we refer the reader to any treatise on infinite-dimensional control theory, for instance, [7] or [21].

We will consider a control system on a time interval of length $r > 0$, with initial condition $h \in H$, formally:

$$y'(t) = Ay(t) + Bu(t), \quad t \in [T-r, T], \qquad y(T-r) = h. \tag{3.2}$$

Here T is an arbitrary fixed real number; the choice of the time interval $[T-r, T]$ allows us to conform with some notation that will be introduced later, when dealing with the unbounded time interval $(-\infty, T]$.

When referring to (3.2), we take as the space of controls the space $L^2(T-r, T; U)$ of all Borel measurable functions $u : [T-r, T] \to U$ satisfying $\int_{T-r}^{T} |u(t)|^2 dt < \infty$. $L^2(T-r, T; U)$ will be endowed with its usual Hilbert norm. The solution $y : [T-r, T] \to H$ of (3.2), or the trajectory corresponding to a control u, is defined as

$$y(t) = e^{(t-T+r)A}h + \int_{T-r}^{t} e^{(t-s)A}Bu(s)\, ds, \qquad t \in [T-r, T],$$

and the space $\mathcal{K}_r$ of states reachable from zero in time r (in short, the space of reachable states) is defined as the set of all elements $y(T)$, as u spans the space of controls and $h = 0$. The space $\mathcal{K}_r$ therefore coincides with the image of the so-called controllability operator $\mathcal{L}_r : L^2(T-r, T; U) \to H$ defined by

$$\mathcal{L}_r u = \int_{T-r}^{T} e^{(T-s)A}Bu(s)\, ds, \qquad u \in L^2(-\infty, T; U).$$

The gramian operator is defined by

$$Q_r h = \int_0^r e^{tA}BB^* e^{tA^*} h\, dt, \qquad h \in H,$$

and we have $\mathcal{L}_r \mathcal{L}_r^* = Q_r$ and $\mathcal{K}_r = im\, \mathcal{L}_r = im\, Q_r^{1/2}$. Clearly, $\mathcal{K}_s \subset \mathcal{K}_r$ for $0 < s < r$.

In the case of a finite time interval considered so far, the assumption (3.1) is not needed, and all the indicated integrals are Bochner integrals.

Now, for fixed $T \in \mathbb{R}$, we consider the following control system in H on the unbounded time interval $(-\infty, T]$:

$$y'(t) = Ay(t) + Bu(t), \quad t \in (-\infty, T], \qquad y(-\infty) = 0. \tag{3.3}$$

This expression is only formal, but now we proceed to giving rigorous definitions. We define a control as an element of $L^2(-\infty, T; U)$, i.e., a Borel measurable function $u : (-\infty, T] \to U$ satisfying $\int_{-\infty}^{T} |u(t)|^2 dt < \infty$. $L^2(-\infty, T; U)$ will be endowed with its usual Hilbert norm. A trajectory of the control system (3.3), corresponding to the control u, is by definition the function

$$y(t) = \int_{-\infty}^{t} e^{(t-s)A}Bu(s)\, ds, \qquad t \in (-\infty, T]. \tag{3.4}$$

We note that, for every $h \in H$,

$$\begin{aligned}
\int_{-\infty}^{t} |\langle e^{(t-s)A}Bu(s), h\rangle|\, ds &\le \int_{-\infty}^{t} |u(s)|\, |B^* e^{(t-s)A^*} h|\, ds \\
&\le \left(\int_{-\infty}^{t} |u(s)|^2\, ds\right)^{1/2} \left(\int_{-\infty}^{t} |B^* e^{(t-s)A^*} h|^2\, ds\right)^{1/2} \\
&= \left(\int_{-\infty}^{t} |u(s)|^2\, ds\right)^{1/2} \left(\int_{0}^{\infty} |B^* e^{sA^*} h|^2\, ds\right)^{1/2},
\end{aligned}$$

and the right-hand side is finite by the assumption (3.1). It follows that $y(t)$ is well defined as a weak integral by Lemma 2.1.

We define the space of reachable states $\mathcal{K}_\infty$ as the set of all elements $y(T)$, as u spans the space of controls. The space $\mathcal{K}_\infty$ therefore coincides with the image of the operator $\mathcal{L}_\infty : L^2(-\infty, T; U) \to H$ defined as

$$\mathcal{L}_\infty u = \int_{-\infty}^{T} e^{(T-s)A} B u(s)\, ds, \qquad u \in L^2(-\infty, T; U).$$

We may call $\mathcal{L}_\infty$ the extended controllability operator. The adjoint operator $\mathcal{L}_\infty^* : H \to L^2(-\infty, T; U)$ is easily computed: for $h \in H$, $\mathcal{L}_\infty^* h$ is the function

$$\mathcal{L}_\infty^* h(s) = B^* e^{(T-s)A^*} h, \qquad s \in (-\infty, T].$$

In particular, we find

$$\mathcal{L}_\infty \mathcal{L}_\infty^* h = \int_{-\infty}^{T} e^{(T-s)A} B B^* e^{(T-s)A^*} h\, ds = \int_0^\infty e^{tA} B B^* e^{tA^*} h\, ds,$$

which is a well-defined weak integral. So we can define the extended gramian operator

$$Q_\infty h = \int_0^\infty e^{tA} B B^* e^{tA^*} h\, dt, \qquad h \in H,$$

and conclude that

$$\mathcal{L}_\infty \mathcal{L}_\infty^* = Q_\infty,$$

which implies that Q_∞ is a bounded nonnegative self-adjoint operator in H and, by Lemma 2.2, that the space $\mathcal{K}_\infty$ coincides with the image of $Q_\infty^{1/2}$.

The reader may note that these are immediate extensions of the corresponding notions for systems on a finite time interval.

4. Perturbed systems: Main result. In this section Hypothesis 3.1 is still in force. Our main concern will be to investigate the behavior of the space of reachable states when the system (3.3) is perturbed by replacing the operator A with $A - \alpha I$, where $\alpha \geq 0$. Note that (3.1) holds with A replaced by $A - \alpha I$ since $\alpha \geq 0$. We still keep $L^2(-\infty, T; U)$ as the space of controls.

Thus we are considering the family of control systems written formally

$$y_\alpha'(t) = (A - \alpha I) y_\alpha(t) + B u(t), \quad t \in (-\infty, T], \qquad y_\alpha(-\infty) = 0. \tag{4.1}$$

We note that $A - \alpha I$ is the generator of the semigroup $\{e^{-\alpha t} e^{tA},\ t \geq 0\}$. According to the previous definitions, for the perturbed control system (4.1) the space of reachable states $\mathcal{K}_\infty^\alpha$ is defined as the image of the operator $\mathcal{L}_\infty^{(\alpha)} : L^2(-\infty, T; U) \to H$ given by

$$\mathcal{L}_\infty^{(\alpha)} u = \int_{-\infty}^{T} e^{-\alpha(T-s)} e^{(T-s)A} B u(s)\, ds, \qquad u \in L^2(-\infty, T; U).$$

The extended gramian operator for (4.1) is

$$Q_\infty^{(\alpha)} h = \int_0^\infty e^{-2\alpha t} e^{tA} B B^* e^{tA^*} h\, dt, \qquad h \in H, \tag{4.2}$$

and we have $\mathcal{L}_\infty^{(\alpha)}(\mathcal{L}_\infty^{(\alpha)})^* = Q_\infty^{(\alpha)}$ and $\mathcal{K}_\infty^\alpha = im\ \mathcal{L}_\infty^{(\alpha)} = im\ (Q_\infty^{(\alpha)})^{1/2}$ as before. Note that $\mathcal{K}_\infty^0 = \mathcal{K}_\infty$, $\mathcal{L}_\infty^{(0)} = \mathcal{L}_\infty$, $Q_\infty^{(0)} = Q_\infty$. We also note that for $0 \le \alpha < \beta$ the inequality $Q_\infty^{(\beta)} \le Q_\infty^{(\alpha)}$ is obvious, and by Lemma 2.2 this implies $\mathcal{K}_\infty^\beta = im\ (Q_\infty^{(\beta)})^{1/2} \subset im\ (Q_\infty^{(\alpha)})^{1/2} = \mathcal{K}_\infty^\alpha$.

We may also consider the perturbed system on a bounded time interval; formally, for $r > 0$,

$$y'_\alpha(t) = (A - \alpha I)y_\alpha(t) + Bu(t), \quad t \in [T-r, T], \qquad\qquad y_\alpha(T-r) = 0.$$

However, the space of reachable states for this system clearly does not depend on α and is equal to the image of the operators $\mathcal{L}_r$ introduced above, since the exponential shift, described in the introduction, leaves the space $L^2(T-r, T; U)$ unchanged.

So far the following inclusions have been (trivially) verified: for $0 \le \alpha < \beta$ and $0 < s < r$ we have

$$\mathcal{K}_\infty^\alpha \supset \mathcal{K}_\infty^\beta \supset \mathcal{K}_r \supset \mathcal{K}_s.$$

Our main result is the following.

THEOREM 4.1. *Assume Hypothesis* 3.1. *Then one and only one of the following situations occurs.*

(i) *We have*

$$\mathcal{K}_\infty^\alpha = \mathcal{K}_\infty^\beta \qquad \textit{whenever} \qquad 0 \le \alpha < \beta.$$

Moreover, there exists $r > 0$ such that $\mathcal{K}_r = \mathcal{K}_\infty^\alpha$ for every $\alpha \ge 0$.

(ii) *There exists $\alpha_0 \in [0, \infty]$ such that*

$$\mathcal{K}_\infty^{\gamma_1} \supsetneq \mathcal{K}_\infty^{\gamma_2} \supsetneq \mathcal{K}_\infty^{\alpha_0} \supsetneq \mathcal{K}_\infty^{\beta_1} = \mathcal{K}_\infty^{\beta_2}$$

whenever $0 \le \gamma_1 < \gamma_2 < \alpha_0 < \beta_1 < \beta_2$.

Moreover, given $\alpha \ge 0$, there exists $r > 0$ such that $\mathcal{K}_r = \mathcal{K}_\infty^\alpha$ if and only if $\alpha_0 < \infty$ and $\alpha > \alpha_0$.

Remark 4.2. In the extreme cases $\alpha_0 = 0$ and $\alpha_0 = +\infty$, part (ii) of Theorem 4.1 is understood as follows:

(a) If $\alpha_0 = 0$, then $\mathcal{K}_\infty^0 \supsetneq \mathcal{K}_\infty^{\beta_1} = \mathcal{K}_\infty^{\beta_2}$ for $0 < \beta_1 < \beta_2 < \infty$.

(b) If $\alpha_0 = +\infty$, then $\mathcal{K}_\infty^{\gamma_1} \supsetneq \mathcal{K}_\infty^{\gamma_2}$ for $0 \le \gamma_1 < \gamma_2 < \infty$.

We remark that $\mathcal{K}_\infty^\alpha$ is not defined for $\alpha = +\infty$.

We may rephrase the statement of the theorem by saying that there are only two mutually exclusive cases: (1) the reachability spaces $\mathcal{K}_\infty^\alpha$ are all equal for all values of the parameter $\alpha \ge 0$; (2) the reachability spaces $\mathcal{K}_\infty^\alpha$ coincide for α larger than some critical value α_0, but $\mathcal{K}_\infty^\alpha$ are all distinct for $\alpha \in [0, \alpha_0]$ (and strictly larger than $\mathcal{K}_\infty^{\alpha_0}$, $\alpha > \alpha_0$). Moreover, in case 1 the states reachable in an unbounded time interval can also be reached in a finite time interval, independently of the value of α. In case 2 the states in $\mathcal{K}_\infty^\alpha$ can be reached in a finite time interval if and only if $\alpha > \alpha_0$.

In section 5 we will give examples to show that cases (i) and (ii) of Theorem 4.1 may occur and, in case (ii), each of the possibilities $\alpha_0 = 0$, $0 < \alpha_0 < \infty$, $\alpha_0 = \infty$ may occur.

In section 4.1 we collect some preliminary results for the proof of Theorem 4.1; the proof is presented in section 4.2.

4.1. Some auxiliary operators. The proof of Theorem 4.1 is based on some properties of the operators

$$S_0(t) = Q_\infty^{-1/2} e^{tA} Q_\infty^{1/2}, \qquad t \geq 0,$$

that we are going to study in this subsection. These operators have been introduced independently in [11] and [3]. The following result, proved in these papers, shows in particular that $S_0(t)$ are everywhere defined bounded linear operators on H, with norm less or equal to 1. We report the proof for completeness, with some simplifications contained in [12].

PROPOSITION 4.3. *Assume Hypothesis* 3.1. *For every* $t > 0$ *we have*

$$im\,(e^{tA} Q_\infty^{1/2}) \subset im\, Q_\infty^{1/2} \qquad \text{and} \qquad \|Q_\infty^{-1/2} e^{tA} Q_\infty^{1/2}\| \leq 1. \tag{4.3}$$

Moreover, the following conditions are equivalent:

(i) $\|S_0(t)\| < 1$;

(ii) $im\, Q_\infty^{1/2} = im\, Q_t^{1/2}$.

Proof. By Lemma 2.2, in order to prove (4.3), it suffices to show

$$|Q_\infty^{1/2} e^{tA^*} x|^2 \leq |Q_\infty^{1/2} x|^2, \quad x \in H,$$

i.e., $e^{tA} Q_\infty e^{tA^*} \leq Q_\infty$. This follows from the definition of Q_∞, since

$$e^{tA} Q_\infty e^{tA^*} = \int_0^\infty e^{(t+s)A} BB^* e^{(t+s)A^*}\, ds = \int_t^\infty e^{rA} BB^* e^{rA^*}\, dr = Q_\infty - Q_t \leq Q_\infty. \tag{4.4}$$

Equation (4.3) is now proved.

First note that $Q_t \leq Q_\infty$, so by Lemma 2.2 the inclusion $im\, Q_t^{1/2} \subset im\, Q_\infty^{1/2}$ always holds. By Lemma 2.2 again, (i) holds if and only if there exists $\alpha \in (0,1)$ such that

$$|Q_\infty^{1/2} e^{tA^*} x|^2 \leq \alpha |Q_\infty^{1/2} x|^2, \qquad x \in H.$$

Now note that (4.4) implies

$$|Q_\infty^{1/2} e^{tA^*} x|^2 = \langle (Q_\infty - Q_t)x, x\rangle,$$

so (i) holds if and only if

$$\langle (Q_\infty - Q_t)x, x\rangle \leq \alpha \langle Q_\infty x, x\rangle, \qquad x \in H,$$

or

$$\langle Q_\infty x, x\rangle \leq (1-\alpha)^{-1} \langle Q_t x, x\rangle, \qquad x \in H.$$

By Lemma 2.2 this is equivalent to $im\, Q_\infty^{1/2} \subset im\, Q_t^{1/2}$. □

The following proposition is the main step toward the proof of Theorem 4.1.

PROPOSITION 4.4. *Assume Hypothesis* 3.1.

(i) *If there exists* $r > 0$ *such that* $\|S_0(r)\| < 1$, *then for every* $\alpha > 0$ *there exists* $c > 0$ *such that* $cQ_\infty \leq Q_\infty^{(\alpha)}$.

(ii) *If there exist* $c > 0$ *and* $\alpha > 0$ *such that* $cQ_\infty \leq Q_\infty^{(\alpha)}$, *then there exists* $r > 0$ *such that* $\|S_0(r)\| < 1$.

Proof. We first note the identity, for $h \in H$ and $r > 0$,

$$\int_0^r |B^* e^{tA^*} h|^2 dt = |Q_\infty^{1/2} h|^2 - |Q_\infty^{1/2} e^{rA^*} h|^2, \tag{4.5}$$

which follows from the computation

$$\begin{aligned} \int_0^r |B^* e^{tA^*} h|^2 dt &= \int_0^\infty |B^* e^{tA^*} h|^2 dt - \int_r^\infty |B^* e^{tA^*} h|^2 dt \\ &= |Q_\infty^{1/2} h|^2 - \int_0^\infty |B^* e^{tA^*} e^{rA^*} h|^2 dt \\ &= |Q_\infty^{1/2} h|^2 - |Q_\infty^{1/2} e^{rA^*} h|^2. \end{aligned}$$

Let us prove (i). We have

$$\langle Q_\infty^{(\alpha)} h, h \rangle = \int_0^\infty e^{-2\alpha t} |B^* e^{tA^*} h|^2 dt \geq e^{-2\alpha r} \int_0^r |B^* e^{tA^*} h|^2 dt,$$

and using (4.5) we obtain

$$|(Q_\infty^{(\alpha)})^{1/2} h|^2 \geq e^{-2\alpha r} (|Q_\infty^{1/2} h|^2 - |Q_\infty^{1/2} e^{rA^*} h|^2).$$

Since $Q_\infty^{1/2} e^{rA^*} h = S_0(r)^* Q_\infty^{1/2} h$, we arrive at

$$|(Q_\infty^{(\alpha)})^{1/2} h|^2 \geq e^{-2\alpha r} (1 - \|S_0(r)\|^2) |Q_\infty^{1/2} h|^2,$$

and the conclusion follows with $c = e^{-2\alpha r}(1 - \|S_0(r)\|^2)$.

Now we prove (ii). We can assume $c < 1$. We have

$$\begin{aligned} |Q_\infty^{1/2} e^{rA^*} h|^2 &= \int_0^\infty |B^* e^{tA^*} e^{rA^*} h|^2 dt = \int_r^\infty |B^* e^{tA^*} h|^2 dt \geq e^{2r\alpha} \int_r^\infty e^{-2t\alpha} |B^* e^{tA^*} h|^2 dt \\ &= e^{2r\alpha} \int_0^\infty e^{-2t\alpha} |B^* e^{tA^*} h|^2 dt - e^{2r\alpha} \int_0^r e^{-2t\alpha} |B^* e^{tA^*} h|^2 dt \\ &= e^{2r\alpha} |(Q_\infty^{(\alpha)})^{1/2} h|^2 - e^{2r\alpha} \int_0^r e^{-2t\alpha} |B^* e^{tA^*} h|^2 dt \\ &\geq e^{2r\alpha} |(Q_\infty^{(\alpha)})^{1/2} h|^2 - e^{2r\alpha} \int_0^r |B^* e^{tA^*} h|^2 dt. \end{aligned}$$

By assumption we have $|(Q_\infty^{(\alpha)})^{1/2} h|^2 \geq c |Q_\infty^{1/2} h|^2$, and using (4.5) we obtain

$$|Q_\infty^{1/2} e^{rA^*} h|^2 \geq c\, e^{2r\alpha} |Q_\infty^{1/2} h|^2 - e^{2r\alpha} (|Q_\infty^{1/2} h|^2 - |Q_\infty^{1/2} e^{rA^*} h|^2)$$

or

$$|Q_\infty^{1/2} e^{rA^*} h|^2 \leq (1 - c)\, e^{2r\alpha} (e^{2r\alpha} - 1)^{-1} |Q_\infty^{1/2} h|^2.$$

Choosing $r > 0$ so large that $\gamma := (1 - c)\, e^{2r\alpha} (e^{2r\alpha} - 1)^{-1} < 1$, and noting again that $Q_\infty^{1/2} e^{rA^*} h = S_0(r)^* Q_\infty^{1/2} h$, we obtain

$$|S_0(r)^* Q_\infty^{1/2} h|^2 \leq \gamma |Q_\infty^{1/2} h|^2, \qquad h \in H.$$

This proves that $|S_0(r)^* k|^2 \leq \gamma |k|^2$ for $k \in im\, Q_\infty^{1/2}$. Since, however, by its definition, $S_0(r) k = 0$ for k in the kernel of $Q_\infty^{1/2}$, which is the orthogonal subspace to $im\, Q_\infty^{1/2}$,

it follows that $|S_0(r)^*k|^2 \le \gamma|k|^2$ for every $k \in H$, which proves that $\|S_0(r)\|^2 \le \gamma < 1$. □

Replacing A with $A - \alpha I$ in the definition of $S_0(t)$, we get a different family of operators that we will denote $S_0^{(\alpha)}(t)$, namely,

$$S_0^{(\alpha)}(t) = e^{-\alpha t}(Q_\infty^{(\alpha)})^{-1/2} e^{tA}(Q_\infty^{(\alpha)})^{1/2}, \qquad t \ge 0,\ \alpha \ge 0. \tag{4.6}$$

Note that $S_0^{(0)}(t) = S_0(t)$. As an immediate consequence of the properties enjoyed by $S_0(t)$, we obtain the following result.

COROLLARY 4.5. *Assume Hypothesis* 3.1 *and let* $\alpha \ge 0$.

(i) *If there exists* $r > 0$ *such that* $\|S_0^{(\alpha)}(r)\| < 1$, *then for every* $\beta > \alpha$ *there exists* $c > 0$ *such that* $cQ_\infty^{(\alpha)} \le Q_\infty^{(\beta)}$.

(ii) *If there exist* $c > 0$ *and* $\beta > \alpha$ *such that* $cQ_\infty^{(\alpha)} \le Q_\infty^{(\beta)}$, *then there exists* $r > 0$ *such that* $\|S_0^{(\alpha)}(r)\| < 1$.

(iii) *If there exist* $\alpha \ge 0$ *and* $r > 0$ *such that* $\|S_0^{(\alpha)}(r)\| < 1$, *then for every* $\beta > \alpha$ *there exists* $s > 0$ *such that* $\|S_0^{(\beta)}(s)\| < 1$.

Proof. To prove points (i) and (ii) it suffices to apply Proposition 4.4, replacing A and α with $A - \alpha I$ and $\beta - \alpha$, respectively.

To prove point (iii) take $\beta_1 > \beta$. Then by point (i) (with β_1 instead of β) there exists $c > 0$ such that $cQ_\infty^{(\alpha)} \le Q_\infty^{(\beta_1)}$. By point (ii) (with β instead of α) there exists $s > 0$ such that $\|S_0^{(\beta)}(s)\| < 1$. □

LEMMA 4.6. *Assume Hypothesis* 3.1 *and suppose that* $\|S_0^{(\alpha)}(r)\| < 1$ *for some* $\alpha > 0$ *and* $r > 0$. *Then for all* $\gamma \in (0, \alpha)$ *sufficiently close to* α *we have* $\|S_0^{(\gamma)}(r)\| < 1$.

Proof. We claim that if $\gamma \in (0, \alpha)$ is sufficiently close to α, then there exists $c > 0$ such that

$$Q_\infty^{(\gamma)} \le c\, Q_\infty^{(\alpha)}. \tag{4.7}$$

Assume the claim for a moment. Since the opposite inequality $Q_\infty^{(\gamma)} \ge Q_\infty^{(\alpha)}$ is obvious, we conclude by Lemma 2.2 that $im\ (Q_\infty^{(\alpha)})^{1/2} = im\ (Q_\infty^{(\gamma)})^{1/2}$. Since we assume $\|S_0^{(\alpha)}(r)\| < 1$, we can apply Proposition 4.3 (replacing A with $A - \alpha I$) and conclude that $im\ (Q_\infty^{(\alpha)})^{1/2} = im\ (Q_r^{(\alpha)})^{1/2}$. Since the latter space is clearly identical to $im\ (Q_r^{(\gamma)})^{1/2}$, we also have $im\ (Q_r^{(\gamma)})^{1/2} = im\ (Q_\infty^{(\gamma)})^{1/2}$, and applying Proposition 4.3 again (replacing A with $A - \gamma I$) we conclude that $\|S_0^{(\gamma)}(r)\| < 1$, and the lemma is proved.

It remains to prove (4.7). For $0 \le \gamma < \alpha$ and $h \in H$,

$$\langle e^{tA} Q_\infty^{(\alpha)} e^{tA^*} h, h\rangle = \int_0^\infty |B^* e^{(t+s)A^*} h|^2 e^{-2\alpha s} ds = e^{2\alpha t} \int_t^\infty |B^* e^{sA^*} h|^2 e^{-2\alpha s} ds.$$

Integrating by parts, we have, for $T > 0$,

$$\begin{aligned}
&\int_0^T e^{-2\gamma t} \langle e^{tA} Q_\infty^{(\alpha)} e^{tA^*} h, h\rangle\, dt \\
&\quad = \int_0^T e^{2(\alpha-\gamma)t} \int_t^\infty |B^* e^{sA^*} h|^2 e^{-2\alpha s}\, ds\, dt \\
&\quad = \frac{e^{2(\alpha-\gamma)T}}{2(\alpha-\gamma)} \int_T^\infty |B^* e^{sA^*} h|^2 e^{-2\alpha s} ds - \frac{1}{2(\alpha-\gamma)} \langle Q_\infty^{(\alpha)} h, h\rangle \\
&\qquad\qquad + \frac{1}{2(\alpha-\gamma)} \int_0^T e^{-2\gamma t} |B^* e^{tA^*} h|^2 dt.
\end{aligned}$$

Since

$$e^{2(\alpha-\gamma)T}\int_T^\infty |B^*e^{sA^*}h|^2e^{-2\alpha s}ds \le \int_T^\infty |B^*e^{sA^*}h|^2e^{-2\gamma s}ds \to 0$$

as $T\to\infty$, we arrive at the identity

$$2(\alpha-\gamma)\int_0^\infty e^{-2\gamma t}\langle e^{tA}Q_\infty^{(\alpha)}e^{tA^*}h,h\rangle\,dt = \langle Q_\infty^{(\gamma)}h,h\rangle - \langle Q_\infty^{(\alpha)}h,h\rangle.$$

We note that $\langle e^{tA}Q_\infty^{(\alpha)}e^{tA^*}h,h\rangle = e^{2\alpha t}\langle (Q_\infty^{(\alpha)})^{1/2}S_0^{(\alpha)}(t)S_0^{(\alpha)}(t)^*(Q_\infty^{(\alpha)})^{1/2}h,h\rangle$, so if we define

$$Rh = 2(\alpha-\gamma)\int_0^\infty e^{2(\alpha-\gamma)t}S_0^{(\alpha)}(t)S_0^{(\alpha)}(t)^*h\,dt, \qquad h\in H,$$

and we assume for a moment that R is a well-defined bounded linear operator, we conclude that

$$\langle (Q_\infty^{(\alpha)})^{1/2}R(Q_\infty^{(\alpha)})^{1/2}h,h\rangle = \langle Q_\infty^{(\gamma)}h,h\rangle - \langle Q_\infty^{(\alpha)}h,h\rangle,$$

and consequently $Q_\infty^{(\gamma)} \le (1+\|R\|)Q_\infty^{(\alpha)}$, which proves (4.7). To show that R is well defined, we first note that the family $\{S_0^{(\alpha)}(t),\ t\ge 0\}$ is a contraction semigroup, and since we assume that $\|S_0^{(\alpha)}(r)\|<1$ for some $r>0$, it follows easily that there exist $M,\omega>0$ such that $\|S_0^{(\alpha)}(t)\| \le M\,e^{-\omega t}$ for every $t>0$. If $\alpha-\gamma<\omega$, then the integral defining R is convergent (as a Bochner integral). This finishes the proof. □

4.2. Proof of Theorem 4.1. During the proof we repeatedly use the following properties proved above: for $0\le\alpha<\beta$ and $r>0$

$$\mathcal{K}_\infty^\alpha = im\ (Q_\infty^{(\alpha)})^{1/2}, \qquad \mathcal{K}_r = im\ Q_r^{1/2}, \qquad \mathcal{K}_\infty^\alpha \supset \mathcal{K}_\infty^\beta \supset \mathcal{K}_r. \tag{4.8}$$

We consider two mutually exclusive cases (i) and (ii).

Case (i). Suppose that we have $\|S_0(r)\|<1$ for some $r>0$.

By Proposition 4.4 (i) and Lemma 2.2 we have $im\ Q_\infty^{1/2} \subset im\ (Q_\infty^{(\alpha)})^{1/2}$ for all $\alpha\ge 0$ or, equivalently, $\mathcal{K}_\infty \subset \mathcal{K}_\infty^\alpha$ so that in fact $\mathcal{K}_\infty = \mathcal{K}_\infty^\alpha$ for all $\alpha\ge 0$. Applying Proposition 4.3, we also have $\mathcal{K}_\infty = im\ Q_\infty^{1/2} = im\ Q_r^{1/2} = \mathcal{K}_r$.

Case (ii). Suppose $\|S_0(r)\|=1$ for every $r>0$.

We define

$$J = \{\alpha\ge 0\ :\ \|S_0^{(\alpha)}(r)\|=1 \text{ for every } r>0\}, \qquad \alpha_0 = \sup J.$$

Note that the set J contains at least $\alpha=0$, so α_0 is well defined and $0\le\alpha_0\le\infty$. By Corollary 4.5 (iii) only the following cases can occur:

(a) $\alpha_0=0$, $J=\{0\}$;
(b) $0<\alpha_0<\infty$, $J=[0,\alpha_0]$;
(c) $\alpha_0=\infty$, $J=[0,\infty)$.

Note that the case $0<\alpha_0<\infty$, $J=[0,\alpha_0)$ is impossible by Lemma 4.6.

Suppose $0\le\gamma_1<\gamma_2<\alpha_0\le\infty$. Assume by contradiction that $\mathcal{K}_\infty^{\gamma_1} = \mathcal{K}_\infty^{\gamma_2}$ (respectively, that $\alpha_0<\infty$ and $\mathcal{K}_\infty^{\gamma_2} = \mathcal{K}_\infty^{\alpha_0}$). Then by (4.8) and Lemma 2.2 we have $cQ_\infty^{(\gamma_1)} \le Q_\infty^{(\gamma_2)}$ (respectively, $cQ_\infty^{(\gamma_2)} \le Q_\infty^{(\alpha_0)}$) for some $c>0$, and Corollary 4.5 (ii)

implies that there exists $r > 0$ such that $\|S_0^{(\gamma_1)}(r)\| < 1$ (respectively, $\|S_0^{(\gamma_2)}(r)\| < 1$), which contradicts the definition of α_0. Next note that the equality $\mathcal{K}_\infty^{\gamma_2} = \mathcal{K}_r$ (respectively, $\alpha_0 < \infty$ and $\mathcal{K}_\infty^{\alpha_0} = \mathcal{K}_r$) cannot hold for any $r > 0$ since by Proposition 4.3, applied to $A - \gamma_2 I$ instead of A (respectively, applied to $A - \alpha_0 I$ instead of A), it would imply that $\|S_0^{(\gamma_2)}(r)\| < 1$ (respectively, $\|S_0^{(\alpha_0)}(r)\| < 1$), which is impossible.

Now suppose that $0 \le \alpha_0 < \beta_1 < \beta_2 < \infty$. Assume by contradiction that $\mathcal{K}_\infty^{\alpha_0} = \mathcal{K}_\infty^{\beta_1}$. Then by (4.8) and Lemma 2.2 we have $cQ_\infty^{(\alpha_0)} \le Q_\infty^{(\beta_1)}$, and Corollary 4.5 (ii) implies that there exists $r > 0$ such that $\|S_0^{(\alpha_0)}(r)\| < 1$, which is impossible. Next note that by the definition of α_0 there exists $r > 0$ such that $\|S_0^{(\beta_1)}(r)\| < 1$. By Corollary 4.5 (i) and Lemma 2.2 we have $im\ (Q_\infty^{(\beta_1)})^{1/2} \subset im\ (Q_\infty^{(\beta_2)})^{1/2}$ or, equivalently, $\mathcal{K}_\infty^{\beta_1} \subset \mathcal{K}_\infty^{\beta_2}$ so that in fact $\mathcal{K}_\infty^{\beta_1} = \mathcal{K}_\infty^{\beta_2}$. Finally, applying Proposition 4.3 (with $A - \beta_1 I$ instead of A), we also have $\mathcal{K}_\infty^{\beta_1} = im\ (Q_\infty^{(\beta_1)})^{1/2} = im\ Q_r^{1/2} = \mathcal{K}_r$.

Theorem 4.1 is now completely proved. □

This proof also describes a criterion to decide which case in Theorem 4.1 occurs, as well as the value of α_0.

COROLLARY 4.7. *Assume Hypothesis* 3.1. *Let* $Q_\infty^{(\alpha)}$ *and* $S_0^{(\alpha)}$ *be defined by* (4.2) *and* (4.6), *respectively. Case* (ii) *in Theorem* 4.1 *occurs if and only if the set*

$$J = \{\alpha \ge 0 \ :\ \|S_0^{(\alpha)}(r)\| = 1 \text{ for every } r > 0\}$$

is nonempty, and then $\alpha_0 = \sup J \in [0, \infty]$.

4.3. Approximately reachable states. If the concept of reachability used so far is replaced by approximate reachability, then the behavior of the perturbed system (4.1) is completely different. This is a well-known fact, but we prefer to report a direct proof in this short section since it is often used in a rather implicit way and it is not easy to find a reference (see, however, Chapter 9 of [19]).

First, let us consider the control system (3.2) and define $\mathcal{H}_r$, the space of approximately reachable states from zero in time r, as the set of all elements $k \in H$ such that for all $\epsilon > 0$ there exists a control $u \in L^2(T - r, T; U)$ such that $|y(T) - k| < \epsilon$, where y denotes the corresponding trajectory of system (3.2) with $h = 0$.

In an analogous way we define the space $\mathcal{H}_\infty$ as the space of approximately reachable states over the unbounded time interval $(-\infty, T]$ for system (3.3), corresponding to controls $u \in L^2(-\infty, T; U)$.

For the perturbed system

$$y'_\alpha(t) = (A - \alpha I)y_\alpha(t) + Bu(t), \tag{4.9}$$

with $\alpha \ge 0$, the spaces of approximately reachable states will be denoted by $\mathcal{H}_r^\alpha$ and $\mathcal{H}_\infty^\alpha$, respectively, for the case of a bounded time interval and an unbounded one. Clearly, $\mathcal{H}_r^\alpha = \mathcal{H}_r$.

We claim that $\mathcal{H}_\infty^\alpha = \cup_r \mathcal{H}_r$ for all α, so, in particular, $\mathcal{H}_\infty^\alpha$ is independent of α.

To prove the claim, let us take an element $k \in \mathcal{H}_\infty^\alpha$; we want to prove that there exists a suitable r such that $k \in \mathcal{H}_r^\alpha = \mathcal{H}_r$. In fact, by definition of the trajectory of a control system (compare with formula (3.4)) and by assumption (3.1), we have that for all $\epsilon > 0$ there exists $r > 0$ so large that $|y_\alpha(T) - \tilde{y}_\alpha(T)| < \epsilon$, where $\tilde{y}_\alpha$ denotes the trajectory of the system (4.9), starting from zero at time $T - r$, driven by the control $\tilde{u} \in L^2(T - r, T; U)$ which coincides with u on $[T - r, T]$. By triangular inequality we conclude that $k \in \mathcal{H}_r^\alpha$, and the claim is proved.

5. Examples and remarks. The aim of this section is to show that all the situations in the statement of Theorem 4.1 may occur, namely, case (i) or case (ii) with $\alpha_0 = 0$ or $0 < \alpha_0 < \infty$ or $\alpha_0 = \infty$. The value of α_0 is difficult to compute in general, since it is defined in terms of the semigroup $S_0^{(\alpha)}$ (compare with Corollary 4.7) and not in terms of A and B. Nevertheless, in some interesting cases we do have explicit solutions to the problem. We also discuss the relevance of Theorem 4.1 to special classes of systems—for instance, the finite-dimensional systems or the null controllable ones.

5.1. Case (i) of Theorem 4.1.

5.1.1. The finite-dimensional case.

COROLLARY 5.1. *Assume that Hypothesis* 3.1 *holds and that* $\dim H < \infty$. *Then case* (i) *of Theorem* 4.1 *occurs and* $\mathcal{K}_\infty^\alpha = \mathcal{K}_t$ *for every* $t > 0$ *and* $\alpha \geq 0$.

Thus every state reachable in an unbounded time interval (no matter what the value of α is) can be reached in an arbitrarily small time $t > 0$.

Proof. We claim that

$$im\, Q_\infty^{1/2} = im\, Q_t^{1/2}, \qquad t > 0. \tag{5.1}$$

Since $im\, Q_\infty^{1/2} = \mathcal{K}_\infty$, case (i) of Theorem 4.1 occurs, and the reachability spaces $\mathcal{K}_\infty^\alpha$ coincide for all $\alpha \geq 0$ and in fact they coincide with $\mathcal{K}_t$ for every $t > 0$.

To prove the claim we note that if $\dim H < \infty$, then (5.1) is equivalent to $\ker Q_\infty^{1/2} = \ker Q_t^{1/2}$ (ker denotes of course the kernel of an operator). Clearly, $\ker Q_\infty^{1/2} \subset \ker Q_t^{1/2}$. Conversely, if $Q_t^{1/2} h = 0$ for some $h \in H$, then $Q_t h = 0$ and $Q^{1/2} e^{sA} h = 0$ for every $s \in [0, t]$. By analyticity, $Q^{1/2} e^{sA} h = 0$ for every $s \geq 0$. This implies $Q_\infty h = 0$ and consequently $Q_\infty^{1/2} h = 0$. □

5.1.2. Exactly null controllable systems.

COROLLARY 5.2. *Assume that Hypothesis* 3.1 *holds and that there exists* $r > 0$ *such that*

$$im\, e^{rA} \subset im\, Q_r^{1/2}. \tag{5.2}$$

Then case (i) *of Theorem* 4.1 *occurs and in fact* $\mathcal{K}_\infty^\alpha = \mathcal{K}_r$ *for every* $\alpha \geq 0$.

It is well known that (5.2) is equivalent to the exact null controllability property in time r, defined as follows: for every $h \in H$ there exists a control $u \in L^2([T - r, T]; U)$ such that the trajectory of the control system

$$y'(t) = Ay(t) + Bu(t), \quad t \in [T - r, T], \qquad\qquad y(T - r) = h,$$

satisfies $y(T) = 0$ (see, e.g., [7], [21]). The pair (A, B) is then called a null controllable pair (in time r). It is also well known that this property holds if $B = I$ (for every $r > 0$).

Proof. We claim that (5.2) implies

$$im\, Q_\infty^{1/2} = im\, Q_r^{1/2}. \tag{5.3}$$

Then it follows from Proposition 4.3 that $\|S_0(r)\| < 1$ and by Corollary 4.7 we conclude that case (i) in Theorem 4.1 occurs.

The claim is proved in [8, Theorem 11.13], but we nevertheless include the following simpler proof. First note that the inequality $Q_r \leq Q_\infty$ implies that $im\, Q_r^{1/2} \subset$

im $Q_\infty^{1/2}$ by Lemma 2.2. Next note that the inclusion (5.2) implies that the operator $\Gamma_r := Q_r^{-1/2}e^{rA}$ is everywhere defined, and since it is closed, it is also bounded, by the closed graph theorem. From the definition of Q_∞ it is easy to obtain the identity $e^{rA}Q_\infty e^{rA^*} + Q_r = Q_\infty$, and it follows that $Q_r^{1/2}(\Gamma_r Q_\infty \Gamma_r^* + I)Q_r^{1/2} = Q_\infty$, which implies $|Q_\infty^{1/2}x|^2 \le (1 + \|\Gamma_r Q_\infty \Gamma_r^*\|)|Q_r^{1/2}x|^2$, $x \in H$, and so, by Lemma 2.2, *im* $Q_\infty^{1/2} \subset$ *im* $Q_r^{1/2}$, and (5.3) is proved. □

5.1.3. A special case.

COROLLARY 5.3. *Suppose that Hypothesis* 3.1 *holds and that, setting* $Q = BB^*$, *we have*

$$\text{(5.4)} \qquad \textit{im}\ e^{tA}Q^{1/2} \subset \textit{im}\ Q^{1/2} \qquad \textit{and} \qquad \|Q^{-1/2}e^{tA}Q^{1/2}\| \le Me^{-\beta t}, \quad t \ge 0,$$

for some constants $\beta > 0$, $M > 0$. *Then case* (i) *of Theorem* 4.1 *occurs and* $\mathcal{K}_\infty^\alpha = \mathcal{K}_r$ *for every* $\alpha \ge 0$ *and* $r > 0$.

We remark that *im* $B =$ *im* $Q^{1/2}$ by Lemma 2.2. Moreover, the inclusion in (5.4) implies that the linear operators $Q^{-1/2}e^{tA}Q^{1/2}$ are everywhere defined and therefore, being obviously closed, they are also continuous.

We note that all the assumptions of the corollary hold true if B is a bounded linear operator from U to H and A is the infinitesimal generator of an exponentially stable, strongly continuous semigroup of operators that commute with $Q = BB^*$. Thus there are many control systems satisfying these assumptions.

Proof. We define $\widehat{S}(t) = Q^{-1/2}e^{tA}Q^{1/2}$ for $t > 0$, $\widehat{S}(0) = I$. Clearly, $\widehat{S}$ is a semigroup of bounded linear operators on H, satisfying $\|\widehat{S}(t)\| \le Me^{-\beta t}$, $t \ge 0$. We note that for $x \in H$, $y \in$ *im* $Q^{1/2}$,

$$\text{(5.5)} \qquad \langle \widehat{S}(t)x, y\rangle = \langle e^{tA}Q^{1/2}x, Q^{-1/2}y\rangle \to \langle x, y\rangle, \qquad \text{as } t \to 0.$$

Since $\widehat{S}$ is bounded in the operator norm, it follows that (5.5) holds for every $x \in H$ and $y \in \overline{\textit{im}\ Q^{1/2}}$ (the closure of *im* $Q^{1/2}$ in H). Since $\widehat{S}(t)x = Q^{-1/2}e^{tA}Q^{1/2}x$ is orthogonal to $\ker Q^{1/2} = \overline{\textit{im}\ Q^{1/2}}$, (5.5) holds for all $x, y \in H$. Thus, for every $x \in H$, $\widehat{S}(t)x \to x$ weakly in H; it is well known that this implies that $\widehat{S}$ is a strongly continuous semigroup.

Let $\widehat{A}$ denote the infinitesimal generator of $\widehat{S}$. We consider the pair $(\widehat{A}, I)$, and we define the corresponding controllability operators

$$\widehat{Q}_\infty x = \int_0^\infty \widehat{S}(t)\widehat{S}(t)^* x\, dt, \quad \widehat{Q}_r x = \int_0^r \widehat{S}(t)\widehat{S}(t)^* x\, dt, \qquad x \in H,\ r > 0.$$

Since the pair $(\widehat{A}, I)$ is null controllable in time r for every $r > 0$, by the results of the previous paragraph we conclude that there exists a constant $C_r > 0$ such that $|\widehat{Q}_\infty^{1/2}x|^2 \le C_r|\widehat{Q}_r^{1/2}x|^2$. Since $Q^{1/2}\widehat{S}(t) = e^{tA}Q^{1/2}$, it follows that

$$Q^{1/2}\widehat{Q}_\infty Q^{1/2}x = \int_0^\infty Q^{1/2}\widehat{S}(t)\widehat{S}(t)^*Q^{1/2}x\, dt = \int_0^\infty e^{tA}Qe^{tA^*}x\, dt = Q_\infty x, \qquad x \in H,$$

and, similarly, $Q^{1/2}\widehat{Q}_r Q^{1/2} = Q_r$. Therefore,

$$|Q_\infty^{1/2}x|^2 = |\widehat{Q}_\infty^{1/2}Q^{1/2}x|^2 \le C_r|\widehat{Q}_r^{1/2}Q^{1/2}x|^2 = C_r|Q_r^{1/2}x|^2, \qquad x \in H,$$

which implies *im* $Q_\infty^{1/2} \subset$ *im* $Q_r^{1/2}$. Since the opposite inclusion is obvious, we have *im* $Q_\infty^{1/2} =$ *im* $Q_r^{1/2}$ for $r > 0$ and the result follows. □

5.1.4. Exactly reachable systems.

COROLLARY 5.4. *Suppose that Hypothesis* 3.1 *holds and that A is exponentially stable. If $\mathcal{K}_\infty = H$, then case* (i) *of Theorem* 4.1 *occurs and consequently $\mathcal{K}_r = H$ for $r > 0$ sufficiently large.*

Systems satifying the condition $\mathcal{K}_\infty = H$ may be called *exactly reachable on* $(-\infty, T]$. Thus, if A is exponentially stable, exact reachability on $(-\infty, T]$ implies exact reachability on a bounded interval $[T - r, T]$.

Proof. By a standard duality argument one can check that the equality $\mathcal{K}_\infty = H$ is equivalent to the following condition: there exists $\kappa > 0$ such that

$$\int_0^\infty |B^* e^{tA^*} x|^2 \, dt \geq \kappa |x|^2, \qquad x \in H. \tag{5.6}$$

Condition (5.6) is called *exact observability* for the pair (A^*, B^*). Since A is exponentially stable, it follows from Proposition 2.8 in [18] that (5.6) holds if and only if there exist $r > 0$ and $\kappa_r > 0$ such that

$$\int_0^r |B^* e^{tA^*} x|^2 \, dt \geq \kappa_r |x|^2, \qquad x \in H,$$

and this is equivalent to the equality $\mathcal{K}_r = H$, again by duality. □

5.2. Case (ii) of Theorem 4.1 with $\alpha_0 = 0$. The example in this section was invented by Gołdys for a different purpose [14]. Let $H = U$ be a Hilbert space and $\{h_k,\ k \geq 1\}$ an orthonormal basis of H. Define the operators A and B setting

$$Ah_k = -\frac{1}{k}\, h_k, \qquad Bh_k = \frac{1}{k^{3/2}}\, h_k.$$

Note that A and B are commuting bounded self-adjoint operators, and A is nonpositive, but, in contrast to section 5.1.3, A does not have bounded inverse. Hypothesis 3.1 is easy to verify. We have $S_0(t) = e^{tA}$ and $e^{tA} h_k = e^{-t/k} h_k$, $k \geq 1$, so that $\|S_0(t)\| = 1$ for every $t > 0$, whereas $\|S_0^{(\alpha)}(t)\| = e^{-\alpha t} \|S_0(t)\| < 1$ for every $t > 0$. Corollary 4.7 shows that in this example we have $\alpha_0 = 0$.

Remark 5.5. In this example the semigroup (e^{tA}) is not exponentially stable. In Remark 5.6 below we will give another example, where $\alpha_0 = 0$ and (e^{tA}) is exponentially stable.

5.3. Case (ii) of Theorem 4.1 with $0 < \alpha_0 < \infty$. We take $H = U = L^2(\mathbb{R})$. Let the operators e^{tA} be the shift operators

$$e^{tA} f(x) = f(x - t), \qquad t \geq 0,\ x \in \mathbb{R},$$

and let the operator B be multiplication by the function $e^{-|x|}$: $Bf(x) = e^{-|x|} f(x)$, $x \in \mathbb{R}$. Then one finds with simple calculations

$$e^{tA^*} f(x) = f(x + t), \qquad e^{tA} B B^* e^{tA^*} f(x) = e^{-2|x-t|} f(x), \qquad t \geq 0,\ x \in \mathbb{R},$$

so that $Q_\infty^{(\alpha)}$, $\alpha \geq 0$, is the multiplication operator by the function g_α:

$$Q_\infty^{(\alpha)} f(x) = g_\alpha(x) f(x), \qquad g_\alpha(x) = \int_0^\infty e^{-2\alpha t} e^{-2|x-t|} \, dt, \qquad x \in \mathbb{R}.$$

It is immediate to verify (3.1), and so Hypothesis 3.1 holds. Elementary computations show that for $\alpha \neq 1$

$$g_\alpha(x) = \begin{cases} \dfrac{e^{-2x}}{2(\alpha-1)}\left(1 - \dfrac{2}{\alpha+1}\, e^{-2(\alpha-1)x}\right) & \text{for} \quad x > 0, \\ \dfrac{e^{2x}}{2(\alpha+1)} & \text{for} \quad x \leq 0, \end{cases}$$

whereas

$$g_1(x) = \begin{cases} e^{-2x}\left(x + \dfrac{1}{4}\right) & \text{for} \quad x > 0, \\ \dfrac{e^{2x}}{4} & \text{for} \quad x \leq 0. \end{cases}$$

For $0 \leq \alpha < \beta$, the inclusion $\mathcal{K}_\infty^\alpha = im\ (Q_\infty^{(\alpha)})^{1/2} \supset im\ (Q_\infty^{(\beta)})^{1/2} = \mathcal{K}_\infty^\beta$ always holds, so we have equality if and only if $im\ (Q_\infty^{(\alpha)})^{1/2} \subset im\ (Q_\infty^{(\beta)})^{1/2}$. Since, clearly, $(Q_\infty^{(\alpha)})^{1/2}$ is the multiplication operator by the function $g_\alpha^{1/2}$, equality holds if and only if $\sup_{x\in\mathbb{R}} g_\alpha(x)/g_\beta(x) < \infty$. Taking into account the previous formulae, one concludes that this holds if and only if $1 < \alpha < \beta$. Therefore, in this example the spaces $\mathcal{K}_\infty^\alpha$ are all equal for $\alpha > 1$, whereas they are all distinct for $0 \leq \alpha \leq 1$ (and $\mathcal{K}_\infty^1 \supsetneq \mathcal{K}_\infty^\alpha$ if $\alpha > 1$).

The number α_0 in the statement of Theorem 4.1 is equal to 1.

Remark 5.6. Let us change the definition of the operator A by subtracting the identity operator; namely, we define the semigroup (e^{tA}) setting

$$e^{tA} f(x) = e^{-t}\, f(x-t), \qquad t \geq 0,\ x \in \mathbb{R}.$$

Then (e^{tA}) clearly satisfies $\|e^{tA}\| \leq e^{-t}$ and so it is exponentially stable. In this case the value of α_0 is changed to $\alpha_0 = 0$.

5.4. Case (ii) of Theorem 4.1 with $\alpha_0 = \infty$. We start with some preliminary considerations. Inspection of the statement of Theorem 4.1 shows that if case (i) occurs or if case (ii) occurs with $\alpha_0 < \infty$, then there exist $\alpha > 0$ and $r > 0$ such that $\mathcal{K}_\infty^\alpha = \mathcal{K}_r$. Since, as already noted, $\mathcal{K}_r \subset \mathcal{K}_s \subset \mathcal{K}_\infty^\alpha$ for $r < s$, it follows that in these cases the spaces $\mathcal{K}_t$ coincide for all values of t large enough. Therefore, in order to find a situation where $\alpha_0 = \infty$, it suffices to construct an example where $\mathcal{K}_t$ are all distinct for $t > 0$.

The example that follows was communicated to us by Zabczyk.

We take $H = L^2([0,\infty))$ and let the operators e^{tA} be the right shift operators

$$e^{tA} f(x) = \begin{cases} f(x-t), & t \geq 0,\ x \geq t, \\ 0, & t > 0,\ x < t. \end{cases}$$

Next we denote by b the characteristic function of the interval $[0,1]$ (note that $b \in H$); we take $U = \mathbb{R}$ and define the operator B as the rank one operator: $Bv = bv$, $v \in \mathbb{R}$. Since, for every $t > 0$,

$$\mathcal{K}_t = im\ \mathcal{L}_t = \left\{ \int_0^t e^{(t-s)A} b\, u(s)\, ds \ :\ u \in L^2([0,t]) \right\},$$

it can be easily verified that the closure of $\mathcal{K}_t$ in H, denoted $\overline{\mathcal{K}_t}$, is the closure of the linear span of

$$\{e^{rA} b \ :\ r \in [0,t]\}.$$

It follows that elements of $\overline{\mathcal{K}_t}$ have supports contained in the interval $[0, 1+t]$ and, moreover, $\overline{\mathcal{K}_t}$ contains functions which are nonzero in a left neighborhood of t. Therefore, the spaces $\overline{\mathcal{K}_t}$ are distinct for different values of t, and so the spaces $\mathcal{K}_t$, $t > 0$, are also all distinct.

6. Applications to the Ornstein–Uhlenbeck process. As explained in the introduction, one of the motivations for studying the reachability spaces $\mathcal{K}_\infty$ introduced above is their probabilistic interpretation. This is the subject of the present section.

Let us consider the following stochastic equation:

$$\begin{cases} dX(t) = AX(t)\,dt + B\,dW(t), & t \geq 0, \\ X(0) = X_0. \end{cases} \tag{6.1}$$

We assume that H and U are real separable Hilbert spaces, A is the generator of a strongly continuous semigroup $\{e^{tA},\ t \geq 0\}$ of bounded linear operators in H, and B is a bounded linear operator from U to H. We assume that we are also given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, endowed with a filtration $\{\mathcal{F}_t,\ t \geq 0\}$ satisfying the usual conditions, and a cylindrical $(\mathcal{F}_t)$-Wiener process $\{W(t)\ t \geq 0\}$ defined on Ω with values in U; see, e.g., [8]. X_0 is an $\mathcal{F}_0$-measurable random variable with values in H. We also assume that the gramian operators

$$Q_t h = \int_0^t e^{sA} B B^* e^{sA^*} h\,ds, \qquad h \in H,\ t > 0,$$

introduced above, have finite trace.

Under these assumptions, the solution of (6.1) is defined as the stochastic process with values in H:

$$X(t) = e^{tA} X_0 + \int_0^t e^{(t-s)A} B\,dW(s), \qquad t \geq 0. \tag{6.2}$$

The integral occurring in (6.2) is the Itô stochastic integral and, for every t, it defines a random variable with values in H because of the finite trace condition Trace $Q_t < \infty$. Moreover, if X_0 is gaussian, then the process X is also gaussian. It is called the (nonsymmetric) Ornstein–Uhlenbeck process. For a detailed exposition of these facts, we refer to [8].

A basic problem is to investigate properties of invariant measures for the process X, i.e., Borel probability measures ν on H such that, if X_0 has distribution ν, then also $X(t)$ has distribution ν for every t. Invariant measures are known to exist if and only if

$$\sup_{t>0} \text{Trace } Q_t < \infty. \tag{6.3}$$

In this case, one invariant measure is the gaussian measure μ on H having zero mean and covariance equal to the extended gramian operator Q_∞ introduced above:

$$Q_\infty h = \int_0^\infty e^{sA} B B^* e^{sA^*} h\,ds, \qquad h \in H.$$

Under some additional assumptions, for instance if A is exponentially stable, μ is the unique invariant measure. For the proof of these facts, we still refer the reader to [8].

Remark 6.1. If (6.3) holds, then it is easy to show that condition (3.1) also holds (in particular, Hypothesis 3.1 is satisfied) and the operator Q_∞ has finite trace.

Associated to the centered gaussian measure μ is the so-called Cameron–Martin space (see, e.g., [1]), which coincides with the image of $Q_\infty^{1/2}$ and hence with the space of reachable states $\mathcal{K}_\infty$ introduced in the previous sections.

We address the problem to study the behavior of the Cameron–Martin space under perturbation of the Ornstein–Uhlenbeck process. Thus, for $\alpha \geq 0$, we consider the processes X_α, solution of

$$\begin{cases} dX_\alpha(t) = (A - \alpha I)X_\alpha(t)\, dt + B\, dW(t), & t \geq 0, \\ X_\alpha(0) = X_0, \end{cases} \tag{6.4}$$

and we consider the centered gaussian measures μ_α with covariance operator $Q_\infty^{(\alpha)}$ defined in (4.2). It follows from the previous discussion that μ_α is an invariant measure for X_α (and it is unique if $A-\alpha$ is exponentially stable) and that the Cameron–Martin space of μ_α coincides with $\mathcal{K}_\infty^\alpha$.

The following proposition is the main result of this section. It is a direct consequence of Theorem 4.1.

PROPOSITION 6.2. *Assume Hypothesis* 3.1 *and assume* (6.3). *Let* μ_α, $\alpha \geq 0$, *be the centered gaussian invariant measures introduced above, and let* $\mathcal{K}_\infty^\alpha$ *be their Cameron–Martin spaces. Then all the conclusions of Theorem* 4.1 *hold true.*

Remark 6.3. The occurrence of the various possibilities described in Theorem 4.1 depends on the inequality $\|S_0^{(\alpha)}(t)\| < 1$ for various values of $\alpha \geq 0$ and $t > 0$ or the equivalent one $im\ (Q_\infty^{(\alpha)})^{1/2} = im\ Q_t^{1/2}$ (compare with Proposition 4.3). These conditions play an important role in connection with various regularity properties of the transition semigroup of the Markov process X_α.

For instance, assume that Hypothesis 3.1 and (6.3) are satisfied, take $\alpha = 0$, and assume that $\ker Q_\infty = \{0\}$. (This simplifies some of the statements.) The transition semigroup of the Ornstein–Uhlenbeck process, denoted $\{R_t,\ t \geq 0\}$, can be considered as a strongly continuous contraction semigroup on each space $L^p(H,\mu)$, $p \in [1,\infty)$ (the space of Borel measurable functions $\phi : H \to \mathbb{R}$ such that $\int_H |\phi(x)|^p \mu(dx) < \infty$, endowed with its usual norm). It is proved in [3, Theorem 2] that the stronger property

$$\begin{aligned} \phi \in L^p(H,\mu),\ p \in (1,\infty) \implies & R_t\phi \in L^q(H,\mu) \text{ and} \\ & \|R_t\phi\|_{L^q(H,\mu)} \leq \|\phi\|_{L^p(H,\mu)} \end{aligned} \tag{6.5}$$

for some $q > p$ holds if and only if $im\ Q_\infty^{1/2} = im\ Q_t^{1/2}$. (The value of q depends on t and p.) Property (6.5) is called hypercontractivity (at time $t > 0$). Thus, by Proposition 4.3 and Corollaries 4.5 and 4.7, hypercontractivity holds at some $t > 0$ if and only if case (i) of Theorem 4.1 occurs.

Similar considerations relate the inequalities $\|S_0^{(\alpha)}(t)\| < 1$ (or the equalities $im\ (Q_\infty^{(\alpha)})^{1/2} = im\ Q_t^{1/2}$) to various regularity properties of the transition semigroup R, such as compactness, differentiability, and smoothing properties in appropriate function spaces. For further details, we refer the reader to [3, 4, 5, 6, 11, 13, 14].

Acknowledgments. We wish to thank Professors Luciano Pandolfi, Roberto Triggiani, and Jerzy Zabczyk for useful discussions. We also thank the anonymous referees for indicating improvements in section 5.

REFERENCES

[1] V. I. Bogachev, *Gaussian Measures*, Math. Surveys Monogr. 62, AMS, Providence, RI, 1998.

[2] N. Bouleau and F. Hirsch, *Dirichlet Forms and Analysis on Wiener Space*, de Gruyter Stud. Math. 14, de Gruyter, Berlin, 1991.

[3] A. Chojnowska-Michalik and B. Gołdys, *Nonsymmetric Ornstein-Uhlenbeck semigroup as second quantized operator*, J. Math. Kyoto Univ., 36 (1996), pp. 481–498.

[4] A. Chojnowska-Michalik and B. Gołdys, *Nonsymmetric Ornstein-Uhlenbeck generators*, in Infinite Dimensional Stochastic Analysis (Amsterdam, 1999), Verh. Afd. Natuurkd. 1. Reeks. K. Ned. Akad. Wet. 52, R. Neth. Acad. Arts Sci., Amsterdam, 2000, pp. 99–116.

[5] A. Chojnowska-Michalik and B. Gołdys, *Generalized Ornstein-Uhlenbeck semigroups: Littlewood-Paley-Stein inequalities and the P. A. Meyer equivalence of norms*, J. Funct. Anal., 182 (2001), pp. 243–279.

[6] A. Chojnowska-Michalik and B. Gołdys, *On regularity properties of nonsymmetric Ornstein-Uhlenbeck semigroup in L^p spaces*, Stochastics Stochastics Rep., 59 (1996), pp. 183–209.

[7] R. F. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Texts Appl. Math. 21, Springer-Verlag, New York, 1995.

[8] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Encyclopedia Math. Appl. 44, Cambridge University Press, Cambridge, UK, 1992.

[9] J. Diestel and J. J. Uhl, *Vector Measures*, Mathematical Surveys 15, AMS, Providence, RI, 1977.

[10] N. Dunford and J. T. Schwartz, *Linear Operators* 1: *General Theory*, Interscience, New York, 1958.

[11] M. Fuhrman, *Hypercontractivity properties of nonsymmetric Ornstein-Uhlenbeck semigroups in Hilbert spaces*, Stochastic Anal. Appl., 16 (1998), pp. 241–260.

[12] M. Fuhrman, *Logarithmic derivatives of invariant measure for stochastic differential equations in Hilbert spaces*, Stochastics Stochastics Rep., 71 (2001), pp. 269–290.

[13] B. Gołdys, *On analyticity of Ornstein-Uhlenbeck semigroups*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl., 10 (1999), pp. 131–140.

[14] B. Gołdys, *On Bilinear Forms Related to Ornstein-Uhlenbeck Semigroup on Hilbert Space*, manuscript.

[15] S. Hansen and G. Weiss, *New results on the operator Carleson measure criterion*, IMA J. Math Control Inform. 14 (1997), pp. 3–32.

[16] B. Jacob and H. J. Zwart, *Disproof of Two Conjectures of George Weiss*, Memorandum Faculteit TW 1546, Universiteit Twente, Enschede, The Netherlands, 2000; also available online from http://www.math.utwente.nl/publications/2000/1546.pdf.

[17] Z. M. Ma and M. Röckner, *Introduction to the Theory of (Non-symmetric) Dirichlet Forms*, Springer-Verlag, New York, 1992.

[18] D. L. Russell and G. Weiss, *A general necessary condition for exact observability*, SIAM J. Control Optim., 32 (1994), pp. 1–23.

[19] O. J. Staffans, *Well-posed Linear Systems*, Cambridge University Press, Cambridge, UK, to appear; also available online from http://www.abo.fi/∼staffans/publ.htm.

[20] G. Weiss, *A powerful generalization of the Carleson measure theorem*, in Open Problems in Mathematical Systems and Control Theory, V. D. Blondel, E. D. Sontag, M. Vidyasagar, and J. C. Willems, eds., Springer-Verlag, London, 1999, pp. 267–272.

[21] J. Zabczyk, *Mathematical Control Theory: An Introduction*, Systems Control Found. Appl., Birkhäuser Boston, Boston, 1982.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1795–1812

SINGULAR PERTURBATIONS IN INFINITE-DIMENSIONAL CONTROL SYSTEMS*

T. D. DONCHEV† AND A. L. DONTCHEV‡

Abstract. In this paper, we consider a singularly perturbed control system involving differential inclusions in Banach spaces with slow and fast solutions. Using the averaging approach, we obtain sufficient conditions for the Hausdorff convergence of the set of slow solutions in the supremum norm. We present applications of the theorem to prove convergence of the fast solutions in terms of invariant measures and convergence of equi-Lipschitz solutions.

Key words. control systems, differential inclusions, singular perturbations, time scales, order reduction, convergence of solutions

AMS subject classifications. 93C70, 34A60, 34E15, 34C29, 49K40

DOI. 10.1137/S0363012902416080

1. Introduction. In this paper, we consider a control system described by two differential inclusions—a "slow" one and a "fast" one, the latter indicated by a "small" parameter ε multiplying the derivative. The problem has the following form:

$$\begin{array}{ll} \dot{x}(t) \in F(x, y, u(t)), & x(0) = x^0, \\ \varepsilon \dot{y}(t) \in G(x, y, u(t)), & y(0) = y^0, \end{array} \tag{1}$$

where $F : E_1 \times E_2 \times U \rightrightarrows E_1$ and $G : E_1 \times E_2 \times U \rightrightarrows E_2$ are set-valued mappings, E_1 and E_2 are Banach spaces with uniformly convex duals E_1^* and E_2^*, respectively, U is a compact metric space, and $t \in I := [0, 1]$. Throughout the paper, we assume that for every $(x, y, u) \in E_1 \times E_2 \times U$, the values $F(x, y, u)$ and $G(x, y, u)$ are nonempty, convex, compact subsets of E_1 and E_2, respectively, and that F and G are upper semicontinuous (USC) in $E_1 \times E_2 \times U$ and lower semicontinuous with respect to u in U for fixed $(x, y) \in E := E_1 \times E_2$. The parameter ε is a positive scalar which may become zero, thus changing the system to a differential-algebraic inclusion, a change justifying the name "singular perturbations." A *control* $u(\cdot)$ is defined as a Lebesgue measurable function with values $u(t) \in U$ for almost every (a.e.) $t \in T$. A *solution* of (1) is any absolutely continuous (AC) function $(x(\cdot), y(\cdot))$ for which there exists a control $u(\cdot)$ such that the relations (1) hold for a.e. $t \in I$. For a fixed $\varepsilon > 0$ we denote by $Z(\varepsilon)$ the set of all solutions to (1). The slow x-part of $Z(\varepsilon)$ is denoted by $X(\varepsilon)$, that is, $X(\varepsilon) = \{x \mid (x, y) \in Z(\varepsilon)\}$, and the fast y-part by $Y(\varepsilon)$. In this paper, we study the asymptotic behavior of the family of sets $X(\varepsilon)$ as $\varepsilon \to 0$.

If F and G are single-valued and U consists of a single point (there is no control) and also $E_1 = \mathbb{R}^n, E_2 = \mathbb{R}^m$ are real Euclidean spaces, (1) becomes a system of singularly perturbed differential equations. The grounds for the modern theory of such equations were already laid in the 1940s, centered around the question of the dependence of solutions on the parameter ε. The earliest publication on this subject seems to be due to Levinson [24]; it was apparently a predecessor of the work of

*Received by the editors October 14, 2002; accepted for publication (in revised form) April 1, 2003; published electronically November 14, 2003.
http://www.siam.org/journals/sicon/42-5/41608.html
†Department of Mathematics, University of Architecture and Civil Engineering, 1 Hr. Smirnenski Str., 1046 Sofia, Bulgaria (tdd51us@yahoo.com).
‡Mathematical Reviews, Ann Arbor, MI 48107-8604 (ald@ams.org).

Tikhonov in [25], [26], [27], which evolved in a basic result, now known as Tikhonov's theorem. Assuming that F and G in (1) are continuous functions acting in Euclidean spaces and there is no control, Tikhonov's theorem can be roughly stated in the following way. For a fixed x and $\varepsilon > 0$, let $\varepsilon\dot{y} = G(x, y)$ have a unique solution $y_\varepsilon(x)$ which converges as $\varepsilon \to 0$ to the unique solution $y(x)$ of the algebraic equation $0 = G(x, y)$ obtained for $\varepsilon = 0$, and, moreover, let the solution of the *associate* equation $\dot{y} \in G(x, y)$ on $[0, \infty)$ be asymptotically stable uniformly in x. Then for any solution (x_0, y_0) of the differential-algebraic equation obtained from (1) for $\varepsilon = 0$, there is a solution $(x_\varepsilon, y_\varepsilon)$ of (1) for $\varepsilon > 0$ which converges to (x_0, y_0) uniformly in I for x and on $[\delta, 1]$ for y as $\varepsilon \to 0$, for any δ satisfying $1 > \delta > 0$. One should note that in the original work of Tikhonov, the uniform in x stability of the associate equation was not assumed; the necessity of such a condition and a complete version of Tikhonov's theorem was given by Hoppensteadt [19], [20].

It was first shown in [12] that a formal generalization of Tikhonov's theorem for differential inclusions, in particular for control systems, is false in the sense that the set of values of the solutions at a given "time" t (the reachable set at t) is not convergent, as $\varepsilon \to 0$, to the set of values of the reduced inclusion obtained for $\varepsilon = 0$. This is mainly caused by the possibility for "rapid" switchings of the control function $u(\cdot)$, with frequency which is proportional to the reciprocal to the value of the parameter ε. A variety of effects may occur when set-valued mappings are involved, to which different "Tikhonov-type" theorems correspond. Since the early 1980s, the literature on Tikhonov-type theorems for differential inclusions and control systems has grown rapidly. Major advances has been made by Artstein [1], Gaitsgory [15], [16], Grammel [17], [18], Donchev and Slavov [9], [10], and Veliov [28], [29]. In particular, Veliov [29] found a Tikhonov-type theorem for differential inclusions in finite dimensions from which the classical theorem of Tikhonov follows directly when the inclusion is an equation.

The present paper is, to the authors' knowledge, a first attempt to obtain a Tikhonov-type theorem for control systems described by differential inclusions in infinite-dimensional spaces. The passage from finite to infinite dimensions requires overcoming several not only technical difficulties, most of which are related to the lack of compactness in the state space. Here, we follow the *averaging approach* to singular perturbations which in finite dimensions has been developed by Gaitsgory [15], [16] and Grammel [17], [18]. Specifically, along with the system (1) we consider the associate system

$$\begin{aligned} &x = \text{const}, \\ &\dot{y}(\tau) \in G(x, y(\tau), u(\tau)) \text{ for } \tau > 0, \quad y(0) = y^0, \end{aligned} \tag{2}$$

and define

$$\hat{V}(x, S, y^0) = \operatorname{cl} \bigcup_{v(\cdot)} \left\{ \frac{1}{S} \int_0^S F(x, y(\tau), v(\tau))\, d\tau \mid y(\cdot) \in \tilde{Y}(x, S, y^0, v(\cdot)) \right\}.$$

Here the union is with respect to all controls $v(\cdot)$ defined on the interval $[0, S]$, that is, measurable functions on $[0, S]$ with $v(\tau) \in U$ for a.e. $\tau \in [0, S]$, and $\tilde{Y}(x, S, y^0, v(\cdot))$ is the set of all solutions $y(\cdot)$ of (2) on $[0, S]$ corresponding to the control $v(\cdot)$ and the initial condition y^0. In our main Theorem 1 stated below, we give conditions under which for every $x \in E_1$ the Hausdorff limit $\lim_{S\to\infty} \hat{V}(x, S, y^0)$ exists and is the same for every fixed $y^0 \in E_2$; that is, this limit depends on x only—we denote it by $V(x)$.

Moreover, the set of slow solutions $X(\varepsilon)$ of (1) has a Hausdorff limit for $\varepsilon \to 0$ in the supremum norm on I, and this limit, denoted X_0, coincides with the set of solutions of the differential inclusion

$$\dot{x}(t) \in V(x(t)), \quad x(0) = x^0. \tag{3}$$

Throughout the paper, $\|\cdot\|$ denotes any norm. Let $\mathcal{X}$ be a Banach space with uniformly convex dual $\mathcal{X}^*$. For $x \in \mathcal{X}, y \in \mathcal{X}^*$, $\langle y, x\rangle$ denotes the value of y at x, and the dual mapping is defined as $J(x) := \{y \in \mathcal{X}^* \mid \|y\| = \|x\| \text{ and } \langle y, x\rangle = \|x\|^2\}$. It is known that for the space $\mathcal{X}$ the dual mapping J is single-valued and uniformly continuous on bounded sets. The support function is defined as $\sigma(y, A) := \sup_{x\in A} \langle x, y\rangle$, and the Hausdorff distance between the sets A, B is denoted by $D_H(A, B)$. The closed hull of a set A is denoted by $\operatorname{cl} A$ and the convex closed hull by $\overline{\operatorname{co}}\, A$. Let Γ be a set-valued mapping acting between Banach spaces which is nonempty-, convex-, and compact-valued and continuous and bounded on bounded sets. For positive reals δ and N we define

$$\omega_\Gamma(\delta, N) = \sup\big\{ D_H(\Gamma(z_1), \Gamma(z_2)) \mid \ \|z_1 - z_2\| \le \delta, \max\{\|z_1\|, \|z_2\|\} \le N\big\}.$$

When it is clear from the context what N is, we write simply $\omega_\Gamma(\delta)$. If Γ is uniformly continuous on bounded sets, then ω_Γ is bounded and $\omega_\Gamma(\delta, N) \to 0$ as $\delta \to 0$. Our main result follows.

THEOREM 1. *Assume the following:*

(i) *There exist positive constants A, B, C, μ such that for all (x_1, y_1), $(x_2, y_2) \in E = E_1 \times E_2$, $u \in U$ the following inequalities hold:*

$$\sigma(J(x_1 - x_2), F(x_1, y_1, u)) - \sigma(J(x_1 - x_2), F(x_2, y_2, u)) \le A\|x_1 - x_2\|^2 + B\|y_1 - y_2\|^2;$$

$$\sigma(J(y_1 - y_2), G(x_1, y_1, u)) - \sigma(J(y_1 - y_2), G(x_2, y_2, u)) \le C\|x_1 - x_2\|^2 - \mu\|y_1 - y_2\|^2.$$

(ii) *For any positive reals m and r,*

$$\sup_{\|x\|\le r, u\in U} \int_0^\infty \omega_{F(x,\cdot,u)}(m \exp(-\mu\tau), r)\, d\tau < \infty,$$

where μ is the constant in the second condition in (i).

Then a mapping $V : E_1 \rightrightarrows E_1$ can be constructed as in (3) *such that, if $X(\varepsilon) = \big\{ x(\cdot) \mid (x(\cdot), y(\cdot))$ is a solution of* (1)$\big\}$ *and X_0 is the set of solutions of* (3)*, we have*

$$\lim_{\varepsilon\to 0} D_H(X(\varepsilon), X_0) = 0,$$

where the Hausdorff distance is defined with the supremum norm on I.

Remark 1. The conditions in (i) for F and G are specific forms of a general property called the *relaxed one-sided Lipschitz* (ROSL) condition. Let $\mathcal{X}$ be a Banach space. A set-valued mapping H from $\mathcal{X}$ into bounded subsets of $\mathcal{X}$ is said to be *one-sided Lipschitz* (OSL) when there exists a constant L such that for every $u, v \in \mathcal{X}$

$$\big\langle J(u - v), h_u - h_v\big\rangle \le L\|u - v\|^2 \text{ for every } h_u \in H(u) \text{ and every } h_v \in H(v).$$

The mapping H is said to be ROSL when there exists a constant L such that for every $u, v \in \mathcal{X}$,

$$\sigma(J(u - v), H(u)) - \sigma(J(u - v), H(v)) \le L\|u - v\|^2.$$

Note that in the assumption (i) the ROSL condition for G is with a negative constant with respect to y.

The OSL condition has been used for quite some time as a dissipative condition; see [22]. The ROSL condition was introduced by the first author in [4] under a different name. Examples and applications of this condition can be found in [6], [7], [8], [23]. In particular, the mapping $\tilde{x} \mapsto \overline{\text{co}}\, f(\tilde{x}, U)$, where $f(\tilde{x}, U) = \bigcup_{u \in U} f(\tilde{x}, u)$, is ROSL with a constant L when $f(\cdot, u)$ is ROSL (e.g., Lipschitz continuous) with a constant L for every $u \in U$.

Remark 2. The mapping ω_Γ is a modulus of continuity of the set-valued mapping Γ. The condition (ii) holds, for example, when $F(x, \cdot, u)$ is hölderian; that is, there exist constants $\alpha \in (0, 1)$ and $\beta > 0$ such that $D_H(F(x, y, u), F(x, z, u)) \leq \beta(\|y - z\|^\alpha + \|y - z\|^{1+\alpha})$, as pointed out in [10].

When F and G are single-valued and Lipschitz continuous with respect to (x, y), the Hausdorff continuity of $X(\cdot)$ at $\varepsilon = 0$ was proved in [16] and in [18] under different assumptions. A related result was also proved in [10] for F and G that are ROSL and continuous. Under weaker assumptions, [13] proved the weaker inclusion $\limsup_{\varepsilon \to 0} X(\varepsilon) \subset X(0)$. We refer also to [14], where upper and lower approximations of singularly perturbed systems are comprehensively studied and interesting examples are considered.

All papers cited above deal with finite-dimensional spaces, and important steps in the proofs in these papers use the compactness of the unit ball. In this paper, we go around such arguments and manage to tackle the infinite-dimensional singularly perturbed differential inclusion of the form (1). It may look specific, but actually it covers a lot of territory. In particular, we cover the case of a control system where the control acting on both x and y appears explicitly while the (different) controls acting separately on x and y are inscribed in the set-valued character of the mappings F and G. The partial differential inclusion described in [21, section I.7] can serve as an example of application of our results.

We do not know whether it would be possible to obtain a Tikhonov theorem of the type of Theorem 1 for a general singularly perturbed infinite-dimensional differential inclusion, such as the inclusion (9) in section 5 of this paper.

Our interests in considering infinite-dimensional systems stems from the natural question of whether and how the well-developed theory in finite dimensions can be extended to tackle infinite-dimensional problems. However, our ultimate goal is broader—not only to generalize but also to understand the meaning of the abstract results for specific infinite-dimensional models such as systems described by differential equations with delay or partial differential equations. In this paper, we chose to present applications of our result to other theoretical works in the area, leaving the exploration of applications to, e.g., distributed parameter systems for future research.

In the next section, we present preliminary results that set the stage for a proof of Theorem 1. Section 3 contains this proof. In section 4, we demonstrate an application of Theorem 1 to establish convergence for the fast subsystem in terms of invariant measures, in the sense of Artstein [1]. Section 5 contains another application of Theorem 1 to obtain an infinite-dimensional version of a result of the authors' with I. Slavov [11], regarding the convergence of sequences of both x and y trajectories that are equi-Lipschitz continuous.

2. Preliminaries. A mapping Γ from a topological space $\mathcal{X}$ into a topological space $\mathcal{Y}$ is said to be upper-semicontinuous (USC) at a given $x \in \mathcal{X}$ if for every open $V \supset \Gamma(x)$ there exists a neighborhood $W \ni x$ such that $V \supset \Gamma(y)$ for $y \in W$. If $\Gamma(\cdot)$ is USC at every point of its domain, we call it USC. When $\mathcal{X}$ and $\mathcal{Y}$ are metric spaces and Γ is compact-valued, Γ is USC iff for every $\delta > 0$ there exists $\lambda > 0$ such that

$\Gamma(B(x,\delta)) \subset B(\Gamma(x),\lambda)$ for any $x \in \mathcal{X}$, where $B(A,\delta) = \{x \in \mathcal{X} \mid \operatorname{dist}(x,A) < \delta\}$. A mapping $\Gamma : I \times \mathcal{X} \rightrightarrows \mathcal{Y}$ is said to be almost USC when for any $\varepsilon > 0$ there exists a compact set $I_\varepsilon \subset I$ with Lebesgue measure $\operatorname{meas}(I_\varepsilon) > 1 - \varepsilon$ such that $\Gamma(\cdot,\cdot)$ is USC at every point of $I_\varepsilon \times \mathcal{X}$. We refer to [3] for all the concepts used in the paper but not explicitly defined. We denote the closed unit balls in E_1 and E_2 by $\mathbb{B}_1$ and $\mathbb{B}_2$, respectively, where E_1 and E_2 are the spaces where the basic inclusion (1) is defined.

In the first two lemmas of this section, we assume that the condition (i) in Theorem 1 holds. These lemmas are based on standard arguments; for the first lemma, see, e.g., the proofs of Lemma 3.1 in [9], while for the second, see [5] or [10].

LEMMA 1. *There exist constants M and $N > 0$ such that for every solution $(\tilde{x}_\varepsilon(\cdot), \tilde{y}_\varepsilon(\cdot))$ of the initial value problem*

$$\begin{aligned} \dot{x}(t) &\in \overline{\operatorname{co}}\, F(x + \mathbb{B}_1, y + \mathbb{B}_2, U), \qquad x(0) = x^0, \\ \varepsilon \dot{y}(t) &\in \overline{\operatorname{co}}\, G(x + \mathbb{B}_1, y + \mathbb{B}_2, U), \qquad y(0) = y^0, \end{aligned}$$

and for every $t \in I$ we have

$$\|\tilde{x}_\varepsilon(t)\| + \|\tilde{y}_\varepsilon(t)\| \le M$$

and

$$\|F(\tilde{x}_\varepsilon(t) + \mathbb{B}_1, \tilde{y}_\varepsilon(t) + \mathbb{B}_2, U)\| + \|G(\tilde{x}_\varepsilon(t) + \mathbb{B}_1, \tilde{y}_\varepsilon(t) + \mathbb{B}_2, U)\| \le N.$$

LEMMA 2. *For every $y^1, y^2 \in E_2$, any $x \in E_1$, and any control $u(\cdot)$, if $y_1(\cdot)$ is a solution of the associate system* (2) *with $y(0) = y^1$, then there exists a solution $y_2(\cdot)$ of* (2) *with $y(0) = y^2$ such that*

$$\|y_1(\tau) - y_2(\tau)\| \le \exp(-\mu\tau)\|y^1 - y^2\| \quad \text{for all } \tau > 0,$$

where μ is the constant in condition (i).

In the remaining part of this section, we assume that both conditions (i) and (ii) of Theorem 1 hold with $m = 2M$ and $r = M$, where M is the constant in Lemma 1. The supremum in (ii) is denoted by L_1. The following corollary will be useful later.

COROLLARY 1. *For compact sets $P \subset E_1$ and $Q \subset E_2$, consider the following differential inclusion:*

$$\dot{y}(\tau) \in \overline{\operatorname{co}}\, G(P, y(\tau), U), \quad y(0) \in Q. \tag{4}$$

There exists a compact set $\hat{K} \subset E_2$ such that $y(t) \in \hat{K}$ for every solution $y(\cdot)$ of (4) *and every $t \ge 0$.*

Proof. First, observe that the set-valued mapping $H(x) := \overline{\operatorname{co}}\, G(P, x, U)$ is ROSL with a constant $-\mu$. Consequently, $\sigma(J(x(t) - 0), H(x(t))) - \sigma(J(x(t) - 0), H(0)) \le -\mu|x(t)|^2$. Therefore, for every solution $y(\cdot)$ of (4) we have

$$\langle J(x(t)), \dot{x}(t) \rangle \le -\mu|x(t)|^2 + |x(t)||H(0)|.$$

Using a standard argument, one can show that

$$\frac{d}{dt}|x(t)| \le -\mu|x(t)| + |H(0)|.$$

Let $m = |H(0)|$. Then either $|x(t)| \le m$ or $\frac{d}{dt}|x(t)| < 0$. Hence $|x(t)| \le \max\{|Q|, m\} = \hat{M}$. From Theorem 1 in [5] we know that the solution set of (4) is nonempty

and precompact in $C([0,T],E_2)$ for every $T>0$. Let $K(t)$ be the set of values $y(t)$ at t of all trajectories $y(\cdot)$ of (4) (the reachable set at t). Denote $N(s,t)=\max\{\exp(-s\mu),\exp(-t\mu)\}$. Then $D_H(K(t),K(s))\le 2N(s,t)\hat{M}$, thanks to Lemma 2. Hence the net $\{K(t)\}_{t\ge 0}$ is a Cauchy net. Therefore, there exists a compact set $K=\lim_{t\to\infty}K(t)$. Since the multimap $t\rightrightarrows K(t)$ is continuous, one has that $\hat{K}=K\bigcup(\bigcup_{t\ge 0}K(t))$ is compact. □

LEMMA 3. *Let $y^1,y^2\in M\mathbb{B}_2$, where M is the constant in Lemma 1. For every solution $y_1(\cdot)$ of (2) with $y_1(0)=y^1$ there exists a solution $y_2(\cdot)$ of (2) with $y_2(0)=y^2$ such that*

$$\frac{1}{S}\int_0^S D_H\big(F(x,y_1(\tau),u(\tau)),F(x,y_2(\tau),u(\tau))\big)\,d\tau\le\frac{L_1}{S}.$$

Proof. It is sufficient to observe that

$$\int_0^S D_H\big(F(x,y_1(\tau),u(\tau)),F(x,y_2(\tau),u(\tau))\big)\,d\tau \le\int_0^\infty \omega_{F(x,\cdot,u)}(2M\exp(-\mu\tau),M)\,d\tau.$$ □

LEMMA 4. *For every $S>0$ the map $\hat{V}(\cdot,S):=\cup_{y\in M\mathbb{B}_2}\hat{V}(\cdot,S,y)$ is ROSL with a constant independent of $S>0$.*

Proof. Let x_1,x_2 be given points in E_1. Let $y_1(\cdot)$ be a solution of (2) with x replaced by x_1 and corresponding control $u(\cdot)$. Define the mapping

$$(\tau,v)\mapsto\Gamma(\tau,v)=\{w\in G(x_2,v,u(\tau))\mid\langle J(y_1(\tau)-v),\dot{y}_1(\tau)-w\rangle \le C\|x_1-x_2\|^2-\mu\|y_1(\tau)-v\|\},$$

where C is from condition (i). The mapping $\Gamma(\cdot,\cdot)$ is almost USC with nonempty, convex, and compact values. Therefore, the inclusion $\dot{y}(t)\in\Gamma(t,y)$, $y(0)=y_0$ admits a solution; see, e.g., Theorem 1 in [5]. That is, there exists $y(\cdot)$ such that

$$\langle J(y_1(t)-y(t)),\dot{y}_1(t)-\dot{y}(t)\rangle\le C\|x_1-x_2\|^2-\mu\|y_1(t)-y(t)\|^2.$$

Hence

$$\|y_1(t)-y(t)\|^2\le\exp(-2\mu t)\int_0^t\exp(2\mu\tau)\{2C\|x_1-x_2\|^2\}\,d\tau =\exp(-2\mu t)\left(\frac{C}{\mu}\|x_1-x_2\|^2\right)[\exp(2\mu t)-1]<\frac{C}{\mu}\|x_1-x_2\|^2.$$

Therefore,

$$\sigma\left(J(x_1-x_2),\frac{1}{S}\int_0^S F(x_1,y_1(\tau),u(\tau))\,d\tau\right) -\sigma\left(J(x_1-x_2),\frac{1}{S}\int_0^S F(x_2,y_2(\tau),u(\tau))\,d\tau\right) \le\frac{1}{S}\int_0^S\Big[A\|x_1-x_2\|^2+B\|x_1-x_2\|\|y_1-y_2\|\Big]\,d\tau\le A\|x_1-x_2\|^2+\frac{BC}{\mu}\|x_1-x_2\|^2.$$

Hence $\hat{V}(\cdot,S,M\mathbb{B}_2)$ is ROSL with a constant $A+BC/\mu$. □

PROPOSITION 1. *There exist a mapping $V : E_1 \rightrightarrows E_1$ and a constant L_2 such that for every $x \in M\mathbb{B}_1$, every $y \in M\mathbb{B}_2$, and every sufficiently large S,*

$$D_H(V(x), \overline{\text{co}}\, \hat{V}(x, S, y)) \le \frac{L_2}{S}.$$

Furthermore, $V(\cdot)$ is USC and ROSL, and for any $x \in M\mathbb{B}_1$

$$\lim_{S\to\infty} D_H(V(x), \hat{V}(x, S, y)) = 0 \quad \textit{uniformly in } y \in M\mathbb{B}_2.$$

Proof. The proof is similar to the proof of Proposition 3.2 of Grammel [17] and will be only sketched. Let $x \in M\mathbb{B}_1$. If for a $y' \in M\mathbb{B}_2$ the set $\overline{\text{co}}\, \hat{V}(x, S, y')$ converges in the Hausdorff sense to some $V(x)$, then it also converges for any other $y \in M\mathbb{B}_2$ to *the same* $V(x)$. We fix $y(0) = y' \in M\mathbb{B}_2$, and, by using the argument in the proof of Proposition 3.2 of [17] which also works in infinite dimensions, we obtain the following.

Claim 1. $D_H(\hat{V}(x, (k+h)S, y'), \hat{V}(x, kS, y')) \le 2N/k$, where N is as in Lemma 1.

Indeed, for all k, all $h \in [0, 1)$, and all $S > 0$, we have

$$\hat{V}(x, (k+h)S, y') = \text{cl} \bigcup_{u(\cdot)} \bigcup_{y(\cdot)\in Y(x,S,y',u(\cdot))} \left\{ \frac{1}{(k+h)S} \left(\int_0^{kS} + \int_{kS}^{(k+h)S} \right) F(x, y(\tau), u(\tau))\, d\tau \right\}.$$

Hence

$$\hat{V}(x, (k+h)S, y') \subset \frac{k}{k+h} \hat{V}(x, kS, y') + \frac{h}{k+h} N\mathbb{B}_1 \subset \hat{V}(x, kS, y') + \frac{2Nh}{k+h}\mathbb{B}_1.$$

Analogously,

$$\hat{V}(x, kS, y') \subset \hat{V}(x, (k+h)S, y') + \frac{2Nh}{k+h}\mathbb{B}_1.$$

Claim 2. For all $k = 1, 2, \ldots$ and for all $S > 0$, we have

$$D_H \left(\hat{V}(x, kS, y), \frac{1}{k} \sum_{i=1}^{k} \hat{V}(x, S, y) \right) \le \frac{L_1}{S}.$$

The proof of Claim 2 is by induction and closely follows the proof of an analogous result in Grammel [17]. Since $D_H(\overline{\text{co}}\, A, \overline{\text{co}}\, B) \le D_H(A, B)$, applying the triangle inequality, one obtains

$$D_H\big(\overline{\text{co}}\, \hat{V}(x, (k+h)S, y), \overline{\text{co}}\, \hat{V}(x, S, y)\big) \le \frac{2N}{k} + \frac{L_1}{S} \tag{5}$$

for all $k = 1, 2, \ldots, S > 0$ and $h \in [0, 1]$. Fix $\varepsilon > 0$. Let $S_1 > 4L_1/\varepsilon$ and $k > 4N/\varepsilon$. Then

$$D_H(\overline{\text{co}}\, \hat{V}(x, (k+h)S_1, y'), \overline{\text{co}}\, \hat{V}(x, S_1, y')) \le \frac{\varepsilon}{2}.$$

Consequently, for $\hat{S}$ and $\tilde{S} > kS_1$ one has $D_H(\overline{\text{co}}\, \hat{V}(x, \hat{S}, y'), \overline{\text{co}}\, \hat{V}(x, \tilde{S}, y')) \le \varepsilon$. Thus $\{\overline{\text{co}}\, \hat{V}(x, S, y')\}_{S>0}$ is a Cauchy net, and therefore it has a limit $V(x)$. Due to Lemma 2

the limit does not depend on $y(0) = y' \in M\mathbb{B}_2$. Furthermore, $V(\cdot)$ is ROSL thanks to Lemma 4. The claim that V is USC follows from the USC property of $\hat{V}$. By virtue of (5) we have

$$D_H(V(x), \overline{\text{co}}\, \hat{V}(x, S, y')) \leq \frac{L_2}{S}.$$

To complete the proof, we use the following fact: Given a compact set A in an arbitrary Banach space with unit ball $\mathbb{B}$, for every $\varepsilon > 0$ there exists n_ε such that

$$D_H\left(\overline{\text{co}}\, A, \frac{1}{k}\sum_{i=1}^{k} A\right) < \varepsilon \quad \text{for} \quad k > n_\varepsilon.$$

Indeed, for $\varepsilon > 0$ there exist a finite number of points $a_i \in \overline{\text{co}}\, A, i = 1, \ldots, n_\varepsilon$, such that if A_ε is the set of those points, then $D_H(\overline{\text{co}}\, A, A_\varepsilon) < \delta$. Further, since $\overline{\text{co}}\, A$ is the closure of the set of all finite convex combinations of elements of A, one has that for every $a_i \in A_\varepsilon$ there exists a finite convex combination of points $a^j \in A, j = 1, 2, \ldots, k_i$, such that

$$\left\| a_i - \sum_{j=1}^{k_i} \lambda_j a^j \right\| < \varepsilon.$$

On the other hand, there exists a natural number N such that for every $n > N$ every such λ_j can be approximated with accuracy $1/n$ by a rational number having the form p_j/n, where $\sum_{j=1}^{k_j} p_j/n = 1$. Hence

$$\left\| a_i - \sum_{j=1}^{k_i} \frac{p_j}{n} a^j \right\| \leq \varepsilon + \frac{k_i}{n} q,$$

where $q \in \mathbb{R}_+$ such that $A \subset q\mathbb{B}$. Therefore,

$$D_H\left(\overline{\text{co}}\, A, \frac{1}{n}\sum_{i=1}^{n} A\right) \leq \varepsilon + \frac{k}{n} q,$$

where $k = \max_i\{k_i\}$. From the arguments above it is clear that for a closed set B with $D_H(A, B) < \varepsilon$ there exists $k(\varepsilon)$ such that $D_H(B_n, \overline{\text{co}}\, A) < 2\varepsilon$ for every $n > k(\varepsilon)$, where $B_n = \frac{1}{n}\sum_{i=1}^{n} B$. Thus one obtains that for every $S > 0$ and every $\delta > 0$ there exists n such that

$$D_H\big(\overline{\text{co}}\, \hat{V}(x, S, y'), \hat{V}(x, (k+h)S, y')\big) < \delta \quad \text{for} \quad k > n.$$

Consequently, $\lim_{S\to\infty} D_H(V(x), \hat{V}(x, S, y)) = 0$. The proof is complete. □

We will also use the following (refined) version of Plis's lemma (see Lemma 8.3 of [3] and [28]).

LEMMA 5. *Given a control $u(\cdot)$ and functions $f : I \to I$ and $g : I \to I$, if (x, y) is AC and such that*

$$\dot{x}(t) \in F(x(t) + f(t)\mathbb{B}_1, y(t) + g(t)\mathbb{B}_2, u(t)),$$
$$\varepsilon\dot{y}(t) \in G(x(t) + f(t)\mathbb{B}_1, y(t) + g(t)\mathbb{B}_2, u(t)),$$

then there exist constants K_x and K_y and a solution (w,z) of (1) with

$$\|x(t)-w(t)\| \le \sqrt{r(t)}, \quad \|y(t)-z(t)\| \le \sqrt{s(t)},$$

where

$$\begin{aligned}
&\dot r(t) = Ar + Bs + K_x(f(t)+g(t)+\omega_J(f(t))), \quad r(0) = \|x^0 - w(0)\|^2,\\
&\varepsilon\dot s(t) = Cr - \mu s + K_y(f(t)+g(t)+\omega_J(g(t))), \quad s(0) = \|y^0 - z(0)\|^2.
\end{aligned}$$

Proof. Consider the mappings

$$\begin{aligned}
&(t,w,z,u) \mapsto \tilde F(t,w,z,u) = \mathrm{cl}\{\alpha \in F(w,z,u) \mid \big\langle J(x(t)-w+f(t)), \dot x(t)-\alpha\big\rangle\\
&\qquad \le A\|x(t)-w+f(t)\|^2 + B\|y(t)-z+g(t)\|^2\},\\
&(t,w,z,u) \mapsto \tilde G(t,w,z,u) = \mathrm{cl}\{\beta \in G(w,z,u) \mid \varepsilon\big\langle J(y(t)-z+g(t)), \dot y(t)-\beta\big\rangle\\
&\qquad \le C\|x(t)-w+f(t)\|^2 - \mu\|y(t)-z+g(t)\|^2\}.
\end{aligned}$$

The so-defined $\tilde F$ and $\tilde G$ are almost USC with nonempty, convex, compact values. Furthermore, the system

$$\begin{aligned}
&\dot x(t) \in \tilde F(t,x,y,u), \quad x(0) = w(0),\\
&\varepsilon\dot y(t) \in \tilde G(t,x,y,u), \quad y(0) = z(0),
\end{aligned}$$

has a solution $(w(t), z(t))$ such that

$$\big\langle J(x(t)-w(t)+f(t)), \dot x(t)-\dot w(t)\big\rangle \le A\|x(t)-w(t)+f(t)\|^2 + B\|y(t)-z(t)+g(t)\|^2 := \mathcal{P}$$

and

$$\varepsilon\big\langle J(y(t)-z(t)+g(t)), \dot y(t)-\dot z(t)\big\rangle \le C\|x(t)-w(t)+f(t)\|^2 - \mu\|y(t)-z(t)+g(t)\|^2.$$

Consequently,

$$\begin{aligned}
&\big\langle J(x(t)-w(t)), \dot x(t)-\dot w(t)\big\rangle \le \mathcal{P}\\
&\quad + \|\big\langle J(x(t)-w(t)+f(t)), \dot x(t)-\dot w(t)\big\rangle - \big\langle J(x(t)-w(t)), \dot x(t)-\dot w(t)\big\rangle\|\\
&\quad \le \mathcal{P} + \|\dot x(t)-\dot w(t)\|\|J(x(t)-w(t)) - J(x(t)-w(t)+f(t))\| \le \mathcal{P} + 2N\omega_J(f(t)).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\big|\|x(t)-w(t)+f(t)\|^2 - \|x(t)-w(t)\|^2\big|\\
&\quad = \big|\|x(t)-w(t)+f(t)\| - \|x(t)-w(t)\|\big| \cdot \big|\|x(t)-w(t)+f(t)\| + \|x(t)-w(t)\|\big|\\
&\quad \le f(t)(2\|x(t)\| + 2\|w(t)\| + f(t)) \le \|f(t)\|(4M+1).
\end{aligned}$$

Hence

$$\big\langle J(x-w), \dot x-\dot w\big\rangle \le A\|x-w\|^2 + B\|y-z\|^2 + 2N\omega_J(f(t)) + (4M+1)Af(t) + (4M+1)Bg(t)).$$

Analogously

$$\begin{aligned}
\varepsilon\big\langle J(y-z), \dot x-\dot w\big\rangle \le{}& C\|x-w\|^2 - \mu\|y-z\|^2 + 2N\varepsilon\omega_J(g(t))\\
&+ (4M+1)Cf(t) + (4M+1)\|\mu\|g(t)).
\end{aligned}$$

Therefore, $\|x(t)-w(t)\|^2 \le r(t)$, $\|y(t)-z(t)\|^2 \le s(t)$, where r and s are defined in the statement of the lemma. ∎

3. Proof of Theorem 1. We divide the interval $[0, \varepsilon^{-1}]$ on subintervals $[t_i, t_{i+1}]$ with lengths $\varepsilon S_\varepsilon > 0$ such that $\lim_{\varepsilon\to 0} S_\varepsilon = \infty$, $\lim_{\varepsilon\to 0}(\varepsilon S_\varepsilon) = 0$. Let $t_j = j\varepsilon S_\varepsilon$, $\tau_j = jS_\varepsilon$ for $j = 0, 1, \dots, E_\varepsilon$, where E_ε is the largest integer less or equal to $(\varepsilon S_\varepsilon)^{-1}$. We also let $t_{E_\varepsilon+1} = 1$.

Step I. Let $(x_\varepsilon(\cdot), y_\varepsilon(\cdot))$ be a solution of (1) for a control $u_\varepsilon(\cdot)$ and for a sufficiently small ε. We will show that there exists a solution $z(\cdot)$ of (3) the uniform in I distance from which to $x_\varepsilon(\cdot)$ can be made arbitrary small.

Denote $\Delta = \varepsilon S_\varepsilon$ and $x_j = x_\varepsilon(t_j)$. We will apply Proposition 1 for $x_\varepsilon(t_j)$, but first we have to adjust $y_\varepsilon(\cdot)$. Define on $[t_j, t_{j+1}]$ the mapping $(t, v, w) \mapsto P_\varepsilon(t, v, w)$, where

$$P_\varepsilon(t,v,w) = \operatorname{cl}\big\{g \in G(v,w,u_\varepsilon(t)) \mid \big\langle J(y_\varepsilon(t)-w), \varepsilon\dot{y}_\varepsilon(t) - g\big\rangle \\ \le C\|x_\varepsilon(t)-v\|^2 - \mu\|y_\varepsilon(t)-w\|^2\big\}.$$

The mapping P_ε is almost USC with nonempty, convex, and compact values. Let $y_z(\cdot)$ be a solution of the differential inclusion

$$\varepsilon\dot{y}(t) \in P_\varepsilon(t, x_j, y(t)), \quad y(t_j) = y_j := \lim_{t\to t_j} y(t), \text{ and } y_z(0) = y^0.$$

Since $\bar{y}(\tau) = y_z(\varepsilon(\tau + \tau_j))$ is a solution of the associate system (2) with $\bar{y}(0) = y_z(\tau_j)$ on $[0, s_\varepsilon]$, we have

$$\varepsilon\big\langle J(y_\varepsilon(t) - y_z(t)), \dot{y}_\varepsilon(t) - \dot{y}_z(t)\big\rangle \le C\|x_\varepsilon(t) - x_j\|^2 - \mu\|y_\varepsilon(t) - y_z(t)\|^2,$$

that is,

$$\varepsilon\big\langle J(y_\varepsilon(t) - y_z(t)), \dot{y}_\varepsilon(t) - \dot{y}_z(t)\big\rangle \le N^2C\|t - t_j\|^2 - \mu\|y_\varepsilon(t) - y_z(t)\|^2.$$

Thus

$$\varepsilon\frac{d}{dt}\|y_\varepsilon(t) - y_z(t)\|^2 \le r(t),$$

where

$$\varepsilon\dot{r}(t) \le 2CN^2\Delta^2 - 2\mu r(t), \quad r(0) = 0.$$

Furthermore,

$$e^{-2\mu t/\varepsilon}\int_0^t e^{2\mu\tau/\varepsilon} 2CN^2\Delta^2/\varepsilon\, d\tau = \frac{CN^2\Delta^2}{\mu}(1 - e^{-2\mu/\varepsilon}).$$

Hence there exists a constant C_1 such that

$$\|y_z(t) - y_\varepsilon(t)\| \le C_1\Delta.$$

Now we want to approximate $x_\varepsilon(\cdot)$ with a function with a "piecewise" constant derivative. For $t \in [t_j, t_{j+1}]$, $j = 0, 1, \dots, E_\varepsilon$, consider the mapping

$$W_\varepsilon(t) = \operatorname{cl}\big\{w \in F(h_j, y_z(t), u_\varepsilon(t)) \mid \big\langle J(x_\varepsilon(t) - h_j), \dot{x}_\varepsilon(t) - w\big\rangle \\ \le A\|x_\varepsilon(t) - h_j\|^2 + B\|y_\varepsilon(t) - y_z(t)\|^2\big\}.$$

Define $h(t) = h_j + \int_{t_j}^{t} w(\tau)\,d\tau$, where $w(t) \in W_\varepsilon(t)$ is strongly measurable. Since $\|w(\tau)\| \le N$ (Lemma 1), we have

$$\big\langle J(x_\varepsilon(t)-h(t)), \dot{x}_\varepsilon(t)-\dot{h}(t)\big\rangle \le 2N\omega_J(N\Delta)+BC_1\Delta^2+A\big|\|x_\varepsilon(t)-h_j\|^2-\|x_\varepsilon(t)-h(t)\|^2\big|.$$

Thus

$$\frac{d}{dt}\|x_\varepsilon(t)_h(t)\|^2 \le 4N\omega_J(N\Delta) + BC_1\Delta^2 + 4MA\Delta + 2A\|x_\varepsilon(t) - h(t)\|^2.$$

Hence $\|x_\varepsilon(t)-h(t)\| \le C_1\sqrt{\omega_J(N\Delta)+M\Delta}$ for some constant C_2. Since $\lim_{\varepsilon\to 0}\Delta(\varepsilon) = 0$, one has that $\|x_\varepsilon(t) - h(t)\| \le \lambda(t)$, where $\lim_{\varepsilon\to 0}\lambda(\varepsilon) = 0$. Due to Proposition 1, there exists $v_j \in V(x_j)$, $j = 1, 2, \ldots, E_\varepsilon - 1$, with

$$\left\| \frac{1}{S_\varepsilon}\int_{\tau_j}^{\tau_{j+1}} \big(\dot{h}(\tau) - v_j\big)\,d\tau \right\| \le \frac{L_2}{S_\varepsilon}$$

(where L_2 is from Proposition 1). The function $m(t) = m(t_j) + (t - t_j)v_j$ for $t \in [t_j, t_{j+1}]$, $j = 0, 1, \ldots, E_\varepsilon$, and $m(0) = x_0$ satisfies

$$\|h(t) - m(t)\| \le \frac{L_2}{S_\varepsilon} \quad \text{for all } t \in I.$$

Clearly,

$$\|m(t) - x_\varepsilon(t)\| \le \|h(t) - m(t)\| + \|h(t) - x_\varepsilon(t)\| \le 2M\Delta + 2C\lambda(\varepsilon).$$

Furthermore, $\dot{m}(t) \in V(x_j)$ and $\|x_j - x_\varepsilon(t)\| \le N\Delta$. Hence

$$\dot{m}(t) \in V\big(m(t) + \big(2N\Delta + 2M\Delta + 2C\lambda(\varepsilon)\big)\mathbb{B}_1\big).$$

Consequently, there exists $\nu(\varepsilon) > 0$ with $\lim_{\varepsilon\to 0}\nu(\varepsilon) = 0$ such that $\dot{m}(t) \in V(m(t) + \nu(\varepsilon)\mathbb{B}_1)$. Therefore, we can apply Theorem 1 of [5], obtaining that there exists a solution $z(\cdot)$ of (3) such that

$$\|z(t) - m(t)\| \le D_1\sqrt{\nu(\varepsilon)} + \omega_J(C_2\sqrt{\nu(\varepsilon)}).$$

Step II. We will show now that for every solution $z(\cdot)$ of (3) there exists a solution $(x_\varepsilon, y_\varepsilon)$ of (1) such that $\|x_\varepsilon(t) - z(t)\| \to 0$ as $\varepsilon \to 0$. To this end we employ the idea of partition used in Step I.

The solution set X_0 of (3) is compact. Therefore, the set $\mathcal{L} = \bigcup_{t\in I}\{z(t) \mid\ z(\cdot) \in X_0\}$ is compact and hence for every $\lambda > 0$ there exists a finite λ-net $\{\tilde{z}_i\}_{i=1}^{n_\lambda}$ of it. Observe that

$$\mathcal{A} = \bigcup_{i=1}^{n_\lambda}(\tilde{z}_i + \lambda\mathbb{B}_1) \supset \mathcal{L} + \nu\mathbb{B}_1$$

for some $\nu > 0$ because $\mathcal{A}$ is open and $\mathcal{L} \subset \mathcal{A}$ is compact. Furthermore, for every $\delta > 0$ there exists S_δ such that

$$D_H\left(V(\tilde{z}_i), \frac{1}{S}\bigcup_{v(\cdot)}\left\{\int_0^S F(\tilde{z}_i, y(\tau), v(\tau))\,d\tau \mid y \in Y(\tilde{z}_i, S, M\mathbb{B}_2, v(\cdot))\right\}\right) < \delta$$

for every i and every $S > S_\delta$, where the union is with respect to the feasible controls $v(\cdot)$ and

$$Y(x, S, M\mathbb{B}_2, v(\cdot)) := \bigcup_{y^0 \in M\mathbb{B}_2} \tilde{Y}(x, S, y^0, v(\cdot)).$$

If $z(\cdot)$ is a solution of (3), then for $j = 1, 2, \dots, E_\varepsilon$, one has

$$\dot{z}(t) \in V\big(z(t_j) + N\Delta\mathbb{B}_1\big).$$

In order to apply Proposition 1, we find an approximate solution of (3), which is near $z(\cdot)$. Iteratively, we define a function $w(\cdot)$ and then a mapping W of the form

$$\begin{aligned}(t,p) \mapsto W(t,p) = \operatorname{cl}\{v \in V(w(t_j)) \mid \big\langle J(z(t)-p), \dot{z}(t) - v\big\rangle \\ \le (L+1)(\|z(t) + w(t_j)\|^2 + \|J(w(t_j)) - J(w(t))\|^2)\},\end{aligned}$$

where L is the ROSL constant of V. For $j = 0$ we set $w(0) = x^0$. The mapping W is nonempty-, convex-, and compact-valued and almost USC. Therefore, there exists a solution $w(t)$ on $[t_j, t_{j+1}]$ of

$$\dot{w}(t) \in W(t, w), \quad w(t_j) = \lim_{t \to t_j} w(t).$$

Thus

$$\|w(t) - z(t)\| \le \sqrt{r(t)}, \qquad \text{where} \qquad \dot{r}(t) = Lr(t) + \omega_J(N\Delta),\ r(0) = 0.$$

If $0 < \delta < \nu$, we can choose Δ so small that $\|w(t) - z(t)\| < \delta$. Let $z_j = \Delta^{-1} \int_{t_j}^{t_{j+1}} \dot{w}(t)\, dt$. By the definition of $w(\cdot)$ we have $z_j \in V(w(t_j))$. Now we will use Proposition 1 and the change of time scale $\tau = t/\varepsilon$. Due to Proposition 1 there exists $\varepsilon(\delta) > 0$ such that for every $0 < \varepsilon < \varepsilon(\delta)$ there exists a control $u(\cdot)$ on $(\tau_j, \tau_{j+1}]$ which satisfies

$$\operatorname{dist}\left(\tilde{z}_j, \frac{1}{S_\varepsilon}\left\{ \int_{\tau_j}^{\tau_{j+1}} F(w(\tau_j), v(\tau), u(\tau))\, d\tau \mid \ v(\cdot) \in Y(w(\tau_j), S_\varepsilon; M\mathbb{B}_2, u(\cdot)) \right\}\right) < \delta$$

for some $\tilde{z}_j$ with $\|z_j - \tilde{z}_j\| < \delta$. By the definition of $Y(w(\tau_j), S_\varepsilon; M\mathbb{B}_2, u(\cdot))$ there exists an AC function $y_w(\cdot)$ defined as follows: $y_w(0) = y_0$, $\dot{y}_w(\tau) \in G(w(\tau_j), y_w(\tau), u(\tau))$, with the convention $v(\tau_j) = \lim_{t \to \tau_j - 0} v(t)$, on $[\tau_j, \tau_{j+1}] = [jS_\varepsilon, (j+1)S_\varepsilon]$, such that

$$\operatorname{dist}\left(\tilde{z}_j, \frac{1}{S_\varepsilon}\left\{ \int_{\tau_j}^{\tau_{j+1}} F(w(\tau_j), y_w(\tau), u(\tau))\, d\tau \right\}\right) < \delta.$$

Let $\dot{h}(\cdot)$ be a measurable selection of $F(w(t_j), y_w(\cdot), u(\cdot))$ such that

$$\left\| \tilde{z}_j - \frac{1}{\Delta} \int_{t_j}^{t_{j+1}} \dot{h}(t)\, dt \right\| \le \delta.$$

Now we define an approximate solution $h(\cdot)$ of (1) which is near $w(\cdot)$. Let $h(t) = \int_0^t \dot{h}(\tau)\, d\tau + x_0$. Then $\|h(t) - w(t)\| \le 2\delta + N\Delta/2$. Therefore,

$$\dot{h}(t) \in F(h(t) + (2\delta + N\Delta/2)\mathbb{B}_1, y_w(t), u(t)).$$

Let $(x_\varepsilon, y_\varepsilon)$ be a solution of

$$\dot{x}(t) \in \tilde{F}(t, x(t), y(t), u(t)),\ x(0) = x^0,$$
$$\varepsilon\dot{y}(t) \in \tilde{G}(t, x(t), y(t), u(t)),\ y(0) = y_0,$$

where $\tilde{F}$ and $\tilde{G}$ are defined on $[t_j, t_{j+1}]$ for $j = 0, 1, \ldots, E_\varepsilon$ as follows:

$$\tilde{F}(t, \alpha, \beta, u(t)) = \{p \in F(\alpha, \beta, u(t)) \mid \langle J(\alpha - h_1(t)), p - \dot{h}(t)\rangle$$
$$\leq A\|\alpha - h_1(t)\|^2 + B\|\beta - y_w(t)\|^2\},\ \text{where } \|h_1(t) - h(t)\| \leq 2\delta + \frac{N\Delta}{2};$$
$$\tilde{G}(t, \alpha, \beta, u(t)) = \{q \in G(\alpha, \beta, u(t)) \mid \langle J(\beta - y_w(t_j)), q - \varepsilon\dot{y}_w(t)\rangle$$
$$\leq C\|\alpha - w(t)\|^2 - \mu\|\beta - y_w(t)\|^2.$$

The mappings $\tilde{F}, \tilde{G}$ are almost USC with nonempty, convex, compact values. Furthermore,

$$|\langle J(\alpha - h_1(t)), p - \dot{h}(t)\rangle - \langle J(\alpha - h(t)), p - \dot{h}(t)\rangle| \leq \omega_J(\|h_1(t) - h(t)\|)\|p - \dot{h}(t)\|$$
$$\leq 2N\omega_J\left(2\delta + \frac{N\Delta}{2}\right)2N,$$

$$A\big|\|\alpha - h_1(t)\|^2 - \|\alpha - h(t)\|^2\big| \leq A\left(2\delta + \frac{N\Delta}{2}\right)4M$$

and

$$C\big|\|\alpha - w(t_j)\|^2 - \|\alpha - h(t)\|^2\big| \leq C\|h(t) - w(t_j)\|(2\|\alpha\| + \|w(t_j)\| + \|h(t)\|)$$
$$\leq 4CM\left(2\delta + \frac{N\Delta}{2}\right).$$

Due to Lemma 5

$$\|x_\varepsilon - z(t)\|^2 \leq m(t) \quad \text{and} \quad \|y_\varepsilon(t) - y_w(t)\|^2 \leq s(t),$$

where

$$\dot{m}(t) \leq Am + Bs + 4\Delta^2 + C_1\left(\omega_J\left(2\delta + \frac{N\Delta}{2}\right) + 2\delta\right), \quad m(0) = 0,$$
$$\varepsilon\dot{s}(t) \leq Cm - \mu s + 4\Delta^2 + 4CM\left(2\delta + \frac{N\Delta}{2}\right), \quad s(0) = 0.$$

Therefore,

$$\|h(t) - x_\varepsilon(t)\| \leq \kappa\sqrt{\omega_J\left(2\delta + \frac{N\Delta}{2}\right) + \delta}$$

and

$$\|y_w(t) - y_\varepsilon(t)\| \leq \kappa_1\sqrt{\delta},$$

where κ and κ_1 are constants. Applying the triangle inequality

$$\|z(t) - x_\varepsilon(t)\| \leq \|z(t) - w(t)\| + \|w(t) - h(t)\| + \|h(t) - x_\varepsilon(t)\|,$$

we complete the proof. □

As a side result of the proof of Theorem 1, we obtain compactness of the set of values of the trajectories, which we state as the following corollary.

COROLLARY 2. *On the assumptions of Theorem* 1*, there exists a compact set* K *such that every solution* $(x_\varepsilon(\cdot), y_\varepsilon(\cdot))$ *of* (1) *satisfies* $(x_\varepsilon(t), y_\varepsilon(t)) \in K$ *for every* $(\varepsilon, t) \in (0, 1] \times I$.

Proof. It is clear that the reachable set $X_\varepsilon(t)$ of the slow part of (1) at the moment t is compact. Also, the reachable set of (3), $X_0(t) = \lim_{\varepsilon \to 0} X_\varepsilon(t)$, is compact. Moreover, the map $(\varepsilon, t) \to X_\varepsilon(t)$ is continuous. Hence $\mathcal{K} = \mathrm{cl} \cup_{t \in I} \cup_{\varepsilon \in [0,1]} X_\varepsilon(t) \subset E_1$ is a compact set. Due to Corollary 1, there exists a compact set $\hat{K}$ such that $y(t) \in \hat{K}$ for every t and every solution $y(\cdot)$ of (4) with P replaced by $\mathcal{K}$. Therefore, for every solution $(x_\varepsilon(\cdot), y_\varepsilon(\cdot))$ one has that $x_\varepsilon(t) \in \mathcal{K}$ and $y_\varepsilon(t) \in \hat{K}$. Consequently, the set $K = \mathcal{K} \times \hat{K}$ is the one we need. □

4. Convergence of the fast trajectories. Let $f : I \times E_2 \to \mathbb{R}$ be almost continuous and integrably bounded. The set of all such functions becomes a Banach space with the norm

$$\|f\|_2 = \int_I \sup_{\|y\| \le M} |f(t, y)|\, dt.$$

Denote this space by $\tilde{E}$. Every continuous function $y : I \to E_2$ can be considered as an element of $\tilde{E}^*$. The sequence $\{\nu_i\}_{i=1}^\infty \subset \tilde{E}^*$ is said to converge to ν in $\tilde{E}^*$-weak* topology if it converges to ν for every $f(\cdot, \cdot) \in \tilde{E}$.

Consider the elements of $\tilde{E}$ restricted on the compact set $\hat{K} \subset E_2$ in Corollary 1. The Banach space of all such elements is isometrically isomorphic to $L^1(I, C(\hat{K}))^*$. With a measurable function $f : \mathbb{R} \to E_2$ and an interval $[a, b] \subset \mathbb{R}$, we associate the probability measure

$$D(f(\cdot), [a, b], \omega) = \frac{1}{b - a} \mathrm{meas}\{t \in [a, b] \mid f(t) \in \omega\},$$

where $\omega \subset E_2$ is a Borel set and meas is the Lebesgue measure. Consider the differential inclusion

$$\dot{y}(\tau) \in S(y(\tau)), \quad y(t) = y^\tau,\ \tau \in [t, \infty). \tag{6}$$

Denote by $P(E)$ the set of all probability measures on E. For $A \subset E_2$ define $A^\eta = \{x \in E_2 \mid \mathrm{dist}(x, A) \le \eta\}$. If ν_1, ν_2 are probability measures in the Borel σ algebra of E_2, we define the Prochorov's distance as

$$\rho_P(\nu_1, \nu_2) = \inf\{\eta > 0 \mid\ \nu_1(A) \le \nu_2(A^\eta) + \eta;\ \nu_2(A) \le \nu_1(A^\eta) + \eta \text{ for every Borel set } A \subset E_2\}.$$

DEFINITION 1. *A probability measure* ν *is a limit invariant measure of* (6) *if there exists a solution* $y(\cdot)$ *defined on* $[t, \infty)$ *such that* $D(y(\cdot), [t, T_k])$ *converges to* ν *in* $P(E_2)$ *for some sequence* $T_k \to \infty$ *with respect to Prochorov's distance. If* γ *is in the closed convex hull of the limit measures, then it is called an invariant measure of* (6).

Remark 3. The above definition is an equivalent form (cf. [2]) of Definition 3.1 of [1].

The following lemma is a consequence of Theorem 1.

LEMMA 6. *Under the conditions of Theorem* 1, *there exists a limit measure* ν *of*

$$\dot{y}(\tau) \in \overline{\text{co}}\, G(x, y(\tau), U), \quad y(0) = y^0 \in \hat{K},\ x \in \mathcal{K}, \tag{7}$$

where $\mathcal{K}$ *is as defined in the proof of Corollary* 1.

The following theorem extends Theorem 4 of [10] and Theorem 7.4 of [1] to infinite dimensions.

THEOREM 2. *Under the conditions of Theorem* 1, *the solution set* $Z(\varepsilon)$ *of* (1) *has a limit denoted by* $Z(0)$ *with respect to* $C(I, E_1) \times [L^1(I, C(\hat{K}))]^*$*-weak* topology. Furthermore, for every* $(x(\cdot), \nu(\cdot)) \in Z(0)$ *one has* $\dot{x} \in V(x)$, $x(0) = x_0$, *and* $\nu(t)$ *is an invariant measure of* (7) *(with* 0 *replaced by* t *and* x *by* $x(t)$*) almost everywhere in* I.

Proof. Let $\varepsilon \to 0$, and let $\{(x_\varepsilon, y_\varepsilon)\}_{\varepsilon>0}$ be a net of solutions of (1). Due to Corollary 1 there exists a compact K with $(x_\varepsilon(t), y_\varepsilon(t)) \in K$. The net $\{x_\varepsilon\}_{\varepsilon>0}$ is bounded and equicontinuous. Furthermore, $\{y_\varepsilon\}_{\varepsilon>0}$ is $[L^1(I, C(\hat{K}))]^*$ bounded; i.e., there exists a point of density (x_0, ν_0) of this sequence. Denote by Z_0 the set of all such points of density for all sequences $\{(x_\varepsilon, y_\varepsilon)\}_{\varepsilon>0}$ of solutions of (1). We will prove that $Z_0 = Z(0)$.

Let $(x, \nu) \in Z_0$; i.e., there exist a sequence $\{\varepsilon_i\}_{i=1}^\infty$ and a corresponding sequence $\{(x_i, y_i)\}_{i=1}^\infty$ of solutions of (1) converging to (x, ν) in the considered topology. It remains to show the existence of $(x_\varepsilon, y_\varepsilon) \in Z(\varepsilon)$ such that $x_\varepsilon \to x$ in $C(I, E_1)$ and $y_\varepsilon \to \nu$ in $[L^1(I, C(K, E_2))]^*$-weak*. To this end consider the following systems (denoted by (SS) for convenience):

$$\begin{aligned}
&\dot{x}(t) \in F(x, y, u),\ x(0) = x^0,\\
&\dot{w}_1(t) \in f_1(t, y(t)),\ w_1(0) = 0,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
&\dot{w}_k(t) \in f_k(t, y(t)),\ w_k(0) = 0,\\
&\varepsilon\dot{y}(t) \in G(x, y, u),\ y(0) = y^0.
\end{aligned}$$

Here $\{f_i\}_{i=1}^\infty$ is a dense in $L^1(I, C(\hat{K}))$ sequence of functions that are almost continuous in t and Lipschitz in y. It is not difficult to see that Theorem 1 is valid also for systems of the form given above. Therefore, the slow part $(X(\varepsilon), W^k(\varepsilon))$ of the solution set of (SS) has a limit in $C(I, E_1) \times C(I, \mathbb{R}^k)$ topology. Let

$$\big(\dot{z}, \dot{w}^k\big) \in V^k(z, w^k), \quad z(0) = x_0, \quad W^k(0) = 0 \text{ for } w^k = (w_1, \ldots, w_k)$$

be corresponding to (SS) differential inclusion (1). Fix k, and let $\delta > 0$ be given. By Theorem 1, there exists $\varepsilon_k(\delta)$ such that

$$D_H\big((X(\varepsilon), W^k(\varepsilon)), (X(0), W^k(0))\big) < \delta \quad \text{for } 0 < \varepsilon \le \varepsilon_k(\delta).$$

Choose a sequence $\{\delta_m\}_{m=1}^\infty$ with $\delta_m > \delta_{m+1}$ and $\lim_{m\to\infty} \delta_m = 0$. Then there exists $\varepsilon_m \to 0$ with

$$D_H\big(\big(X(\varepsilon), W^m(\varepsilon)\big), \big(X(0), W^m(0)\big)\big) < \delta_m \text{ for all positive } \varepsilon < \varepsilon_m.$$

Hence, for every $\varepsilon \in [\varepsilon_m, \varepsilon_{m+1})$ there exists $(x_\varepsilon, y_\varepsilon) \in Z(\varepsilon)$ such that

$$\|x_\varepsilon(t) - x(t)\| \le \delta_m \qquad \text{and} \qquad \left\| f_j(\cdot, y(\cdot)) - \int_K f_j(\cdot)\nu(\cdot)(\,dy) \right\|_{L_1} < \delta_m \tag{8}$$

for $j = 1, 2, \dots, m$. Let $\tilde{Z}(\varepsilon)$ be the set of all $(x_\varepsilon, y_\varepsilon) \in Z(\varepsilon)$ satisfying (8) for $\varepsilon \in [\varepsilon_m, \varepsilon_{m+1})$. Then $\tilde{Z}(\varepsilon)$ is nonempty and $C(I, E_1 \times R^m)$ is compact for every $\varepsilon > 0$. For $(x_\varepsilon, y_\varepsilon) \in \tilde{Z}(\varepsilon)$, we have $(x_\varepsilon, y_\varepsilon) \to (x, \nu)$ with respect to the $C(I, E_1)$-strong $\times$ $[L^1(I, C(\hat{K}))]^*$-weak* topology.

In what follows, we follow the proofs of Theorem 7.4 and Corollary 8.3 of [1] to obtain that $\nu(\cdot)$ is an invariant measure. Denote by $\mathcal{M}(x(t))$ the set of all invariant measures of (7) with 0 replaced by t and x by $x(t)$. The last set is nonempty thanks to Lemma 6. From Lemma 5.4 of [1] we know that $\mathcal{M}(\cdot)$ has closed graph. Furthermore, $\mathcal{M}(x(\cdot))$ is measurable (see the proof of Lemma 8.2 of [1]). Let $\tilde{y}_{\varepsilon_j}(\cdot))$ converge to $\nu(\cdot)$ on $0 \le t < t + \Delta \le 1$. From Egorov's theorem we know that for every $\delta > 0$ there exists an open set $E_\delta \subset I$ with $\text{meas}(E_\delta) < \delta$ such that $\Delta(\varepsilon_j) \to 0$ as $j \to \infty$ and $D(\tilde{y}_j(\cdot), [t, t + \Delta_j]) \to \mathcal{M}(x(t))$ uniformly on $t \in I \setminus E_\delta$. Then $\nu(t) \in \mathcal{M}(x(t))$ for a.e. $t \in I$. □

5. Lipschitz continuous solutions. In this section we consider the differential inclusion

$$\begin{pmatrix} \dot{x}(t) \\ \varepsilon \dot{y}(t) \end{pmatrix} \in H(x(t), y(t)), \quad x(0) = x^0, y(0) = y^0, \tag{9}$$

where H is a multivalued mapping from $E := E_1 \times E_2$ into E. Clearly, (9) is more general than (1) in the context of the analysis of this paper. However, the convergence result we are able to obtain below is somewhat weaker than Theorem 1.

For $L > 0$ denote by $Z_L(\varepsilon)$ the set of all $(x_\varepsilon(\cdot), y_\varepsilon(\cdot)) \in Z(\varepsilon)$ that are Lipschitz continuous with a constant L on I for x and on $[\sqrt{\varepsilon}, 1]$ for y. Denote by $H_1(x, y)$ and $H_2(x, y)$ the projections of $H(x, y)$ on E_1 and on E_2, respectively. The following theorem is a generalization of the main result in [11] to infinite-dimensional spaces.

THEOREM 3. *Let the conditions of Theorem* 1 *hold with F and G replaced by H_1 and H_2, respectively. Then there exists a constant $L > 0$ such that $Z_L(\varepsilon) \neq \emptyset$ for every $\varepsilon > 0$, and if $\delta > 0$ is fixed, then $Z_L(\cdot)$ is USC at $\varepsilon = 0$ with respect to $C(I, E_1) \times C([\delta, 1], E_2)$.*

Proof. Fix $\varepsilon > 0$, and let n be a natural number. We set $h = 1/n$. For $k = 1, 2, \dots, n-1$ and a.e. $t \in [kh, (k+1)h]$, we construct successively a solution $(x_\varepsilon^h(\cdot), y_\varepsilon^h(\cdot))$ of (9) such that

$$\varepsilon \frac{d}{dt} \|y_\varepsilon^h(t) - y_\varepsilon^h(t-h)\| \le -\mu \|y_\varepsilon^h(t) - y_\varepsilon^h(t-h)\| + C \|x_\varepsilon^h(t) - x_\varepsilon^h(t-h)\|.$$

If $(x_\varepsilon^h, y_\varepsilon^h)$ is already defined on $[0, kh]$ for $(x_\varepsilon(\cdot), y_\varepsilon(\cdot)) \in Z(\varepsilon)$ and $t \in [kh, (k+1)h]$, then we define

$$\Gamma_\varepsilon^{kh}(t, x, y) = \Big\{ (p, q) \in H(x, y) \mid \varepsilon \Big\langle J(y_\varepsilon^h(t-h) - y), \dot{y}_\varepsilon^h(t-h) - \frac{q}{\varepsilon} \Big\rangle \le -\mu \|y_\varepsilon^h(t-h) - y\|^2 + C \|x_\varepsilon^h(t) - x\|^2 \Big\}.$$

The mapping Γ_ε^{kh} is almost USC with nonempty, convex, and compact values. Therefore, for $t \in [kh, (k+1)h]$ the system

$$\begin{pmatrix} \dot{x}(t) \\ \varepsilon \dot{y}(t) \end{pmatrix} \in \Gamma_\varepsilon^{kh}(t, x(t), y(t)), \quad x(kh) = x_\varepsilon(kh),\ y(kh) = y_\varepsilon(kh)$$

has a solution $(x_\varepsilon(\cdot), y_\varepsilon(\cdot))$. From Lemma 1, $\|x_\varepsilon^h(t+h) - x_\varepsilon^h(t)\| \le Nh$. Furthermore,

$$\varepsilon\big\langle J(y_\varepsilon^h(t) - y_\varepsilon^h(t-h)), \dot{y}_\varepsilon^h(t) - \dot{y}_\varepsilon^h(t-h)\big\rangle \le -\mu\|y_\varepsilon^h(t) - y_\varepsilon^h(t-h)\|^2 + CN^2h^2.$$

Thus

$$\|y_\varepsilon^h(t)-y_\varepsilon^h(t-h)\|^2 \le 2\exp\left(-2\mu\frac{t-h}{\varepsilon}\right)\left[\|y_\varepsilon^h(h)-y_\varepsilon^h(0)\|^2+\int_h^t \exp\left(2\mu\frac{s-h}{\varepsilon}\right)\frac{CN^2h^2}{\varepsilon}ds\right]$$
$$\le C_2h^2\left(1+\frac{1}{\varepsilon}\exp\left(-2\mu\frac{t-h}{\varepsilon}\right)\right)$$

for a constant $C_2 > 0$. Let $h \to 0$. From Theorem 1 of [5] we know that $Z(\varepsilon)$ is compact in the space $C(I, E)$. Hence there exists $(x_\varepsilon(t), y_\varepsilon(t)) = \lim_{h\to 0}(x_\varepsilon^h(t), y_\varepsilon^h(t))$. Also,

$$\|y_\varepsilon(t+\tau) - y_\varepsilon(t)\| \le C_2\left(1 + \frac{1}{\varepsilon}\exp\left(-2\mu\frac{t}{\varepsilon}\right)\right)\tau.$$

Then there exists a constant N_1 such that $\|y_\varepsilon(t+\tau)-y_\varepsilon(t)\| \le N_1\tau$ for $t \in [\sqrt{\varepsilon}, 1-\tau]$. Indeed,

$$\lim_{\varepsilon\to 0}\max_{t\in[\sqrt{\varepsilon},1]}\frac{1}{\varepsilon}\exp\left(-2m\frac{t}{\varepsilon}\right) = 0.$$

Denote $L = \max(N, N_1)$. Let $\{x_\varepsilon(\cdot), y_\varepsilon(\cdot)\}_{\varepsilon>0} \subset Z_L(\varepsilon)$ be the net just defined. It follows from Theorem 1 that there exists a subnet, say, $x_\varepsilon(\cdot)$, converging uniformly on I to $x_0(\cdot)$. Moreover, since $z_\varepsilon(t) = (x_\varepsilon(t), y_\varepsilon(t))$ is in a compact set K (Corollary 2), we obtain that there exists a subnet, say, $(x_\varepsilon(\cdot), y_\varepsilon(\cdot))$, such that $x_\varepsilon(t) \to x_0(t)$ uniformly on I and $y_\varepsilon(t) \to y_0(t)$ uniformly on $[\delta, 1]$ for every $\delta > 0$. Clearly, the limit satisfies

$$\begin{pmatrix}\dot{x}_0(t)\\ 0\end{pmatrix} \in H(x_0(t), y_0(t)) \qquad \text{for a.e. } t \in I, \qquad x_0(0) = x^0,$$

where $y_0(\cdot)$ is Lipschitz continuous with a constant L on the interval I. □

Note that, without the restriction to equi-Lipschitz trajectories, the trajectory map $Z(\cdot)$ may be not USC at $\varepsilon = 0$, even in finite dimensions.

REFERENCES

[1] Z. Artstein, *Invariant measures of differential inclusions applied to singular perturbations*, J. Differential Equations, 152 (1999), pp. 289–307.

[2] Z. Artstein, *Invariant measures of set-valued maps*, J. Math. Anal. Appl., 252 (2000), pp. 696–706.

[3] K. Deimling, *Multivalued Differential Equations*, de Gruyter Ser. Nonlinear Anal. Appl. 1, Walter de Gruyter & Co., Berlin, 1992.

[4] T. Donchev, *Functional differential inclusions with monotone right hand side*, Nonlinear Anal., 16 (1991), pp. 543–552.

[5] T. Donchev, *Semicontinuous differential inclusions*, Rend. Sem. Mat. Univ. Padova, 101 (1999), pp. 147–160.

[6] T. Donchev and E. Farkhi, *Stability and Euler approximation of one-sided Lipschitz differential inclusions*, SIAM J. Control Optim., 36 (1998), pp. 780–796.

[7] T. Donchev and E. Farkhi, *Approximation of one-sided Lipschitz differential inclusions with discontinuous right-hand sides*, in Calculus of Variations and Differential Equations, Chapman & Hall/CRC Res. Notes Math. 410, A. Ioffe, S. Reich, and I. Shafrir, eds., Chapman & Hall/CRC, Boca Raton, FL, 1999, pp. 101–118.

[8] T. Donchev and R. Ivanov, *On the existence of solutions of differential inclusions in uniformly convex Banach spaces*, Math. Balkanica, 6 (1992), pp. 13–24.
[9] T. Donchev and I. Slavov, *Singularly perturbed functional differential inclusions*, Set-Valued Anal., 3 (1995), pp. 113–128.
[10] T. Donchev and I. Slavov, *Averaging method for one-sided Lipschitz differential inclusions with generalized solutions*, SIAM J. Control Optim., 37 (1999), pp. 1600–1613.
[11] A. Dontchev, T. Donchev, and I. Slavov, *A Tikhonov-type theorem for singularly perturbed differential inclusions*, Nonlinear Anal., 26 (1996), pp. 1547–1554.
[12] A. Dontchev and V. Veliov, *Singular perturbation in Mayer's problem for linear systems*, SIAM J. Control Optim., 21 (1983), pp. 566–581.
[13] O. Filatov and M. Hapaev, *Averaging of differential inclusions with fast and slow variables*, Math. Notes, 47 (1990), pp. 102–109.
[14] O. Filatov and M. Hapaev, *Averaging of Systems of Differential Inclusions*, Moskow University Press, Moskow, 1998 (in Russian).
[15] V. Gaitsgory, *Use of averaging method in control problems*, Differ. Equ., 22 (1986), pp. 1290–1299.
[16] V. Gaitsgory, *Suboptimization of singularly perturbed control systems*, SIAM J. Control Optim., 30 (1992), pp. 1228–1249.
[17] G. Grammel, *Averaging of singularly perturbed systems*, Nonlinear Anal., 28 (1997), pp. 1851–1865.
[18] G. Grammel, *Singularly perturbed differential inclusions: An averaging approach*, Set-Valued Anal., 4 (1996), pp. 361–374.
[19] F. Hoppensteadt, *Stability in systems with parameter*, J. Math. Anal. Appl., 18 (1967), pp. 129–134.
[20] F. Hoppensteadt, *Properties of solutions of ordinary differential equations with small parameters*, Comm. Pure Appl. Math., 24 (1971), pp. 807–840.
[21] S. Hu and N. Papageorgiou, *Handbook of Multivalued Analysis*, *Vol.* II, Kluwer, Dodrecht, The Netherlands, 2000.
[22] V. Lakshmikantham and S. Leela, *Nonlinear Differential Equations in Abstract Spaces*, Pergamon, Oxford, UK, 1981.
[23] F. Lempio and V. Veliov, *Discrete approximations of differential inclusions*, Bayreuth. Math. Schr., 54 (1998), pp. 149–232.
[24] N. Levinson, *Perturbations of discontinuous solutions of nonlinear systems of differential equations*, Proc. Natl. Acad. Sci. USA, 33 (1947), pp. 214–218.
[25] A. Tikhonov, *On the dependence of the solutions of differential equations on a small parameter*, Mat. Sb. (N.S.), 22 (1948), pp. 193–204.
[26] A. Tikhonov, *On systems of differential equations containing parameters*, Mat. Sb. (N.S.), 27 (1950), pp. 147–156.
[27] A. Tikhonov, *Systems of differential equations containing small parameters in the derivatives*, Mat. Sb. (N.S.), 31 (1952), pp. 575–586.
[28] V. Veliov, *Differential inclusions with stable subinclusions*, Nonlinear Anal., 23 (1994), pp. 1027–1038.
[29] V. Veliov, *A generalization of Tikhonov theorem for singularly perturbed differential inclusions*, J. Dynam. Control Systems, 3 (1997), pp. 1–28.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1813–1833

A VARIATIONAL APPROACH TO NONLINEAR ESTIMATION*

SANJOY K. MITTER† AND NIGEL J. NEWTON‡

Abstract. We consider estimation problems, in which the estimand, X, and observation, Y, take values in measurable spaces. Regular conditional versions of the forward and inverse Bayes formula are shown to have dual variational characterizations involving the minimization of *apparent information* and the maximization of *compatible information*. These both have natural information-theoretic interpretations, according to which Bayes' formula and its inverse are optimal information processors. The variational characterization of the forward formula has the same form as that of Gibbs measures in statistical mechanics. The special case in which X and Y are diffusion processes governed by stochastic differential equations is examined in detail. The minimization of apparent information can then be formulated as a stochastic optimal control problem, with cost that is quadratic in both the control and observation fit. The dual problem can be formulated in terms of infinite-dimensional deterministic optimal control. Local versions of the variational characterizations are developed which quantify information *flow* in the estimators. In this context, the information conserving property of Bayesian estimators coincides with the Davis–Varaiya martingale stochastic dynamic programming principle.

Key words. Bayesian inference, information theory, Legendre-type transforms, nonlinear filtering, stochastic optimal control

AMS subject classifications. 93E11, 93E20, 94A15, 62F15, 60E10, 60G35

DOI. 10.1137/S0363012901393894

1. Introduction.

This article investigates a variational formulation of Bayesian estimation with a natural information-theoretic interpretation. The two "directions" of an abstract Bayes formula (likelihood function to posterior distribution and vice-versa) are given variational representations. The forward representation involves the minimization of *apparent information* of probability measures on the space of the estimand. This apparent information is made up of two parts: the information gain of the measure over the prior distribution for the estimand and a *residual* term representing the information value of the observation, complementary to this. The apparent information of probability measures is greater than or equal to the total information in the observation, with equality if and only if the measure is the posterior distribution of the estimand. Thus the (forward) Bayes formula can be thought of as an optimal "information processor" in that it balances input and output information. Suboptimal processors appear to have access to more information than there is in the observation. The variational representation of the inverse Bayes formula involves the maximization

*Received by the editors August 20, 2001; accepted for publication (in revised form) May 31, 2003; published electronically December 17, 2003. This work was started while the second author was on a period of Study Leave visiting LIDS, supported by EPSRC grant GR/M97039. It continued during his visits to the INRIA research groups Omega and Sigma2. Further support was provided by the Army Research Office under the MURI grant: Data Fusion in Large Arrays of Microsensors DAAD19-00-1-0466; under Center for Imaging Science subcontract, grant DAAD19-99-1-0012 (Johns Hopkins University; the MURI grant: Vision Strategies and ATR Performance subcontract 654-21256 (Brown University); and by an INTEL grant. This publication is also an output from a research project funded by the Cambridge–MIT Institute (CMI). CMI is funded in part by the U.K. Government. The research was carried out for CMI by Massachusetts Institute of Technology. CMI accepts no responsibility for any information provided or views expressed.

http://www.siam.org/journals/sicon/42-5/39389.html

†Department of Electrical Engineering and Computer Science, and Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 (mitter@mit.edu).

‡Department of Electronic Systems Engineering, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK (njn@essex.ac.uk).

of *compatible information* of likelihood functions on the space of the estimand. This is defined to be the difference between the information in an unspecified observation associated with the likelihood function and that part of this information complementary to the (given) posterior distribution. The compatible information of likelihood functions is less than or equal to the information gain of the posterior distribution over the prior, with equality if and only if the likelihood function is equivalent to that provided by the inverse Bayes formula. Once again, the inverse Bayes formula can be thought of as an optimal processor, balancing input and output information. However, in this case, rather than appearing to have an additional source of information, suboptimal processors lose (or fail to make use of) part of the input information.

In section 2, the estimand, X, and the observation, Y, of the Bayesian problem are supposed to take values in Borel spaces $(\mathbf{X}, \mathcal{X})$ and $(\mathbf{Y}, \mathcal{Y})$, respectively. The starting point is a "regular conditional" version of the Bayes formula. In section 3, the results are specialized to the estimation of diffusion processes with partial observations. In that context, the regular conditional probability distribution can be chosen to be continuous in the observations. It also has the key property of being Markovian. This means that the family of measures over which apparent information is minimized can be restricted to the distributions of the process X when a "finite energy" feedback control is applied through the drift coefficient. Thus, in this case, the minimization of apparent information can be interpreted in terms of a problem in stochastic optimal control. This is explored in section 4.

The dual variational problem for diffusion processes is developed in section 5. One interpretation of it is as a problem in infinite-dimensional deterministic optimal control. The optimal trajectory of the dual problem is a "likelihood filter" for the process X in reversed time, from which the corresponding nonlinear filter can be found. This gives a new interpretation to a connection between an optimal control problem in one time direction and a nonlinear filter in the other which was made for nondegenerate diffusions in [6] via the Hopf transformation and used to give existence and uniqueness results for the unnormalized conditional density equation with unbounded observations. The results of sections 3–5 are established under fairly weak conditions. In particular, they include the case of degenerate diffusions.

In the context of estimators for diffusion processes, there is a "local" version of the variational formulations which characterizes flow rates of information and shows that Bayesian processors are conservative in the sense that they balance input and output flow rates. This is the subject of section 6.

A variational representation of the Fokker–Planck equation for diffusion processes is discussed in [10]. This involves the minimization of the "energy" of drift coefficients over those that give rise to a particular set of marginal densities. There, as here, the modification of the drift coefficient can be interpreted as the application of a *control* term, which re-expresses the variational problem as one in optimal control. The two problems are somewhat different though. In particular, the controls admitted in [10] give rise to mutually singular transition probabilities, which are certainly not permitted in the present context.

A preliminary account of some of the results herein was reported in [11].

2. A variational formulation of Bayesian estimation. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $(\mathbf{X}, \mathcal{X})$ and $(\mathbf{Y}, \mathcal{Y})$ Borel spaces, and $X : \Omega \to \mathbf{X}$ and $Y : \Omega \to \mathbf{Y}$ measurable mappings with distributions P_X, P_Y, and P_{XY} on $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{X} \times \mathcal{Y}$,

respectively. Suppose that

(H1) there exists a σ-finite (reference) measure, λ_Y, on $\mathcal{Y}$ such that $P_{XY} \ll P_X \otimes \lambda_Y$. (This could be P_Y itself.)

Let $Q : \mathbf{X} \times \mathbf{Y} \to [0, \infty)$ be a version of the associated Radon–Nikodym derivative, and

$$\bar{\mathbf{Y}} = \left\{ y \in \mathbf{Y} : 0 < \int_{\mathbf{X}} Q(x, y) P_X(dx) < \infty \right\}; \tag{2.1}$$

then $\bar{\mathbf{Y}} \in \mathcal{Y}$ and $P_Y(\bar{\mathbf{Y}}) = 1$. Let $H : \mathbf{X} \times \mathbf{Y} \to (-\infty, +\infty]$ be defined by

$$\begin{aligned} H(x, y) &= -\log(Q(x, y)) \qquad \text{if } y \in \bar{\mathbf{Y}}, \\ &\quad 0 \qquad \text{otherwise;} \end{aligned} \tag{2.2}$$

then $P_{X|Y} : \mathcal{X} \times \mathbf{Y} \to [0, 1]$, defined by

$$P_{X|Y}(A, y) = \frac{\int_A \exp(-H(x, y)) P_X(dx)}{\int_{\mathbf{X}} \exp(-H(x, y)) P_X(dx)}, \tag{2.3}$$

is a *regular conditional probability distribution* for X given Y; i.e.,

$P_{X|Y}(\,\cdot\,, y)$ is a probability measure on $\mathcal{X}$ for each y,

$P_{X|Y}(A, \,\cdot\,)$ is $\mathcal{Y}$-measurable for each A, and

$P_{X|Y}(A, Y) = P(X \in A \,|\, Y)$ a.s.

Equations (2.1)–(2.3) constitute an "outcome-by-outcome" abstract Bayes formula, yielding a posterior probability distribution for X for each outcome of Y. Of course, for any y belonging to a set of P_Y-measure zero, $P_{X|Y}(\,\cdot\,, y)$ depends on the choice of version of the Radon–Nikodym derivative Q. However, in particular examples, we can often find a version such that $P_{X|Y}(A, \,\cdot\,)$ is continuous for each $A \in \mathcal{X}$.

Let $\mathcal{P}(\mathcal{X})$ be the set of probability measures on $(\mathbf{X}, \mathcal{X})$ and $\mathcal{H}(\mathbf{X})$ the set of $(-\infty, +\infty]$-valued, measurable functions on the same space. For $\tilde{P}_X, \hat{P}_X \in \mathcal{P}(\mathcal{X})$, and $\tilde{H} \in \mathcal{H}(\mathbf{X})$, we define

$$\begin{aligned} h(\tilde{P}_X \,|\, \hat{P}_X) &= \int_{\mathbf{X}} \log\left(\frac{d\tilde{P}_X}{d\hat{P}_X}\right) d\tilde{P}_X \quad \text{if } \tilde{P}_X \ll \hat{P}_X \text{ and the integral exists,} \\ &\quad +\infty \qquad \text{otherwise;} \end{aligned} \tag{2.4}$$

$$\begin{aligned} i(\tilde{H}) &= -\log\left(\int_{\mathbf{X}} \exp(-\tilde{H}) dP_X\right) \quad \text{if } 0 < \int_{\mathbf{X}} \exp(-\tilde{H}) dP_X < \infty, \\ &\quad -\infty \qquad \text{otherwise;} \end{aligned} \tag{2.5}$$

$$\begin{aligned} \langle \tilde{H}, \tilde{P}_X \rangle &= \int_{\mathbf{X}} \tilde{H} d\tilde{P}_X \qquad \text{if the integral exists,} \\ &\quad +\infty \qquad \text{otherwise.} \end{aligned} \tag{2.6}$$

It is well known that the relative entropy $h(\tilde{P}_X \,|\, \hat{P}_X)$ can be interpreted as the *information gain* of the probability measure $\tilde{P}_X$ over $\hat{P}_X$. In fact, any version of $-\log(d\tilde{P}_X / d\hat{P}_X)$ is a generalization of the Shannon information for X. For almost all x, it is a measure of the "relative degree of surprise" in the outcome $X = x$ for the two distributions $\tilde{P}_X$ and $\hat{P}_X$. Thus $h(\tilde{P}_X \,|\, \hat{P}_X)$ is the average *reduction* in the degree of surprise in this outcome arising from the acceptance of $\tilde{P}_X$ as the distribution for X, rather than $\hat{P}_X$.

If we interpret $\exp(-\tilde{H})$ as a likelihood function for X, associated with some (unspecified) observation, then $\tilde{H}(x)$ is the "residual degree of surprise" in that observation if we already know that $X = x$, and $i(\tilde{H})$ is the "total degree of surprise" in that observation, i.e., the information in the unspecified observation, if all we know about X is its prior P_X. In what follows we shall call $\tilde{H}(X)$ the *X-conditional information* in the unspecified observation and $i(\tilde{H})$ the information in that observation. (Of course, $H(X, y)$ and, respectively, $i(H(\,\cdot\,, y))$ are the X-conditional information and the information in the observation that $Y = y$.)

PROPOSITION 2.1. *Suppose that* (H1) *is satisfied, and H and $P_{X|Y}$ are as defined above. Then for any y such that*

$$-\int_{\mathbf{X}} H(x,y)\exp(-H(x,y))P_X(dx) < \infty, \quad (\textit{where } +\infty\exp(-\infty) = 0), \tag{2.7}$$

$$\text{(i)} \qquad i(H(\,\cdot\,,y)) = \min_{\tilde{P}_X \in \mathcal{P}(\mathcal{X})} \left\{ h(\tilde{P}_X \,|\, P_X) + \langle H(\,\cdot\,,y), \tilde{P}_X \rangle \right\}; \tag{2.8}$$

$$\text{(ii)} \qquad h(P_{X|Y}(\,\cdot\,,y) \,|\, P_X) = \max_{\tilde{H} \in \mathcal{H}(\mathbf{X})} \left\{ i(\tilde{H}) - \langle \tilde{H}, P_{X|Y}(\,\cdot\,,y) \rangle \right\}; \tag{2.9}$$

(iii) *$P_{X|Y}(\,\cdot\,, y)$ is the unique minimizer in* (2.8)*;*
(iv) *if H^* is a maximizer in* (2.9)*, then there exists a real constant K such that*

$$H^*(X) = H(X,y) + K \quad a.s.$$

Proof. If $y \in \bar{\mathbf{Y}}$ and (2.7) holds, then $h(P_{X|Y}(\,\cdot\,,y) \,|\, P_X) < \infty$, $i(H(\,\cdot\,,y)) > -\infty$, and $H(\,\cdot\,,y) \in L_1(P_{X|Y}(\,\cdot\,,y))$. This is also true if $y \notin \bar{\mathbf{Y}}$ since, in that case, $H(\,\cdot\,,y) = 0$ and $P_{X|Y}(\,\cdot\,,y) = P_X$. Thus it is clear that the minimum in (2.8) is less than $+\infty$, and the maximum in (2.9) is greater than $-\infty$.

Suppose that, for $\tilde{P}_X \in \mathcal{P}(\mathcal{X})$, $h(\tilde{P}_X \,|\, P_X) < \infty$ and $H(\,\cdot\,,y) \in L_1(\tilde{P}_X)$. It readily follows that $\tilde{P}_X \ll P_{X|Y}(\,\cdot\,,y)$, so that

$$h(\tilde{P}_X \,|\, P_X) = \int_{\mathbf{X}} \left(\log\left(\frac{d\tilde{P}_X}{dP_{X|Y}}(x,y) \right) + \log\left(\frac{dP_{X|Y}}{dP_X}(x,y) \right) \right) \tilde{P}_X(dx),$$

and

$$h(\tilde{P}_X \,|\, P_X) + \langle H(\,\cdot\,,y), \tilde{P}_X \rangle = i(H(\,\cdot\,,y)) + h(\tilde{P}_X \,|\, P_{X|Y}(\,\cdot\,,y)). \tag{2.10}$$

It is easy to show that, for any $\tilde{P}_X \in \mathcal{P}(\mathcal{X})$, the relative entropy functional $h(\,\cdot\,|\,\tilde{P}_X)$ is nonnegative, evaluates to zero at $\tilde{P}_X$, and is strictly convex on the subset of $\mathcal{P}(\mathcal{X})$ for which it is finite. This establishes parts (i) and (iii).

Suppose now that, for $\tilde{H} \in \mathcal{H}(\mathbf{X})$, $i(\tilde{H}) > -\infty$ and $\tilde{H} \in L_1(P_{X|Y}(\,\cdot\,,y))$. Let $\tilde{P}_X$ be defined by (2.3) with $\tilde{H}$ replacing $H(\,\cdot\,,y)$. It readily follows that $P_{X|Y}(\,\cdot\,,y) \ll \tilde{P}_X$, and so

$$\begin{aligned} i(\tilde{H}) - \tilde{H}(X) &= \log\left(\frac{d\tilde{P}_X}{dP_X}(X) \right) \\ &= \log\left(\frac{dP_{X|Y}}{dP_X}(X,y) \right) - \log\left(\frac{dP_{X|Y}}{d\tilde{P}_X}(X,y) \right). \end{aligned}$$

Thus

$$i(\tilde{H}) - \langle \tilde{H}, P_{X|Y}(\,\cdot\,, y)\rangle = h(P_{X|Y}(\,\cdot\,, y) \,|\, P_X) - h(P_{X|Y}(\,\cdot\,, y) \,|\, \tilde{P}_X). \tag{2.11}$$

Suppose that there is a set $A \in \mathcal{X}$ for which $P_{X|Y}(A, y) = 0$ but $\tilde{P}_X(A) > 0$. Let $\tilde{P}'_X$ be defined by

$$\tilde{P}'_X(B) = \left(\tilde{P}_X(A^C)\right)^{-1} \tilde{P}_X(A^C \cap B) \qquad \text{for all } B \in \mathcal{X}.$$

Then $h(P_{X|Y}(\,\cdot\,, y) \,|\, \tilde{P}'_X) < h(P_{X|Y}(\,\cdot\,, y) \,|\, \tilde{P}_X)$, and so any maximizer in (2.11) must be absolutely continuous with respect to $P_{X|Y}(\,\cdot\,, y)$. It is easy to show that, for any $\tilde{P}_X \in \mathcal{P}(\mathcal{X})$, the relative entropy functional $h(\tilde{P}_X \mid \cdot\,)$ is nonnegative, evaluates to zero at $\tilde{P}_X$, and is strictly convex on the subset of $\mathcal{P}(\mathcal{X})$ consisting of measures that are absolutely continuous with respect to $\tilde{P}_X$. This establishes parts (ii) and (iv). ∎

Remark 1. If the mutual information between X and Y is finite,

$$\int_{\mathbf{X}\times\mathbf{Y}} \log\left(\frac{dP_{XY}}{d(P_X \otimes P_Y)}\right) dP_{XY} < \infty, \tag{2.12}$$

then there exists a version of Q for which (2.7) is satisfied for all y.

Remark 2. Proposition 2.1 is a special case of an energy-entropy duality that plays a major role in statistical physics and in the theory of large deviations. More general results of this nature are widely available in the literature. (See, for example, [5].) Our aim in this section is to provide an information-theoretic interpretation of the result in the Bayesian context. The simple proof we provide here makes use of the special nature of that context.

Parts (i) and (ii) of Proposition 2.1 both concern the processing of information over and above that in the prior P_X. In part (i), the source of additional information is the observation that $Y = y$. The abstract Bayes formula extracts the part of this information pertinent to X, $h(P_{X|Y}(\,\cdot\,, y) \,|\, P_X)$, and leaves the residual information, $\langle H(\,\cdot\,, y), P_{X|Y}(\,\cdot\,, y)\rangle$. One can think of the input information as being held in the likelihood function, $\exp(-H(\,\cdot\,, y))$, and the extracted information as being held in the distribution, $P_{X|Y}(\,\cdot\,, y)$. An arbitrary estimation procedure that postulates $\tilde{P}_X$ as a "postobservation" distribution for X appears to have access to additional information, in that it yields an information gain on X of $h(\tilde{P}_X \,|\, P_X)$, and a residual information of $\langle H(\,\cdot\,, y), \tilde{P}_X\rangle$. The sum of these two terms (the term in brackets on the right-hand side of (2.8)) is strictly greater than the actual information available, $i(H(\,\cdot\,, y))$, unless $\tilde{P}_X = P_{X|Y}(\,\cdot\,, y)$. We shall call it the *apparent information* of the estimator $\tilde{P}_X$. (Implicit in the interpretation of $h(\tilde{P}_X \,|\, P_X)$ as an information gain is the assumption that $\tilde{P}_X$ represents a rational belief about X given the prior and some additional knowledge, such as an observation.)

In part (ii), the source of additional information is the posterior distribution, $P_{X|Y}(\,\cdot\,, y)$. The aim now is to postulate an observation (with likelihood function $\exp(-\tilde{H})$) which would give rise to this distribution. The input information here, $h(P_{X|Y}(\,\cdot\,, y) \,|\, P_X)$, is merged with the residual information of the postulated observation, $\langle \tilde{H}, P_{X|Y}(\,\cdot\,, y)\rangle$, and the result is greater than or equal to the total information in the postulated observation, $i(\tilde{H})$, with equality if and only if the observation is compatible with $P_{X|Y}(\,\cdot\,, y)$ in the sense of part (iv) of the proposition. The term in brackets on the right-hand side of (2.9) can be thought of as that part of the information in the postulated observation compatible with $P_{X|Y}(\,\cdot\,, y)$. We shall call it the

compatible information of the likelihood function $\exp(-\tilde{H})$. Another interpretation is that the input information, $h(P_{X|Y}(\cdot\,, y)\,|\,P_X)$, is processed to produce compatible information resulting in a net loss of information except when the processor is optimal.

Throughout the rest of the paper, the apparent information and compatible information will be denoted by $F(\tilde{P}_X, y)$ and $G(\tilde{H}, y)$, i.e.,

$$F(\tilde{P}_X, y) = h(\tilde{P}_X \,|\, P_X) + \langle H(\cdot\,, y), \tilde{P}_X \rangle, \tag{2.13}$$

$$G(\tilde{H}, y) = i(\tilde{H}) - \langle \tilde{H}, P_{X|Y}(\cdot\,, y) \rangle. \tag{2.14}$$

As (2.10) and (2.11) show, the minimization of F is equivalent to the minimization of the *information excess* of the estimator $\tilde{P}_X$, $h(\tilde{P}_X \,|\, P_{X|Y}(\cdot\,, y))$, and the maximization of G is equivalent to the minimization of the *information deficit* of the likelihood function $\exp(-\tilde{H})$, $h(P_{X|Y}(\cdot\,, y)\,|\,\tilde{P}_X)$. In fact (as was pointed out by an anonymous referee), these interpretations still hold in the absence of (2.7). However, in not identifying the source information or the extracted information, they do not show the information processing aspects of Bayesian estimation in quite the same way as the quantities F and G. Moreover, F and G make clear the compromises involved in Bayesian estimation. Part (i) of the proposition shows how $P_{X|Y}(\cdot\,, y)$ compromises between being close to the prior P_X and fitting with the observation $Y = y$, whereas part (ii) shows how $H(\cdot\,, y)$ (or its equivalents) compromise between holding a lot of information but not too much residual information.

Of course it is possible to give other variational characterizations of $P_{X|Y}(\cdot\,, y)$. For example, one could consider it as the minimizer of the total variation norm of the difference measure $\tilde{P}_X - P_{X|Y}(\cdot\,, y)$. However, such characterizations lack the information-theoretic interpretation discussed above: F and G are natural error measures for suboptimal estimation procedures. The characterization (2.8) could be used as a basis for approximations. For example, we may wish to approximate a posterior distribution by a discrete law on a finite partition of $\mathbf{X}$. The size of the partition may be fixed, but we may be able to choose the law and the details of the partition by means of a finite number of parameters. The characterization (2.8) could form the basis of an optimization with respect to this set of parameters. Similarly, the characterization (2.9) could be used as a basis for the study of modeling errors, in that it shows the information loss arising from the use of an incorrect likelihood function. Since the use of an incorrect prior, P_X^e (with $P_X^e \ll P_X$), with a Bayesian procedure is equivalent to the use of the incorrect likelihood function

$$\exp(-H^e(\cdot\,, y)) = \exp(-H(\cdot\,, y)) \frac{dP_X^e}{dP_X},$$

(2.9), with $\tilde{H} = H^e(\cdot\,, y)$, also shows the information loss arising through the use of an incorrect prior. Furthermore, if there were any uncertainty in the likelihood function or the prior, the resulting information loss could be studied by means of game-theoretic methods.

Proposition 2.1 is an instance of a Legendre-type transform between the relative entropy of probability measures and the logarithm of the exponential moment of real-valued random variables. A similar transform occurs in the characterization of Gibbs measures in statistical mechanics [8]. In that context, $(\mathbf{X}, \mathcal{X})$ is the *configuration space* of a physical system (the Cartesian product of a number, N, of identical spaces), H is a *Hamiltonian* representing the energies of the configurations, and F is the *free energy*

of the probability measure $\tilde{P}_X$ with respect to the reference measure, P_X, and H. A Gibbs measure represents a thermodynamic state of the system in thermodynamic equilibrium. If N is finite, then there is only one Gibbs measure, and it takes the form (2.3). Gibbs theory comes into its full richness only when N is infinite, in which case there may be multiple Gibbs measures, and formulae such as (2.3) are no longer appropriate. However, variational characterizations are. We note that the Bayesian estimator can be seen to compromise between being close to the prior and fitting with the observation in exactly the same way that a thermodynamic system in equilibrium compromises between maximizing entropy and minimizing average energy.

3. Path estimators. The techniques of section 2 are specialized here for the case in which the estimand, X, and observation, Y, are, respectively, continuous $\mathbb{R}^n$- and $\mathbb{R}^d$-valued processes governed by the following Itô integral equations:

$$\begin{aligned} &X_t = X_0 + \int_0^t b(X_s, s)\, ds + \int_0^t \sigma(X_s, s)\, dV_s \quad \text{for } 0 \le t \le T, \\ &X_0 \sim \mu, \end{aligned} \tag{3.1}$$

$$Y_t = \int_0^t g(X_s)\, ds + W_t \quad \text{for } 0 \le t \le T, \tag{3.2}$$

where $X_t, V_t \in \mathbb{R}^n$, μ is a law on $(\mathbb{R}^n, \mathcal{B}^n)$, $Y_t, W_t \in \mathbb{R}^d$, and b, σ, and g are measurable mappings. Under suitable regularity conditions, these equations will be unique in law and have a weak solution $(\Omega, \mathcal{F}, (\mathcal{F}_t), P, (V, W), (X, Y))$, i.e., a filtered probability space supporting an $(n + d)$-dimensional Brownian motion (V, W) and an $(n + d)$-dimensional semimartingale (X, Y) such that (3.1) and (3.2) are satisfied for all t. The abstract spaces $(\mathbf{X}, \mathcal{X})$ and $(\mathbf{Y}, \mathcal{Y})$ of section 2 now become the spaces $(C([0, T]; \mathbb{R}^n), \mathcal{B}_T)$ and $(C([0, T]; \mathbb{R}^d), \mathcal{B}_T)$ of continuous functions, topologized by the uniform norm. We continue to use the notation $(\mathbf{X}, \mathcal{X})$ and $(\mathbf{Y}, \mathcal{Y})$, though, for the sake of brevity.

Let λ_Y be Wiener measure on $(\mathbf{Y}, \mathcal{Y})$. Under suitable conditions on μ, b, σ, and g, we might expect (H1) to be satisfied and the mutual information, $\mathbf{E} \log(dP_{XY}/d(P_X \otimes \lambda_Y)(X, Y))$, to be finite. This will allow us to proceed as in section 2 to construct a function H on $X \times Y$, and a corresponding regular conditional probability, $P_{X|Y}$, such that (2.7) holds for all y. Furthermore, if we can show that $P_{X|Y}(\cdot\,, y) \sim P_X$, then we shall be able to construct a continuous strictly positive martingale M_y on Ω such that

$$M_{y,t} = \mathbf{E}\left(\frac{dP_{X|Y}(\cdot\,, y)}{dP_X}(X) \,\middle|\, \mathcal{F}_t^X \right) \quad \text{for } 0 \le t \le T,$$

where $(\mathcal{F}_t^X)$ is the filtration generated by the process X. It will then follow from the Cameron–Martin–Girsanov theory that

$$M_{y,t} = M_{y,0} \exp\left(\int_0^t U'_{y,s}\, (dX_s - b(X_s, s)\, ds) - \frac{1}{2} \int_0^t |\sigma(X_s, s)' U_{y,s}|^2\, ds \right) \tag{3.3}$$

for some progressively measurable $\mathbb{R}^n$-valued process U_y. $P_{X|Y}(\cdot\,, y)$ will then be the distribution of a *controlled* process, X_y, satisfying an equation like (3.1), but with a different initial law and with a control term, $\sigma\sigma'(X_s, s)U_{y,s}$, entering the drift coefficient. The use of the progressively measurable control $\tilde{U}$ instead of U_y will result in a process $\tilde{X}$ having a distribution whose apparent information relative to

$(P_X, H(\cdot, y))$ is greater than or equal to that of X_y. Thus, at least in part, the variational characterization of section 2 will become a problem in stochastic optimal control.

We might also expect $P_{X|Y}(\cdot, y)$ to be Markov (at least for almost all y), in which case it will be appropriate to restrict admissible controls $\tilde{U}$ to *feedback* controls of the form $u(\tilde{X}_t, t)$. It should also then be possible to define regular conditional *transition* probabilities for $P_{X|Y}$. With this in mind, let $(\chi_t, 0 \le t \le T)$ be the coordinate process on $\mathbf{X}$, and

$$\mathcal{X}_s^t = \sigma(\chi_r, s \le r \le t) \quad \text{for } 0 \le s \le t \le T. \tag{3.4}$$

We should be able to construct regular conditional probabilities

$$P_{X|Y}^{s+} : \mathcal{X}_s^T \times \mathbb{R}^n \times C([s,T]; \mathbb{R}^d) \to [0,1]$$

such that, for all $A \in \mathcal{X}_s^T$,

$$P_{X|Y}(A, y) = \int_{R^n} P_{X|Y}^{s+}(A, z, (y_t - y_s, s \le t \le T))\, P_{X|Y}(\chi_s^{-1}(dz), y). \tag{3.5}$$

These will have variational characterizations in terms of the corresponding regular conditional probabilities for the prior, P_X, and appropriately constructed likelihood functions. This will lead toward a *localized* version of the results of section 2.

In what follows, we develop the above ideas in a rigorous manner. We do this by placing constraints on b and σ such that (3.1) has a *strong* solution and then use the techniques of stochastic flows. This has the advantage that we are able to include problems with degenerate diffusion coefficients, which are important in many areas of application. (In fact our approach also applies to some problems not satisfying a hypoellipticity condition.)

The constraints we place on μ, b, σ, and g also fit well with Clark's *robustness* ideas (see [2]). These lead to an explicit function H and corresponding regular conditional probability, $P_{X|Y}$, that is Markov for every y. They also admit unbounded observation functions g, which are needed in the linear case.

We suppose that μ, b, σ, and g satisfy the following technical conditions:

(H2) there exists an $\epsilon > 0$ such that

$$\int_{\mathbb{R}^n} \exp\left(\epsilon |z|^2\right) \mu(dz) < \infty;$$

(H3) σ is bounded, and b and σ are uniformly Lipschitz continuous on compact sets and differentiable with respect to the components of z, the derivatives being continuous and bounded;

(H4) g has continuous first, second, and third derivatives, and there exist $C < \infty$ and $\alpha < \infty$ such that for all $z \in \mathbb{R}^n$

$$\sum_i \left| \frac{\partial g}{\partial z_i}(z) \right| \le C,$$

$$\sum_{i,j} \left| \frac{\partial^2 g}{\partial z_i \partial z_j}(z) \right| \le C(1 + |z|),$$

$$\text{and} \quad \sum_{i,j,k} \left| \frac{\partial^3 g}{\partial z_i \partial z_j \partial z_k}(z) \right| \le C(1 + |z|^\alpha).$$

It follows from (H3) that (3.1) has a *strong* solution $\Phi : \mathbb{R}^n \times \mathbf{X} \to \mathbf{X}$, so that on the probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P, X_0, (V, W))$ supporting an $\mathbb{R}^n$-valued random variable X_0 with distribution μ, and $(n+d)$-dimensional vector Brownian motion (V, W), independent of X_0, $(X_t = \Phi_t(X_0, V), \mathcal{F}_t; 0 \le t \le T)$ is a continuous semimartingale satisfying (3.1). (See, for example, [15].)

It follows from (H2)–(H4) that $\mathbf{E}\int_0^T |g(X_t)|^2\,dt < \infty$, and from this and the independence of X and W it follows by standard results (see, for example, [9]) that (H1) is satisfied when the reference measure λ_Y is the Wiener measure and the Radon–Nikodym derivative takes the form

$$\frac{dP_{XY}}{d(P_X \otimes \lambda_Y)}(X, Y) = \exp\left(\int_0^T g(X_t)'\,dY_t - \frac{1}{2}\int_0^T |g(X_t)|^2\,dt\right). \tag{3.6}$$

In order to develop the representations of Proposition 2.1, we first need a version of this that is well defined for all y. Under (H2)–(H4) the process $(g(X_t), \mathcal{F}_t, 0 \le t \le T)$ is a semimartingale, and so it is possible to "integrate by parts" in (3.6) and define Q as any measurable function such that, for each y,

$$Q(X, y) = \exp\left(y_T' g(X_T) - \int_0^T y_t'\,dg(X_t) - \frac{1}{2}\int_0^T |g(X_t)|^2\,dt\right). \tag{3.7}$$

(See [2] and [3].) It can also be shown (see, for example, [13], [14]) that the resulting regular conditional probability, $P_{X|Y}$, is continuous in y in the sense of the topology associated with the convergence of means of bounded measurable functions, that

$$0 < \mathbf{E}Q(X, y) < \infty \qquad \text{for all } y, \tag{3.8}$$

and that

$$\mathbf{E}Q(X, y)\log(Q(X, y)) \le \mathbf{E}Q(X, y)^2 < \infty. \tag{3.9}$$

Thus the set $\bar{\mathbf{Y}}$ of (2.1) can be taken to be the entire space $\mathbf{Y}$ in this case, and (2.7) is satisfied for all y. Proposition 2.1 can thus be applied for each y, and $H = -\log(Q)$.

We can now split the path estimation problem as suggested by (3.5). For any $z \in \mathbb{R}^n$ and any $0 \le s \le T$, let $(X_t^{z,s}; s \le t \le T)$ be the solution of (3.1) on the interval $s \le t \le T$ with "initial condition" $X_s^{z,s} = z$, and let

$$H_p : [0, T] \times [0, T] \times \mathbb{R}^n \times \mathbf{X} \times \mathbf{Y} \to \mathbb{R}$$

be a measurable function such that

$$\begin{aligned} H_p(s, t, z, X^{z,s}, y) = {} & -y_t' g(X_t^{z,s}) + y_s' g(z) + \int_s^t y_r'\,dg(X_r^{z,s}) \\ & + \frac{1}{2}\int_s^t |g(X_r^{z,s})|^2\,dr \quad \text{for } 0 \le s \le t \le T. \end{aligned} \tag{3.10}$$

The fact that such a function exists follows from the "strong solution" hypothesis (H3), as does the decomposition

$$H(X, y) = H_p(0, s, X_0, X, y) + H_p(s, T, X_s, (X_t, s \le t \le T), y). \tag{3.11}$$

$H_p(s,t,z,\cdot\,,\cdot\,)$ is the equivalent of H for the problem of estimating the path $(X_r^{z,s}, s \le r \le t)$ given the observation $(Y_r^{z,s}, s \le r \le t)$, where

$$Y_t^{z,s} = \int_s^t g(X_r^{z,s})\,dr + W_t - W_s \quad \text{for } s \le t \le T.$$

In particular, $H_p(s,T,z,\cdot\,,\cdot\,)$ is the equivalent of H for the problem of estimating $X^{z,s}$ given $Y^{z,s}$. Let $v(z,s,y)$ be the minimum apparent information for this problem; then, according to Proposition 2.1 (i),

$$v(z,s,y) = -\log\left(\mathbf{E}\exp(-H_p(s,T,z,X^{z,s},y))\right). \tag{3.12}$$

It now follows that, for any $A \in \mathcal{X}_0^s$,

$$P_{X|Y}(A,y) = \frac{\mathbf{E}\mathbf{1}_A(X)\exp\left(-H_p(0,s,X_0,X,y) - v(X_s,s,y)\right)}{\mathbf{E}\exp\left(-H_p(0,s,X_0,X,y) - v(X_s,s,y)\right)}, \tag{3.13}$$

and from Jensen's inequality and (3.9) it follows that $H_p(0,s,\chi_0,\cdot\,,y) + v(\chi_s(\cdot\,),s,y)$ satisfies (2.7) for all s. So, from Proposition 2.1, the path measure $P_{X|Y}$ restricted to $\mathcal{X}_0^s$ is the unique probability measure on $\mathcal{X}_0^s$ that minimizes the apparent information

$$F_s(\tilde{P}_{X,s},y) = h(\tilde{P}_{X,s}\,|\,P_{X,s}) + \langle H_p(0,s,\chi_0,\cdot\,,y)\,,\,\tilde{P}_{X,s}\rangle + \langle v(\chi_s,s,y)\,,\,\tilde{P}_{X,s}\rangle, \tag{3.14}$$

where $P_{X,s}$ is the restriction of P_X to $\mathcal{X}_0^s$. It also easily follows that the minimum apparent information in (3.14) does not depend on s.

These arguments show that the variational form of the path estimation problem (3.1), (3.2) can be interpreted in terms of dynamic programming, with value function v. For each s we can split the problem into two subproblems: the estimation of $X^{z,s}$ for each z (resulting in a minimum apparent information of $v(z,s,y)$), followed by the estimation of $(X_t, 0 \le t \le s)$, where $v(X_s,s,y)$ plays a part in the likelihood function. $v(X_s,s,y)$ summarizes that part of the likelihood function associated with increments of Y after time s. The first subproblem can be interpreted in terms of stochastic optimal control, where the cost is the apparent information of the controlled process. This is developed in the next section.

4. A stochastic control formulation. We consider the first variational subproblem discussed above with $s = 0$. In keeping with the comments above on dynamic programming, it turns out that we need consider only feedback controls. Also, because controls are intended to produce a change in measure of the form (3.3), it is appropriate to let the control enter the drift through the map $z \mapsto az$, where $a = \sigma\sigma'$.

Consider the following controlled equation:

$$\tilde{X}_t = \theta + \int_0^t \Big(b(\tilde{X}_s,s) + a(\tilde{X}_s,s)u(\tilde{X}_s,s)\Big)\,ds + \int_0^t \sigma(\tilde{X}_s,s)\,d\tilde{V}_s, \tag{4.1}$$

where the initial condition, θ, is nonrandom. Let $\mathbf{U}$ be the set of measurable functions $u : \mathbb{R}^n \times [0,T] \to \mathbb{R}^n$ with the following properties:

(U1) u is continuous,

(U2) $\mathbf{E}\Gamma^u = 1$, where

$$\Gamma^u = \exp\left(\int_0^T u'\sigma(X_t^{\theta,0},t)\,dV_t - \frac{1}{2}\int_0^T |\sigma' u(X_t^{\theta,0},t)|^2\,dt\right), \tag{4.2}$$

and $(\Omega,\mathcal{F},P)$, V, and $X^{z,s}$ are as defined in section 3.

LEMMA 4.1. *If b and σ satisfy* (H3) *and $u \in \mathbf{U}$, then* (4.1) *has a weak solution and is unique in law.*

Proof. From (H3) and (U1) it follows that

$$P\left(\int_0^T \left|\sigma' u(X_t^{\theta,0}, t)\right|^2 \, dt < \infty\right) = 1.$$

This, together with (U2) and Girsanov's theorem, shows that V^u, defined by

$$V_t^u = V_t - \int_0^t \sigma' u(X_s^{\theta,0}, s)\, ds, \tag{4.3}$$

is a standard Brownian motion under the probability measure P^u, defined by

$$\frac{dP^u}{dP} = \Gamma^u. \tag{4.4}$$

This shows that $(\Omega, \mathcal{F}, (\mathcal{F}_t), P^u, X^{\theta,0}, V^u)$ is a weak solution of (4.1).

Next, suppose that $(\tilde{\Omega}, \tilde{\mathcal{F}}, (\tilde{\mathcal{F}}_t), \tilde{P}, \tilde{X}, \tilde{V})$ is a weak solution of (4.1), and, for each natural number N, let $\tau_N : \mathbf{X} \to [0, T]$ be defined by

$$\tau_N(x) = \inf\{t \geq 0 : |x_t| \geq N\} \wedge T.$$

Since $\tilde{X}$ is continuous, $\tilde{P}(\tau_N(\tilde{X}) \to T) = 1$. Also, since u satisfies (U1),

$$\tilde{\mathbf{E}} \exp\left(\frac{1}{2}\int_0^{\tau_N(\tilde{X})} \left|\sigma' u(\tilde{X}_s, s)\right|^2 \, ds\right) < \infty,$$

and so, from a standard variation of Novikov's theorem (see, for example, Theorem 6.1 in [9]), it follows that $(M_t, \tilde{\mathcal{F}}_t, 0 \leq t \leq T)$, where

$$M_t = \exp\left(-\int_0^t u'\sigma(\tilde{X}_s, s)\, d\tilde{V}_s - \frac{1}{2}\int_0^t \left|\sigma' u(\tilde{X}_s, s)\right|^2 \, ds\right) \tag{4.5}$$

is a local martingale with respect to the sequence of stopping times $(\tau_N(\tilde{X});\ N = 1, 2, \ldots)$. Let

$$\tilde{V}_t^N = \tilde{V}_t + \int_0^{t\wedge\tau_N(\tilde{X})} \sigma' u(\tilde{X}_s, s)\, ds;$$

then, by Girsanov's theorem, $\tilde{V}^N$ is a standard Brownian motion under the probability measure $\tilde{P}^N$, defined by $d\tilde{P}^N = M_{\tau_N(\tilde{X})} d\tilde{P}$. Let $(\mathcal{X}_t;\ 0 \leq t \leq T)$ be the filtration on $(\mathbf{X}, \mathcal{X})$ generated by the coordinate process (χ_t). Since

$$\tilde{X}_{t\wedge\tau_N(\tilde{X})} = \Phi_{t\wedge\tau_N(\tilde{X})}(\theta, \tilde{V}^N) \qquad \text{for } 0 \leq t \leq T,$$

where Φ is the strong solution to (3.1), the law of $\tilde{X}$ restricted to $\mathcal{X}_{\tau_N}$ is identical to that of $X^{\theta,0}$ under P^u, restricted to the same sigma-field. Finally, for any $A \in \mathcal{X}$,

$$\begin{aligned}\tilde{P}(\tilde{X} \in A, \tau_N(\tilde{X}) = T) &= \tilde{P}(\tilde{X} \in A) - \tilde{P}(\tilde{X} \in A, \tau_N(\tilde{X}) < T)\\ &\to \tilde{P}(\tilde{X} \in A),\end{aligned}$$

and so, since the events on the left-hand side each belong to one of $(\mathcal{X}_{\tau_N};\ N = 1, 2, \ldots)$, the law of $\tilde{X}$ on $\mathcal{X}$ is identical to that of $X^{\theta,0}$ under P^u. □

Let $(\tilde{\Omega}, \tilde{\mathcal{F}}, (\tilde{\mathcal{F}}_t), \tilde{P}, \tilde{X}, \tilde{V})$ be a weak solution of (4.1) for some $u \in \mathbf{U}$. We define the cost for controls in $\mathbf{U}$ as the apparent information of the resulting distribution of $\tilde{X}$, $\tilde{P}_X$. This is measured relative to the prior $P_X^{\theta,0}$ (the distribution of $X^{\theta,0}$) and $H_p(0, T, \theta, \cdot\,, y)$ (as defined in (3.10)).

$$
\begin{aligned}
J(u, \theta, y) &= h(\tilde{P}_X \mid P_X^{\theta,0}) + \langle H_p(0, T, \theta, \cdot\,, y), \tilde{P}_X \rangle \\
&= \frac{1}{2}\tilde{\mathbf{E}} \int_0^T |\sigma' u(\tilde{X}_t, t)|^2 \, dt - y_T' g(\theta) + \frac{1}{2}\tilde{\mathbf{E}} \int_0^T |g(\tilde{X}_t)|^2 \, dt \\
&\quad -\tilde{\mathbf{E}} \int_0^T (y_T - y_t)'(\mathcal{L}g + \mathcal{D}gau)(\tilde{X}_t, t) \, dt \qquad \text{if the integrals exist,} \\
&\quad +\infty \qquad \text{otherwise,}
\end{aligned} \tag{4.6}
$$

where $\mathcal{L}$ is the differential operator associated with X,

$$
\mathcal{L} = \sum_i b_i \frac{\partial}{\partial z_i} + \frac{1}{2} \sum_{i,j} a_{i,j} \frac{\partial^2}{\partial z_i \partial z_j},
$$

and $\mathcal{D}$ is the row-vector jacobian operator, $\mathcal{D} = [\partial/\partial z_1 \ \partial/\partial z_2 \cdots \partial/\partial z_n]$. The cost functional has a more appealing form in the special case that the observation path, y, is everywhere differentiable:

$$
J(u, \theta, y) = \frac{1}{2}\tilde{\mathbf{E}} \int_0^T \left(|\sigma' u(\tilde{X}_t, t)|^2 + |\dot{y}_t - g(\tilde{X}_t)|^2 \right) dt - \frac{1}{2} \int_0^T |\dot{y}_t|^2 \, dt. \tag{4.7}
$$

This involves an "energy" term for the control and a "least-squares" term for the observation path fit. These correspond to the two terms in Bayes' formula representing the degrees of match with the prior distribution and the observation path. The optimal control problem (4.1), (4.7) can be thought of as a type of energy-constrained *tracking* problem. The optimal control, under which the distribution of $\tilde{X}$ is the regular conditional probability distribution $P_{X|Y}(\cdot\,, y)$, is derived in the following theorem.

THEOREM 4.2. *Suppose that b, σ, and g satisfy* (H3) *and* (H4)*, and let the function $u_* : \mathbb{R}^n \times [0, T] \times \mathbf{Y} \to \mathbb{R}^n$ be defined by*

$$
u_* = -(\mathcal{D}v)', \tag{4.8}
$$

where v is as defined in (3.12)*. Then, for each $y \in \mathbf{Y}$, $u_*(\cdot\,, \cdot\,, y)$ belongs to $\mathbf{U}$, and for all $\theta \in \mathbb{R}^n$, $y \in \mathbf{Y}$, and $\tilde{P}_X \in \mathcal{P}(\mathcal{X})$ (not necessarily the distribution of a controlled process),*

$$
J(u_*(\cdot\,, \cdot\,, y), \theta, y) \le h(\tilde{P}_X \mid P_X^{\theta,0}) + \langle H_p(0, T, \theta, \cdot\,, y), \tilde{P}_X \rangle. \tag{4.9}
$$

Proof. The proof is in three parts. The first uses the methods of stochastic flows to establish a stochastic representation formula for u_*, (4.20). The second proves the statement of the theorem for nondegenerate systems with bounded coefficients. Finally, a truncation argument is used to extend this result to the general case. Only the time-homogeneous case (b and σ not dependent on t) is treated in order to avoid excessive notation. The arguments extend in an obvious way to the general case.

Standard moment bounding arguments (see, for example, Theorem 4.6 in [9]) show that for each natural number m there exists a $C_m < \infty$, not depending on z or s, such that

$$\sup_{s\le t\le T} \mathbf{E}|X_t^{z,s}|^{2m} \le C_m\left(1+|z|^{2m}\right) \tag{4.10}$$

$$\text{and} \quad \sup_{s\le t\le T} \mathbf{E}\,\|\Psi_t^{z,s}\|^{2m} \le C_m, \tag{4.11}$$

where $(\Psi_t^{z,s} \in \mathbb{R}^{n\times n}; s \le t \le T)$ is the solution of the equation of first-order variation associated with $X^{z,s}$,

$$\Psi_t^{z,s} = I + \int_s^t \mathcal{D}b(X_r^{z,s})\Psi_r^{z,s}\,dr + \sum_i \int_s^t \mathcal{D}\sigma_i(X_r^{z,s})\Psi_r^{z,s}\,dV_{i,r}. \tag{4.12}$$

Here and in what follows, σ_i is the ith column of σ, and $V_{i,t}$ is the ith component of V_t. For any $z, \tilde z \in \mathbb{R}^n$ and any $0 \le s \le t \le T$

$$X_t^{z,s} - X_t^{\tilde z,s} = (z-\tilde z) + \int_s^t (b(X_r^{z,s}) - b(X_r^{\tilde z,s}))\,dr + \int_s^t (\sigma(X_r^{z,s}) - \sigma(X_r^{\tilde z,s}))\,dV_r,$$

and so for any natural number m there exists a $C_m < \infty$, not depending on s, t, z, or $\tilde z$, such that

$$\begin{aligned}\mathbf{E}\sup_{s\le r\le t}\left|X_r^{z,s} - X_r^{\tilde z,s}\right|^{2m} &\le 3^{2m-1}\Bigg(|z-\tilde z|^{2m} + \mathbf{E}\sup_{s\le r\le t}\left|\int_s^r (b(X_q^{z,s}) - b(X_q^{\tilde z,s}))\,dq\right|^{2m} \\ &\qquad + \mathbf{E}\sup_{s\le r\le t}\left|\int_s^r (\sigma(X_q^{z,s}) - \sigma(X_q^{\tilde z,s}))\,dV_q\right|^{2m}\Bigg) \\ &\le C_m\left(|z-\tilde z|^{2m} + \int_s^t \mathbf{E}\sup_{s\le q\le r}\left|X_q^{z,s} - X_q^{\tilde z,s}\right|^{2m}\,dr\right),\end{aligned}$$

where we have used Doob's submartingale inequality, (4.10), (H3), and standard bounds for the moments of stochastic integrals. It thus follows from the Gronwall lemma that

$$\mathbf{E}\sup_{s\le t\le T}|X_t^{z,s} - X_t^{\tilde z,s}|^{2m} \le C_m\exp(C_m T)|z-\tilde z|^{2m} \qquad \text{for all } (z,\tilde z,s). \tag{4.13}$$

Similarly,

$$\mathbf{E}\sup_{s\le t\le T}|X_t^{z,s}|^{2m} \le C_m(1+|z|^{2m}) \qquad \text{for all } (z,s), \tag{4.14}$$

and so for any $\epsilon > 0$ and any bounded set $A \subset \mathbb{R}^n$ there exists a $C < \infty$ such that

$$P\left(\sup_{s\le t\le T}|X_t^{z,s}| > C\right) < \epsilon/4 \qquad \text{for all } (z,s) \in A\times[0,T].$$

From (H3) and (H4) it follows that $\mathcal{D}(\mathcal{L}g)$ is uniformly continuous on compacts, and so for any $\eta > 0$ there exists a $\delta > 0$ such that if $z, \tilde z \in A$ and $|z-\tilde z| < \delta$,

$$P\left(\sup_{s\le t\le T}\left\|\mathcal{D}(\mathcal{L}g)(X_t^{z,s}) - \mathcal{D}(\mathcal{L}g)(X_t^{\tilde z,s})\right\| > \eta,\ \sup_{s\le t\le T}(|X_t^{z,s}| \vee |X_t^{\tilde z,s}|) \le C\right) < \epsilon/2,$$

so that

$$P\left(\sup_{s\le t\le T}\left\|\mathcal{D}(\mathcal{L}g)(X_t^{z,s})-\mathcal{D}(\mathcal{L}g)(X_t^{\tilde z,s})\right\|>\eta\right)<\epsilon. \tag{4.15}$$

The polynomial growth of $\mathcal{D}(\mathcal{L}g)$ together with (4.14) and the Vallée–Poussin theorem shows that, for any $0<p<\infty$, the family

$$\left\{\sup_{s\le t\le T}\|\mathcal{D}(\mathcal{L}g)(X_t^{z,s})\|^p\,;z\in A,0\le s\le T\right\}$$

is uniformly integrable. This and (4.15) show that for any $0<p<\infty$

$$\mathbf{E}\sup_{s\le t\le T}\left\|\mathcal{D}(\mathcal{L}g)(X_t^{z,s})-\mathcal{D}(\mathcal{L}g)(X_t^{\tilde z,s})\right\|^p=o(|z-\tilde z|^0) \tag{4.16}$$

uniformly on $A\times[0,T]$. Similar arguments show that $\mathcal{D}g(X_t^{z,s})$, $\mathcal{D}b(X_t^{z,s})$, $\mathcal{D}\sigma_i(X_t^{z,s})$, and $\mathcal{D}(\mathcal{D}g\sigma_i)(X_t^{z,s})$ for $i=1,2,\dots,n$ have the same property.

It follows from the mean-value theorem that

$$\begin{aligned}X_t^{z,s}-X_t^{\tilde z,s}&=(z-\tilde z)+\int_s^t\mathcal{D}b\left(\alpha_{0,r}X_r^{z,s}+(1-\alpha_{0,r})X_r^{\tilde z,s}\right)(X_r^{z,s}-X_r^{\tilde z,s})\,dr\\&\quad+\sum_i\int_s^t\mathcal{D}\sigma_i\left(\alpha_{i,r}X_r^{z,s}+(1-\alpha_{i,r})X_r^{\tilde z,s}\right)(X_r^{z,s}-X_r^{\tilde z,s})\,dV_{i,r},\end{aligned}$$

where $0<\alpha_{i,r}<1$ and $\alpha_{i,r}$ is $\mathcal{F}_r$-measurable for each i. The above continuity properties, Hölder's inequality, and techniques similar to those used to prove (4.13) now show that for any $0<p<\infty$

$$\mathbf{E}\sup_{s\le t\le T}\left|X_t^{z,s}-X_t^{\tilde z,s}-\Psi_t^{z,s}(z-\tilde z)\right|^p=o(|z-\tilde z|^p), \tag{4.17}$$

and

$$\mathbf{E}\,|\Theta(z,s,y)-\Theta(\tilde z,s,y)-\xi(z,s,y)\Theta(z,s,y)(z-\tilde z)|^p=o(|z-\tilde z|^p), \tag{4.18}$$

both uniformly on $A\times[0,T]$, where

$$\Theta(z,s,y)=\exp\left(-H_p(s,T,z,X^{z,s},y)\right)$$

and

$$\begin{aligned}\xi(z,s,y)&=(y_T-y_s)'\mathcal{D}g(z)+\sum_i\int_s^T(y_T-y_t)'\mathcal{D}(\mathcal{D}g\sigma_i)(X_t^{z,s})\Psi_t^{z,s}\,dV_{i,t}\\&\quad+\int_s^T(y_T-y_t)'\mathcal{D}(\mathcal{L}g)(X_t^{z,s})\Psi_t^{z,s}\,dt-\int_s^T g'(X_t^{z,s})\mathcal{D}g(X_t^{z,s})\Psi_t^{z,s}\,dt.\end{aligned}$$

Thus $\mathcal{D}\rho=\mathbf{E}\xi\Theta$, where $\rho=\mathbf{E}\Theta$. Now, Jensen's inequality shows that

$$\inf_{z\in A,0\le s\le T}\rho(z,s,y)\ge\inf_{z\in A,0\le s\le T}\exp(\mathbf{E}\log(\Theta(z,s,y)))>0, \tag{4.19}$$

and so

$$u_*(z,s,y)=\frac{\mathbf{E}\xi(z,s,y)\Theta(z,s,y)}{\mathbf{E}\Theta(z,s,y)}. \tag{4.20}$$

We now consider the special case in which y is differentiable with Hölder continuous derivative, b and g are bounded, and there exists an $\epsilon > 0$ such that

$$\tilde{z}' a(z) \tilde{z} \geq \epsilon |\tilde{z}|^2 \qquad \text{for all } z, \tilde{z} \in \mathbb{R}^n. \tag{4.21}$$

In this case ρ is continuously differentiable with respect to s, is twice continuously differentiable with respect to z, and by a standard extension of the Feynman–Kac formula satisfies the following partial differential equation (see, for example, [7]):

$$\frac{\partial \rho}{\partial s} + \mathcal{L}\rho + \left(\dot{y} - \frac{1}{2}g\right)' g\rho = 0 \quad \text{on } \mathbb{R}^n \times (0,T), \quad \rho(\,\cdot\,, T, y) = 1. \tag{4.22}$$

Since $v = -\log(\rho)$, the value function v satisfies

$$\frac{\partial v}{\partial s} + \mathcal{L}v - \frac{1}{2}\mathcal{D}va(\mathcal{D}v)' - \left(\dot{y} - \frac{1}{2}g\right)' g = 0 \quad \text{on } \mathbb{R}^n \times (0,T), \quad v(\,\cdot\,, T, y) = 0. \tag{4.23}$$

Now, because of (4.10), (4.11), and the boundedness of g and $\mathcal{D}g$, $u_*(\,\cdot\,,\,\cdot\,, y)$ is also bounded and, by Novikov's theorem, satisfies (U2). We have thus shown that in this special case $u_*(\,\cdot\,,\,\cdot\,, y) \in \mathbf{U}$. Let V^* and P^* be abbreviations for $V^{u_*(\,\cdot\,,\,\cdot\,,y)}$ and $P^{u_*(\,\cdot\,,\,\cdot\,,y)}$, respectively, where, for $u \in \mathbf{U}$, V^u and P^u are as defined by (4.3) and (4.4). Then Itô's rule and (4.23) show that

$$\begin{aligned} 0 = v(X_T^{\theta,0}, T, y) = v(\theta, 0, y) &+ \int_0^T \left(\left(\dot{y}_t - \frac{1}{2}g\right)' g - \frac{1}{2}|\sigma' u_*|^2 \right) (X_t^{\theta,0}, t, y)\, dt \\ &- \int_0^T (u_*'\sigma)(X_t^{\theta,0}, t, y)\, dV_t^*. \end{aligned}$$

As was pointed out in the proof of Lemma 4.1, $(\Omega, \mathcal{F}, (\mathcal{F}_t), P^*, X^{\theta,0}, V^*)$ is a weak solution of (4.1), and so, since g, $u_*(\,\cdot\,,\,\cdot\,, y)$ and σ are bounded,

$$\begin{aligned} v(\theta, 0, y) &= \mathbf{E}^* \int_0^T \left(\frac{1}{2}|\sigma' u_*| - \left(\dot{y}_t - \frac{1}{2}g\right)' g \right) (X_t^{\theta,0}, t, y)\, dt \\ &= J(u_*(\,\cdot\,,\,\cdot\,, y), \theta, y). \end{aligned}$$

By definition, $v(\theta, 0, y)$ is the minimum apparent information, and so we have established (4.9) in this special case. A consequence of (4.9), and the uniqueness of the measure minimizing apparent information, is that the distribution of $\tilde{X}$ when $u = u_*(\,\cdot\,,\,\cdot\,, y)$ is the regular conditional distribution of $X^{\theta,0}$ given that $Y = y$. Thus, in this special case,

$$\Gamma^{u_*(\,\cdot\,,\,\cdot\,,y)} = \frac{\Theta(\theta, 0, y)}{\rho(\theta, 0, y)} \text{a.s.}$$

Next, suppose that the additional constraints placed on y, b, g, and σ are removed. For any natural number N, let

$$\begin{aligned} b_N(z) &= b(z) \exp(-|z|^2/N), \\ g_N(z) &= g(z) \exp(-|z|^2/N), \\ \sigma_N(z) &= \begin{bmatrix} \sigma & N^{-1} I \end{bmatrix} \qquad (\text{an } n \times 2n \text{ matrix}), \end{aligned}$$

and let y^N be a sequence of differentiable elements of $\mathbf{Y}$ with Hölder continuous derivatives such that $\|y - y^N\| \to 0$. Then b_N and g_N are bounded and σ_N satisfies (4.21), b_N, σ_N, and g_N satisfy (H3) and (H4) uniformly in N, and b_N, σ_N, g_N, $\mathcal{D}b_N$, $\partial\sigma_N/\partial z_i$, and $\mathcal{D}g_N$ converge to b, $[\sigma\, 0]$, g, $\mathcal{D}b$, $[\partial\sigma/\partial z_i\, 0]$, and $\mathcal{D}g$ (respectively) uniformly on compacts. We add the subscript (or superscript) N to X, Ψ, Θ, etc. to indicate that y, b, g, and σ have been replaced by y^N, b_N, g_N, and σ_N in the various definitions and that V has been replaced by the $2n$-dimensional Brownian motion, (V_t, B_t). Now

$$X_t^{z,s} - X_t^{N,z,s} = \int_s^t \left(b_N(X_r^{z,s}) - b_N(X_r^{N,z,s})\right)\, dr + \int_s^t \left(\sigma(X_r^{z,s}) - \sigma(X_r^{N,z,s})\right)\, dV_r + \int_s^t \left(b(X_r^{z,s}) - b_N(X_r^{z,s})\right)\, dr - N^{-1}(B_t - B_s).$$

Arguments similar to those used to prove (4.13), (4.17), and (4.18) show that, for any natural number m and any bounded set $A \subset \mathbb{R}^n$,

$$\mathbf{E} \sup_{s\le t\le T} \left|X_t^{z,s} - X_t^{N,z,s}\right|^{2m} \to 0, \tag{4.24}$$

$$\mathbf{E} \sup_{s\le t\le T} \left\|\Psi_t^{z,s} - \Psi_t^{N,z,s}\right\|^{2m} \to 0,$$

$$\mathbf{E}\left|\Theta(z,s,y) - \Theta_N(z,s,y^N)\right|^{2m} \to 0, \tag{4.25}$$

$$\text{and} \quad \mathbf{E}\left|\xi(z,s,y) - \xi_N(z,s,y^N)\right|^{2m} \to 0,$$

all uniformly on $A \times [0,T]$. This, Hölder's inequality, and (4.19) show that

$$u_{*N}(\cdot\,,\cdot\,,y^N) \to u_*(\cdot\,,\cdot\,,y) \qquad \text{uniformly on } A\times[0,T]. \tag{4.26}$$

Thus $u_*(\cdot\,,\cdot\,,y)$ satisfies (U1). It follows from (4.24) and (4.26) that

$$\sup_{0\le t\le T}\left|u_*(X_t^{\theta,0},t,y) - u_{*N}(X_t^{N,\theta,0},t,y^N)\right| \to 0 \qquad \text{in probability,}$$

so that

$$\Gamma_N^{u_{*N}(\cdot\,,\cdot\,,y^N)} \to \Gamma^{u_*(\cdot\,,\cdot\,,y)} \qquad \text{in probability.} \tag{4.27}$$

It also follows from (4.25) and (4.19) that

$$\Gamma_N^{u_{*N}(\cdot\,,\cdot\,,y^N)} = \frac{\Theta_N(\theta,0,y^N)}{\rho_N(\theta,0,y^N)} \to \frac{\Theta(\theta,0,y)}{\rho(\theta,0,y)} \qquad \text{in probability,} \tag{4.28}$$

and so $u_*(\cdot\,,\cdot\,,y)$ satisfies (U2), and the unique distribution of $\tilde{X}$ under this control coincides with the regular conditional distribution of X given that $Y = y$. This establishes (4.9) in the general case. □

We return now to the path estimator with initial distribution μ. The minimization of apparent information can be expressed in terms of the following controlled process with random initial condition:

$$\begin{aligned} &\tilde{X}_t = \tilde{X}_0 + \int_0^t \left(b(\tilde{X}_s,s) + a(\tilde{X}_s,s)u(\tilde{X}_s,s)\right)\, ds + \int_0^t \sigma(\tilde{X}_s,s)\, d\tilde{V}_s, \\ &\tilde{X}_0 \sim \tilde{\mu}. \end{aligned} \tag{4.29}$$

A simple variant of Lemma 4.1 shows that, if u is continuous and satisfies (U2) for all $\theta \in \mathbb{R}^n$, then this equation is unique in law and has a weak solution for any initial law, $\tilde{\mu}$. Let $\tilde{P}_X$ be the distribution of $\tilde{X}$ corresponding to the pair $(\tilde{\mu}, u)$; it follows from (3.14) and the subsequent discussion that

$$F(\tilde{P}_X, y) = F_0(\tilde{\mu}, y) = h(\tilde{\mu} \,|\, \mu) + \langle J(u, \cdot\,, y), \tilde{\mu} \rangle, \tag{4.30}$$

and this is minimized by the choice $u = u_*(\cdot\,, \cdot\,, y)$ and $\tilde{\mu} = \mu_Y(\cdot\,, y)$, where for $B \in \mathcal{B}^n$

$$\mu_Y(B, y) = P_{X|Y}(\chi_0^{-1}(B), y). \tag{4.31}$$

Thus, for each y, the regular conditional probability distribution $P_{X|Y}(\cdot\,, y)$ is Markovian with "initial" marginal $\mu_Y(\cdot\,, y)$ and differential operator

$$\mathcal{L}_y = \sum_i (b + a u_*(\cdot\,, \cdot\,, y))_i \frac{\partial}{\partial z_i} + \frac{1}{2} \sum_{i,j} a_{i,j} \frac{\partial^2}{\partial z_i \partial z_j}. \tag{4.32}$$

Of course, the nonlinear filter and interpolator for the process X can be found from the marginals of this path space measure.

5. The inverse problem. The variational characterization of the inverse problem (parts (ii) and (iv) of Proposition 2.1) can also be applied to the path estimator. This involves choosing a likelihood function to be compatible with the (given) regular conditional probability distribution, $P_{X|Y}(\cdot\,, y)$. In section 4, we minimized apparent information over probability measures corresponding to weak solutions of (4.29). Here, we maximize compatible information over (negative) log-likelihood functions, $\tilde{H}$, that give rise to posterior distributions of this type.

Let $(\Omega, \mathcal{F}, P)$, μ, V, and X be as defined in section 3. For each probability measure on $\mathbb{R}^n$, $\tilde{\mu}$, with $\tilde{\mu} \ll \mu$, and each continuous u satisfying (U2) for all θ, let $\tilde{H}$ be a measurable function such that

$$\begin{aligned} \tilde{H}(X) &= -\log\left(\frac{d\tilde{P}_X}{dP_X}(X)\right) + K \\ &= -\log\left(\frac{d\tilde{\mu}}{d\mu}(X_0)\right) - \int_0^T u'\sigma(X_t, t)\, dV_t + \frac{1}{2}\int_0^T |\sigma' u(X_t, t)|^2\, dt + K, \end{aligned} \tag{5.1}$$

where $K \in \mathbb{R}$ and $\tilde{P}_X$ is as defined following (4.29). We shall assume that $\mu_Y(\cdot\,, y) \ll \tilde{\mu}$. If this is not so, then, as shown in the proof of Proposition 2.1, we can always choose another $\tilde{\mu}$ resulting in more compatible information, for which it is. The term K in (5.1) is the information in the associated (unspecified) observation.

Integral log-likelihood functions of the form (5.1) can be thought of as being associated with observations that are "distributed in time," in that information from them gradually becomes available as t increases.

The characterization of $P_{X|Y}$ in terms of stochastic control can be used to express the compatible information corresponding to $\tilde{H}$ as follows:

$$\begin{aligned} G(\tilde{H}, y) &= K - \langle \tilde{H}\,, P_{X|Y}(\cdot\,, y) \rangle \\ &= K + h(\mu_Y(\cdot\,, y) \,|\, \mu) - h(\mu_Y(\cdot\,, y) \,|\, \tilde{\mu}) \\ &\qquad + \int_0^T \int_{\mathbb{R}^n} \left(u_* - \frac{1}{2}u\right)' a u(z, t, y) P_{X|Y}(\chi_t^{-1}(dz), y)\, dt. \end{aligned} \tag{5.2}$$

Log-likelihood functions of the form (5.1) could come from many different types of observation. The only constraints placed on u here are that it be continuous and that it satisfy (U2) for all θ. We could further constrain it to take the form

$$u(z,s) = -(\mathcal{D}\tilde{v})'(z,s,\tilde{y}),$$

where

$$\tilde{v}(z,s,\tilde{y}) = -\log \mathbf{E} \exp\left(\int_s^T \left(\dot{\tilde{y}}_t - \frac{1}{2}\tilde{g}(X_t^{z,s}) \right)' \tilde{g}(X_t^{z,s})\, dt \right)$$

for appropriate $\tilde{g}$ and $\tilde{y}$. This would correspond to observations of the "signal-plus-white-noise" variety similar to (3.2) but with "controlled" observation function and path, $\tilde{g}$ and $\tilde{y}$. This would show the effects of errors in the observation function or approximations of the observation path. Under appropriate regularity conditions, $\tilde{v}$ will satisfy the following partial differential equation:

$$-\frac{\partial \tilde{v}}{\partial t} = \mathcal{L}\tilde{v} - \frac{1}{2}\mathcal{D}\tilde{v}a(\mathcal{D}\tilde{v})' - \left(\dot{\tilde{y}}_t - \frac{1}{2}\tilde{g} \right)' \tilde{g}; \quad \tilde{v}(\,\cdot\,, T) = 0. \tag{5.3}$$

Thus one interpretation of the inverse problem involves an infinite-dimensional deterministic optimal control problem in reversed time, with control $(\tilde{g}, \tilde{y})$, and payoff

$$\Pi(\tilde{g},\tilde{y}) = \int_0^T \int_{\mathbb{R}^n} \mathcal{D}\tilde{v}a \left(u_* - \frac{1}{2}(\mathcal{D}\tilde{v})' \right)(z,t,y) P_{X|Y}(\chi_t^{-1}(dz), y)\, dt. \tag{5.4}$$

The optimal trajectory for this dual problem, $v(\,\cdot\,,\,\cdot\,,y)$ is a time-reversed likelihood filter for X given Y, and the measure $\exp(-v(z,s,y))P_X(\chi_s^{-1}(dz))$ is an unnormalized regular conditional probability distribution for X_s given observations $(Y_t - Y_s, s \le t \le T)$, which coincides with that provided by the Zakai equation for the time-reversed problem. This provides an information-theoretic explanation of the connection between nonlinear filtering and stochastic optimal control used in [6] as well as widening its scope. For a somewhat different problem involving optimization over observation functions, see [16].

6. Information flow and localization. The results of section 2 concerning the information conserving properties of Bayesian estimators can be localized in the context of the diffusion problem (3.1), (3.2). Proposition 2.1 can be applied to provide variational characterizations of various conditional probabilities of the path measure $P_{X|Y}$, including transition probabilities, and these can be used to characterize the *flow* of information at a given time and in a given state.

For any initial law $\tilde{\mu} \ll \mu$ and any control u satisfying (U1) and (U2) for all θ, let $(\tilde{\Omega}, \tilde{\mathcal{F}}, (\tilde{\mathcal{F}}_t), \tilde{P}, \tilde{X}, \tilde{V})$ be a weak solution of (4.29), let $\tilde{P}_X$ be the distribution of $\tilde{X}$, and let $P_{X,s}$, $\tilde{P}_{X,s}$, and $P_{X,s|Y}(\,\cdot\,, y)$ be the restrictions of P_X, $\tilde{P}_X$, and $P_{X|Y}(\,\cdot\,, y)$ to $\mathcal{X}_0^s$ (as defined in (3.4)). It follows from the results of section 4 that $P_{X,s|Y}(\,\cdot\,, y)$ coincides with $\tilde{P}_{X,s}$ when $\tilde{\mu} = \mu_Y(\,\cdot\,, y)$ and $u(\,\cdot\,, t) = u_*(\,\cdot\,, t, y)$ for $0 \le t \le s$. As shown in the discussion following (3.13), this is the unique probability measure on $\mathcal{X}_0^s$ minimizing the apparent information (3.14). The sum of the first two terms on the right-hand side of (3.14) is the apparent information of $\tilde{P}_{X,s}$ in the context of estimators of $(X_t, 0 \le t \le s)$ given observations $(Y_t, 0 \le t \le s)$, which we can think of as being the apparent information *up to* time s. The third term on the right-hand side

of (3.14) is the information in the observations $(Y_t - Y_s, s \le t \le T)$, which we can think of as being the information *remaining* in the observations Y at time s. As s increases, the estimator corresponding to $(\tilde{\mu}, u)$ progressively converts observation information into apparent information. If $u = u_*(\cdot\,, \cdot\,, y)$, then this process is conservative, in that $F_s(\tilde{P}_{X,s}, y)$ does not change with s. However, if u is not optimal, then the apparent information can increase faster than the observation information decreases.

We can refine this argument as follows. Let

$$\tilde{I}_s = \log\left(\frac{d\tilde{P}_{X,s}}{dP_{X,s}}(\tilde{X})\right) + H_p(0, s, \tilde{X}_0, \tilde{X}, y) + v(\tilde{X}_s, s, y) \quad \text{for } 0 \le s \le T, \tag{6.1}$$

where H_p is defined in (3.10). Then it follows from (3.11) that, for all $0 \le s \le t \le T$,

$$\begin{aligned}\tilde{I}_t = \tilde{I}_s + \log\left(\frac{d\tilde{P}_{X,t}}{dP_{X,t}} \times \frac{dP_{X,s}}{d\tilde{P}_{X,s}}(\tilde{X})\right) + H_p(s, t, \tilde{X}_s, (\tilde{X}_r, s \le r \le T), y)\\ + v(\tilde{X}_t, t, y) - v(\tilde{X}_s, s, y).\end{aligned} \tag{6.2}$$

Let $\tilde{Q}_X$ and Q_X be, respectively, the distributions of $(X_r^{z,s}, s \le r \le t)$ (as defined in section 3) with and without the application of the control $(u(X_r^{z,s}, r), s \le r \le t)$. The apparent information of $\tilde{Q}_X$ in the context of estimators for $(X_r^{z,s}, s \le r \le t)$ given $Y^{z,s}$ is

$$\begin{aligned}F_{s,t}(z, \tilde{Q}_X, y) &= h(\tilde{Q}_X \,|\, Q_X) + \langle H_p(s, t, z, \cdot\,, y), \tilde{Q}_X\rangle + \langle v(\chi_t, t, y), \tilde{Q}_X\rangle,\\ &= v(z, s, y) + \frac{1}{2}\int_s^t \int_{\mathbb{R}^n} |\sigma'(u - u_*(\tilde{z}, r, y))|^2\, \tilde{Q}_X(\chi_r^{-1}(d\tilde{z}))\, dr,\end{aligned} \tag{6.3}$$

where we have used (2.10). It now follows that

$$\tilde{\mathbf{E}}(\tilde{I}_t \,|\, \tilde{\mathcal{F}}_s) = \tilde{I}_s + \frac{1}{2}\int_s^t \tilde{\mathbf{E}}\left(|\sigma'(u - u_*)(\tilde{X}_r, r, y)|^2 \,\middle|\, \tilde{\mathcal{F}}_s\right) dr.$$

Thus $(\tilde{I}_t, \tilde{\mathcal{F}}_t)$ is a submartingale and a martingale if $u = u_*(\cdot\,, \cdot\,, y)$. This is the Davis–Varaiya [4] characterization of the optimal control for the problem of section 4.

Setting $t = s + \delta s$ in (6.3), we obtain the following local information quantities:

$$h(\tilde{Q}_X | Q_X) = \frac{1}{2}|\sigma' u(z, s)|^2 \delta s + o(\delta s), \tag{6.4}$$

$$\langle H_p(s, s + \delta s, z, \cdot\,, y), \tilde{Q}_X\rangle = -g(z)'\delta y + \frac{1}{2}|g(z)|^2 \delta s + o(\delta s), \tag{6.5}$$

$$\begin{aligned}\langle v(\chi_{s+\delta s}, s + \delta s, y), \tilde{Q}_X\rangle &= v(z, s, y) + g(z)'\delta y\\ &\quad - \left(\left(u - \frac{1}{2}u_*\right)' a u_* + \frac{1}{2}|g|^2\right)(z, s, y)\delta s + o(\delta s).\end{aligned} \tag{6.6}$$

Equation (6.4) shows the local increase in information gain of the distribution of the process (4.29) over P_X, (6.5) shows the local increase in the residual information of the estimator $\tilde{P}_X$, and (6.6) shows the local decrease in the average information remaining in the observation after time s. If y is differentiable at s, then there is a local rate of increase of apparent information of $|\sigma' u(z, s)|^2/2 - (\dot{y}_s - g/2)' g(z)$ and a local rate of

decrease of the remaining observation information of $(u - u_*/2)'au^*(z,s,y) - (\dot{y}_s - g/2)'g(z)$. The former exceeds the latter unless the control is optimal.

The dual problem can also be localized in this way. For u as above, let $\tilde{H}_p$ be a measurable function such that

$$\tilde{H}_p(s,t,z,X^{z,s}) = -\int_s^t u'\sigma(X_r^{z,s},r)\,dV_r + \frac{1}{2}\int_s^t |\sigma' u(X_r^{z,s},r)|^2\,dr + (K_s - K_t), \tag{6.7}$$

where K is differentiable and $K_T = 0$. This can be thought of as being the equivalent of $H_p(s,t,z,X^{z,s},y)$ for an unspecified time-distributed observation such that at time s the remaining information in the observation is K_s. (This corresponds to $\tilde{H}(X)$ of (5.1) with $K = K_0$.) Let Q_X^* be the distribution of $(X_r^{z,s}, s \le r \le t)$ when it is controlled by the optimal control. Taking expectation with respect to Q_X^* in (6.7) and taking the limit as $t \downarrow s$, we obtain a local rate of decrease of compatible information of $(u_* - u/2)'au(z,s,y)$. The local rate of increase of the information gain of $P_{X|Y}(\,\cdot\,,y)$ is, of course, $|\sigma' u_*(z,s,y)|^2/2$. The latter exceeds the former unless u is optimal.

In the global dual problem (5.1), the regular conditional probability $P_{X|Y}(\,\cdot\,,y)$ is the source of information. At time s the information in this source is

$$S_s = h(\tilde{\mu}|\mu) + \frac{1}{2}\int_0^s \int_{\mathbb{R}^n} |\sigma' u_*(z,t,y)|^2 P_{X|Y}(\chi_t^{-1}(dz),y)\,dt.$$

At time T there is no information in the observation and no residual information—all the information is still in the source. As s decreases, information flows out of the source at a rate $\dot{S}_s$; it is merged with residual information and flows into the observation at a rate $\dot{K}_s$. If u is optimal, then the flow is conservative, whereas more generally information is lost.

Let $\mathcal{H}_{z,s}$ be the Hilbert space of n-vectors of reals with inner product

$$\langle \alpha, \beta \rangle_{z,s} = \alpha' a(z,s)\beta.$$

The developments above show that the regular conditional probability $P_{X|Y}(\,\cdot\,,y)$ is locally characterized at the point (z,s) by the diffusion coefficients $a(z,s)$ and $(b(z,s) + a(z,s)\alpha_*)$, where α_* minimizes

$$\frac{1}{2}\|\alpha\|_{z,s}^2 - \langle \alpha, u_*(z,s,y)\rangle_{z,s}, \tag{6.8}$$

whereas the optimal trajectory in the dual problem (5.3) is locally characterized in that its negative gradient at the point (z,s), β_* maximizes

$$\langle \beta, u_*(z,s,y)\rangle_{z,s} - \frac{1}{2}\|\beta\|_{z,s}^2. \tag{6.9}$$

The local balance of the Bayesian path estimator is thus characterized by the Legendre transform pair (6.8), (6.9). Of course, this is the characterization of the optimal control problem of section 4 provided by the stochastic maximum principle, the adjoint process being the gradient of the optimal dual state, $v(\,\cdot\,,\,\cdot\,,y)$, evaluated at $(\tilde{X}_t, t)$.

7. Conclusions. This article has developed dual variational characterizations of Bayesian estimation, in which the "cost" functionals have particular information-theoretic meaning. These characterizations provide a natural framework for the study of modeling and approximation errors in estimators such as nonlinear filters. They also link such issues with a broader theory of "stochastic dissipativeness" (see [1]), on which the ideas and techniques of statistical physics can be brought to bear. We believe that this will have a number of advantages, for example, in the study of the long-term behavior of stochastic systems. For a recent development of this type see [12]. The characterizations also provide a framework for the representation of estimators, in a broader context, as apparent information minimizers and compatible information maximizers. These issues will be explored elsewhere.

Acknowledgments. The authors would like to thank one of the referees and the Associate Editor for their comments and suggestions which have led to substantial improvement in the paper. The second author would like to thank Denis Talay and Francois LeGland for their hospitality during his visits to the INRIA research groups Omega and Sigma2.

REFERENCES

[1] V. S. Borkar and S. K. Mitter, *A note on stochastic dissipativeness*, in Directions in Mathematical Systems Theory and Optimization, A. Rantzer and C. I. Byrnes, eds., Springer-Verlag, New York, 2002, pp. 41–49.

[2] J. M. C. Clark, *The design of robust approximations to the stochastic differential equations of nonlinear filtering*, in Communication Systems and Random Process Theory, NATO Advanced Study Inst. Ser., E: Appl. Sci. 25, J. K. Skwirzynski, ed., Sijthoff and Noordhoff, Alphen aan den Rijn, 1978, pp. 721–734.

[3] M. H. A. Davis, *A pathwise solution of the equations of nonlinear filtering*, Theory Probab. Appl., 27 (1983), pp. 167–175.

[4] M. H. A. Davis and P. P. Varaiya, *Dynamic programming conditions for partially observable stochastic systems*, SIAM J. Control, 11 (1973), pp. 226–261.

[5] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, Wiley, New York, 1997.

[6] W. H. Fleming and S. K. Mitter, *Optimal control and nonlinear filtering for nondegenerate diffusion processes*, Stochastics, 8 (1982), pp. 63–77.

[7] A. Friedman, *Stochastic Differential Equations and Applications, Vol.* 1, Academic Press, New York, 1975.

[8] H.-O. Georgii, *Gibbs Measures and Phase Transitions*, de Gruyter, Berlin, 1988.

[9] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes* 1*–General Theory*, Springer-Verlag, New York, 1977.

[10] T. Mikami, *Dynamical systems in the variational formulation of the Fokker–Planck equation by the Wasserstein metric*, Appl. Math. Optim., 42 (2000), pp. 203–227.

[11] S. K. Mitter and N. J. Newton, *The duality between estimation and control*, in Optimal Control and Partial Differential Equations, J. L. Menaldi, E. Rofman, and A. Sulem, eds., IOS Press, Amsterdam, 2000.

[12] S. K. Mitter and N. J. Newton, *Information flow and entropy production in the Kalman–Bucy filter*, submitted.

[13] N. J. Newton, *Observation sampling and quantisation for continuous-time estimators*, Stochastic Process. Appl., 87 (2000), pp. 311–337.

[14] J. Picard, *Robustesse de la solution des problemes de filtrage avec bruit blanc independant*, Stochastics, 13 (1984), pp. 229–245.

[15] L. C. G. Rogers and D. Williams, *Diffusions, Markov Processes and Martingales: Part* 2*—Itô Calculus*, Wiley, New York, 1986.

[16] B. M. Miller and W. J. Runggaldier, *Optimization of observations: A stochastic control approach*, SIAM J. Control Optim., 35 (1997), pp. 1030–1052.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1834–1853

NOISE ASSISTED HIGH-GAIN STABILIZATION: ALMOST SURELY OR IN SECOND MEAN*

HANS CRAUEL†, IAKOVOS MATSIKIS‡, AND STUART TOWNLEY‡

Abstract. For a linear control system with multiplicative white noise, we develop (asymptotic) formulas for the dependence of almost sure and second mean exponential growth rates on a high-gain parameter k. We show that if the diffusion matrix is skew-symmetric so that the noise enters in a purely skew-symmetric way, then the function g, where $g(p)/p$ denotes the exponential growth rate of the pth mean, converges to a straight line, uniformly for $p \in [0, 2]$, as $k \to \infty$. We use these formulas to show that the feedback control system in Stratonovich form is high-gain stabilizable even if the zero-dynamics are unstable, provided that the noise is strong enough. This contrasts with the noise free case, where we need the zero-dynamics to be exponentially stable.

We then consider a class of systems where the diffusion matrix is not skew-symmetric and show that the almost sure and pth mean growth rates have different limiting behavior as $k \to \infty$.

Key words. Lyapunov exponents, high-gain feedback, moment exponents, second mean exponents, Furstenberg–Khasminskii formula

AMS subject classifications. Primary, 93E15; Secondary, 37H15, 60H10, 93D15, 93D21

DOI. 10.1137/S0363012901393900

1. Introduction.

The dependence of dynamical properties of systems on parameters is central to many problems in control theory and in dynamical systems. For example, consider a linear single-input single-output control system of the form

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx, \end{aligned} \tag{1}$$

where $x \in \mathbf{R}^d$, A is a real $d \times d$-matrix, and $B, C^T \in \mathbf{R}^d$. A basic control design technique is to study the root locus for (1), i.e., the eigenvalues of $A - kBC$ as k varies. The root locus technique is used, for example, to show that if $CB > 0$ and if (A, B, C) is minimum phase, then high-gain feedback control $u = -ky$ is stabilizing in the sense that the eigenvalues of $A - kBC$ are in the left half-plane for all k sufficiently large.

The parametric dependence of dynamical properties for linear stochastic differential equations (LSDEs) has been investigated with great intensity during the last decades. For a survey on approaches which are closest in spirit to the present work, as well as for further references, see Wihstutz [10]. Note also that the bifurcation behavior of a noisy Duffing–van der Pol-oscillator has been studied in Arnold [1, Chapter 9]. Motivated by this, Imkeller and Lederer [7, 8] performed a more detailed study of the dependence on parameters of Lyapunov exponents for a noisy damped harmonic oscillator.

*Received by the editors August 17, 2001; accepted for publication (in revised form) May 23, 2003; published electronically December 17, 2003.

http://www.siam.org/journals/sicon/42-5/39390.html

†Institut für Mathematik, Technische Universität Ilmenau, Weimarer Straße 25, 98693 Ilmenau, Germany (hans.crauel@tu-ilmenau.de). The work of this author was financed by EPSRC grant GR/M98357.

‡School of Mathematical Sciences, Laver Building, University of Exeter, Exeter EX4 4QE, UK (imat@maths.ex.ac.uk, townley@maths.ex.ac.uk). The work of Iakovos Matsikis was supported by a University of Exeter Ph.D. studentship.

We want to investigate dynamical properties of certain LSDEs obtained by applying proportional feedback. More precisely, we consider a noisy version of (1) with proportional output feedback, i.e.,

$$dx = (A - kBC)x\,dt + \sum_{j=1}^{m} A_j x * dW_j(t). \tag{2}$$

We are interested in the dependence of exponential growth rates for (2) on k and, in particular, on asymptotic formulas, valid for large k. While for deterministic systems this can be derived by studying the dependence of eigenvalues on parameters, for LSDEs the situation is more complicated. Indeed, there are several competing notions of growth rates (for example, in the pth mean or almost surely). Furthermore, the growth rates depend on which notion of solution of (2) is adopted—Itô or Stratonovich.

The paper is organized as follows. In section 2 we recall some basic facts about LSDEs and their exponential growth rates, specifically Lyapunov exponents and moment exponents. In section 3 we consider almost sure growth rates, and we characterize under which conditions (2) can be stabilized almost surely by high-gain feedback. This means that the leading Lyapunov exponent becomes negative for all sufficiently high gain. We restrict ourselves to the simplest possible nontrivial case, which is the case of a 2×2-system. To calculate the almost sure growth rate, i.e., the leading Lyapunov exponent, we use the Furstenberg–Khasminskii formula, which yields a closed, albeit complicated, formula. These results are in the spirit of those obtained by Imkeller and Lederer [7, 8], who obtained explicit formulas for Lyapunov exponents of certain linear two-dimensional systems arising from the linearization of the Duffing–van der Pol-oscillator around zero.

In section 4 we consider the second mean growth rates and characterize under which conditions the system can be stabilized in the second mean by high-gain feedback. Here we use the classical technique of considering the norm induced by a positive definite matrix P (thus defining a Lyapunov function), applying the Itô formula to obtain estimates for the exponential growth of second moments, and then choosing the matrix P in an appropriate way. As a result we obtain that when the noise enters the system in a purely skew-symmetric way, then for high gain the exponential growth rate of the second moment approaches twice the Lyapunov exponent, the almost sure exponential growth rate. When the noise is entering nonskew-symmetrically, however, the growth rates have different limiting behavior as $k \to \infty$. We prove the main result, Theorem 3.2, in an appendix.

2. Notions of growth rates for LSDEs. To introduce the notions of growth rate for (2), consider a general LSDE

$$dx = Ax\,dt + \sum_{j=1}^{m} A_j x * dW_j(t), \tag{3}$$

where A and A_j are $d \times d$-matrices, W_j are independent standard Wiener processes, $1 \le j \le m$, and $*$ stands for an interpretation of (3) either as an Itô or as a Stratonovich equation. Denote by $x(t, x_0)$ the solution of (3) at time t, with initial condition x_0 at time $t = 0$.

There are several nonequivalent notions of exponential growth rates for (3). The

leading *Lyapunov exponent* of (3) is defined as

$$\lambda = \lim_{t\to\infty} \frac{1}{t} \log\Big(\sup_{\|x_0\|=1} \|x(t;x_0)\| \Big) \qquad P\text{-a. s.}, \tag{4}$$

where $\|\cdot\|$ is any norm on $\mathbf{R}^d$. Existence of the almost sure limit and the fact that the limit is constant almost surely follows from the subadditive ergodic theorem of Kingman [9]. The LSDE (3) is said to be *almost surely exponentially stable* if $\lambda < 0$.

The exponential growth rate in the pth mean is $g(p)/p$, where $g(p)$ is given by

$$g(p) = \lim_{t\to\infty} \frac{1}{t} \log E\Big(\sup_{\|x_0\|=1} \|x(t;x_0)\|^p \Big). \tag{5}$$

In (5), E denotes expectation. The function $p \mapsto g(p)$ is a real analytic function and convex, and $g'(0) = \lambda$, provided certain nondegeneracy conditions are satisfied; see Arnold, Oeljeklaus, and Pardoux [5] (and also Arnold, Kliemann, and Oeljeklaus [4] for the case of colored noise). Consequently,

$$\frac{g(p)}{p} \geq \frac{g(q)}{q} \geq g'(0) = \lambda \tag{6}$$

for $p \geq q > 0$. In particular, the growth rate in the pth mean is greater than or equal to the almost sure growth rate for $p > 0$. We are interested in the dependence of the growth rates given by (4) and (5) for the LSDE (2) as we vary the feedback gain k. For example, how does the feedback gain influence the difference between $g(2)$ and 2λ?

As mentioned earlier in the introduction, the dependence of growth rates of LSDEs on parameters has been of interest. One of the earliest results of this type, of relevance here, goes back to Arnold, Crauel, and Wihstutz [2]. Consider a Stratonovich equation of the form

$$dx = Ax\,dt + \sum_{j=1}^{m} u_j B_j x \circ dW_j(t), \tag{7}$$

where B_j, $1 \leq j \leq m$, form a basis for the $d(d-1)/2$-dimensional linear space of skew symmetric $d \times d$-matrices and, as before, the W_j are independent standard Wiener processes. In [2], the following limiting behavior for the leading Lyapunov exponent is obtained:

$$\lambda \to \frac{1}{d} \operatorname{tr} A = \frac{1}{d} \sum_{i=1}^{d} a_{ii} \qquad \text{as} \quad \min u_j \to \infty.$$

This result has been subsequently generalized by Arnold, Eizenberg, and Wihstutz [3] to allow for weaker conditions on the (B_j). One consequence of this result is the surprising observation that we can "stabilize" (7) by high-intensity noise if $\operatorname{tr} A < 0$. So, given a deterministic system $\dot{x} = Ax$ with $\operatorname{tr} A < 0$, it suffices to agitate the system by noise as in (7) and increase the intensity u until almost sure stability is achieved.

While high-intensity noise would seem impractical, this result motivates us to determine if noise can enhance more traditional stabilization by proportional feedback.

Second mean feedback stabilization of a general controlled Itô LSDE

$$\left\{\begin{array}{rcl} dx &=& (Ax + Bu_0)\,dt + \displaystyle\sum_{j=1}^{m}(A_j x + B_j u_j)\,dW_j(t), \\ y &=& Cx \end{array}\right. \tag{8}$$

has been considered by Damm [6]. Here $u_0, \dots, u_m$ are controls, and A, A_j, B, B_j, and C are matrices of suitable dimensions. This problem is quite involved. It turns out that stabilization in the second mean is equivalent to the existence of positive definite solutions to certain generalized linear matrix inequalities. We are not aware of any similar results for almost sure stabilization. Rather than pursuing this general feedback problem, we limit our interest to the parametric dependence of the growth rates λ and $g(p)$ in a set-up less general than (8). We restrict our attention to the 2×2 case of system (2). This is the simplest nontrivial case, and it allows for a neat and clear formulation of the main results. Higher dimensional cases would also be possible, but the technical conditions are considerably more involved, and we currently do not have a complete picture.

3. Almost sure exponential growth rates. As already mentioned, we will investigate the simplest case possible. That is, we investigate the dependence of the almost sure exponential growth rate, i.e., the Lyapunov exponent, on the parameters k and σ for the Stratonovich equation

$$dx = \begin{pmatrix} a-k & b \\ c & d \end{pmatrix} x\,dt + \sigma \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} x \circ dW(t). \tag{9}$$

In (9), $a, b, c, d \in \mathbf{R}$ are fixed parameters. The drift matrix $\left(\begin{smallmatrix} a-k & b \\ c & d \end{smallmatrix}\right)$ in (9) arises via a change of coordinates from (1) in the case $u = -ky$, and $CB > 0$, if the dimension is 2. We show that $\lim_{k\to\infty} \lambda_{k,\sigma} = \frac{1}{2}(2d - \sigma^2)$, so that (9) is high-gain almost surely exponentially stabilizable if and only if $d < \frac{1}{2}\sigma^2$.

Remark 3.1. Without high-gain k, we have from Arnold, Crauel, and Wihstutz [2] that $\lim_{\sigma\to\infty} \lambda = \frac{1}{2}(a+d)$, and so almost sure stability if mixing of the negative trace by the noise is strong enough. With high gain but no noise, high-gain stabilization is possible only if the zero-dynamics are exponentially stable, i.e., if $d < 0$, while the noise term has a stabilizing effect and with high enough gain k the intensity of the noise has to be only strong enough to overcome the influence of possibly "unstable zero-dynamics."

While the results are simply stated, their derivation, as with analogous results obtained by Imkeller and Lederer [7, 8], is quite involved. To proceed we first recall some details about the well-known method for calculating Lyapunov exponents of an LSDE

$$dx = Ax\,dt + \sum_{j=1}^{m} A_j x \circ dW_j(t) \tag{10}$$

via the Furstenberg–Khasminskii formula. Here A, A_j are $d \times d$-matrices, and W_j are independent Wiener processes, $1 \le j \le m$, and the equation is interpreted in the Stratonovich sense.

Projection of (10) from $\mathbf{R}^d \setminus \{0\}$ onto the unit sphere $S^{d-1} = \{v \in \mathbf{R}^d : |v| = 1\}$

by $x \mapsto x/|x| =: s$ gives the (nonlinear) SDE

$$ds = g_A(s)\,dt + \sum_{j=1}^{m} g_{A_j}(s) \circ dW_j(t) \tag{11}$$

on S^{d-1}, where $g_A(s) = As - (s, As)s$. The SDE (11) defines a random dynamical system on S^{d-1}. Associated to every Lyapunov exponent there exists an invariant measure for this random dynamical system, which is supported by the Oseledets space associated with this Lyapunov exponent; for details see Arnold [1, Chapters 3 and 4]. Furthermore, the maximal Lyapunov exponent (almost sure exponential growth rate) is given by

$$\lambda = \int_{S^{d-1}} \left((s, As) + \sum_{j=1}^{m} \left(\frac{1}{2}((A_j + A_j^*)s, A_j s) - (s, A_j s)^2 \right) \right) dp(s), \tag{12}$$

where p is a (suitable) invariant measure for the Markov semigroup induced by (11). Equation (12) is the Furstenberg–Khasminskii formula in the form it takes for a linear system induced by the LSDE (10). In particular, if (11) is sufficiently nondegenerate, then there exists a unique invariant Markov measure p, and this measure has a smooth density with respect to the Lebesgue measure on S^{d-1}, which we denote by p again. Nondegeneracy means that a certain hypoellipticity condition is satisfied, whose precise form is not of interest here. The density p is given as a suitably normalized solution of the associated Fokker–Planck equation. See Arnold [1] or Imkeller and Lederer [7, 8].

In our particular case we have the Stratonovich LSDE

$$dx = \begin{pmatrix} a-k & b \\ c & d \end{pmatrix} x\,dt + \sigma \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} x \circ dW(t),$$

and (11) becomes

$$ds = g_A(s)\,dt + g_B(s) \circ dW(t)$$

with $g_A(s) = As - (s, As)s$,

$$\begin{aligned} (s, As) &= (a-k)s_1^2 + (b+c)s_1 s_2 + d s_2^2, \\ g_A(s) &= \begin{pmatrix} (a-k)s_1 + b s_2 - (a-k)s_1^3 - (b+c)s_1^2 s_2 - d s_2^2 s_1 \\ c s_1 + d s_2 - (a-k)s_1^2 s_2 - (b+c)s_1 s_2^2 - d s_2^3 \end{pmatrix}, \\ g_B(s) &= \sigma \begin{pmatrix} -s_2 \\ s_1 \end{pmatrix}, \end{aligned}$$

and $s = (s_1, s_2)^T$. In polar coordinates $s = (\cos\varphi, \sin\varphi)^T$ with $\varphi \in [-\pi/2, \pi/2]$, this becomes

$$\begin{aligned} d\varphi &= \frac{ds_2}{\cos\varphi} \\ &= \frac{c\cos\varphi + d\sin\varphi - (a-k)\cos^2\varphi\sin\varphi - (b+c)\cos\varphi\sin^2\varphi - d\sin^3\varphi}{\cos\varphi}\,dt \\ &\quad + \sigma\,dW(t) \\ &= \left(c\cos^2\varphi + (d-a+k)\cos\varphi\sin\varphi - b\sin^2\varphi \right) dt + \sigma\,dW(t). \end{aligned}$$

To determine the invariant measure for the associated Markov semigroup we first note that the SDE for the angle φ is elliptic. Consequently, there is a unique invariant measure for the Markov semigroup, and this invariant measure has a C^∞ density $\varphi \mapsto p_{k,\sigma}(\varphi)$, which is a solution of the Fokker–Planck equation. In the present case the Fokker–Planck equation results in the ordinary differential equation

$$-\left(\frac{1}{2}\sigma^2 p\right)'' + \Big(\big(c\cos^2\varphi + (d-a+k)\cos\varphi\sin\varphi - b\sin^2\varphi\big)p\Big)' = 0$$

with periodic boundary conditions on $[-\pi/2, \pi/2)$. This gives

$$p' = \frac{2}{\sigma^2}(c + (d-a+k)\cos\varphi\sin\varphi - (b+c)\sin^2\varphi)p + \gamma,$$

where γ has to be chosen such that p is periodic. Rewriting in 2φ-terms gives

$$\begin{aligned} p' &= \frac{2}{\sigma^2}\left(c + \frac{1}{2}(d-a+k)\sin 2\varphi - \frac{1}{2}(b+c)(1-\cos 2\varphi)\right)p + \gamma \\ &= \frac{1}{\sigma^2}\Big((c-b) + (d-a+k)\sin 2\varphi + (b+c)\cos 2\varphi\Big)p + \gamma. \end{aligned}$$

Using standard trigonometric identities and setting

$$\begin{aligned} R(\varphi,\eta) = \frac{1}{\sigma^2}\Bigg((c-b)(\varphi-\eta) + \frac{1}{2}(d-a+k)(\cos 2\eta - \cos 2\varphi) \\ + \frac{1}{2}(b+c)(\sin 2\varphi - \sin 2\eta)\Bigg), \end{aligned}$$

we obtain that the general, still to be normalized, solution p is given by

$$p(\varphi) = \mathrm{e}^{R(\varphi,-\frac{\pi}{2})}\, p\left(-\frac{\pi}{2}\right) + \gamma\int_{-\pi/2}^{\varphi} \mathrm{e}^{R(\varphi,\eta)}\, d\eta. \tag{13}$$

In (13) γ has to be chosen such that $p(\frac{\pi}{2}) = p(-\frac{\pi}{2})$. For the leading Lyapunov exponent we then obtain

$$\lambda = \int_{-\pi/2}^{\pi/2}\Big((a-k)\cos^2\varphi + (b+c)\cos\varphi\sin\varphi + d\sin^2\varphi\Big)\, p(\varphi)\, d\varphi. \tag{14}$$

Starting from (13) and (14), we obtain the following theorem on the dependence of the Lyapunov exponent on high-gain feedback.

THEOREM 3.2. *The Lyapunov exponent $\lambda = \lambda_{k,\sigma}$ of the Stratonovich LSDE* (9) *satisfies*

$$\lambda_{k,\sigma} = d - \frac{\sigma^2}{2} + O(k^{-1})$$

for k large. In particular,

$$\lim_{k\to\infty} \lambda_{k,\sigma} = d - \frac{\sigma^2}{2}$$

for every a, b, c, d.

The proof is given in the appendix.

Remark 3.3. (i) We used the explicit formula for the leading Lyapunov exponent $\lambda = \lambda_{k,\sigma}$ given by (14), invoking the density $p = p(k)$ given explicitly by (13). Use of such explicit formulas is not possible for higher dimensional LSDEs. For $d > 2$ one might invoke the more systematic approach as described, for example, by Wihstutz [10]. This involves putting $\varepsilon = k^{-\alpha}$ for suitable $\alpha > 0$ and denoting the generator induced by the projected SDE (11) by L_ε to obtain an expansion of $\lambda = \lambda(\varepsilon)$ around $\varepsilon = 0$ by expanding the expression $\lambda(\varepsilon) = \int q_\varepsilon\, p_\varepsilon$ given by (12), where p_ε is determined by $L_\varepsilon p_\varepsilon = 0$ (which is the Fokker–Planck equation). Performing this asymptotic expansion formally for the two-dimensional case considered here yields the limiting behavior $\lambda = \lambda_{k,\sigma} + o(1)$ for k large in a more intuitive manner than the direct approach which we adopt. However, it would seem from the developments in [10] that justification of this more systematic asymptotic expansion approach needs arguments which are at least as complicated as those used in the direct approach.

(ii) Using a slightly modified approach to the one invoked in the proof of Theorem 3.2 gives more information, specifically allowing a further expansion of the Lyapunov exponent λ in k. This gives

$$\lambda_{k,\sigma} = d - \frac{\sigma^2}{2} + \left(bc - \frac{\sigma^4}{4}\right) k^{-1} + O(k^{-2}). \tag{15}$$

Alternatively we obtain the same result by using the formal expansion approach as described above in (i).

4. Exponential growth rates in the second mean. In the calculation of Lyapunov exponents in the previous section we adopted a Stratonovich interpretation. To calculate the exponential growth rate of the second mean, i.e., $g(2)$, it is more suitable to work with an Itô interpretation. So in order to compare results between the two notions of growth rate, we need to transform (9) from Stratonovich to Itô form. A general Stratonovich LSDE

$$dx = Ax\, dt + \sum_{j=1}^{m} A_j x \circ dW_j(t)$$

is equivalent to the Itô LSDE

$$dx = \left(A + \frac{1}{2}\sum_{j=1}^{m} A_j^2\right) x\, dt + \sum_{j=1}^{m} A_j x\, dW_j(t).$$

In our case this transforms (9) to

$$dx = \begin{pmatrix} a - k - \frac{1}{2}\sigma^2 & b \\ c & d - \frac{1}{2}\sigma^2 \end{pmatrix} x\, dt + \sigma \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} x\, dW(t).$$

For an Itô equation

$$dx = A_0 x\, dt + A_1 x\, dW(t),$$

$g(2)$ coincides with the smallest $\eta \in \mathbf{R}$ such that there exists a symmetric positive definite matrix P with

$$A_0^T P + P A_0 + A_1^T P A_1 \leq \eta P. \tag{16}$$

Indeed this follows from (16) by adopting $\|x\| = \langle x, Px\rangle^{1/2}$ as a norm in the definition of $g(2)$ and by using the Itô formula applied to $V(t) = \langle x(t), Px(t)\rangle$.

PROPOSITION 4.1. *The exponential growth rate of the second mean of the Stratonovich LSDE* (9)*, which is* $g(2)/2$ *with* $g(2) = g_{k,\sigma}(2)$ *given by* (5)*, satisfies*

$$g_{k,\sigma}(2) = (2d - \sigma^2) + O(k^{-1}).$$

Hence, for σ *fixed,*

$$\lim_{k\to\infty} g_{k,\sigma}(2) = 2d - \sigma^2$$

for every a, b, c, d.

Proof. From (6) we already know that $g_{k,\sigma}(2) \geq 2\lambda_{k,\sigma}$ for every k, σ; hence Theorem 3.2 yields

$$g_{k,\sigma}(2) \geq (2d - \sigma^2) + O(k^{-1}) \tag{17}$$

for $k \to \infty$. It remains to show that $g_{k,\sigma}(2) \leq (2d - \sigma^2) + O(k^{-1})$, i.e., that there exist η and $P > 0$ such that

$$\begin{bmatrix} a-k-\frac{1}{2}\sigma^2 & c \\ b & d-\frac{1}{2}\sigma^2 \end{bmatrix} P + P \begin{bmatrix} a-k-\frac{1}{2}\sigma^2 & b \\ c & d-\frac{1}{2}\sigma^2 \end{bmatrix} + \sigma^2 \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} P \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} < \eta P \tag{18}$$

and $\eta \leq (2d-\sigma^2)+O(k^{-1})$. Without loss of generality, we choose $P = \left[\begin{smallmatrix} 1 & q \\ q & p \end{smallmatrix}\right]$. Denoting by $Q(P)$ the left-hand side of (18), we obtain

$$Q(P) = \begin{bmatrix} 2a - 2k - (1-p)\sigma^2 + 2cq & (a - k + d - 2\sigma^2)q + b + cp \\ (a - k + d - 2\sigma^2)q + b + cp & 2dp + (1-p)\sigma^2 + 2bq \end{bmatrix}.$$

Choosing $p = p(k) = 1 + \frac{k}{\sigma^2}$, we obtain

$$Q(P) = \begin{bmatrix} 2a - k + 2cq & (a + d - k - 2\sigma^2)q + b + c(1 + \frac{k}{\sigma^2}) \\ (a + d - k - 2\sigma^2)q + b + c(1 + \frac{k}{\sigma^2}) & (2d - \sigma^2 + \frac{\sigma^2}{p} + \frac{2bq}{p})p \end{bmatrix},$$

where we left p unresolved in the lower right entry. Now choose, for k sufficiently large, $q = q(k) = \frac{b+cp}{k+2\sigma^2-a-d}$. For this choice of $P = P(k)$ we obtain

$$\begin{aligned} Q(P) &= \begin{bmatrix} 2a - k + 2cq & 0 \\ 0 & ((2d - \sigma^2) + \frac{\sigma^2}{p} + \frac{2bq}{p})p \end{bmatrix} \\ &\leq \left((2d - \sigma^2) + O(k^{-1})\right) P. \end{aligned} \tag{19}$$

Combining (17) and (19), we conclude that for $k \to \infty$ the second mean exponent $g_{k,\sigma}(2)$ satisfies

$$g_{k,\sigma}(2) = 2d - \sigma^2 + O(k^{-1})$$

as required. ∎

Remark 4.2. Further expansion of the second mean growth rate $g(2)$ in terms of k gives

$$g_{k,\sigma}(2) = 2d - \sigma^2 + \left(2bc + \frac{\sigma^4}{2}\right) k^{-1} + O(k^{-2}) = 2\lambda_{\kappa,\sigma} + \sigma^4 k^{-1} + O(k^{-2}).$$

This shows that the difference between almost sure and second mean exponents disappears asymptotically with order $1k$.

Until now we have assumed that the noise enters in a skew-symmetric way. For the following slightly more general case the arguments carry over in a rather straightforward manner.

THEOREM 4.3. *Consider the Stratonovich LSDE*

$$dx = \begin{pmatrix} a-k & b \\ c & d \end{pmatrix} x\,dt + \begin{pmatrix} \gamma & -\sigma \\ \sigma & \gamma \end{pmatrix} x \circ dW(t) \tag{20}$$

with $\gamma \in \mathbf{R}$. Then the following hold.

(i) *The almost sure exponential growth rate of* (20)*, the Lyapunov exponent $\lambda_{k,\sigma}$, is independent of γ. In particular,*

$$\lambda_{k,\sigma} = d - \frac{\sigma^2}{2} + O(k^{-1}), \qquad \text{and therefore} \quad \lim_{k\to\infty} \lambda_{k,\sigma} = d - \frac{\sigma^2}{2}$$

for every $\gamma \in \mathbf{R}$.

(ii) *The exponential growth rate of the second mean, $g_{k,\sigma,\gamma}(2)/2$, satisfies*

$$g_{k,\sigma,\gamma}(2) = 2d + 2\gamma^2 - \sigma^2 + O(k^{-1})$$

for large k. In particular,

$$\lim_{k\to\infty} g_{k,\sigma,\gamma}(2) = 2d + 2\gamma^2 - \sigma^2.$$

Note. The theorem can be rephrased as $\frac{g(2)}{2} = \lambda + \gamma^2$ asymptotically.

Proof. It is straightforward to see that the leading Lyapunov exponent of (20) is independent of γ. Indeed, since $\gamma\,\mathrm{Id}$ commutes with all matrices, solutions of (20) and (9) can be transformed into each other as follows. If $x(t;x_0)$ is a solution of (9), then $y(t;x_0) = x(t;x_0)\mathrm{e}^{\gamma\sigma W(t)}$ is a solution of (20). Since $\lim W(t)/t = 0$ almost surely for $t \to \infty$, the Lyapunov exponents of (20) are the same as those of (9). Consequently, application of Theorem 3.2 to (20) for $\gamma = 0$ proves (i).

Concerning the exponential growth rate of the second mean, denote by I the identity matrix, and put $J = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$. Transforming the Stratonovich SDE $dx = A_0 x\,dt + A_1 x \circ dW(t)$ with $A_1 = \gamma I + \sigma J$ to Itô form gives

$$dx = \left(A_0 + \frac{1}{2}A_1^2\right) x\,dt + \sigma A_1 x\,dW(t),$$

where $A_1^2 = (\gamma^2 - \sigma^2)I + 2\gamma\sigma J$. In order to determine $g(2)$ we have to find the smallest $\eta \in \mathbf{R}$ for which there exists a positive definite P such that

$$\begin{aligned} \left(A_0 + \frac{1}{2}(\gamma^2-\sigma^2)I + \gamma\sigma I\right)^T P + P\left(A_0 + \frac{1}{2}(\gamma^2-\sigma^2)I + \gamma\sigma I\right) \\ + \big(\gamma I + \sigma J\big)^T P\big(\gamma I + \sigma J\big) \le \eta P. \end{aligned} \tag{21}$$

Now the left-hand side of (21) equals

$$\left(A_0 + \left(\gamma^2 - \frac{\sigma^2}{2}\right) I + 2\gamma\sigma J\right)^T P + P\left(A_0 + \left(\gamma^2 - \frac{\sigma^2}{2}\right) I + 2\gamma\sigma J\right) + \sigma^2\, J^T P J,$$

which is simply (18) for the Itô LSDE

$$dx = \left(A_0 + \left(\gamma^2 - \frac{1}{2}\sigma^2\right) I + \gamma\sigma J\right) x\, dt + \sigma J x\, dW(t),$$

corresponding to the Stratonovich equation

$$dx = (A_0 + \gamma^2 I + \gamma\sigma J)x\, dt + \sigma J x \circ dW(t). \tag{22}$$

Since (22) has the same form as (9) and, in particular,

$$A_0 + \gamma^2 I + \gamma\sigma J = \begin{pmatrix} \tilde{a} - k & \tilde{b} \\ \tilde{c} & \tilde{d} \end{pmatrix}$$

with $\tilde{d} = d + \gamma^2$, it follows by applying Proposition 4.1 that

$$g(2) = 2d + 2\gamma^2 - \sigma^2 + O(k^{-1})$$

for $k \to \infty$, proving (ii). □

Remark 4.4. For the case of a skew-symmetric diffusion matrix we see from the calculations of λ and $g(2)$ that while

$$g(2) > 2\lambda$$

(the inequality is strict here—see Arnold, Oeljeklaus, and Pardoux [5]), in the limit, for high-gain k tending to infinity,

$$\lim_{k\to\infty} g_{k,\sigma}(2) = 2 \lim_{k\to\infty} \lambda_{k,\sigma} = 2d - \sigma^2.$$

This means that the strictly convex functions $p \mapsto g_{k,\sigma}(p)$ converge with k to infinity to the linear $p \mapsto (d - \frac{1}{2}\sigma^2)p$, uniformly in $p \in [0, 2]$, for every $\sigma > 0$. So the high-gain feedback leads to an asymptotic degeneracy in $p \mapsto g(p)$.

Remark 4.5. A similar degeneracy occurs in the case of the high-intensity noise problem of Arnold, Crauel, and Wihstutz [2]. Recall that for the Stratonovich equation

$$dx = \begin{pmatrix} a & b \\ c & d \end{pmatrix} x\, dt + \sigma \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} x \circ dW(t)$$

we have

$$\lim_{\sigma\to\infty} \lambda = \frac{1}{2}(a + d).$$

Transforming into the equivalent Itô equation

$$dx = \begin{pmatrix} a - \frac{1}{2}\sigma^2 & b \\ c & d - \frac{1}{2}\sigma^2 \end{pmatrix} x\, dt + \sigma \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} x\, dW(t)$$

and invoking the characterization of $g(2)$ given around (16), one obtains

$$\lim_{\sigma\to\infty} g(2) = (a + d).$$

To see this, we again need to find η so that

$$\tilde{Q}(P) := \begin{bmatrix} 2a-(1-p)\sigma^2+2cq & (a+d-\sigma^2)q+b+cp \\ (a+d-\sigma^2)q+b+cp & 2dp+(1-p)\sigma^2+2bq \end{bmatrix} \leq \eta \begin{bmatrix} 1 & q \\ q & p \end{bmatrix}.$$

Put $p = 1 + \frac{d-a}{\sigma^2}$. Then we need

$$\begin{bmatrix} a+d+2cq & (a+d-2\sigma^2)q+b+c(1+\frac{d-a}{\sigma^2}) \\ (a+d-2\sigma^2)q+b+c(1+\frac{d-a}{\sigma^2}) & a+d-d\frac{(a-d)}{\sigma^2}+2bq \end{bmatrix}$$
$$\leq \eta \begin{bmatrix} 1 & q \\ q & p \end{bmatrix}.$$

Choosing

$$q = \frac{b+c(1+(\frac{d-a}{\sigma^2}))}{2\sigma^2-a-d} = O\left(\frac{1}{\sigma^2}\right)$$

for $2\sigma^2 > a+d$ gives

$$\tilde{Q}(P) = \begin{pmatrix} a+d+O(\frac{1}{\sigma^2}) & 0 \\ 0 & a+d+O(\frac{1}{\sigma^2}) \end{pmatrix} \leq \left((a+d)+O\left(\frac{1}{\sigma^2}\right)\right) \begin{pmatrix} 1 & q \\ q & p \end{pmatrix},$$

so that $g_\sigma(2) \leq (a+d)+O(\sigma^{-2})$. Since $g_\sigma(2) \geq 2\lambda_\sigma$ and

$$\lim_{\sigma\to\infty} \lambda_\sigma = (a+d)/2,$$

we conclude that

$$\lim_{\sigma\to\infty} g_\sigma(2) = a+d.$$

Remark 4.6. In the case of a more general diffusion matrix considered in Theorem 4.3, the degeneration described in Remark 4.4 does not occur. In fact, here noise assisted high-gain stabilization may take place with respect to almost sure stability but not with respect to second mean stability. In particular, with a diffusion matrix of the form $\sigma \left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$ we have convergence of the almost sure exponential growth rate $\lambda_{k,\sigma}$ to $d-\sigma^2/2$ for $k \to \infty$, whereas $g_{k,\sigma}(2)/2$ converges to $d+\sigma^2/2$ for $k \to \infty$.

5. Concluding remarks. We have considered the dependence of growth rates on the feedback gain k for the simplest case of an LSDE arising from proportional feedback applied to a second order relative degree one control system. We have obtained explicit formulas for the Lyapunov exponents and asymptotics of the growth rates valid for large enough k. We have shown, in particular, that in case of a purely skew-symmetric noise the asymptotic dependence on k for k large is the same whether we consider λ, the leading Lyapunov exponent, or whether we consider the growth rate in the pth mean for any $p \in [0,2]$. This contrasts with the situation where for any fixed k we would have that

$$\lambda < \frac{g(p)}{p} < \frac{g(q)}{q}$$

for all $p < q$.

As an application of these asymptotic estimates we see that the Stratonovich equation (9) is high-gain stabilizable (almost surely or in the pth mean, $0 < p \leq 2$) if and only if

$$\sigma^2 > 2d.$$

In particular, for $\sigma \neq 0$ we can allow $d > 0$, i.e., unstable zero-dynamics, whereas if $\sigma = 0$, we need $d < 0$ i.e., the zero-dynamics to be exponentially stable. We see in this simple example that the noise has a *stabilizing effect*.

When the noise enters in a certain nonskew-symmetric way, then the same comments apply to almost sure growth rates, so the noise is still *stabilizing with respect to almost sure stability*. However, the same noise is *destabilizing with respect to stability in the second mean*.

Appendix. In this appendix we give the proof of Theorem 3.2. We first consider a special case.

The case of a symmetric drift matrix. We first treat the case $b = c$. This means that the drift matrix A in (9) is symmetric. Then $\varphi \mapsto R(\varphi, \eta)$ is periodic for every η, and therefore (13) yields that periodicity of p holds if and only if $\gamma = 0$. This gives

$$p(\varphi) = \exp\left[-\frac{1}{2\sigma^2}\Big((d-a+k)(1+\cos 2\varphi) - 2b\sin 2\varphi\Big)\right] p(-\pi/2).$$

In order to make p the density of a probability measure, we have to choose

$$p\left(-\frac{\pi}{2}\right) = \left(\int_{-\pi/2}^{\pi/2} \exp\left[-\frac{1}{2\sigma^2}\Big((d-a+k)(1+\cos 2\varphi) - 2b\sin 2\varphi\Big)\right] d\varphi\right)^{-1}.$$

Note that $p(-\frac{\pi}{2}) > 0$. Having determined the dependence of the invariant density p on k and σ, we now turn to the calculation of the leading Lyapunov exponent as given by (14), which here takes the form

$$\lambda = \int_{-\pi/2}^{\pi/2} \Big((a-k)\cos^2\varphi + 2b\cos\varphi\sin\varphi + d\sin^2\varphi\Big)\, p(\varphi)\, d\varphi. \tag{23}$$

We first ignore the normalizing factor $p(-\frac{\pi}{2})$ and consider

$$I := \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \big((a-k)\cos^2\varphi + 2b\cos\varphi\sin\varphi + d\sin^2\varphi\big) \times \exp\left(-\frac{1}{2\sigma^2}\big((d-a+k)(1+\cos 2\varphi) - 2b\sin 2\varphi\big)\right) d\varphi.$$

Defining $R_k = \sqrt{(k+d-a)^2 + (2b)^2}$ and ψ_k via $\tan\psi_k = \frac{k+d-a}{2b}$ and using the formula

$$s\cos 2\varphi + t\sin 2\varphi = R\sin(2\varphi + \psi)$$

with $R = \sqrt{s^2 + t^2}$ and $\tan\psi = s/t$, we obtain

$$I = \exp\left(\frac{a-k-d}{2\sigma^2}\right) \int_{-\pi/2}^{\pi/2} \Big[(a-k)\cos^2\varphi + 2b\cos\varphi\sin\varphi + d\sin^2\varphi\Big] \times \exp\left(-\frac{1}{2\sigma^2} R_k \sin(2\varphi + \psi_k)\right) d\varphi.$$

Rearranging

$$\begin{aligned}
&(a-k)\cos^2\varphi + 2b\cos\varphi\sin\varphi + d\sin^2\varphi \\
&= (a-k-d)\cos^2\varphi + 2b\cos\varphi\sin\varphi + d \\
&= \frac{a-k-d}{2}(\cos 2\varphi + 1) + b\sin 2\varphi + d \\
&= \frac{1}{2}\Big[(a-k-d)\cos 2\varphi + 2b\sin 2\varphi + a + d - k\Big]
\end{aligned}$$

and using periodicity of sine and cosine together with the identities $\sin(\beta + \pi/2) = \cos\beta$ and $\cos(\beta + \pi/2) = -\sin\beta$, we obtain

$$\begin{aligned}
I &= \frac{1}{2}\exp\Big(\frac{a-k-d}{2\sigma^2}\Big) \\
&\quad \times \int_{-\pi/2}^{\pi/2} \Big[(a-k-d)\cos(2\varphi - \psi_k) + 2b\sin(2\varphi - \psi_k) + a + d - k\Big] \\
&\qquad \times \exp\Big(-\frac{R_k}{2\sigma^2}\sin 2\varphi\Big)\, d\varphi \\
&= \frac{1}{2}\exp\Big(\frac{a-k-d}{2\sigma^2}\Big) \\
&\quad \times \int_{-\pi/2}^{\pi/2} \Big[-(a-k-d)\sin(2\varphi - \psi_k) + 2b\cos(2\varphi - \psi_k) + a + d - k\Big] \\
&\qquad \times \exp\Big(-\frac{R_k}{2\sigma^2}\cos 2\varphi\Big)\, d\varphi \\
&= \frac{1}{2}\exp\Big(\frac{a-k-d}{2\sigma^2}\Big)\exp\Big(\frac{R_k}{2\sigma^2}\Big) \\
&\quad \times \int_{-\pi/2}^{\pi/2} \Big[(k+d-a)\big(\sin 2\varphi\cos\psi_k - \cos 2\varphi\sin\psi_k\big) \\
&\qquad + 2b\big(\cos 2\varphi\cos\psi_k + \sin 2\varphi + \sin 2\varphi\sin\psi_k\big) + a + d - k\Big] \\
&\qquad \times \exp\Big(-\frac{R_k}{\sigma^2}\cos^2\varphi\Big)\, d\varphi.
\end{aligned}$$

Integrating over a symmetric interval around zero, we keep only integrals over even functions. This gives

$$\begin{aligned}
I &= \exp\Big(\frac{a-k-d+R_k}{2\sigma^2}\Big) \\
&\times \int_0^{\pi/2} \Big[(a-d-k)\cos 2\varphi\sin\psi_k + 2b\cos 2\varphi\cos\psi_k + a + d - k\Big] \\
&\qquad \times \exp\Big(-\frac{R_k}{\sigma^2}\cos^2\varphi\Big)\, d\varphi.
\end{aligned}$$

Now $\cos\psi_k = 2b/R_k$ and $\sin\psi_k = (k+d-a)/R_k$, so

$$\begin{aligned} I &= \exp\Big(\frac{a-k-d+R_k}{2\sigma^2}\Big) \\ &\quad\times \int_o^{\pi/2} \left[\frac{4b^2-(k+d-a)^2}{R_k}\cos 2\varphi + a + d - k\right] \exp\Big(-\frac{R_k}{\sigma^2}\cos^2\varphi\Big)\, d\varphi \\ &= \exp\Big(\frac{a-k-d+R_k}{2\sigma^2}\Big) \\ &\quad\times \int_0^{\pi/2} \left[a+d-k+\frac{(k+d-a)^2-4b^2}{R_k} - \frac{(k+d-a)^2-4b^2}{R_k} 2\cos^2\varphi\right] \\ &\qquad\times \exp\Big(-\frac{R_k}{\sigma^2}\cos^2\varphi\Big)\, d\varphi. \end{aligned}$$

From the theory of hypergeometric functions we know that

$$H := \int_0^{\pi/2} \exp\Big(-\frac{R_k}{\sigma^2}\cos^2\varphi\Big)\, d\varphi \;=\; \frac{\pi}{2}\mathcal{H}\left(\left[\frac{1}{2}\right], [1], \frac{R_k}{\sigma^2}\right)$$

and

$$K := \int_0^{\pi/2} \exp\Big(-\frac{R_k}{\sigma^2}\cos^2\varphi\Big)\cos^2\varphi\, d\varphi \;=\; \frac{\pi}{4}\mathcal{H}\left(\left[\frac{3}{2}\right], [2], \frac{R_k}{\sigma^2}\right),$$

where $\mathcal{H}$ denotes hypergeometric functions. This gives

$$\begin{aligned} I &= \exp\Big(\frac{a-k-d+R_k}{2\sigma^2}\Big) \\ &\quad\times \left(\left[a+d-k+\frac{(k+d-a)^2-4b^2}{\sqrt{(k+d-a)^2+4b^2}}\right] H - 2\left[\frac{(k+d-a)^2-4b^2}{\sqrt{(k+d-a)^2+4b^2}}\right] K\right). \end{aligned}$$

Using similar calculations, we obtain for the normalizing factor

$$\begin{aligned} p\left(-\frac{\pi}{2}\right) &= \pi\exp\Big(\frac{a-d-k+R_k}{2\sigma^2}\Big)\mathcal{H}\left(\left[\frac{1}{2}\right], [1], \frac{R_k}{\sigma^2}\right) \\ &= 2\exp\Big(\frac{a-d-k+R_k}{2\sigma^2}\Big) H. \end{aligned}$$

Now

$$\begin{aligned} a+d-k+\frac{(k+d-a)^2-4b^2}{\sqrt{(k+d-a)^2+4b^2}} &= 2d + e_1(k), \\ \frac{(k+d-a)^2-4b^2}{\sqrt{(k+d-a)^2+4b^2}} &= k + e_2(k), \\ H &= \frac{\pi}{2}\Big(\frac{\sigma^2}{R_k}\Big)^{1/2} + e_3(k), \\ K &= \frac{\pi}{4}\Big(\frac{\sigma^2}{R_k}\Big)^{3/2} + e_4(k), \quad \text{and} \\ R_k = \sqrt{(k+d-a)^2+(2b)^2} &= k + e_5(k), \end{aligned}$$

where $|e_1(k)| \le Mk^{-1}$, $|e_2(k)| \le M$, $|e_3(k)| \le Mk^{-3/2}$, $|e_4(k)| \le Mk^{-5/2}$, and $|e_5(k)| \le Mk^{-1/2}$ for a suitable constant M. In the following we will denote constants independent of k by M without always noting when their value changes. This gives

$$I = \pi \left(d - \frac{\sigma^2}{2} \right) \exp\left(\frac{a-d}{2\sigma^2} \right) \frac{\sqrt{\sigma^2}}{k^{1/2}} + e_6(k) \tag{24}$$

and, introducing the notion J for later reference,

$$J := p\left(-\frac{\pi}{2}\right) = \pi \exp\left(\frac{a-d}{2\sigma^2} \right) \frac{\sqrt{\sigma^2}}{k^{1/2}} + e_7(k), \tag{25}$$

where $|e_6(k)| \le Mk^{-3/2}$ and $|e_7(k)| \le Mk^{-3/2}$.

From (23) we obtain

$$\lambda = \lambda_{k,\sigma} = \frac{I}{J} = d - \frac{\sigma^2}{2} + e_8(k)$$

with $|e_8(k)| \le Mk^{-1}$.

This proves the following proposition.

PROPOSITION A.1. *In the case of a symmetric drift matrix, i.e., in the case $b = c$, the leading Lyapunov exponent $\lambda = \lambda_{k,\sigma}$ of the LSDE* (9) *satisfies*

$$\lambda_{k,\sigma} = d - \frac{\sigma^2}{2} + O(k^{-1})$$

for k large. In particular,

$$\lim_{k\to\infty} \lambda_{k,\sigma} = d - \frac{\sigma^2}{2}.$$

The case of a general drift matrix. Having assumed $b = c$ in the drift matrix for the calculations in the previous subsection, we now use these calculations to obtain a result for the case $b \neq c$.

Proof of Theorem 3.2. Having established the result for the case $b = c$ in the previous section, we proceed by showing that the general case is close to the symmetric case for k sufficiently large.

First note that (13), (14), $p(-\frac{\pi}{2}) = p(\frac{\pi}{2})$, and $\int p(\varphi)\, d\varphi = 1$ give $\lambda = I/J$ with

$$I = \int_{-\pi/2}^{\pi/2} q(\varphi) m(\varphi) g(\varphi)\, d\varphi \qquad \text{and} \quad J = \int_{-\pi/2}^{\pi/2} m(\varphi) g(\varphi)\, d\varphi,$$

where

$$\begin{aligned} q(\varphi) &= (a-k)\cos^2\varphi + (b+c)\cos\varphi\sin\varphi + d\sin^2\varphi \\ &= (a-d-k)\cos^2\varphi + (b+c)\cos\varphi\sin\varphi + d, \\ m(\varphi) &= \exp\Big(\frac{1}{2\sigma^2}[(b+c)\sin 2\varphi - (d-a+k)(\cos 2\varphi + 1)] \Big), \\ g(\varphi) &= \exp\left(\frac{c-b}{\sigma^2}\left(\varphi + \frac{\pi}{2}\right)\right) \\ &\quad + \left[1 - \exp\left(\frac{c-b}{\sigma^2}\pi\right)\right] \left[\exp\left(\frac{c-b}{\sigma^2}\left(\varphi - \frac{\pi}{2}\right)\right)\right] f(\varphi) \end{aligned}$$

and f is given by

$$f(\varphi) = \frac{\displaystyle\int_{-\pi/2}^{\varphi} \exp\left(\tfrac{1}{2\sigma^2}\left[-2(c-b)\eta + (d-a+k)\cos 2\eta - (b+c)\sin 2\eta\right]\right)\, d\eta}{\displaystyle\int_{-\pi/2}^{\pi/2} \exp\left(\tfrac{1}{2\sigma^2}\left[-2(c-b)\eta + (d-a+k)\cos 2\eta - (b+c)\sin 2\eta\right]\right)\, d\eta}.$$

In the symmetric case we had $g(\varphi) \equiv 1$ on $[-\pi/2, \pi/2]$. We will show that I and J are close to the corresponding integrals with $g \equiv 1$. While the integrals I and J look at first glance rather intractable, there are several simplifications if k is large. In particular, for k large we have $f(\varphi) \approx 0$ for $-\pi/2 \le \varphi \le -\pi/4$ and $f(\varphi) \approx 1$ for $\pi/4 \le \varphi \le \pi/2$. We will make this more precise below.

First note that $0 \le f \le 1$; hence there exists $M > 0$, independent of k, such that $|g(\varphi)| \le M$ for all φ. Next consider the behavior of $m(\varphi)$ for large k. Here the relevant term is $-(d-a+k)(\cos 2\varphi + 1)$. Note that $\cos 2\varphi + 1 \ge 0$ for all φ. Define intervals I_k and J_k close to $-\pi/2$ and $\pi/2$, respectively, such that $1 + \cos 2\varphi \ge \frac{1}{\sqrt{k}}$ for all $\varphi \notin I_k \cup J_k$ by

$$I_k = \left[-\pi/2, -\cos^{-1}\frac{1}{\sqrt{2\sqrt{k}}}\right] \qquad \text{and} \quad J_k = \left[\cos^{-1}\frac{1}{\sqrt{2\sqrt{k}}}, \pi/2\right].$$

For $\varphi \notin I_k \cup J_k$,

$$-(d-a+k)(1+\cos 2\varphi) \le -\frac{\sqrt{k}}{3}$$

for k large. Noting that, for k large,

$$|q(\varphi)| = |(a-k)\cos^2\varphi + (b+c)\cos\varphi\sin\varphi + d\sin^2\varphi| \le 2k,$$

we obtain

$$\left|\int_{[-\pi/2,\pi/2]\setminus(I_k\cup J_k)} q(\varphi)m(\varphi)g(\varphi)\, d\varphi\right| \le k\, M\, \exp\left(-\frac{\sqrt{k}}{3}\right)$$

for some $M > 0$. Since $k\exp\left(-\frac{\sqrt{k}}{3}\right) \le \exp\left(-\frac{\sqrt{k}}{4}\right)$ for k large, we thus obtain

$$\left|\int_{[-\pi/2,\pi/2]\setminus(I_k\cup J_k)} q(\varphi)m(\varphi)g(\varphi)\, d\varphi\right| \le M\exp\left(-\frac{\sqrt{k}}{4}\right),$$

and so

$$\left|\int_{-\pi/2}^{\pi/2} q(\varphi)m(\varphi)g(\varphi)\, d\varphi - \int_{I_k\cup J_k} q(\varphi)m(\varphi)g(\varphi)\, d\varphi\right| \le M\exp\left(-\frac{\sqrt{k}}{4}\right). \tag{26}$$

By similar arguments one obtains

$$\left|\int_{-\pi/2}^{\pi/2} q(\varphi)m(\varphi)\, d\varphi - \int_{I_k\cup J_k} q(\varphi)m(\varphi)\, d\varphi\right| \le M\exp\left(-\frac{\sqrt{k}}{4}\right). \tag{27}$$

It follows from (26) and (27) that we need only to deal with the difference between I and $\int q(\varphi)m(\varphi)\,d\varphi$ on $I_k \cup J_k$.

Loosely speaking, we have $g(\varphi) \approx 1$ on I_k and

$$g(\varphi) \approx \exp\left[\frac{c-b}{\sigma^2}\pi\right] + \left(1 - \exp\left[\frac{c-b}{\sigma^2}\pi\right]\right) = 1$$

on J_k, both uniformly in k.

To be more precise, consider first $f(\varphi)$ for $\varphi \in I_k$. For $\varphi \in I_k$ and k sufficiently large we have $\cos 2\varphi \le -1/2$, and for $\varphi \in [-\pi/6, \pi/6]$ we have $\cos 2\varphi \ge 1/2$. It follows that, for $\varphi \in I_k$,

$$\begin{aligned} 0 \le f(\varphi) &\le \frac{\int_{I_k} \exp\left(\frac{1}{2\sigma^2}\left[-2(c-b)\eta + (d-a+k)\cos 2\eta - (b+c)\sin 2\eta\right]\right)\,d\eta}{\int_{-\pi/6}^{\pi/6} \exp\left(\frac{1}{2\sigma^2}\left[-2(c-b)\eta + (d-a+k)\cos 2\eta - (b+c)\sin 2\eta\right]\right)\,d\eta} \\ &\le M\exp(-k), \end{aligned}$$

whence

$$\left| g(\varphi) - \exp\Big(\frac{c-b}{\sigma^2}(\varphi + \tfrac{\pi}{2})\Big)\right| \le M\exp(-k).$$

Expanding $\varphi \mapsto \exp\big(\frac{c-b}{\sigma^2}(\varphi + \frac{\pi}{2})\big)$ in Taylor series around $-\pi/2$, we obtain

(28)
$$g(\varphi) = 1 + \frac{c-b}{\sigma^2}(\varphi + \pi/2) + \frac{1}{2}\Big(\frac{c-b}{\sigma^2}(\varphi + \pi/2)\Big)^2 + \frac{1}{6}\Big(\frac{c-b}{\sigma^2}(\varphi + \pi/2)\Big)^3 + e_1(k)$$

with $|e_1(k)| \le Mk^{-1}$ for all $\varphi \in I_k$.

Similarly, for $\varphi \in J_k$ we have

$$\begin{aligned} 1 \ge f(\varphi) &= 1 - \frac{\int_{\varphi}^{\pi/2} \exp\left(\frac{1}{2\sigma^2}\left[-2(c-b)\eta + (d-a+k)\cos 2\eta - (b+c)\sin 2\eta\right]\right)\,d\eta}{\int_{-\pi/2}^{\pi/2} \exp\left(\frac{1}{2\sigma^2}\left[-2(c-b)\eta + (d-a+k)\cos 2\eta - (b+c)\sin 2\eta\right]\right)\,d\eta} \\ &\ge 1 - M\exp(-k), \end{aligned}$$

and expanding the exp-term in $\varphi \mapsto g(\varphi)$ in Taylor series around $\pi/2$ gives

(29)
$$g(\varphi) = 1 + \frac{c-b}{\sigma^2}(\varphi - \pi/2) + \frac{1}{2}\Big(\frac{c-b}{\sigma^2}(\varphi - \pi/2)\Big)^2 + \frac{1}{6}\Big(\frac{c-b}{\sigma^2}(\varphi - \pi/2)\Big)^3 + e_2(k)$$

with $|e_2(k)| \le Mk^{-1}$ for all $\varphi \in J_k$.

Based on these estimates and expansions, we would expect, loosely speaking, that

$$\int_{I_k \cup J_k} q(\varphi)m(\varphi)g(\varphi)\,d\varphi \approx \int_{I_k \cup J_k} q(\varphi)m(\varphi)\,d\varphi.$$

Actually, this is not as simple as it looks. The problem is that the right-hand side is $O(1/\sqrt{k})$, and so we have to be careful with arguments based on any "errors" being small for $k \to \infty$.

We proceed more carefully: On I_k we obtain, invoking (28),

$$\begin{aligned} &q(\varphi)m(\varphi)g(\varphi) \\ &= q(\varphi)m(\varphi) \\ &\quad \times\Big[1 + \frac{c-b}{\sigma^2}(\varphi + \pi/2) + \frac{1}{2}\Big(\frac{c-b}{\sigma^2}(\varphi + \pi/2)\Big)^2 + \frac{1}{6}\Big(\frac{c-b}{\sigma^2}(\varphi - \pi/2)\Big)^3 + e_1(k)\Big]. \end{aligned}$$

So

$$\begin{aligned}
&q(\varphi)m(\varphi)g(\varphi) - q(\varphi)m(\varphi)\\
&= q(\varphi)\exp\Big(\frac{b+c}{2\sigma^2}\sin 2\varphi\Big)\exp\Big(-\frac{k+d-a}{\sigma^2}\cos^2\varphi\Big)\\
&\quad\times\Big[\frac{c-b}{\sigma^2}(\varphi+\pi/2) + \frac{1}{2}\Big(\frac{c-b}{\sigma^2}(\varphi+\pi/2)\Big)^2 + \frac{1}{6}\Big(\frac{c-b}{\sigma^2}(\varphi+\pi/2)\Big)^3 + e_1(k)\Big].
\end{aligned}$$

Similarly, on J_k we obtain from (29)

$$\begin{aligned}
&q(\varphi)m(\varphi)g(\varphi) - q(\varphi)m(\varphi)\\
&= q(\varphi)\exp\Big(\frac{b+c}{2\sigma^2}\sin 2\varphi\Big)\exp\Big(-\frac{k+d-a}{\sigma^2}\cos^2\varphi\Big)\\
&\quad\times\Big[\frac{c-b}{\sigma^2}(\varphi-\pi/2) + \frac{1}{2}\Big(\frac{c-b}{\sigma^2}(\varphi-\pi/2)\Big)^2 + \frac{1}{6}\Big(\frac{c-b}{\sigma^2}(\varphi-\pi/2)\Big)^3 + e_2(k)\Big].
\end{aligned}$$

We consider the contributions of the terms $\big((a-d-k)\cos^2\varphi\big)$ and $\big((b+c)\cos\varphi\sin\varphi + d\big)$ in $q(\varphi)$ separately. The term $\big((a-d-k)\cos^2\varphi\big)$ contributes

$$\begin{aligned}
&\int_{I_k}(a-d-k)\cos^2\varphi\,\exp\Big(\frac{b+c}{2\sigma^2}\sin 2\varphi\Big)\exp\Big(-\frac{k+d-a}{\sigma^2}\cos^2\varphi\Big)\\
&\times\Big[\frac{c-b}{\sigma^2}(\varphi+\pi/2) + \frac{1}{2}\Big(\frac{c-b}{\sigma^2}(\varphi+\pi/2)\Big)^2 + \frac{1}{6}\Big(\frac{c-b}{\sigma^2}(\varphi+\pi/2)\Big)^3 + e_1(k)\Big]\,d\varphi
\end{aligned}$$

to $\int_{I_k}\big(q(\varphi)m(\varphi)g(\varphi) - q(\varphi)m(\varphi)\big)\,d\varphi$. The corresponding contribution from J_k is

$$\begin{aligned}
&\int_{J_k}(a-d-k)\cos^2\varphi\,\exp\Big(\frac{b+c}{2\sigma^2}\sin 2\varphi\Big)\,\exp\Big(-\frac{k+d-a}{\sigma^2}\cos^2\varphi\Big)\\
&\times\Big[\frac{c-b}{\sigma^2}(\varphi-\pi/2) + \frac{1}{2}\Big(\frac{c-b}{\sigma^2}(\varphi-\pi/2)\Big)^2 + \frac{1}{6}\Big(\frac{c-b}{\sigma^2}(\varphi-\pi/2)\Big)^3 + e_2(k)\Big]\,d\varphi.
\end{aligned}$$

Taking Taylor expansions of $\varphi\mapsto\sin 2\varphi$ and $\varphi\mapsto\cos^2\varphi$ around $-\pi/2$ and $\pi/2$, respectively, changing variables, and combining the contributions of the $(a-d-k)\cos^2\varphi$-terms from I_k and J_k, we arrive at an error term of the form

$$\begin{aligned}
&\int_0^{\frac{1}{\sqrt{2\sqrt{k}}}}(a-d-k)\,x^2\,\exp\Big(-\frac{k+d-a}{\sigma^2}x^2\Big)\\
&\times\Big[\Big(\exp\Big(-\frac{b+c}{\sigma^2}x\Big) - \exp\Big(\frac{b+c}{\sigma^2}x\Big)\Big)\Big(\frac{c-b}{\sigma^2}x + \frac{1}{6}\Big(\frac{c-b}{\sigma^2}\Big)^3\Big)\\
&\quad+\frac{1}{2}\Big(\exp\Big(-\frac{b+c}{\sigma^2}x\Big) + \exp\Big(\frac{b+c}{\sigma^2}x\Big)\Big)\Big(\frac{c-b}{\sigma^2}x\Big)^2\Big]\,dx \;+\; e_3(k)
\end{aligned}\tag{30}$$

with $|e_3(k)|\le Mk^{-3/2}$. Here we used that, after the change of variables around $\pi/2$ and around $-\pi/2$, respectively, $\cos^2\varphi$ gives a term of the order $x^2+O(x^4)$. Using integration by parts, one verifies that the error terms $e_1(k)$ and $e_2(k)$ contribute an $O(k^{-3/2})$-term, and the $O(x^4)$-term from $\cos^2\varphi$ contributes an $O(k^{-2})$-term.

Identifying the exponential terms inside the square brackets as $-2\sinh(\frac{b+c}{\sigma^2}x)$ and $\cosh(\frac{b+c}{\sigma^2}x)$ and expanding $\sinh x$ and $\cosh x$, respectively, around $x=0$, (30) becomes

$$Mk\int_0^{\frac{1}{\sqrt{2\sqrt{k}}}}x^4\exp\left(-\frac{k+d-a}{\sigma^2}x^2\right)\,dx + O(k^{-3/2})$$

for k large. Again using integration by parts this can be seen to grow not faster than $Mk^{-3/2}$ for $k \to \infty$ (recall that M is a variable constant).

Using similar calculations in which we again need to combine contributions from I_k and J_k, we can show that the term $\big((b+c)\cos\varphi\sin\varphi+d\big)$ in $q(\varphi)$ contributes to the integral $\int_{I_k\cup J_k}\big(q(\varphi)m(\varphi)g(\varphi) - q(\varphi)m(\varphi)\big)\,d\varphi$ an error term bounded in magnitude by $Mk^{-3/2}$.

Collecting the above estimates, we obtain

$$\Big|\int_{I_k\cup J_k} q(\varphi)m(\varphi)g(\varphi)\,d\varphi - \int_{I_k\cup J_k} q(\varphi)m(\varphi)\,d\varphi\Big| \le Mk^{-3/2} \qquad \text{and}$$
$$\Big|\int_{I_k\cup J_k} m(\varphi)g(\varphi)\,d\varphi - \int_{I_k\cup J_k} m(\varphi)\,d\varphi\Big| \le Mk^{-3/2}$$

for some positive M and k sufficiently large. In view of (26) and (27), this implies

$$\Big|\int_{-\pi/2}^{\pi/2} q(\varphi)m(\varphi)g(\varphi)\,d\varphi - \int_{-\pi/2}^{\pi/2} q(\varphi)m(\varphi)\,d\varphi\Big| \le Mk^{-3/2} \qquad \text{and} \tag{31}$$
$$\Big|\int_{-\pi/2}^{\pi/2} m(\varphi)g(\varphi)\,d\varphi - \int_{-\pi/2}^{\pi/2} m(\varphi)\,d\varphi\Big| \le Mk^{-3/2}. \tag{32}$$

Now put

$$I_{\text{sym}} = \int_{-\pi/2}^{\pi/2} q(\varphi)m(\varphi)\,d\varphi \qquad \text{and} \quad J_{\text{sym}} = \int_{-\pi/2}^{\pi/2} m(\varphi)\,d\varphi;$$

the subscript "sym" refers to the symmetric case treated in (the proof of) Proposition A.1 From (24) and (25) we obtain

$$I_{\text{sym}} = \pi\left(d - \frac{\sigma^2}{2}\right)\exp\left(\frac{a-d}{2\sigma^2}\right)\frac{\sqrt{\sigma^2}}{k^{1/2}} + \eta_1(k),$$
$$J_{\text{sym}} = \pi\exp\left(\frac{a-d}{2\sigma^2}\right)\frac{\sqrt{\sigma^2}}{k^{1/2}} + \eta_2(k)$$

with both $|\eta_j(k)| \le Mk^{-3/2}$, $j = 1, 2$, where we note that the derivation of (24) and (25) goes through without changes if $2b$ is replaced by $b + c$.

Rewriting (31) and (32) with this notation gives

$$I = I_{\text{sym}} + e_4(k) \qquad \text{and} \quad J = J_{\text{sym}} + e_5(k),$$

where $|e_j(k)| \le Mk^{-\frac{3}{2}}$, $j = 4, 5$. So finally, we obtain

$$\lambda_{k,\sigma} = \frac{I}{J} = \frac{I_{\text{sym}} + e_4(k)}{J_{\text{sym}} + e_5(k)} = d - \frac{\sigma^2}{2} + e_6(k),$$

where, due to the $k^{-1/2}$-term in I_{sym} and J_{sym}, respectively, $|e_6(k)| \le Mk^{-1}$. This proves the claim, finishing the proof of Theorem 3.2. □

Acknowledgments. This paper was prepared while Hans Crauel was with the School of Mathematical Sciences of the University of Exeter. Our thanks go to Ludwig Arnold (Bremen) for pointing out the approach sketched in Remark 3.3 and for discussing it with us.

REFERENCES

[1] L. Arnold, *Random Dynamical Systems*, Springer-Verlag, New York, 1998.

[2] L. Arnold, H. Crauel, and V. Wihstutz, *Stabilization of linear systems by noise*, SIAM J. Control Optim., 21 (1983) pp. 451–461.

[3] L. Arnold, A. Eizenberg, and V. Wihstutz, *Large noise asymptotics of invariant measures, with applications to Lyapunov exponents*, Stochastics Stochastics Rep., 59 (1996) pp. 71–142.

[4] L. Arnold, W. Kliemann, and E. Oeljeklaus, *Lyapunov exponents of linear stochastic systems*, in Lyapunov Exponents, Proceedings, Bremen 1984, Lecture Notes in Math. 1186, L. Arnold and V. Wihstutz, eds., Springer-Verlag, Berlin, 1986, pp. 85–125.

[5] L. Arnold, E. Oeljeklaus, and E. Pardoux, *Almost sure and moment stability for linear Itô equations*, in Lyapunov Exponents, Proceedings, Bremen 1984, Lecture Notes in Math. 1186, L. Arnold and V. Wihstutz, eds., Springer-Verlag, Berlin, 1986, pp. 129–159.

[6] T. Damm, *Generalized Riccati equations and stabilization of stochastic systems*, in Control Applications of Optimization 2000, Proceedings of the 11th IFAC Workshop, Vol. 2, V. Zakharov, ed., Elsevier Science, New York, 2000, pp. 22–27.

[7] P. Imkeller and C. Lederer, *An explicit description of the Lyapunov exponent of the noisy damped harmonic oscillator*, Dyn. Stab. Syst., 14 (1999), pp. 385–405.

[8] P. Imkeller and C. Lederer, *Some formulas for Lyapunov exponents and rotation numbers in two dimensions and the stability of the harmonic oscillator and the inverted pendulum*, Dyn. Syst., 16 (2001), pp. 29–61.

[9] J. F. C. Kingman, *The ergodic theory of subadditive stochastic processes*, J. Roy. Statist. Soc. Ser. B, 30 (1968), pp. 499–510.

[10] V. Wihstutz, *Perturbation methods for Lyapunov exponents*, in Stochastic Dynamics, H. Crauel and M. Gundlach, eds., Springer-Verlag, New York, 1999, pp. 209–239.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1854–1875

EXPLICIT SOLUTION TO A ROBUST QUEUEING CONTROL PROBLEM*

PAUL DUPUIS†

Abstract. We consider the robust optimal control of a law of large numbers approximation of a stochastic network. The robust control problem is formulated as a differential game, with one player choosing the policies that determine service and routing assignments and the other choosing quantities such as the arrival and service rates, subject to constraints. The cost to be minimized by the first player and maximized by the second is the time until the origin is reached. An explicit formula is given for the value function, and some of its basic properties are studied.

Key words. robust control, queueing network, differential game, explicit solutions

AMS subject classifications. 35F30, 91A23, 29L25, 49N70, 90B10

DOI. 10.1137/S0363012901395753

1. Introduction.

This paper considers the problem of robust service and routing control for a network of servers. Consider such a network, and assume that at each station there are a finite number of distinct customer classes, each with its own buffer. In this paper we will work directly with what is sometimes called a "fluid" model for the network [19]. Models of this sort are usually obtained as law of large numbers approximations to more detailed models [6, 17] and are particularly appealing because in many cases related optimization problems admit closed form solutions [23, 24, 12].

Another feature of the networks we consider is model uncertainty, such as uncertainty in the arrival and service rates. To deal with model uncertainty we adapt the differential game formulation of robust control for unconstrained nonlinear systems [15]. Thus we consider a network where there are two players. One player in the game will represent the "true" control (e.g., service assignments and routing decisions). The other player represents the uncertain or poorly modeled aspects of the system (e.g., arrival and service rates). In keeping with existing convention, we will refer to this latter control as "nature." The two players are antagonistic, with the first player attempting to maintain good system performance.

Differential game formulations provide a powerful tool for the design of robust controls [4, 15]. In many situations knowledge of the true system is limited. System parameters (e.g., arrival rates) may drift with time, and statistical properties (e.g., correlations) may also be unstable. There may be aspects of the system that are left unmodeled, either because they cannot be estimated in any reliable way or because they lead to a model that is too complicated to be useful. This is a common occurrence in stochastic networks, where the network to be controlled is often a subnetwork of some larger system, and "full state information" is simply not available to the controller of the subnetwork.

*Received by the editors September 21, 2001; accepted for publication (in revised form) June 2, 2003; published electronically December 17, 2003. This research was supported in part by the National Science Foundation (NSF-DMS-0072004, NSF-ECS-9979250) and by the Army Research Office (DAAD19-99-1-0223).

http://www.siam.org/journals/sicon/42-5/39575.html

†Lefschetz Center for Dynamical Systems, Brown University, Providence, RI 02912 (dupuis@cfm.brown.edu).

In situations like these the use of a single "nominal" model can be problematic. For example, just as in the case of unconstrained systems one can construct examples where controls that are optimal in some sense for the nominal model perform poorly when the model is perturbed even slightly. A differential game formulation allows one to construct controls that perform uniformly well over a class of perturbations of the nominal model, with each choice of nature's control corresponding to a different perturbation of the design model. It is, of course, this insensitivity to model perturbations that warrants the term "robust" control. Variations of different kinds can be accommodated through the choice of the cost structure, and one can carefully balance the pursuit of optimality with respect to a nominal model against the need to provide good performance for a range of models. Indeed, for many current design problems (see, e.g., [16]) one would like the maximum robustness possible given certain guaranteed bounds on performance. The main result of this paper is the explicit solution to a robust control problem for a network. By explicit what we mean is that the value function can be represented in terms of a finite dimensional optimization problem and that from this value function one can obtain controls with specific robust properties.

In formulating the differential game special attention must be paid to the cost applied to nature's control, since this determines the degree to which model perturbations are allowed. Within the realm of "fluid" models there are at least two types of cost structures that are natural. One is a cost that simply imposes a constraint on the model parameters. We will refer to this as the case of a "hard constraint." An alternative is to make nature pay an increasing cost for perturbations away from the nominal model, and we will refer to this type of cost as corresponding to "soft constraints." Hard constraints turn out to be mathematically simpler for fluid models of networks, even though the reverse seems to be true for unconstrained systems. The present paper will focus on the case of hard constraints. Of particular interest among games with soft constraints are those which arise as the limit of "risk-sensitive" control problems [2]. However, for problems of control, until the origin is reached the running cost in these games is not of one sign, and so there is no longer uniqueness for the corresponding PDE [18] (even within the viscosity solution framework). This makes the analysis much more difficult.

Besides the cost that nature must pay, one must also specify the cost that the true control faces. In this paper the cost we consider is the time to move the state of the system from an arbitrary position to zero (i.e., all queues empty). The true control will try to minimize this time, while the opposing control will attempt to delay it as much as possible. This cost seems to be a natural analogue, in the setting of constrained systems, of the familiar quadratic cost for unconstrained systems. In particular, it leads to controls with optimal robust stability (in addition to optimal robust performance), and it also allows for a fairly explicit closed form solution.

An outline of the paper is as follows. In section 2 we give a precise formulation of the game problem and show through examples how various systems can be put into this framework. Section 3 includes the main result of the paper, which is a finite dimensional max/min representation for the value function of the game introduced in section 2. Qualitative properties of the value function (convexity, differentiability, etc.) are also discussed in section 3. The proof of this representation is given in section 4. The concluding section 5 formally discusses how an optimal true control can be constructed in feedback form. A proof of the existence of value for the games we consider is given in an appendix.

2. Formulation of the control problem. In this section we formulate the robust control problem as a constrained deterministic differential game. As discussed in the introduction, the model we use can be viewed as a law of large numbers approximation to a more detailed stochastic model. This connection will be used for interpretive purposes throughout the section. The state space of the process is $\mathbb{R}_+^N$, and one can interpret each of the components as a queue length associated with a specific customer class.

The formulation of the model involves two collections of N-dimensional vectors. The first are the *directions of constraint*, which we designate by $\{d_i, i = 1, \ldots, N\}$. These vectors are used to define the Skorokhod or reflection map, which properly corrects the dynamics of the model when one or more components of the state are zero (i.e., one or more customer classes are empty). The second collection is designated $\{v_{jk}, j = 1, \ldots, J, k = 1, \ldots, K\}$ and is used to define the dynamics of the system away from the boundary. Nature's control takes values in a compact convex set $\mathcal{A} \subset \mathbb{R}^K$, and the index $j \in \{1, \ldots, J\}$ corresponds to one of the possible "pure" service/routing configurations the true controller can select (illustrative examples will be given below). If nature chooses the control $\alpha \in \mathcal{A}$ and the true control is the pure configuration j, then the quantity $\sum_{k=1}^K \alpha_k v_{jk}$ characterizes the (law of large numbers) evolution of the network when the state of the network is away from $\partial\mathbb{R}_+^N$. More general service/routing policies can be obtained by considering convex combinations of the pure controls, in which case the velocity of the system is given by

$$F(\rho, \alpha) \doteq \sum_{j=1}^{J} \sum_{k=1}^{K} \rho_j \alpha_k v_{jk},$$

where

$$\rho = (\rho_1, \ldots, \rho_J) \in \mathcal{S} \doteq \left\{ x \in \mathbb{R}^J : x_j \geq 0, j = 1, \ldots, J, \sum_{j=1}^{J} x_j = 1 \right\},$$

and ρ_j is the fraction of time allocated to the pure configuration j.

We will assume the following condition on the directions of constraint. The condition is by now classical in the study of approximations to queueing networks and is called the Harrison–Reiman condition in [11]. It was first used in [14]. Although the Harrison–Reiman condition is usually associated with single class networks, it also defines the proper Skorokhod problem for many formulations of controlled multiclass networks as well. Note that the condition is the original Harrison–Reiman condition and not the generalization that is also studied in [11].

CONDITION 2.1. *For each* $i \in \{1, \ldots, N\}$

$$(d_i)_i = 1, \text{ and } (d_i)_j \leq 0 \text{ for } j \neq i.$$

Let D *be the matrix whose* i*th column is* d_i*. Then the spectral radius of* $I - D$ *is less than* 1.

The following simple examples illustrate the role these different quantities play. The ith unit basis vector is denoted by e_i. Some of the most difficult aspects in the control of networks are due to feedback and the interactions between different servers. From this perspective, the first two examples are too simple to be of great interest. Also, it should be noted that the game formulation we consider in this paper allows routing only at the "fringes" of the network and not between nodes.

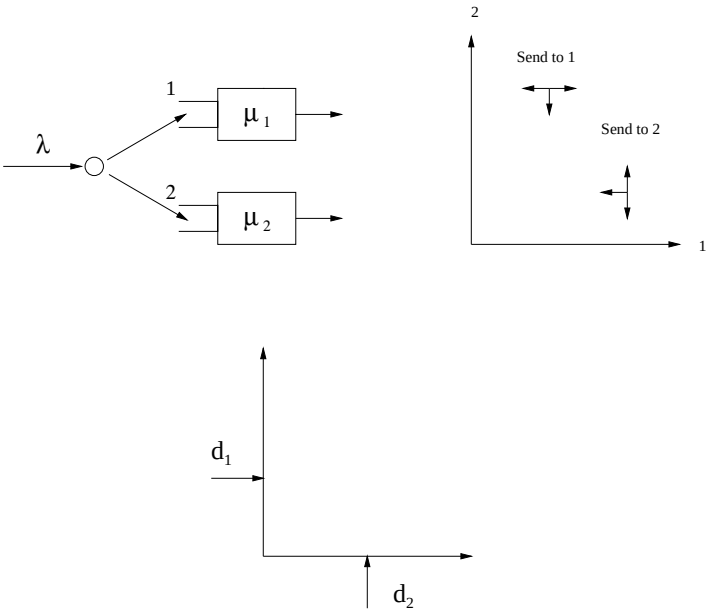


FIG. 1. *Simple routing control.*

Example 1. The first example is a simple routing control problem. The rate of arrivals to the router is $\lambda(t)$, and the service rates of the two servers are $\mu_1(t)$ and $\mu_2(t)$, respectively. The system is illustrated in Figure 1.

This model is put into the framework described above by setting

$$\begin{array}{lll} \alpha_1 = \mu_1, & v_{1,1} = -e_1, & v_{2,1} = -e_1, \\ \alpha_2 = \mu_2, & v_{1,2} = -e_2, & v_{2,2} = -e_2, \\ \alpha_3 = \lambda, & v_{1,3} = e_1, & v_{2,3} = e_2. \end{array}$$

The choice of $\mathcal{A}$ determines the uncertainties and perturbations against which the optimal true control will be robust. For example, if the nominal service and arrival rates are $\bar{\lambda} = 1$ and $\bar{\mu}_i = 1, i = 1, 2$, and if the service rates are well modeled and the arrival rate less so, then one might consider a set of the form

$$\mathcal{A} = [\bar{\mu}_1^L, \bar{\mu}_1^U] \times [\bar{\mu}_2^L, \bar{\mu}_2^U] \times [\bar{\lambda}^L, \bar{\lambda}^U] = [.9, 1.1] \times [.9, 1.1] \times [.5, 1.5].$$

This model is very simple and perhaps too simple to capture any "probabilistic" intuition. For example, there is no constraint on combinations of μ_1 and μ_2. From a probabilistic perspective one might imagine that it is less likely that both of these parameters would equal their minimum value at the same time. The introduction of a constraint to account for this would lead to a set $\mathcal{A}$ with a "curved" boundary.

One might also wish to consider an increasing family of sets $\mathcal{A}(c)$ indexed by $c \in [0, \infty)$ and with $\mathcal{A}(0)$ just the nominal model. The largest c such that a certain robust performance measure can be met (e.g., finiteness of the value function) is an important quantity. In particular, it characterizes the control that is most robust, where the sense of robustness is determined by the shape of $\mathcal{A}$ and the relative uncertainty it assigns to different aspects of the network.

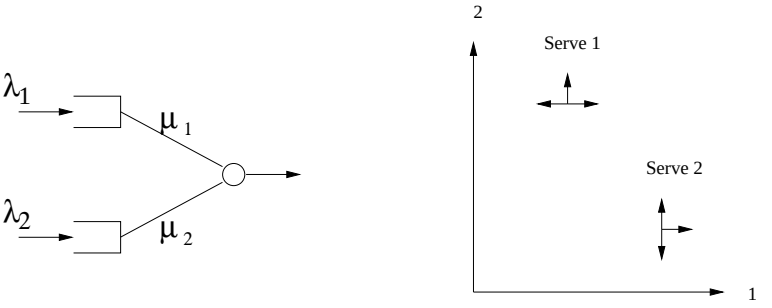


FIG. 2. *Simple service control.*

We return our consideration to the particular example of Figure 1. If a service is attempted at server 1 and the queue is empty, then the proper compensating action is simply to return queue 1 to the level zero. As a consequence, the direction of constraint for the corresponding face is just $d_1 = (1, 0)$. A corresponding remark applies to queue 2.

Example 2. Here we consider a simple service control problem. The system is illustrated in Figure 2.

The Skorokhod problem for this is the same as that for Example 1. This model is put into the standard framework by setting

$$\begin{array}{lll} \alpha_1 = \mu_1, & v_{1,1} = -e_1, & v_{2,1} = 0, \\ \alpha_2 = \mu_2, & v_{1,2} = 0, & v_{2,2} = -e_2, \\ \alpha_3 = \lambda_1, & v_{1,4} = e_1, & v_{2,4} = e_1, \\ \alpha_4 = \lambda_2, & v_{1,4} = e_2, & v_{2,4} = e_2. \end{array}$$

Example 3. This example considers a network of servers, and as a consequence the associated Skorokhod map is more involved. The network is illustrated in Figure 3. Since there are six customer classes, the domain is $\mathbb{R}^6_+$. Suppose the service rate for class i is $\mu_i(t)$ and the arrival rate is $\lambda(t)$.

An example of a pure configuration (labeled, say, j) is to route to class 1 and serve classes 3, 2, and 5. If we let $(\alpha_1, \ldots, \alpha_6, \alpha_7) = (\mu_1, \ldots, \mu_6, \lambda)$, then the vectors v_{jk} for $k = 2, 3, 5, 7$ are

$$v_{j2} = (-e_2 + e_6), v_{j3} = (-e_3 + e_4), v_{j5} = (-e_5 + e_3), v_{j7} = e_1,$$

while $v_{j1} = v_{j4} = v_{j6} = 0$. The velocity of the network under this configuration is $\mu_2 v_{j2} + \mu_3 v_{j3} + \mu_5 v_{j5} + \lambda v_{j7}$.

If a service is attempted for, say, customer class $i = 3$ and the queue is empty, then queue 3 must be returned to zero and in addition queue 4 must be reduced by the same amount. Consequently, the proper direction of constraint for face $i = 3$ is $d_3 = (e_3 - e_4)$. Analogous considerations can be used to identify all other directions of constraint.

Example 4. In some problems there is randomized (uncontrolled) routing. For example, after service a fraction θ_j of the class i customers may become class j customers, and a fraction $\theta_0 = 1 - \sum_{j=1, j\neq i}^{J} \theta_j$ of the customers could leave the system. Let $(d_i)_j = -\theta_j$ if $j \neq i$ and $(d_i)_i = 1$. Then the direction of constraint is d_i on the face $\{x \in \mathbb{R}^N_+ : x_i = 0\}$, and the reason is the same as in the last case: compensating

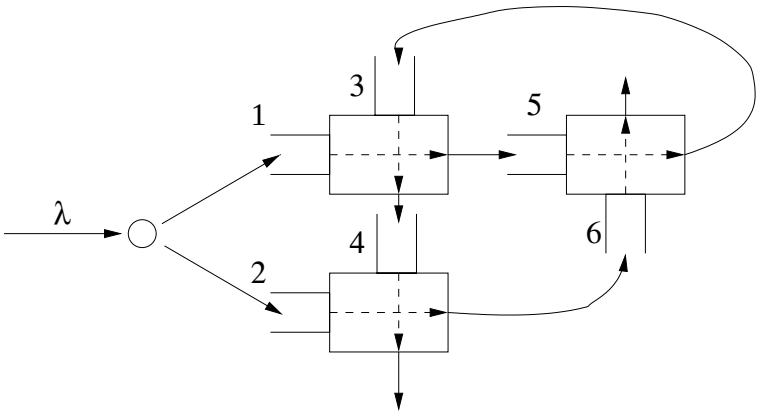


FIG. 3. *Network model.*

for a "fictitious" service of a customer of class i requires a boost to coordinate i and a corresponding decrease in coordinate j with constant of proportionality θ_j [21].

To formulate the robust control problem we must specify the dynamics. Let

$$C_+([0,\infty):\mathbb{R}^N) \doteq \{\psi \in C([0,\infty):\mathbb{R}^N) : \psi(0) \in \mathbb{R}^N_+\},$$

where $C([0,\infty):\mathbb{R}^N)$ is the usual space of continuous functions with the sup norm metric, and suppose that a set of vectors that satisfy Condition 2.1 is given. For each point x on the boundary of $\mathbb{R}^N_+$ let

$$d(x) \doteq \left\{ \sum_{i\in I(x)} a_i d_i : a_i \geq 0, \left\| \sum_{i\in I(x)} a_i d_i \right\| = 1 \right\},$$

where $I(x) \doteq \{i : x_i = 0\}$. The Skorokhod problem assigns to every path $\psi \in C_+([0,\infty):\mathbb{R}^N)$ a path ϕ that starts at $\phi(0) = \psi(0)$ but is constrained to $\mathbb{R}^N_+$ as follows. If ϕ is in the interior of $\mathbb{R}^N_+$, then the evolution of ϕ mimics that of ψ, in that the increments of the two functions are the same until ϕ hits the boundary of $\mathbb{R}^N_+$. When ϕ is on the boundary a constraining "force" is applied to keep ϕ in the domain, and this force can only be applied in one of the directions $d(\phi(t))$ and only for t such that $\phi(t)$ is on the boundary. The precise definition is as follows. For $\eta \in C([0,\infty):\mathbb{R}^N)$ and $t \in [0,\infty)$ we let $|\eta|(t)$ denote the total variation of η on $[0,t]$ with respect to the Euclidean norm on $\mathbb{R}^N$.

DEFINITION 2.1. *Let $\psi \in C_+([0,\infty):\mathbb{R}^N)$ be given. Then (ϕ,η) solves the Skorokhod problem for ψ (with respect to $\mathbb{R}^N_+$ and $d_i, i = 1,\ldots,N$) if $\phi(0) = \psi(0)$, and if for all $t \in [0,\infty)$*

1. $\phi(t) = \psi(t) + \eta(t)$,
2. $\phi(t) \in \mathbb{R}^N_+$,
3. $|\eta|(t) < \infty$,
4. $|\eta|(t) = \int_{[0,t]} 1_{\{\phi(s)\in\partial\mathbb{R}^N_+\}} d|\eta|(s)$,
5. *there exists a Borel measurable function $\gamma : [0,\infty) \to \mathbb{R}^N_+$ such that $d|\eta|$-almost*

everywhere $\gamma(t) \in d(\phi(t))$, *and such that*

$$\eta(t) = \int_{[0,t]} \gamma(s) d|\eta|(s).$$

Note that η changes only when ϕ is on the boundary and only in the directions $d(\phi)$.

Under Condition 2.1 the Skorokhod problem has a solution for all $\psi \in C_+([0,\infty) : \mathbb{R}^N)$. In addition, the mapping $\psi \to \phi$ is Lipschitz continuous [9, 14].

We next define a constrained ODE. As is proved in [9], one can define a projection $\pi : \mathbb{R}^N \to \mathbb{R}^N_+$ that is consistent with the constraint directions $\{d_i, i = 1, \ldots, N\}$, in that $\pi(x) = x$ if $x \in \mathbb{R}^N_+$, and if $x \notin \mathbb{R}^N_+$, then $\pi(x) - x = \alpha r$, where $\alpha \geq 0$, $\pi(x) \in \partial\mathbb{R}^N_+$, and $r \in d(\pi(x))$. Figure 4 illustrates the projection for a two dimensional problem.

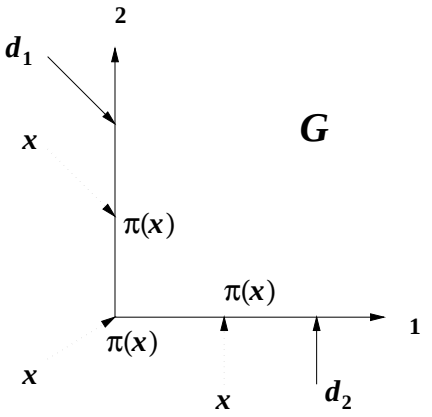


FIG. 4. *The discrete projection.*

With this projection given, we can define for each point $x \in \partial\mathbb{R}^N_+$ and each $v \in \mathbb{R}^N$ the *projected velocity*

$$\pi(x, v) \doteq \lim_{\Delta \downarrow 0} \frac{\pi(x + \Delta v) - \pi(x)}{\Delta}. \tag{1}$$

For details on why this limit is always well defined and further properties of the projected velocity, we refer to [7, section 3 and Lemma 3.8] and [8]. The dynamical model for the game we consider is then given by

$$\dot{\phi}(t) = \pi(\phi(t), F(\rho(t), \alpha(t))), \tag{2}$$

where

$$F(\rho(t), \alpha(t)) = \sum_{j=1}^{J} \sum_{k=1}^{K} \rho_j(t) \alpha_k(t) v_{jk}, \tag{3}$$

and for all $t \in [0,\infty)$ the true control $\rho(t)$ takes values in the set $\mathcal{S}$ and nature's control $\alpha(t)$ takes values in the set $\mathcal{A}$. According to the Skorokhod problem, the velocity $F(\rho, \alpha)$ governs the evolution of the network when all states are positive.

When one or more states are negative, the projection of the velocity provides the proper correction to the dynamics due to nonnegativity constraints.

An absolutely continuous function $\phi : [0,\infty) \to \mathbb{R}^N_+$ is a solution to (2) if the equation is satisfied in an almost everywhere sense in t. By using the regularity properties of the associated Skorokhod map, one can prove that all the standard qualitative properties (existence and uniqueness of solutions, stability with respect to perturbations, etc.) hold [9]. In fact, because of the particularly simple nature of the right-hand side (i.e., $\pi(\phi(t), \beta(t))$ rather than $\pi(\phi(t), b(\phi(t)) + \beta(t))$ for some function b), one can show that ϕ solves (2) if and only if ϕ is the image of $\psi(t) \doteq \int_0^t F(\rho(s), \alpha(s))ds + x$ under the Skorokhod map, in which case all such issues become trivial [9].

The ODE (2) defines the dynamics for the game that we will consider. The cost we consider is the time for the state to reach the origin, which the true control will attempt to minimize and which nature will try to prolong. As usual in differential games, one must deal with the issue of which player has the "information advantage" [13]. In our formulation we will allow the true controller to use "relaxed" controls, and as a consequence the game will have value; i.e., the value function will be the same regardless of who has the information advantage. However, the (upper) value of the game would not be the same if ordinary controls were used. In this sense the game defined as the limit of risk-sensitive control problems is superior, in that the upper value does not depend on whether pure or relaxed controls are used [2].

We use the standard Elliot–Kalton formulation of the game. Define the spaces of (open loop) controls

$$N \doteq \{\rho : [0,\infty) \to \mathcal{S} : \rho \text{ is measurable}\}$$

and

$$M \doteq \{\alpha : [0,\infty) \to \mathcal{A} : \alpha \text{ is measurable}\}.$$

We identify any two controls that are equal almost everywhere. Given $x \in \mathbb{R}^N_+$, the dynamics of the game are given by (2) and (3). Associated with these dynamics is the cost

$$C_x(\rho, \alpha) \doteq \tau_x,$$

where $\tau_x \doteq \inf\{t \geq 0 : \phi(t) = 0\}$. A mapping $\theta : N \to M$ is said to be a *strategy for the maximizing player* if for each $s \geq 0$ and $\rho, \hat{\rho} \in N$

$$\rho(t) = \hat{\rho}(t) \text{ for a.e. } 0 \leq t \leq s$$

implies

$$\theta[\rho](t) = \theta[\hat{\rho}](t) \text{ for a.e. } 0 \leq t \leq s.$$

A *strategy for the minimizing player*, which will be denoted by δ, is defined in an analogous manner. We denote by Θ the set of all maximizing strategies and by Δ the set of all minimizing ones. The *lower value of the game* and the *upper value of the game* are defined by

$$V^-(x) \doteq \inf_{\delta \in \Delta} \sup_{\alpha \in M} C_x(\delta[\alpha], \alpha) \tag{4}$$

and

$$V^+(x) \doteq \sup_{\theta\in\Theta}\inf_{\rho\in N} C_x(\rho,\theta[\rho]), \tag{5}$$

respectively. If $V^-(x) = V^+(x)$, then the game is said to have value.

Let $V : \mathbb{R}^N \to \mathbb{R}$. For points $x \in \mathbb{R}^N$ and directions $w \in \mathbb{R}^N$ for which the limit exists, we let $D_wV(x)$ denote the directional derivative in direction w at x:

$$D_wV(x) \doteq \lim_{a\downarrow 0}\frac{V(x+aw)-V(x)}{a}.$$

We say that V is *radially linear* if $V(ax) = aV(x)$ for all $x \in \mathbb{R}^N$ and $a \in [0,\infty)$.

3. Representation for the value function. For $V^+(x)$ and $V^-(x)$ to be finite we will need some conditions. Define the convex cone

$$\mathcal{C} \doteq \left\{-\sum_{i=1}^N a_i d_i : a_i \geq 0, i = 1,\dots,N\right\},$$

which is the negative of the cone of constraint directions that are allowed at the origin. As observed in [7], this cone can be used to characterize stability conditions for (2).

The following formula gives an explicit representation for the value of the game defined in the last section. The precise statement is given at the end of the section, and the proof is given in section 4. Recall that $F(\rho,\alpha) \doteq \sum_{j=1}^J\sum_{k=1}^K \rho_j\alpha_k v_{jk}$. Then set

$$W(x) \doteq \sup_{\alpha\in\mathcal{A}}\inf_{\rho\in\mathcal{S}}\inf\left\{\sigma : x+\sigma F(\rho,\alpha)\in\mathcal{C}\right\}. \tag{6}$$

We will also make use of

$$W_\alpha(x) \doteq \inf_{\rho\in\mathcal{S}}\inf\left\{\sigma : x+\sigma F(\rho,\alpha)\in\mathcal{C}\right\}. \tag{7}$$

The following condition is necessary and sufficient for $W(x)$ to be finite for all $x \in \mathbb{R}^N$. Let $\mathcal{C}^\circ$ denote the interior of $\mathcal{C}$.

CONDITION 3.1. *For each $\alpha \in \mathcal{A}$ there exists $\rho \in \mathcal{S}$ such that*

$$F(\rho,\alpha) \in \mathcal{C}^\circ.$$

It follows directly from the definition of $W_\alpha(x)$ that under this condition $W_\alpha(x) < \infty$ for all $x \in \mathbb{R}^N$. Since $\mathcal{A}$ is compact, an open covering argument can be used to prove that $W(x) < \infty$ for all $x \in \mathbb{R}^N$.

In order to motivate the representation (6), we first consider (7). In this case there is just "true" control for a fixed set of arrival and service rates. It turns out that W_α equals the minimum time for a control problem that uses the dynamics defined by the Skorokhod problem and stops when the origin is reached. However, from the formula for W_α it is clear that W_α equals the solution to the minimum time problem with the much simpler dynamics $\dot\phi(t) = F(\rho(t),\alpha)$ and the stopping set $\mathcal{C}$. Away from the boundary $\partial\mathbb{R}^N_+$ these two different minimum time problems should satisfy the same Hamilton–Jacobi–Bellman equation. Owing to the constraining dynamics, the first minimum time problem should satisfy a Neumann boundary condition $\langle DW_\alpha(x), d_i\rangle = 0$ for $i \in I(x)$ on $\partial\mathbb{R}^N_+\backslash\{0\}$ (in the viscosity sense). It turns out (under Condition 2.1) that the shape of the stopping set $\mathcal{C}$ in the second minimum

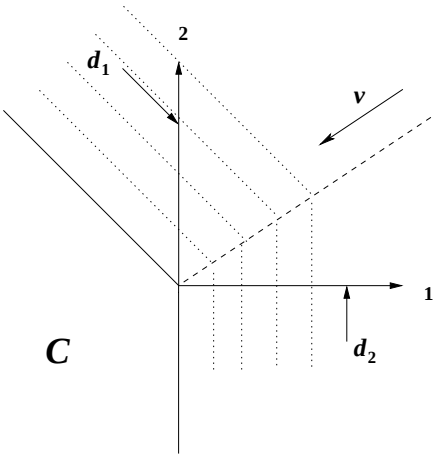


FIG. 5. *Level sets of W_α: Classical boundary conditions.*

time problem produces a function whose gradient satisfies this boundary condition, and so by uniqueness one would expect the two minimum time problems to coincide on $\mathbb{R}^N_+$ [1]. Figure 5 illustrates the situation for a particular two dimensional problem with no control (so that $F(\rho, \alpha) = v$ is a constant). The dotted lines indicate level curves of W_α, and the dashed line indicates the boundary between two regions where the gradient of the value function is constant.

Since the level curves of W_α are parallel to d_i for x near $\{x \in (\mathbb{R}^2_+)\backslash\{0\} : x_i = 0\}$, the gradient is orthogonal to the direction of constraint. Thus the boundary conditions hold, even in a classical sense.

The situation is not always so simple, as indicated by Figure 6. The interpretations of the dotted and dashed lines are the same as in Figure 5. Note that in this case it is only the boundary condition which corresponds to d_1 that holds in the classical sense. Along the boundary that corresponds to d_2 the inner product of the gradient and the direction of constraint is strictly positive. This is due to the fact that on this boundary v points into the interior of $\mathbb{R}^2_+$. One of the important properties of viscosity solutions is that they allow such relaxations of the boundary conditions when they are not physically relevant.

The remarkable fact is that an analogous representation continues to hold even in the game problem, with simply an additional supremization on $\alpha \in \mathcal{A}$. It should be noted that even though the game has value, one cannot permute the $\inf_{\rho\in\mathcal{S}}$ and $\sup_{\alpha\in\mathcal{A}}$ in (6).

In the rest of this section we will prove qualitative properties of W that are needed for the proof that W is the value of the game. W_α turns out to be the value for a control problem, and the identification of W_α as the value function for such problems first appears in [23, 24].

THEOREM 3.1. *Assume that Conditions* 3.1 *and* 2.1 *are satisfied and define $W_\alpha(x)$ for $x \in \mathbb{R}^N$ and $\alpha \in \mathcal{A}$ by* (7)*. The following conclusions hold.*

1. *W_α is finite and radially linear on $\mathbb{R}^N$.*
2. *For each $x \in \mathbb{R}^N$ the infimum in* (7) *is achieved at some probability vector ρ.*
3. *W_α is convex on $\mathbb{R}^N$.*
4. *$W_\alpha(x) > 0$ for $x \in \mathbb{R}^N_+\backslash\{0\}$.*

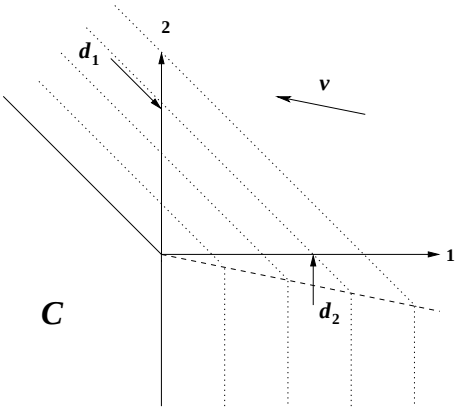


FIG. 6. *Level sets of W_α: Without classical boundary conditions.*

Proof. Under Condition 3.1 it is obvious that the cone $\mathcal{C}$ can be reached from any starting point x, and so $W_\alpha(x) < \infty$, while radially linearity is an immediate consequence of the definition of $W_\alpha(x)$. It follows from the compactness of $\mathcal{S}$ that the infimum is achieved in the definition of $W_\alpha(x)$. Thus the proofs of parts 1 and 2 are complete.

To prove property 3 we first consider points x_1 and x_2 such that $W_\alpha(x_1) = W_\alpha(x_2) \neq 0$. Let c denote the common value, and let ρ^1 and ρ^2 denote minimizing probability vectors in the expression that defines $W_\alpha(x_1)$ and $W_\alpha(x_2)$, respectively. Thus $x_i + cF(\rho^i, \alpha) \in \partial\mathcal{C}$ for $i = 1, 2$. For $s \in [0, 1]$, the convexity of $\mathcal{C}$ implies

$$\begin{aligned} &sx_1 + (1-s)x_2 + scF(\rho^1, \alpha) + (1-s)cF(\rho^2, \alpha) \\ &= (sx_1 + (1-s)x_2) + cF(s\rho^1 + (1-s)\rho^2, \alpha) \in \mathcal{C}. \end{aligned}$$

Since $s\rho^1 + (1-s)\rho^2 \in \mathcal{S}$, it follows that

$$W_\alpha(sx_1 + (1-s)x_2) \leq c = sW_\alpha(x_1) + (1-s)W_\alpha(x_2).$$

We next consider the case of any points x_1 and x_2 such that $W_\alpha(x_1) \neq 0$ and $W_\alpha(x_2) \neq 0$. Let

$$c = sW_\alpha(x_1) + (1-s)W_\alpha(x_2).$$

Since W_α is radially linear,

$$\begin{aligned} &W_\alpha(sx_1 + (1-s)x_2) \\ &= W_\alpha\left(s\frac{W_\alpha(x_1)}{W_\alpha(x_1)}x_1 + (1-s)\frac{W_\alpha(x_2)}{W_\alpha(x_2)}x_2\right) \\ &= W_\alpha\left(\left[\frac{sW_\alpha(x_1)}{c}\frac{x_1}{W_\alpha(x_1)} + \frac{(1-s)W_\alpha(x_2)}{c}\frac{x_2}{W_\alpha(x_2)}\right]c\right) \\ &\leq \frac{sW_\alpha(x_1)}{c}W_\alpha\left(\frac{x_1}{W_\alpha(x_1)}c\right) + \frac{(1-s)W_\alpha(x_2)}{c}W_\alpha\left(\frac{x_2}{W_\alpha(x_2)}c\right) \\ &= sW_\alpha(x_1) + (1-s)W_\alpha(x_2). \end{aligned}$$

The case where $W_\alpha(x_1)$ or $W_\alpha(x_2)$ equals zero is similar and omitted.

Finally we must prove property 4. The Harrison–Reiman condition implies what is called the *completely-$\mathcal{S}$* condition [20, 5], which requires the existence of a vector $v \in \mathbb{R}^N$ satisfying $v_i > 0, i = 1, \ldots, N$, and $\langle v, \gamma \rangle > 0$ for all $\gamma \in d(0)$. Hence if $y \in \mathcal{C} \backslash \{0\}$, then $\langle v, y \rangle < 0$, and so $y \notin \mathbb{R}^N_+$. This shows that

$$\mathcal{C} \cap \mathbb{R}^N_+ = \{0\},$$

and therefore $W_\alpha(x) > 0$ for all $x \in \mathbb{R}^N_+ \backslash \{0\}$. □

THEOREM 3.2. *Assume that Conditions* 3.1 *and* 2.1 *are satisfied and define* $W(x)$ *for* $x \in \mathbb{R}^N$ *by* (6). *The following conclusions hold.*

1. W *is finite and radially linear on* $\mathbb{R}^N$.
2. W *is convex on* $\mathbb{R}^N$.
3. $W(x) > 0$ *for* $x \in \mathbb{R}^N_+ \backslash \{0\}$.

Proof. It follows from Condition 3.1 that for each $x \in \mathbb{R}^N$ $W_\alpha(x)$ is bounded uniformly in $\alpha \in \mathcal{A}$. All the claims then follow from the preceding theorem and $W(x) = \sup_{\alpha \in \mathcal{A}} W_\alpha(x)$. □

Remark. Since W is convex, directional derivatives exist at all points and for all directions.

THEOREM 3.3. *Assume that Conditions* 2.1 *and* 3.1 *are satisfied and define* $W(x), V^-(x)$, *and* $V^+(x)$ *by* (6), (4), *and* (5), *respectively. Then for all* $x \in \mathbb{R}^N_+$

$$W(x) = V^-(x) = V^+(x).$$

4. Proof of the representation. In this section we give the proof of Theorem 3.3. The proof that the differential game has value (i.e., that $V^-(x) = V^+(x)$) is deferred to the appendix. We first prove some preparatory lemmas. Let

$$\rho(x, \alpha) \doteq \{\rho \in \mathcal{S} : x + W_\alpha(x) F(\rho, \alpha) \in \mathcal{C}\}.$$

LEMMA 4.1. *Assume Condition* 3.1. *For each* $x \in \mathbb{R}^N_+$ *and* $\alpha \in \mathcal{A}$ *the set* $\rho(x, \alpha)$ *is nonempty and convex, and moreover the mapping from* $\mathbb{R}^N_+ \times \mathcal{A}$ *to* $\mathcal{S}$ *defined by* $(x, \alpha) \to \rho(x, \alpha)$ *is upper semicontinuous.*

Proof. Fix $x \in \mathbb{R}^N_+$ and $\alpha \in \mathcal{A}$, and let ρ^m come within $1/m$ of the infimum of

$$\inf \{\sigma : x + \sigma F(\rho, \alpha) \in \mathcal{C}\}$$

over $\rho \in \mathcal{S}$. Let σ^m come within $1/m$ of the infimum over σ when $\rho = \rho^m$. By extracting a subsequence, we can assume that $(\rho^m, \sigma^m) \to (\rho^*, W_\alpha(x))$ with $\rho^* \in \mathcal{S}$. We claim that $\rho^* \in \rho(x, \alpha)$. Indeed, we have

$$x + W_\alpha(x) F(\rho^*, \alpha) = \lim_{m \to \infty} (x + \sigma^m F(\rho^m, \alpha)) \in \mathcal{C},$$

which proves that $\rho^* \in \rho(x, \alpha)$ and shows that $\rho(x, \alpha)$ is nonempty. Since $\rho \to F(\rho, \alpha)$ is linear, it follows that $\rho(x, \alpha)$ is also convex.

To prove the upper semicontinuity we first show that $W_\alpha(x)$ is jointly continuous in (x, α). Let $(x^i, \alpha^i) \to (x, \alpha)$ as $i \to \infty$. Under Condition 3.1, for all $\varepsilon > 0$ we can find $\rho \in \mathcal{S}$ such that $x + [W_\alpha(x) + \varepsilon] F(\rho, \alpha) \in \mathcal{C}^\circ$. This implies $\limsup_{i \to \infty} W_{\alpha^i}(x^i) \leq W_\alpha(x) + \varepsilon$, and since $\varepsilon > 0$ is arbitrary $\limsup_{i \to \infty} W_{\alpha^i}(x^i) \leq W_\alpha(x)$. Let $\rho^i \in \rho(x^i, \alpha^i)$. By extracting a subsequence, we can assume that $W_{\alpha^i}(x^i) \to M$ and $\rho^i \to \rho \in \mathcal{S}$. Taking the limit as $i \to \infty$ in

$$x^i + W_{\alpha^i}(x^i) F(\rho^i, \alpha^i) \in \mathcal{C}$$

gives

$$x + MF(\rho, \alpha) \in \mathcal{C},$$

and therefore $\liminf_{i\to\infty} W_{\alpha^i}(x^i) \geq W_\alpha(x)$. We conclude that $W_\alpha(x)$ is jointly continuous in (x, α).

Next let $(x^i, \alpha^i) \to (x, \alpha)$ as $i \to \infty$, and let $\rho^i \in \rho(x^i, \alpha^i)$. We must show that $\rho^i \to \rho^*$ implies $\rho^* \in \rho(x, \alpha)$. Using the continuity of $W_\alpha(x)$,

$$x + W_\alpha(x)F(\rho^*, \alpha) = \lim_{i\to\infty} \left(x^i + W_{\alpha^i}(x^i)F(\rho^i, \alpha^i)\right) \in \mathcal{C}.$$

We conclude that $\rho^* \in \rho(x, \alpha)$, and therefore $(x, \alpha) \to \rho(x, \alpha)$ is upper semicontinuous. □

LEMMA 4.2. *Assume that Conditions* 2.1 *and* 3.1 *are satisfied and define* $W(x)$ *for* $x \in \mathbb{R}^N$ *by* (6). *Consider any point* $x \in \mathbb{R}^N_+ \backslash \{0\}$ *and let* $v \in \mathbb{R}^N$ *be such that* $x + W(x)v \in \mathcal{C}$. *Then*

$$D_v W(x) \leq -1.$$

Proof. Since $x + W(x)v \in \mathcal{C}$, it follows that $W(x + W(x)v) = 0$. The convexity of W then implies that for any $a \in (0, W(x))$

$$\begin{aligned} W(x + av) &\leq \frac{a}{W(x)} W(x + W(x)v) + \left(1 - \frac{a}{W(x)}\right) W(x) \\ &\leq \left(1 - \frac{a}{W(x)}\right) W(x). \end{aligned}$$

It follows that

$$D_v W(x) \doteq \lim_{a\downarrow 0} \frac{W(x + av) - W(x)}{a} \leq -1. \qquad \square$$

Define

$$B(x) \doteq \{v : v = F(\rho, \alpha) \text{ for some } \rho \in \rho(x, \alpha), \alpha \in \mathcal{A}\}.$$

These are the velocities that are optimal (for the true controller) at x for the control problem $W_\alpha(x)$ for some $\alpha \in \mathcal{A}$.

LEMMA 4.3. *Assume that Conditions* 2.1 *and* 3.1 *are satisfied and define* $W(x)$ *for* $x \in \mathbb{R}^N$ *by* (6). *Then for any* $x \in \mathbb{R}^N_+$ *and* $v \in B(x)$ *we have* $x + W(x)v \in \mathcal{C}$.

Proof. Suppose that $v \in B(x)$ corresponds to $\alpha \in \mathcal{A}$. We know that

$$x + W_\alpha(x)v \in \mathcal{C} \text{ and } W_\alpha(x) \leq W(x).$$

If $v \in \mathcal{C}$, then we are done, since $\mathcal{C}$ is a cone with vertex at the origin. Now $x \in \mathbb{R}^N_+$ implies that $v = v_1 + v_2$, where $v_1 \in -\mathbb{R}^N_+$ and $v_2 \in \mathcal{C}$. Thus we need only show $-\mathbb{R}^N_+ \subset \mathcal{C}$. Since $\mathcal{C}$ is a convex cone, to show this it is enough to prove that $-e_i \in \mathcal{C}$ for each $i = 1, \ldots, N$.

Let the vectors $\{d_i^*, i = 1, \ldots, N\}$ be defined by

$$\langle d_i, d_j^* \rangle = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

Condition 2.1 implies the vectors $\{d_i, i = 1, \ldots, N\}$ are linearly independent, and so this is well defined. The vectors $\{d_i^*, i = 1, \ldots, N\}$ provide an external representation for $\mathcal{C}$ in that

$$\mathcal{C} = \left\{ y \in \mathbb{R}^N : \langle y, d_i^* \rangle \leq 0, i = 1, \ldots, N \right\}.$$

Thus $-e_i \in \mathcal{C}$ will follow if we show $\langle e_i, d_j^* \rangle \geq 0$ for $j = 1, \ldots, N$. Let D be the matrix whose ith column is d_i. Then D^{-1} is the matrix whose jth row is d_j^*. We can write $D = I - A$, where A is nonnegative. Under Condition 2.1 the spectral radius of A is less than one, and so we can express D^{-1} as $\sum_{\ell=0}^{\infty} A^\ell$. This shows that D^{-1} is nonnegative and completes the proof of the lemma. □

We recall the definition of the projected velocity given in (1) and $I(x) \doteq \{i : x_i = 0\}$ for $x \in \mathbb{R}_+^N$.

LEMMA 4.4. *Assume that Conditions* 2.1 *and* 3.1 *are satisfied and define* $W(x)$ *for* $x \in \mathbb{R}^N$ *by* (6). *Let* $x \in \mathbb{R}_+^N$ *be given. Let* $y \leq x$ *componentwise, and assume* $y \notin \mathcal{C}$ *(so that* $W(y) > 0$*). Let* $v \in B(y)$*, and suppose there exist* $a_i \geq 0, i \in I(x)$*, such that*

$$\left\langle v + \sum_{i \in I(x)} a_i d_i, e_j \right\rangle = 0, \quad j \in I(x). \tag{8}$$

Let $q = v + \sum_{i \in I(x)} a_i d_i$. *Then*

$$D_q W(y) \leq -1.$$

Proof. By Lemma 4.2 it is enough to show that $y + W(y)q \in \mathcal{C}$. According to the last lemma $y + W(y)v \in \mathcal{C}$, and so we can express $(y/W(y)) + v$ as $-\sum_{i=1}^N \bar{a}_i d_i$ for some constants $\bar{a}_i \geq 0, i = 1, \ldots, N$. To prove

$$\begin{aligned} y + W(y)q &= W(y) \left[(y/W(y)) + v + \sum_{i \in I(x)} a_i d_i \right] \\ &= W(y) \left[-\sum_{i=1}^N \bar{a}_i d_i + \sum_{i \in I(x)} a_i d_i \right] \\ &\in \mathcal{C}, \end{aligned}$$

it is therefore enough to show that $\bar{a}_i \geq a_i$ for $i \in I(x)$.

Since $v = -(y/W(y)) - \sum_{i=1}^N \bar{a}_i d_i$, (8) can be rewritten as

$$\left\langle -(y/W(y)) - \sum_{i=1}^N \bar{a}_i d_i + \sum_{i \in I(x)} a_i d_i, e_j \right\rangle = 0$$

for $j \in I(x)$. Let M denote the cardinality of $I(x)$. We recall that $\langle d_j, e_i \rangle \leq 0$ if $i \neq j$ and $y_i \leq x_i \leq 0$ for $i \in I(x)$. As a consequence, we can rewrite this system of M equations as

$$(I - D_M)r = q,$$

where I is the $M \times M$ identity matrix, D_M is nonnegative with spectral radius less than 1, $r_j = a_j - \bar{a}_j$ for $j \in I(x)$, and $q_i = \sum_{j \notin I(x)} \bar{a}_j \langle d_j, e_i \rangle + (y_i/W(y)) \leq 0$ for

each $i \in I(x)$. Since each component of $r = (\sum_{\ell=0}^{\infty} D_M^\ell)q$ is obviously nonpositive, we conclude that $a_i \le \bar{a}_i$ for $i \in I(x)$. □

In the proof of Theorem 3.3 we will need to construct a nearly optimal strategy for the minimizing player to prove that $V^-(x) \le W(x)$. If W were smooth, then such a strategy would be easy to construct. However, since W is only convex, it must be mollified to construct this policy, and this mollification in turn complicates the construction of the optimal control on the boundary. In the lemma that follows we apply the previous lemma to deal with this issue.

LEMMA 4.5. *Assume that Conditions* 2.1 *and* 3.1 *are satisfied and define* $W(x)$ *for* $x \in \mathbb{R}^N$ *by* (6). *Let* $\gamma > 0$ *be given. Then there exists a convex, continuously differentiable, and radially linear function* $W_\gamma : \mathbb{R}^N \to [0, \infty)$ *such that for all* $x \in \mathbb{R}_+^N$, $\alpha \in \mathcal{A}$, *and* $\rho \in \rho(x, \alpha)$,

$$|W_\gamma(x) - W(x)| \le \gamma W(x) \tag{9}$$

and

$$\langle \pi(x, F(\rho, \alpha)), DW_\gamma(x) \rangle \le -(1 - \gamma). \tag{10}$$

Proof. Fix $\gamma > 0$. We begin by noting a relation between directional derivatives and subdifferentials for convex functions. Fix $x \in \mathbb{R}_+^N$, and let $\partial W(x)$ denote the set of subdifferentials of W at x. Then for any $v \in \mathbb{R}^N$ and any $q \in \partial W(x)$, $\langle q, v \rangle \le D_v W(x)$. According to Lemmas 4.1, 4.2, and 4.3, for each $\alpha \in \mathcal{A}$ and $\rho \in \rho(x, \alpha)$ we have $D_{F(\rho,\alpha)} W(x) \le -1$, and therefore for all such α and ρ

$$\langle q, F(\rho, \alpha) \rangle \le -1 \tag{11}$$

for all $q \in \partial W(x)$.

We next mollify the function W. Define the convex set $G \doteq \{x : W(x) \le 1\}$. For $a > 0$ define the translation $G_a \doteq \{y = x + a(1, \ldots, 1) : x \in G\}$, and for $\delta > 0$ consider the δ-fattening $G_a^\delta \doteq \{y : \|y - x\| \le \delta \text{ for some } x \in G_a\}$. Since $0 \in G^\circ$, we can assume without loss that a is small enough that the origin is contained in the interior of G_a^δ. As we will see, the translation is needed to ensure that the fattening does not interfere with the boundary conditions that are required of the mollification. Finally, let

$$W_a^\delta(x) \doteq \inf\{c \ge 0 : x \in \partial(cG_a^\delta)\}.$$

The construction is illustrated in Figure 7.

It is easy to check that W_a^δ is finite and convex. Also, it is well known that G_a^δ has a C^1 boundary for each $\delta > 0$, and thus W_a^δ is continuously differentiable on $\mathbb{R}_+^N \backslash \{0\}$. We first compute the gradient of W_a^δ. Fix any point $x \in \mathbb{R}_+^N \backslash \{0\}$ and let n be the outward normal to G_a^δ at $y \doteq x/W_a^\delta(x)$. Since W_a^δ is radially linear, the gradient of $W_a^\delta(x)$ must be proportional to n, which means there must be a supporting hyperplane of the form $\langle x, rn \rangle$ to W_a^δ at x (here we use the fact that $W_a^\delta(0) = 0$). Thus using the equality $W_a^\delta(x) = \langle x, rn \rangle$, we find that

$$DW_a^\delta(x) = rn = \left(\frac{W_a^\delta(x)}{\langle x, n \rangle} \right) n = \left(\frac{1}{\langle y, n \rangle} \right) n.$$

Let y' be the unique point in G_a that is exactly distance δ from y, and let $z = y - a(1, \ldots, 1)$. Then n is also an outward normal to G at z, and an analogous

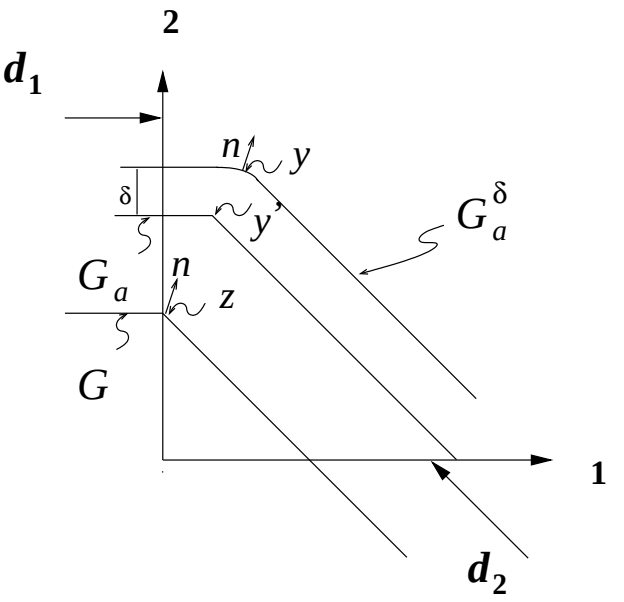


FIG. 7. *Construction of* G_a^δ.

calculation to the one just given shows that for any point of the form bz, $b \in (0, \infty)$, $(1/\langle z, n\rangle)\, n$ is a subdifferential to W at bz. Therefore,

$$DW_a^\delta(x) = \left(\frac{\langle z, n\rangle}{\langle y, n\rangle}\right) q,$$

where q is a subdifferential to W at z. We can make $|y - z|$ as small as desired by choosing $a > 0$ and $\delta > 0$ small. Let $\rho \in \rho(x, \alpha)$.

Since $\langle y, n\rangle$ is uniformly bounded from below away from zero, for all sufficiently small $a > 0$ and $\delta > 0$, (11) implies

$$\langle F(\rho, \alpha), DW_a^\delta(x)\rangle \le -(1 - \gamma).$$

Observe that conditions (8) characterize $\pi(x, F(\rho, \alpha))$. Thus if we knew that $z \le y$ (componentwise), then

$$\langle \pi(x, F(\rho, \alpha)), DW_a^\delta(x)\rangle \le -(1 - \gamma)$$

would also follow from Lemma 4.4. However, $z \le y$ follows easily by fixing $a > 0$ and then choosing $\delta \in (0, a)$. Finally, it is also easy to check that G and G_a^δ can be made arbitrarily close in the Hausdorff topology, which immediately implies

$$|W_a^\delta(x) - W(x)| \le \gamma W(x)$$

when a and δ are small. The lemma now follows by taking $W_\gamma = W_a^\delta$ for suitable $a > 0$ and $\delta > 0$. □

In the proof of Theorem 3.3 we will use a verification argument to show $V^-(x) \le W_\gamma(x)$ plus a small error. The use of feedback controls for the minimizing player would

be problematic. The next lemma will allow the use of piecewise constant controls and thereby simplify the proof. The lemma is an immediate consequence of the continuity of $DW_\gamma(x)$ for $\gamma > 0$ and $x \neq 0$.

LEMMA 4.6. *Assume that Conditions* 2.1 *and* 3.1 *are satisfied and for* $\gamma > 0$ *define* $W_\gamma(x)$ *for* $x \in \mathbb{R}^N$ *by Lemma* 4.5. *Then there is* $\nu > 0$ *such that for all* $z \in \mathbb{R}^N_+$ *with* $\|z\| = 1$, *all* y *with* $\|z - y\| \leq \nu$, *all* $\alpha \in \mathcal{A}$, *and all* $\rho \in \rho(z, \alpha)$,

$$\langle \pi(z, F(\rho, \alpha)), DW_\gamma(y) \rangle \leq -(1 - 2\gamma). \tag{12}$$

Proof of Theorem 3.3. We first prove that $W(x) \leq V^+(x)$. Fix $x \in \mathbb{R}^d_+ \setminus \{0\}$ and $\alpha \in \mathcal{A}$. Let $\rho \in N$ be any open loop control, and let $\tau_x > 0$ be the corresponding first time that the origin is reached by the solution to

$$\dot{\phi}(t) = \pi\left(\phi(t), F(\rho(t), \alpha)\right), \quad \phi(0) = x.$$

If $\tau_x = \infty$, there is nothing to prove, and so we assume $\tau_x < \infty$. Using the definition of the Skorokhod problem, there exist $a_i(t) \geq 0, i = 1, \ldots, N, t \in [0, \tau_x]$, such that

$$\dot{\phi}(t) = F(\rho(t), \alpha) + \sum_{i=1}^{N} a_i(t) d_i$$

for almost every $t \in [0, \tau_x]$. Integrating over $[0, \tau_x]$ and using the definition $\bar{\rho} \doteq \frac{1}{\tau_x} \int_0^{\tau_x} \rho(t) dt$, we find that

$$-x = \tau_x F(\bar{\rho}, \alpha) - \omega$$

for some $\omega \in \mathcal{C}$, and so $x + \tau_x F(\bar{\rho}, \alpha) \in \mathcal{C}$. The definition of $W_\alpha(x)$ then implies $\tau_x \geq W_\alpha(x)$. Since $\theta[\rho](t) = \alpha$ is a legitimate strategy to use in the definition of $V^+(x)$ and $\rho \in N$ is arbitrary, it follows that $V^+(x) \geq W_\alpha(x)$ for all $\alpha \in \mathcal{A}$. Taking the supremum on $\alpha \in \mathcal{A}$ gives $V^+(x) \geq W(x)$.

We next prove $W(x) \geq V^-(x)$. Let $\gamma \in (0, 1/2)$, and let $\nu > 0$ be given according to Lemma 4.6. Fix $x \in \mathbb{R}^d_+ \setminus \{0\}$ and let the open loop control $\alpha \in M$ be given. We recursively construct a strategy $\delta \in \Delta$ as follows. Given a point of the form $x_i \neq 0$ (with $x_0 = x$) and corresponding times τ_i (with $\tau_0 = 0$), we consider the normalized version $z_i = x_i / \|x_i\|$. Let $\rho^*(x, \alpha)$ be any single-valued and measurable selection from $\rho(x, \alpha)$. We define $\delta[\alpha](t)$ for $t \in [\tau_i, \tau_{i+1})$ to be $\rho^*(z_i, \alpha(t))$, where $\tau_{i+1} > \tau_i$ is defined by

$$\inf\{t \geq \tau_i : \|\phi(t)/\|\phi(t)\| - z_i\| \geq \nu\} \wedge \inf\{t \geq \tau_i : \phi(t) = 0\},$$

where

$$\dot{\phi}(t) = \pi(\phi(t), F(\rho^*(z_i, \alpha(t)), \alpha(t))), \quad \phi(\tau_i) = x_i.$$

Since the speed $\|\dot{\phi}(t)\|$ is uniformly bounded from above, it is easy to check that $\inf\{t \geq \tau_i : \|\phi(t)/\|\phi(t)\| - z_i\| \geq \nu\} - \tau_i$ is uniformly bounded away from zero if x_i is in a closed set that does not contain the origin. We will make use of the fact that for any $x \neq 0$ and any v, $\pi(x, v) = \pi(x/\|x\|, v)$. According to Lemma 4.6,

$$\langle DW_\gamma(\phi(t)), \dot{\phi}(t) \rangle = \langle DW_\gamma(\phi(t)), \pi(\phi(t), F(\rho^*(z_i, \alpha(t)), \alpha(t))) \rangle \leq -1 + 2\gamma$$

for almost every t prior to the first time ϕ hits the origin, and therefore for all such times

$$W_\gamma(\phi(t)) - W_\gamma(x) = \int_0^t \langle DW_\gamma(\phi(t)), \dot{\phi}(t)\rangle ds \leq -t(1-2\gamma).$$

We conclude that

$$W_\gamma(\phi(t)) \leq W_\gamma(x) - t(1-2\gamma),$$

and therefore ϕ reaches the origin by time $W_\gamma(x)/(1-2\gamma)$. This implies $V^-(x) \leq W_\gamma(x)/(1-2\gamma)$ and, since $\gamma > 0$ is arbitrary, that $V^-(x) \leq W(x)$.

Thus we have shown that $V^-(x) \leq W(x) \leq V^+(x)$. The proof that $V^-(x) = V^+(x)$ is based on a uniqueness result for the corresponding PDE and is presented in the appendix. This completes the proof of the theorem. □

5. Synthesis of controls. The "true" controls used to prove $W(x) \geq V^-(x)$ in the proof of Theorem 3.3 are not very useful, since they require knowledge of the control that nature applies at all times. In this section we will formally discuss how to construct controls that are optimal (or nearly optimal) and that depend only on the state of the network.

Formally, $W = V^- = V^+$ is the solution to the equation

$$\sup_{\alpha\in\mathcal{A}} \inf_{\rho\in\mathcal{S}} [\langle DW(x), F(\rho,\alpha)\rangle + 1] = 0, \qquad x \in (\mathbb{R}^N_+)^\circ, \tag{13}$$

together with the boundary conditions

$$\langle DW(x), d_i\rangle = 0, \qquad i \in I(x), x \in \partial\mathbb{R}^N_+ \backslash \{0\}, \qquad W(0) = 0. \tag{14}$$

Since F is affine in each variable separately and $\mathcal{A}$ and $\mathcal{S}$ are compact and convex, [22, Corollary 37.6.2] implies that the sup and inf in (13) can be interchanged (i.e., one expects the game to have value).

Since W is not necessarily smooth we cannot expect a classical sense solution to (13)–(14), and so one must consider a weak sense solution, e.g., viscosity solutions. Because W is convex, the set of subdifferentials to W at x (denoted $D^-W(x)$) is never empty. It follows from the characterization of viscosity solutions (see the appendix) that for any $q \in D^-W(x)$ there exists at least one saddle point $(\rho(q), \alpha(q))$ such that

$$\sup_{\alpha\in\mathcal{A}} \langle q, F(\rho(q),\alpha)\rangle \leq -1.$$

Let $R(q)$ denote the set of all points $\rho \in \mathcal{S}$ which have this property. It is easy to check that this set-valued function is upper semicontinuous: $q_n \to q$, $\rho_n \to \rho$, and $\rho_n \in R(q_n)$ implies $\rho \in R(q)$. At each point $x \in \mathbb{R}^N_+$ we define a set of controls $S(x) \subset \mathcal{S}$ by

$$S(x) = \cup_{q\in D^-W(x)} R(q).$$

Note that since $x \to D^-W(x)$ and $q \to R(q)$ are upper semicontinuous, so is the composition $S(x)$, and that the radial linearity of W implies a radial homogeneity of S: $S(ax) = S(x)$ for all $x \in \mathbb{R}^N_+$ and $a \in (0,\infty)$. The set of conjectured controls for x in the interior is then $S(x)$.

However, when on the boundary we must be more careful. As can easily be seen by considering two dimensional examples, there is an important distinction depending on whether the boundary condition holds in a classical sense or not. The following conjectures for the form of the optimal control are based on the analysis of two dimensional examples and have not been verified in any generality. Let us first consider the case of a point x, where $I(x) = i$ for a single value i. In this case the classical sense formulation of the boundary condition is $\langle DW(x), d_i\rangle = 0$. If this condition holds, it means that all optimally controlled trajectories push into the boundary and that any selection from $S(x)$ is optimal. If however $\langle DW(x), d_i\rangle \neq 0$, then even if some elements from $S(x)$ lead to trajectories that push into the boundary, we must restrict ourselves to only those for which the saddle point dynamics do not push strictly into the boundary. If the boundary condition is not valid in the classical sense, then we conjecture that this set is always nonempty. Analogous considerations hold for the points at the intersection of two or more faces. In general, choosing a control for which the saddle point dynamics push into a face is only allowed when the corresponding boundary condition holds in the classical sense.

Appendix. In this appendix we will prove that the game has value, i.e., that $V^+(x) = V^-(x)$ for all $x \in \mathbb{R}^N_+$. A key ingredient is a uniqueness result for the PDE that V^+ and V^- should satisfy. An excellent general reference for the theory of viscosity solutions of first order nonlinear PDEs is the book [3]. The particular results we will need can be found in [1] (see also [10]).

For $q \in \mathbb{R}^N$ define

$$\begin{aligned} H(q) &= \max_{\alpha\in\mathcal{A}} \min_{\rho\in\mathcal{S}} \left[\langle q, F(\rho,\alpha)\rangle + 1\right] \\ &= \min_{\rho\in\mathcal{S}} \max_{\alpha\in\mathcal{A}} \left[\langle q, F(\rho,\alpha)\rangle + 1\right], \end{aligned}$$

where the two expressions on the right-hand side are equal since $F(\rho,\alpha)$ is affine in each variable separately and $\mathcal{S}$ and $\mathcal{A}$ are convex and compact. Consider a Lipschitz continuous function $V : \mathbb{R}^N_+ \to \mathbb{R}$, and for a continuously differentiable function $g : \mathbb{R}^N \to \mathbb{R}$ let y be a local maximum (respectively, minimum) of

$$x \to V(x) - g(x).$$

Then V is called a viscosity subsolution (respectively, viscosity supersolution) to (13) and (14) if

$$H(Dg(y)) \vee \max_{i\in I(y)} \langle Dg(y), d_i\rangle \geq 0, \tag{15}$$

$$\left(H(Dg(y)) \wedge \min_{i\in I(y)} \langle Dg(y), d_i\rangle \leq 0\right), \tag{16}$$

and

$$V(0) \leq 0, \qquad (V(0) \geq 0). \tag{17}$$

We henceforth drop the adjective "viscosity" and note that a function that is both a sub- and supersolution is called a solution.

Recall that $V : \mathbb{R}^N_+ \to \mathbb{R}$ is said to be *radially linear* if $V(ax) = aV(x)$ for all $x \in \mathbb{R}^N_+$ and $a \in [0,\infty)$. According to [1, Theorem 4.3], there is only one function V

satisfying the following conditions: (i) V is a viscosity solution to (15)–(17), (ii) V is Lipschitz continuous and radially linear, and (iii) $V(x) > 0$ for $x \in (\mathbb{R}^N_+)\backslash\{0\}$. Suppose that V^+ and V^- satisfy conditions (ii) and (iii) of the last sentence. Then standard arguments based on dynamic programming can be used to show that (i) holds (see [1, Theorem 3.2] and [3, Chapter VIII]). Thus $V^+ = V^-$ will follow if we can prove that (ii) and (iii) hold for both V^+ and V^-.

Assume for now that V^+ is uniformly bounded on bounded sets. It follows from Theorem 3.3 that $0 \le V^-(x) \le V^+(x) \le \infty$. It is also immediate from the definitions that both $V^+(x)$ and $V^-(x)$ are radially linear and that $V^+(x) \wedge V^-(x) > 0$ for $x \in (\mathbb{R}^N_+)\backslash\{0\}$. Thus if V^+ is uniformly bounded on bounded sets, all that needs to be shown is that V^+ and V^- are Lipschitz continuous. We give the proof for V^+ and note that the proof for V^- is analogous. Let $M \doteq \max_{y:\|y\|=1} V^+(y)$, and assume for now that $M < \infty$. Owing to the radial linearity, $V^+(x) \le M\|x\|$. Fix points $x, y \in \mathbb{R}^N_+$ and $\varepsilon > 0$. Let $K < \infty$ be the Lipschitz constant of the Skorokhod map defined in section 2. We claim that V^+ is Lipschitz continuous with constant MK. The proof adapts a standard argument [3]. Choose $\bar\theta \in \Theta$ such that $V^+(x) \le \inf_{\rho\in N} C_x(\rho, \bar\theta[\rho]) + \varepsilon/2$. Since $\bar\theta$ is suboptimal at y, $V^+(y) \ge \inf_{\rho\in N} C_y(\rho, \bar\theta[\rho])$, and hence there is $\bar\rho$ such that $V^+(y) \ge C_y(\bar\rho, \bar\theta[\bar\rho]) - \varepsilon/2$. Note that also $V^+(x) \le C_x(\bar\rho, \bar\theta[\bar\rho]) + \varepsilon/2$, and hence

$$V^+(x) - V^+(y) \le C_x(\bar\rho, \bar\theta[\bar\rho]) - C_y(\bar\rho, \bar\theta[\bar\rho]) + \varepsilon.$$

If $C_y(\bar\rho, \bar\theta[\bar\rho]) \ge C_x(\bar\rho, \bar\theta[\bar\rho])$ (i.e., it takes longer to reach the origin from y than x), then of course $V^+(x) - V^+(y) \le \varepsilon$. On the other hand, if $C_y(\bar\rho, \bar\theta[\bar\rho]) \le C_x(\bar\rho, \bar\theta[\bar\rho])$, then we can stop the process that was started at x at time $\sigma \doteq C_y(\bar\rho, \bar\theta[\bar\rho])$. If we let $\phi^x(t)$ and $\phi^y(t)$ denote the processes started at the points x and y, then the Lipschitz property of the Skorokhod map implies $\|\phi^x(\sigma) - \phi^y(\sigma)\| \le K\|x - y\|$. Since $\phi^y(\sigma) = 0$, this means that $\|\phi^x(\sigma)\| \le K\|x - y\|$. We can now use dynamic programming to argue that $V^+(x) \le \sigma + V^+(\phi^x(\sigma)) + \varepsilon/2 \le \sigma + MK\|x - y\| + \varepsilon/2$, and thus $V^+(x) - V^+(y) \le MK\|x - y\| + \varepsilon$. Combining the two cases and using that $\varepsilon > 0$ is arbitrary, it follows that $V^+(x) - V^+(y) \le MK\|x - y\|$ for all $x, y \in \mathbb{R}^N_+$.

With the proof that V^+ and V^- are Lipschitz continuous complete, all that remains is to prove that V^+ is uniformly bounded on bounded sets. Under Condition 2.1, it was shown in [9, Lemma 2.1, Theorem 2.1, and p. 60] that there is a compact convex set $B \subset \mathbb{R}^N$ with the following properties:

1. $0 \in B^\circ$,
2. if $z \in \partial B$ and n is an outward normal to B at z, then $|\langle z, e_i\rangle| \le 1$ implies $\langle n, d_i\rangle = 0$,
3. if $z \in \partial B$ and n is an outward normal to B at z, then $\langle z, e_i\rangle\langle n, d_i\rangle \ge 0$.

By considering sets of the form $B^\delta \doteq \{y : \|y - x\| \le \delta \text{ for some } x \in B\}$ (with $\delta > 0$), it is easy to verify that without loss we can assume B has a continuously differentiable boundary. Define the function $R : \mathbb{R}^N \to [0, \infty)$ by

$$R(x) \doteq \inf\{c : x \in \partial(cB)\}.$$

From the convexity, properties 1 and 2 listed above, and the smoothness of ∂B, it follows that R is continuously differentiable save at $x = 0$ and that for $x \in \mathbb{R}^N_+\backslash\{0\}$

$$\langle DR(x), d_i\rangle = 0 \text{ if } i \in I(x). \tag{18}$$

Now fix any point $x \in \mathbb{R}^N_+\backslash\{0\}$, and let $z \in \partial B$ satisfy $z = ax$ for some $a \in (0, \infty)$. If n is the corresponding outward normal to B at z, then $DR(x) = bn$ for some

$b \in (0, \infty)$. According to properties 2 and 3 above, $\langle DR(x), d_i \rangle \geq 0$. Let the vectors $d_i^*, i = 1, \ldots, N$, be defined by $\langle d_i, d_j^* \rangle = \delta_{ij}$, where δ_{ij} is 1 if $i = j$ and 0 otherwise. These vectors are well defined, since Condition 2.1 implies the linear independence of $\{d_i, i = 1, \ldots, N\}$. Writing $DR(x) = \sum_{i=1}^N c_i d_i^*$, it follows from $\langle DR(x), d_i \rangle \geq 0$ that $c_i \geq 0$ for $i = 1, \ldots, N$. We now apply Condition 3.1. It follows from this condition that for each $\alpha \in \mathcal{A}$ there are $\rho \in \mathcal{S}$ and $c > 0$ such that

$$\langle DR(x), F(\rho, \alpha) \rangle \leq -c.$$

Since $\mathcal{A}$ is compact, an open covering argument shows the existence of $c > 0$ such that

$$\max_{\alpha \in \mathcal{A}} \min_{\rho \in \mathcal{S}} \langle DR(x), F(\rho, \alpha) \rangle \leq -c. \tag{19}$$

Finally, the radial linearity of R, the continuity of $DR(x)$, and another open covering argument that uses the compactness of $\partial B \cap \mathbb{R}_+^N$ show that $c > 0$ can be selected so that (19) holds for all $x \in \mathbb{R}_+^N \backslash \{0\}$.

Equations (18) and (19) imply that R/c is a (classical) supersolution to (13) and (14). Standard arguments based on dynamic programming can then be used to show that $V^+(x) \leq R(x)/c$. (See, for example, the proof of Theorem 3.3.) This completes the proof that $V^+(x) = V^-(x)$ for all $x \in \mathbb{R}_+^N$.

REFERENCES

[1] R. Atar and P. Dupuis, *A differential game with constrained dynamics and viscosity solutions of a related HJB equation*, Nonlinear Anal., 51 (2002), pp. 1105–1130.

[2] R. Atar, P. Dupuis, and A. Shwartz, *An escape time criterion for queueing networks: Asymptotic risk-sensitive control via differential games*, Math. Oper. Res., to appear.

[3] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Basel, Switzerland, 1997.

[4] T. Basar and P. Bernhard, *H^∞ Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, Birkhäuser, Basel, Switzerland, 1991.

[5] A. Bernard and A. El Kharroubi, *Régulation de processus dans le premier orthant de $\mathbb{R}^N$*, Stochastics Stochastics Rep., 34 (1991), pp. 149–167.

[6] D. Bertsimas, I. Paschalidis, and J. Tsitsiklis, *Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance*, Ann. Appl. Probab., 4 (1994), pp. 43–75.

[7] A. Budhiraja and P. Dupuis, *Simple necessary and sufficient conditions for the stability of constrained processes*, SIAM J. Appl. Math., 59 (1999), pp. 1686–1700.

[8] M. V. Day, J. Hall, J. Menendez, D. Potter, and I. Rothstein, *Robust optimal service analysis of single-server re-entrant queues*, Comput. Optim. Appl., 22 (2002), pp. 261–302.

[9] P. Dupuis and H. Ishii, *On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications*, Stochastics Stochastics Rep., 35 (1991), pp. 31–62.

[10] P. Dupuis and H. Ishii, *On oblique derivative problems for fully nonlinear second-order elliptic PDEs on domains with corners*, Hokkaido Math J., 20 (1991), pp. 135–164.

[11] P. Dupuis and K. Ramanan, *Convex duality and the Skorokhod problem* II, Probab. Theory Related Fields, 115 (1999), pp. 197–236.

[12] P. Dupuis and K. Ramanan, *An explicit formula for the solution of certain optimal control problems on domains with corners*, Teor. Ĭmovīr. Mat. Stat., 63 (2000), pp. 32–48 (in Russian).

[13] R. J. Elliott and N. J. Kalton, *The Existence of Value in Differential Games*, Memoirs of the American Mathematical Society 126, AMS, Providence, RI, 1972.

[14] J. M. Harrison and M. I. Reiman, *Reflected Brownian motion on an orthant*, Ann. Probab., 9 (1981), pp. 302–308.

[15] J. W. Helton and M. R. James, *Extending H^∞ Control to Nonlinear Systems: Control of Nonlinear Systems to Achieve Performance Objectives*, Adv. Des. Control 1, SIAM, Philadelphia, 1999.

[16] R. Leelahakriengkrai and R. Agrawal, *Scheduling in Multimedia CDMA Wireless Networks*, preprint, University of Wisconsin, Madison, Wisconsin, 1999.
[17] C. Maglaras, *A methodology for dynamic control policy design for stochastic processing networks via fluid models*, in Proceedings of the 36th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1997, pp. 1208–1219.
[18] W. M. McEneaney, *Uniqueness for viscosity solutions of nonstationary Hamilton–Jacobi–Bellman equations under some a priori conditions (with applications)*, SIAM J. Control Optim., 33 (1995), pp. 1560–1576.
[19] S. P. Meyn, *Sequencing and routing in multiclass queueing networks. Part* I: *Feedback regulation*, SIAM J. Control Optim., 40 (2001), pp. 741–776.
[20] M. I. Reiman and R. J. Williams, *A boundary property of semimartingale reflecting Brownian motions*, Probab. Theory Related Fields, 77 (1988), pp. 87–97.
[21] M. I. Reiman, *Open queueing networks in heavy traffic*, Math. Oper. Res., 9 (1984), pp. 441–458.
[22] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[23] G. Weiss, *On optimal draining of re-entrant fluid lines*, in Stochastic Networks, IMA Vol. Math. Appl. 71, F. P. Kelley and R. J. Williams, eds., Springer-Verlag, New York, 1995, pp. 91–103.
[24] G. Weiss, *Optimal draining of re-entrant fluid lines: Some solved examples*, in Stochastic Networks, Theory and Applications, Lecture Notes Ser. 4, F. P. Kelley, S. Zachary, and I. Ziedins, eds., Claredon Press, Oxford, UK, 1996, pp. 19–34.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1876–1894

ZERO-SUM AVERAGE SEMI-MARKOV GAMES: FIXED-POINT SOLUTIONS OF THE SHAPLEY EQUATION*

OSCAR VEGA-AMAYA†

Abstract. This paper deals with zero-sum average semi-Markov games with Borel state and action spaces, unbounded payoffs, and mean holding times. A solution to the Shapley equation is obtained via the Banach fixed-point theorem assuming that the model satisfies a Lyapunov-like condition, a growth hypothesis on the payoff function, and the mean holding times, besides standard continuity and compactness requirements.

Key words. zero-sum semi-Markov games, average payoff criterion, Lyapunov conditions, fixed-point approach

AMS subject classifications. 90D10, 90D20, 93E05

DOI. 10.1137/S0363012902408423

1. Introduction. Several recent papers have used variants of a Lyapunov-like condition to solve an average payoff optimization problem for Markovian systems with unbounded payoff and Borel state and action spaces (see, e.g., [9], [13], [14] for Markov models, [15], [22], [32] for semi-Markov models, [11], [16], [25] for zero-sum Markov games, and [17] for zero-sum semi-Markov games). The key property used in all these papers is that the imposed Lyapunov condition yields the so-called *weighted geometric ergodicity (WGE) property*, which is a generalization of the standard *uniform geometric ergodicity* in Markov chain theory (see [10], [12], and [23] for a detailed discussion of these concepts). Roughly speaking, in these papers the WGE property is combined, explicitly or implicitly, either with the *vanishing discount factor approach* or with some variants of the *policy iteration algorithm* for proving their main results. These facts are the first main differences from the present paper since, in spite of imposing a similar stability condition, we use instead a "fixed-point approach" which does not rely, at least explicitly, on the WGE property.

The fixed-point approach allows us to obtain directly the Shapley equation, which in turn yields the existence of a stationary optimal strategy pair or saddle point—see Theorem 4.7 (a) and (b). In contrast, the approaches followed in [11], [16], [25] first show the existence of a stationary saddle point and then establish the Shapley equation. On the other hand, [22], [21], [15], [17] refer to auxiliary models related to the original one; more precisely, [22] and [21] use the so-called *Schweitzer's data transformation* [29], while the analysis in [15] and [17] relies on certain *perturbed models*.

A second key difference concerns the times between two consecutive decision epochs. In contrast with discrete-time Markov control processes and Markov games, the decision epochs in semi-Markov control processes are random; thus it is necessary to ensure that such processes are *regular*, that is, that they experience only finitely many transitions in each finite time period. This is usually done by means of the Ross condition [28, Prop. 5.1], which assumes that transition times are greater than

*Received by the editors May 25, 2002; accepted for publication (in revised form) June 4, 2003; published electronically December 17, 2003. This research was supported by CONACyT (México) under grant 28309-E.

http://www.siam.org/journals/sicon/42-5/40842.html

†Departamento de Matemáticas, Universidad de Sonora, Luis Encinas y Rosales s/n, Hermosillo, Sonora, 83000, México (ovega@gauss.mat.uson.mx).

some $\gamma > 0$ with a probability of at least $\epsilon > 0$, independently of the state of the system and the control chosen (see Ross [28, Prop. 5.1]). On the other hand, Ross' condition also implies that the mean holding time function is bounded below by a positive constant, which plays a crucial role in the approaches followed in [32], [15], [17], [22] , and [21]; in fact, in the latter four references it is also assumed that the mean holding time function is bounded above by a constant. In the present paper, we do not need either the Ross condition or to assume that the mean holding time function is bounded below by a positive constant as is done in most papers (see, e.g., [2], [5], [26], and their references).

It is important to mention that, as a by-product, the fixed-point approach yields a minimax characterization of a certain solution of the Shapley equation—Theorem 4.7 (c)—which seemingly has not been previously discussed in the literature dealing with zero-sum stochastic games.

We should also mention that the fixed-point approach has been used in several early papers (see, e.g., [7], [12], [19], [27]) but under much stronger ergodicity conditions, which, in particular, exclude the case of unbounded payoffs. The variant of the Lyapunov condition we consider here was recently introduced in [30] for Markov control process and was used in [8] to study minimax problems. In fact, the present paper extends to zero-sum semi-Markov games the results of the two latter references.

For brief surveys of the existing literature on stochastic games with finite or denumerable state space the reader can consult [1], [3], [6], [7], and [20].

The remainder of the paper is organized as follows. The semi-Markov game model and the (ratio) expected average payoff (EAP) criterion are introduced in sections 2 and 3, respectively. The assumptions and main results are stated in section 4. The proofs of all results are given in sections 5 and 6.

2. The game model. Throughout the paper we shall use the following notation. Given a *Borel space* S—that is, a Borel subset of a complete separable metric space—$\mathcal{B}(S)$ denotes the Borel σ-algebra and "measurability" always means measurability with respect to $\mathcal{B}(S)$. The class of all probability measures on S is denoted by $\mathbb{P}(S)$. Given two Borel spaces S and S', a *stochastic kernel* $\varphi(\cdot|\cdot)$ on S given S' is a function such that $\varphi(\cdot|s')$ is in $\mathbb{P}(S)$ for each $s' \in S'$, and $\varphi(B|\cdot)$ is a measurable function on S' for each $B \in \mathcal{B}(S)$. Moreover, $\mathbb{R}_+$ stands for the nonnegative real number subset and $\mathbb{N}$ ($\mathbb{N}_0$, resp.) denotes the positive (nonnegative, resp.) integers subset.

The semi-Markov game model. This paper is concerned with a zero-sum semi-Markov game modeled by

$$(\mathbf{X}, \mathbf{A}, \mathbf{B}, K_A, K_B, Q, F, r),$$

where $\mathbf{X}$ is the *state space*, and the sets $\mathbf{A}$ and $\mathbf{B}$ are the *control spaces* for players 1 and 2, respectively. It is assumed that all these sets are *Borel spaces*. The *constraint sets* K_A and K_B are *Borel* subsets of $\mathbf{X} \times \mathbf{A}$ and $\mathbf{X} \times \mathbf{B}$, respectively. Thus, for each $x \in \mathbf{X}$, the *x-sections*

$$A(x) := \{a \in \mathbf{A} : (x, a) \in K_A\},$$
$$B(x) := \{b \in \mathbf{B} : (x, b) \in K_B\}$$

stand for the sets of *admissible actions or controls* for players 1 and 2, respectively.

Now, let

$$\mathbb{K} := \{(x, a, b) : x \in \mathbf{X}, a \in A(x), b \in B(x)\},$$

which, by [24], is a Borel subset of $\mathbf{X}\times\mathbf{A}\times\mathbf{B}$. The *transition law* $Q(\cdot|\cdot)$ of the system is a stochastic kernel on $\mathbf{X}$ given $\mathbb{K}$. For each $(x,a,b,y)\in\mathbb{K}\times\mathbf{X}$, $F(\cdot|x,a,b,y)$ is a distribution function on $\mathbb{R}_+ := [0,+\infty)$, and $F(t|\cdot)$ is a measurable function on $\mathbb{K}\times\mathbf{X}$ for each $t\in\mathbb{R}_+$. Finally, the payoff r is a measurable function on $\mathbb{K}\times\mathbb{R}_+$.

The game is played over an infinite horizon as follows: at time $t=0$ the game is observed in some state $x_0=x\in\mathbf{X}$ and the players independently choose controls $a_0=a\in A(x_0)$ and $b_0=b\in B(x_0)$. Then, the system remains in state $x_0=x$ for a nonnegative random time δ_1 and player 1 receives the amount $r(x,a,b,\delta_1)$ from player 2. At time δ_1 the system jumps to a new state $x_1=x'\in\mathbf{X}$ according to the probability measure $Q(\cdot|x,a,b)$. The distribution of the random variable δ_1, given that the system has jumped into state x', is $F(\cdot|x,a,b,x')$; that is,

$$F(t|x,a,b,x') = \Pr\left[\delta_1\le t|x_0=x,a_0=a,b_0=b,x_1=x'\right]\quad\forall t\in\mathbb{R}_+.$$

Thus, given that $x_0=x, a_0=a$, and $b_0=b$, the distribution of δ_1 is

$$G(t|x,a,b) := \int_0^{+\infty} F(t|x,a,b,y)Q(dy|x,a,b)\quad\forall t\in\mathbb{R}_+,(x,a,b)\in\mathbb{K},$$

and it is called the *holding time distribution*. Immediately after the transition occurs, the players again choose controls, say, $a_1=a'\in A(x')$ and $b_1=b'\in B(x')$, and the above process is repeated over and over again.

This procedure yields a stochastic processes $\{(x_n,a_n,b_n,\delta_{n+1})\}$, where, for each $n\in\mathbb{N}_0$, x_n is the state of the system, a_n and b_n are the control variables for players 1 and 2, respectively, and δ_{n+1} is the *holding time* at state x_n. The goal of player 1 (player 2, resp.) is to maximize (minimize, resp.) his/her flow rewards (costs, resp.)

$$r(x_0,a_0,b_0,\delta_1), r(x_1,a_1,b_1,\delta_2),\dots$$

over an infinite horizon using an "expected average reward (cost) criterion" defined by (3.1) below.

The functions on $\mathbb{K}$ given as

$$\tau(x,a,b) := \int_0^{+\infty} tG(dt|x,a,b), \tag{2.1}$$

$$R(x,a,b) := \int_0^{+\infty} r(x,a,b,t)G(dt|x,a,b) \tag{2.2}$$

are called the *mean holding time* and the *mean payoff*, respectively.

Strategies. Let $H_0:=\mathbf{X}$ and $H_n := \mathbb{K}\times\mathbb{R}_+\times H_{n-1}$ for $n\in\mathbb{N}$. Then, for each $n\in\mathbb{N}_0$, a generic element of H_n is denoted as

$$h_n := (x_0,a_0,b_0,\delta_1,\dots,x_{n-1},a_{n-1},b_{n-1},\delta_n,x_n),$$

which can be thought of as the history of the game up to the time of the nth transition

$$T_n := T_{n-1}+\delta_n,\quad n\in\mathbf{N}, \tag{2.3}$$

where $T_0:=0$.

Thus a *strategy* for player 1 is a sequence $\pi^1 = \{\pi_n^1\}$ of stochastic kernels π_n^1 on $\mathbf{A}$ given H_n satisfying the constraint

$$\pi_n^1(A(x_n)|h_n) = 1 \quad \forall h_n \in H_n, n \in \mathbb{N}_0.$$

The class of all strategies for player 1 is denoted by Π^1.

For each $x \in \mathbf{X}$, let $\mathbb{A}(x) := \mathbb{P}(A(x))$ and denote by Φ^1 the class of all stochastic kernels φ^1 on $\mathbf{A}$ given $\mathbf{X}$ such that $\varphi^1(\cdot|x) \in \mathbb{A}(x)$ for all $x \in \mathbf{X}$. A policy π^1 is called *stationary* if

$$\pi_n^1(\cdot|h_n) = \varphi^1(\cdot|x_n) \quad \forall h_n \in H_n, n \in \mathbb{N}_0,$$

for some stochastic kernel φ^1 in Φ^1. Following a standard convention, Φ^1 is identified with the class of stationary strategies for player 1. The sets of strategies Π^2 and Φ^2 for player 2 are defined in a similar way but writing $B(x)$ and $\mathbb{B}(x)$ instead of $A(x)$ and $\mathbb{A}(x)$, respectively.

Let $(\Omega, \mathcal{F})$ be the (canonical) measurable space consisting of the sample space $\Omega := (\mathbb{K} \times \mathbb{R}_+)^\infty$ and its product σ-algebra. Thus, for each strategy pair $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$ and each "initial state" $x \in \mathbf{X}$, there exists a probability measure $P_x^{\pi^1,\pi^2}$ defined on $(\Omega, \mathcal{F})$ which governs the evolution of the stochastic process $\{(x_n, a_n, b_n, \delta_{n+1})\}$. The expectation operator with respect to the measure probability $P_x^{\pi^1,\pi^2}$ is denoted as $E_x^{\pi^1,\pi^2}$.

Throughout the paper we shall use the following notation: for a measurable function u on $\mathbb{K}$ and a stationary strategy pair $(\varphi^1, \varphi^2) \in \Phi^1 \times \Phi^2$, let

$$u_{\varphi^1,\varphi^2}(x) := \int_{B(x)} \int_{A(x)} u(x,a,b)\varphi^1(da|x)\varphi^2(db|x) \quad \forall x \in \mathbf{X}. \tag{2.4}$$

Thus, in particular, we shall write

$$R_{\varphi^1,\varphi^2}(x) := \int_{B(x)} \int_{A(x)} R(x,a,b)\varphi^1(da|x)\varphi^2(db|x),$$

$$\tau_{\varphi^1,\varphi^2}(x) := \int_{A(x)} \int_{B(x)} \tau(x,a,b)\varphi^1(da|x)\varphi^2(db|x),$$

and, similarly,

$$Q_{\varphi^1,\varphi^2}(\cdot|x) := \int_{B(x)} \int_{A(x)} Q(\cdot|x,a,b)\varphi^1(da|x)\varphi^2(db|x)$$

for all $x \in \mathbf{X}$.

If the players use a stationary strategy pair, say, (φ^1, φ^2), then the state process $\{x_n\}$ is a Markov chain with transition probability $Q_{\varphi^1,\varphi^2}(\cdot|\cdot)$. In this case, the *n-step transition probability* is denoted by $Q^n_{\varphi^1,\varphi^2}(\cdot|\cdot)$ for each $n \in \mathbb{N}_0$, where $Q^0_{\varphi^1,\varphi^2}(\cdot|x)$ is the Dirac measure at $x \in \mathbf{X}$. Thus, for each $u \in B_W(\mathbf{X})$,

$$Q^n_{\varphi^1,\varphi^2}u(x) := \int_{\mathbf{X}} u(y)Q^n_{\varphi^1,\varphi^2}(dy|x) = E_x^{\varphi^1,\varphi^2}u(x_n) \quad \forall x \in \mathbf{X}, n \in \mathbb{N}_0.$$

3. The expected average payoff criterion. The (ratio) *expected average payoff* (EAP) for the strategy pair $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, given the initial state $x_0 = x \in \mathbf{X}$, is defined as

$$J(\pi^1, \pi^2, x) := \liminf_{n\to\infty} \frac{E_x^{\pi^1,\pi^2} \sum_{k=0}^{n-1} r(x_k, a_k, b_k, \delta_{k+1})}{E_x^{\pi^1,\pi^2} T_n}. \tag{3.1}$$

It is easy to verify using properties of conditional expectation that

$$E_x^{\pi^1,\pi^2} \delta_{k+1} = E_x^{\pi^1,\pi^2} \tau(x_k, a_k, b_k)$$

and also that

$$E_x^{\pi^1,\pi^2} r(x_k, a_k, b_k, \delta_{k+1}) = E_x^{\pi^1,\pi^2} R(x_k, a_k, b_k)$$

for all $x \in \mathbf{X}, (\pi^1, \pi^2) \in \Pi^1 \times \Pi^2, k \in \mathbb{N}_0$. Thus (3.1) can be rewritten as

$$J(\pi^1, \pi^2, x) = \liminf_{n\to\infty} \frac{E_x^{\pi^1,\pi^2} \sum_{k=0}^{n-1} R(x_k, a_k, b_k)}{E_x^{\pi^1,\pi^2} \sum_{k=0}^{n-1} \tau(x_k, a_k, b_k)}. \tag{3.2}$$

Now consider the functions on $\mathbf{X}$ defined as

$$L(x) := \sup_{\pi^1\in\Pi^1} \inf_{\pi^2\in\Pi^2} J(\pi^1, \pi^2, x) \quad \text{and} \quad U(x) := \inf_{\pi^2\in\Pi^2} \sup_{\pi^1\in\Pi^1} J(\pi^1, \pi^2, x), \tag{3.3}$$

which are called the *lower value* and the *upper value* of the game, respectively, for the ratio EAP criterion. In general, $L(\cdot) \le U(\cdot)$, but if $L(\cdot) = U(\cdot)$ holds, the common function is called the *value of the game* and is denoted by $V(\cdot)$.

If the game has a value $V(\cdot)$, a strategy $\pi_*^1 \in \Pi^1$ is said to be EAP-*optimal* for player 1 if

$$\inf_{\pi^2\in\Pi^2} J(\pi_*^1, \pi^2, x) = V(x) \quad \forall x \in \mathbf{X}.$$

Similarly, $\pi_*^2 \in \Pi^2$ is said to be EAP-*optimal* for player 2 if

$$\sup_{\pi^1\in\Pi^1} J(\pi^1, \pi_*^2, x) = V(x) \quad \forall x \in \mathbf{X}.$$

If π_*^i is EAP-optimal for player i $(i = 1, 2)$, then (π_*^1, π_*^2) is called an *EAP-optimal pair* or *saddle point*. Note that (π_*^1, π_*^2) is EAP-optimal if and only if

$$J(\pi^1, \pi_*^2, x) \le J(\pi_*^1, \pi_*^2, x) \le J(\pi_*^1, \pi^2, x) \quad \forall x \in \mathbf{X}, (\pi^1, \pi^2) \in \Pi^1 \times \Pi^2.$$

4. Assumptions and main results. Practically all the related literature assumes that the mean holding time τ is bounded below by a positive constant (see, e.g., [2], [5], [15], [17], [19], [20], [22], [26], [29], [32], and their references). In the present paper it is only assumed that the mean holding time is a positive function, which together with Proposition 4.4 (a) below implies that the processes experience finitely many transitions on each finite time period (see [31]).

ASSUMPTION 4.1. $\tau(x, a, b) > 0$ *for all* $(x, a, b) \in \mathbb{K}$.

The second hypothesis imposes a growth condition both in the mean holding time and the mean payoff.

ASSUMPTION 4.2. *There exists a measurable function $W(\cdot)$ on $\mathbf{X}$ bounded below by a constant $\theta > 0$ such that*

$$\max\{\tau(x,a,b), |R(x,a,b)|\} \leq KW(x) \quad \forall (x,a,b) \in \mathbb{K},$$

for a fixed positive constant K.

To state the third set of hypotheses—as well as several of its consequences—some notation is required. For a measurable function $u(\cdot)$ on $\mathbf{X}$, define the *weighted norm with respect to W* (W-norm) as

$$||u||_W := \sup_{x \in \mathbf{X}} \frac{|u(x)|}{W(x)},$$

and denote by $B_W(\mathbf{X})$ the Banach space of all measurable functions with finite W-norm. Moreover, for a measure $\gamma(\cdot)$ on $\mathbf{X}$ let

$$\gamma(u) := \int_{\mathbf{X}} u(x)\gamma(dx),$$

whenever the integral is well defined.

ASSUMPTION 4.3 (Lyapunov condition). *There exist a nontrivial measure $\nu(\cdot)$ on $\mathbf{X}$, a nonnegative measurable function $S(\cdot)$ on $\mathbb{K}$, and a positive constant $\lambda < 1$ such that the following hold:*

(a) $\nu(W) < \infty$;
(b) $Q(B|x,a,b) \geq \nu(B)S(x,a,b)$ *for all* $B \in \mathcal{B}(\mathbf{X}), (x,a,b) \in \mathbb{K}$;
(c) $\int_{\mathbf{X}} W(y)Q(dy|x,a,b) \leq \lambda W(x) + S(x,a,b)\nu(W)$ *for all* $(x,a,b) \in \mathbb{K}$;
(d) $\nu(S_{\varphi^1,\varphi^2}) > 0$ *for all* $(\varphi^1,\varphi^2) \in \Phi^1 \times \Phi^2$.

As we mentioned in the introduction, Assumption 4.3 allows us to use a fixed-point approach. More precisely, if we define the kernel

$$\widehat{Q}(\cdot|x,a,b) := Q(\cdot|x,a,b) - \nu(\cdot)S(x,a,b) \quad \forall (x,a,b) \in \mathbb{K}, \tag{4.1}$$

which, under Assumption 4.3, is nonnegative, then Assumption 4.3 (c) can be expressed equivalently as

$$\int_{\mathbf{X}} W(y)\widehat{Q}(dy|x,a,b) \leq \lambda W(x) \quad \forall (x,a,b) \in \mathbb{K}, \tag{4.2}$$

which, roughly speaking, means that $\widehat{Q}(\cdot|\cdot)$ satisfies a certain contraction property. This is precisely the contraction property that we shall exploit to prove our main results (Theorems 4.5 and 4.7 below).

Assumption 4.3 was first used in [30], though it is actually a simplified version of the Lyapunov condition introduced in [9]. Specifically, besides the conditions in Assumption 4.3, [9] assumes the existence of a common irreducibility measure for the transition laws induced by the stationary strategies and also that the inequality in Assumption 4.3 (c) holds uniformly, that is, $\inf_{\varphi^1,\varphi^2} \nu(S_{\varphi^1,\varphi^2}) > 0$. However, as it is shown in [30, Thm. 3.3]—see Proposition 4.4 below—the latter condition is not required and the irreducibility condition is redundant.

As in [9], several other papers have used Lyapunov conditions similar to Assumption 4.3 (see, e.g., [13], [14], [15], [16], [17], [25]). The main difference between the mentioned papers and ours is that they rely on the WGE mentioned in the introduction,

while a fixed-point approach is used herein. For instance, in lieu of Assumption 4.3, papers [15], [16], [17], [25] suppose, roughly speaking, that

$$\int_{\mathbf{X}} W(y)Q(dy|x,a,b) \le \lambda W(x) + b\mathbf{I}_C(x) \quad \forall (x,a,b) \in \mathbb{K}, \tag{4.3}$$

where C is a Borel subset of $\mathbf{X}$, b is a positive constant, and $\lambda \in (0,1)$ and also that

$$Q_{\varphi^1,\varphi^2}(\cdot|x) \ge \delta \mathbf{I}_C(x)\nu_{\varphi^1,\varphi^2}(\cdot) \quad \forall x \in \mathbf{X}, (\varphi^1,\varphi^2) \in \Phi^1 \times \Phi^2, \tag{4.4}$$

where each $\nu_{\varphi^1,\varphi^2}(\cdot)$ is a probability measure on $\mathbf{X}$ and δ is a positive constant. A quick glance at the latter conditions shows that they do not lead directly to a contraction property as in (4.2).

On the other hand, recently Küenle and Schurat[1] [18] also used fixed-point arguments under the conditions (4.3)–(4.4). To do this, they first show that these two conditions imply

$$\int_{\mathbf{X}} V(y)Q_{\varphi^1,\varphi^2}(dy|x) \le \lambda' V(x) + \nu_{\varphi^1,\varphi^2}(V)\mathbf{I}_C(x) \quad \forall x \in \mathbf{X}, (\varphi^1,\varphi^2) \in \Phi^1 \times \Phi^2,$$

with $V := W + b$ and $\lambda' := (\lambda + b)/(1 + b)$, which combined with (4.3) yields a contraction property as in (4.2). However, instead of taking advantage of this property, Küenle and Schurat use some "contraction" operators which lead to a parametrized family of functional equations depending on two parameters; they then show some continuity and monotonicity properties of the solutions of such equations, which in turn yield a solution of the Shapley equation. In contrast, with our approach, the solutions to the Shapley equation are obtained directly using the Banach fixed-point theorem for some operator closely related to the Shapley equation.

In the next proposition some important consequences of Assumptions 4.2 and 4.3 are stated, which are proved in [30] using fixed-points arguments too.

PROPOSITION 4.4. *Suppose that Assumption* 4.3 *holds. Then, for each stationary strategy pair* $(\varphi^1,\varphi^2) \in \Phi^1 \times \Phi^2$*, the following hold:*

(a) *The transition law* $Q_{\varphi^1,\varphi^2}(\cdot|x)$ *is positive Harris recurrent. Thus, in particular, there exists a unique invariant probability measure* $\mu_{\varphi^1,\varphi^2}(\cdot)$*, that is,*

$$\mu_{\varphi^1,\varphi^2}(\cdot) = \int_{\mathbf{X}} Q_{\varphi^1,\varphi^2}(\cdot|x)\mu_{\varphi^1,\varphi^2}(dx).$$

Moreover, ν *is an irreducibility measure for* $Q_{\varphi^1,\varphi^2}(\cdot|\cdot)$.

(b) $\mu_{\varphi^1,\varphi^2}(W)$ *is finite; in fact, it holds the bounds*

$$\theta \le \mu_{\varphi^1,\varphi^2}(W) \le \frac{\nu(W)}{(1-\lambda)\nu(X)}. \tag{4.5}$$

Next observe that, under the Assumptions 4.1–4.3, by Proposition 4.4 the constants

$$\rho(\varphi^1,\varphi^2) := \frac{\mu_{\varphi^1,\varphi^2}(R_{\varphi^1,\varphi^2})}{\mu_{\varphi^1,\varphi^2}(\tau_{\varphi^1,\varphi^2})} \quad \forall (\varphi^1,\varphi^2) \in \Phi^1 \times \Phi^2 \tag{4.6}$$

[1]The author thanks one of the referees for bringing his attention to the paper by Küenle and Schurat [18].

are finite. Then, for each $(\varphi^1, \varphi^2) \in \Phi^1 \times \Phi^2$, define on $B_W(\mathbf{X})$ the operator

$$L_{\varphi^1,\varphi^2} u(x) := \overline{R}_{\varphi^1,\varphi^2}(x) + \int_{\mathbf{X}} u(y) Q_{\varphi^1,\varphi^2}(dy|x) \quad \forall x \in \mathbf{X}, \tag{4.7}$$

where

$$\overline{R}_{\varphi^1,\varphi^2}(\cdot) := R_{\varphi^1,\varphi^2}(\cdot) - \rho(\varphi^1, \varphi^2)\tau_{\varphi^1,\varphi^2}(\cdot). \tag{4.8}$$

THEOREM 4.5. *Suppose that Assumptions* 4.1, 4.2, *and* 4.3 *hold. Then for each stationary strategy pair* $(\varphi^1, \varphi^2) \in \Phi^1 \times \Phi^2$, *the following hold:*

(a) *There exists a unique function* $h_{\varphi^1,\varphi^2} \in B_W(\mathbf{X})$, *with* $\nu(h_{\varphi^1,\varphi^2}) = 0$, *that satisfies the (semi-Markov) Poisson equation*

$$\begin{aligned} h_{\varphi^1,\varphi^2}(x) &= L_{\varphi^1,\varphi^2} h_{\varphi^1,\varphi^2}(x) \\ &= \overline{R}_{\varphi^1,\varphi^2}(x) + \int_{\mathbf{X}} h_{\varphi^1,\varphi^2}(y) Q_{\varphi^1,\varphi^2}(dy|x) \quad \forall x \in \mathbf{X}. \end{aligned}$$

(b) *Moreover,* $J(\varphi^1, \varphi^2, \cdot) = \rho(\varphi^1, \varphi^2)$.

Now, we impose some compactness/continuity conditions on the model to ensure the existence of measurable minimizers/maximizers; notice that this can be done in several settings (see, e.g., [10, Thm. 3.5, p. 28] or [8, Lemma 3.5]). Here, for simplicity, we consider the following one.

ASSUMPTION 4.6 (compactness/continuity conditions). *For each* $(x, a, b) \in \mathbb{K}$, *the following hold:*

(a) $A(x)$ *and* $B(x)$ *are nonempty compact subsets.*

(b) $R(x, \cdot, b)$ *is upper semicontinuous on* $A(x)$, *and* $R(x, a, \cdot)$ *is lower semicontinuous on* $B(x)$.

(c) $\tau(x, \cdot, b)$ *and* $\tau(x, a, \cdot)$ *are continuous on* $A(x)$ *and* $B(x)$, *respectively.*

(d) $S(x, \cdot, b)$ *and* $S(x, a, \cdot)$ *are continuous on* $A(x)$ *and* $B(x)$, *respectively.*

(e) *For each bounded measurable function* v *on* $\mathbf{X}$, *the functions*

$$\int_{\mathbf{X}} v(y) Q(dy|x, \cdot, b) \quad \text{and} \quad \int_{\mathbf{X}} v(y) Q(dy|x, a, \cdot)$$

are continuous on $A(x)$ *and* $B(x)$, *respectively.*

(f) *The functions*

$$\int_{\mathbf{X}} W(y) Q(dy|x, \cdot, b) \quad \text{and} \quad \int_{\mathbf{X}} W(y) Q(dy|x, a, \cdot)$$

are continuous on $A(x)$ *and* $B(x)$, *respectively.*

THEOREM 4.7. *Suppose that Assumptions* 4.1, 4.2, 4.3, *and* 4.6 *hold. Then the following hold:*

(a) *There exist a unique function* $h^* \in B_W(\mathbf{X})$ *with* $\nu(h^*) = 0$, *a stationary strategy pair* $(\varphi^1_*, \varphi^2_*) \in \Phi^1 \times \Phi^2$, *and a constant* ρ^* *which satisfy the Shapley*

equation

$$h^*(x) = \min_{\varphi^2 \in \Phi^2} \left\{ R_{\varphi^1_*, \varphi^2}(x) - \rho^* \tau_{\varphi^1_*, \varphi^2}(x) + \int_{\mathbf{X}} h^*(y) Q_{\varphi^1_*, \varphi^2}(dy|x) \right\} \quad \forall x \in \mathbf{X}$$

$$= \max_{\varphi^1 \in \Phi^1} \left\{ R_{\varphi^1, \varphi^2_*}(x) - \rho^* \tau_{\varphi^1, \varphi^2_*}(x) + \int_{\mathbf{X}} h^*(y) Q_{\varphi^1, \varphi^2_*}(dy|x) \right\}$$

$$= R_{\varphi^1_*, \varphi^2_*}(x) - \rho^* \tau_{\varphi^1_*, \varphi^2_*}(x) + \int_{\mathbf{X}} h^*(y) Q_{\varphi^1_*, \varphi^2_*}(dy|x).$$

(b) *The constant ρ^* is the value of the game and $(\varphi^1_*, \varphi^2_*)$ is an EAP-optimal stationary strategy pair. That is, $J(\varphi^1_*, \varphi^2_*, \cdot) = \rho^*$ and*

$$J(\pi^1, \varphi^2_*, \cdot) \le \rho^* \le J(\varphi^1_*, \pi^2, \cdot) \quad \forall (\pi^1, \pi^2) \in \Pi^1 \times \Pi^2.$$

Hence, by Theorem 4.5*,*

$$h^*(\cdot) = h_{\varphi^1_*, \varphi^2_*}(\cdot).$$

(c) *Moreover,*

$$\rho^* = \rho(\varphi^1_*, \varphi^2_*) = \max_{\varphi^1 \in \Phi^1} \min_{\varphi^2 \in \Phi^2} \rho(\varphi^1, \varphi^2) = \min_{\varphi^2 \in \Phi^2} \max_{\varphi^1 \in \Phi^1} \rho(\varphi^1, \varphi^2), \tag{4.9}$$

$$h^*(\cdot) = h_{\varphi^1_*, \varphi^2_*}(\cdot) = \min_{\varphi^2 \in \mathbf{F}^2} h_{\varphi^1_*, \varphi^2}(\cdot) = \max_{\varphi^1 \in \Phi^1} h_{\varphi^1, \varphi^2_*}(\cdot), \tag{4.10}$$

where $\mathbf{F}^i$ stands for the class of all stationary EAP-optimal strategies for player i ($i = 1, 2$).

It is worth mentioning that, to the best of our knowledge, the minimax characterization of the solution $h^*(\cdot)$ of the Shapley equation given in (4.10) has not been discussed in any of the previous papers dealing with zero-sum stochastic games, even for the case of discrete state space.

5. Proof of Theorem 4.5. For the proof of the results in section 4 several preliminary results are needed. The first few are collected in the next lemma.

LEMMA 5.1. *Suppose that Assumptions* 4.1, 4.2*, and* 4.3 *hold. Then the following hold:*

(a) *For each function u in $B_W(\mathbf{X})$,*

$$\lim_{n \to \infty} \frac{1}{n} E_x^{\pi^1, \pi^2} u(x_n) = 0 \quad \forall x \in \mathbf{X}, (\pi^1, \pi^2) \in \Pi^1 \times \Pi^2.$$

(b) *For each stationary strategy pair $(\varphi^1, \varphi^2) \in \Phi^1 \times \Phi^2$, it holds that*

$$\mu_{\varphi^1, \varphi^2}(S_{\varphi^1, \varphi^2}) \ge \frac{(1 - \lambda)\theta}{\nu(W)} > 0.$$

(c) *If in addition Assumptions* 4.1 *and* 4.2 *hold, then*

$$\frac{\mu_{\varphi^1, \varphi^2}(S_{\varphi^1, \varphi^2})}{\mu_{\varphi^1, \varphi^2}(\tau_{\varphi^1, \varphi^2})} \ge \frac{1 - \lambda}{K \nu(W)} > 0.$$

Proof of Lemma 5.1. First note that Assumption 4.3 (b) implies $S(\cdot,\cdot,\cdot) \leq 1/\nu(\mathbf{X})$. Thus Assumption 4.2 and iterations of the inequality in Assumption 4.3 (c) yield

$$\theta \leq E_x^{\pi^1,\pi^2} W(x_n) \leq \lambda^n W(x) + \frac{\nu(W)(1-\lambda^n)}{\nu(\mathbf{X})(1-\lambda)} \quad \forall x \in \mathbf{X}, (\pi^1,\pi^2) \in \Pi^1 \times \Pi^2, n \in \mathbb{N};$$

thus $\lim_{n\to\infty} \frac{1}{n} E_x^{\pi^1,\pi^2} W(x_n) = 0$. Hence the result in part (a) holds for any function $u(\cdot)$ in $B_W(\mathbf{X})$.

Now, from Assumption 4.3 (c), we have

$$\int_{\mathbf{X}} W(y) Q_{\varphi^1,\varphi^2}(dy|x) \leq \lambda W(x) + \nu(W) S_{\varphi^1,\varphi^2}(x) \quad \forall x \in \mathbf{X},$$

for all pairs $(\varphi^1,\varphi^2) \in \Phi^1 \times \Phi^2$. Then, integrating both sides of this inequality with respect to the invariant probability measure $\mu_{\varphi^1,\varphi^2}(\cdot)$, we obtain

$$\mu_{\varphi^1,\varphi^2}(W) \leq \lambda \mu_{\varphi^1,\varphi^2}(W) + \nu(W) \mu_{\varphi^1,\varphi^2}(S_{\varphi^1,\varphi^2}),$$

which combined with Assumption 4.2 yields the result in part (b). Finally, part (c) follows directly from Assumption 4.1 and part (b). □

The following lemma concerns the existence of solutions to the Poisson equation which, in addition to being interesting in itself, plays a key role in our development. In fact, its proof exhibits the way we take advantage of the contraction property (4.2).

LEMMA 5.2. *Suppose Assumptions* 4.2 *and* 4.3 *hold and let* $(\varphi^1,\varphi^2) \in \Phi^1 \times \Phi^2$ *be fixed but arbitrary. Then, for each function* v *in* $B_W(\mathbf{X})$ *there exists a unique function* h_v *in* $B_W(\mathbf{X})$, *with* $\nu(h_v) = 0$, *which satisfies the Poisson equation*

$$h_v(x) = v(x) - \mu_{\varphi^1,\varphi^2}(v) + \int_{\mathbf{X}} h_v(y) Q_{\varphi^1,\varphi^2}(dy|x) \quad \forall x \in \mathbf{X}. \tag{5.1}$$

Thus, from Lemma 5.1(a),

$$\mu_{\varphi^1,\varphi^2}(v) = \lim_{n\to\infty} \frac{1}{n} E_x^{\varphi^1,\varphi^2} \sum_{k=0}^{n-1} v(x_k) \quad \forall x \in \mathbf{X}. \tag{5.2}$$

Proof of Lemma 5.2. Fix a function $v \in B_W(\mathbf{X})$, and let $\mu(\cdot) := \mu_{\varphi^1,\varphi^2}(\cdot)$, $S(\cdot) := S_{\varphi^1,\varphi^2}(\cdot|\cdot)$, and $Q(\cdot|\cdot) := Q_{\varphi^1,\varphi^2}(\cdot|\cdot)$. Next, define

$$\widehat{T}u(x) = v(x) - \mu(v) + \int_{\mathbf{X}} u(y) \widehat{Q}(dy|x) \quad \forall x \in \mathbf{X}, u \in B_W(\mathbf{X}).$$

By Assumption 4.3 (c), it is clear that $\widehat{T}$ maps $B_W(\mathbf{X})$ into itself. Moreover, for any functions $u, w \in B_W(\mathbf{X})$, it holds that

$$|\widehat{T}u(x) - \widehat{T}w(x)| \leq \int_{\mathbf{X}} |u(y) - w(y)| \, \widehat{Q}(dy|x)$$

$$\leq ||u-w||_W \int_{\mathbf{X}} W(y) \widehat{Q}(dy|x) \leq ||u-w||_W \; \lambda W(x)$$

for all $x \in \mathbf{X}$. Hence

$$||\widehat{T}u - \widehat{T}w||_W \leq \lambda \, ||u - w||_W.$$

That is, $\widehat{T}$ is a contraction operator from $B_W(\mathbf{X})$ into itself with modulus λ. Then, by the Banach fixed-point theorem, there exists a unique function $h_v \in B_W(\mathbf{X})$ that satisfies the equation

$$h_v(x) = v(x) - \mu(v) + \int_{\mathbf{X}} h_v(y)\widehat{Q}(dy|x) \quad \forall x \in \mathbf{X}$$

$$= v(x) - \mu(v) + \int_{\mathbf{X}} h_v(y)Q(dy|x) - \nu(h_v)S(x).$$

Now, an integration with respect to the invariant probability measure $\mu(\cdot)$ in both sides of the last equation yields

$$\nu(h_v)\mu(S) = 0,$$

which, by Lemma 5.1 (b), implies that $\nu(h_v) = 0$. Therefore, h_v satisfies the Poisson equation

$$h_v(x) = v(x) - \mu(v) + \int_{\mathbf{X}} h_v(y)Q(dy|x) \quad \forall x \in \mathbf{X},$$

which proves (5.1).

Finally, the property (5.2) is obtained by iteration of the Poisson equation and using Lemma 5.1 (a). □

Now we proceed to prove Theorem 4.5.

Proof of Theorem 4.5. Let $(\varphi^1, \varphi^2) \in \Phi^1 \times \Phi^2$ be fixed but arbitrary. Thus, since the function

$$v(\cdot) := \overline{R}_{\varphi^1,\varphi^2}(\cdot) = R_{\varphi^1,\varphi^2}(\cdot) - \rho(\varphi^1, \varphi^2)\, \tau_{\varphi^1,\varphi^2}(\cdot)$$

is in $B_W(\mathbf{X})$, by Lemma 5.2 there exists a unique function $h_{\varphi^1,\varphi^2} \in B_W(\mathbf{X})$ with $\nu(h_{\varphi^1,\varphi^2}) = 0$ that satisfies the Poisson equation

$$h_{\varphi^1,\varphi^2}(x) = \overline{R}_{\varphi^1,\varphi^2}(x) + \int_{\mathbf{X}} h_{\varphi^1,\varphi^2}(y)Q_{\varphi^1,\varphi^2}(dy|x) \quad \forall x \in \mathbf{X}.$$

This proves part (a) of the theorem.

Next, to prove part (b), first note that iteration of the last equation yields

$$h_{\varphi^1,\varphi^2}(x) = E_x^{\varphi^1,\varphi^2}\left[\sum_{k=1}^{n-1} R_{\varphi^1,\varphi^2}(x_k) - \rho(\varphi^1, \varphi^2)\sum_{k=1}^{n-1}\tau_{\varphi^1,\varphi^2}(x_k)\right]$$

(5.3)

$$+ \int_{\mathbf{X}} h_{\varphi^1,\varphi^2}(y)Q^n_{\varphi^1,\varphi^2}(dy|x)$$

for all $n \in \mathbb{N}$ and $x \in \mathbf{X}$. Moreover, by Assumptions 4.1 and 4.2, applying Lemma 5.2 with $v(\cdot) := \tau_{\varphi^1,\varphi^2}(\cdot)$, we obtain

$$\mu_{\varphi^1,\varphi^2}(\tau_{\varphi^1,\varphi^2}) = \lim_{n\to\infty} \frac{1}{n} E_x^{\varphi^1,\varphi^2} \sum_{k=1}^{n-1} \tau_{\varphi^1,\varphi^2}(x_k) > 0 \quad \forall x \in \mathbf{X},$$

which combined with (5.3) and Lemma 5.1 (a) implies that

$$\rho(\varphi^1,\varphi^2)=\lim_{n\to\infty}\frac{E_x^{\varphi^1,\varphi^2}\sum_{k=0}^{n-1}R_{\varphi^1,\varphi^2}(x_k)}{E_x^{\varphi^1,\varphi^2}\sum_{k=0}^{n-1}\tau_{\varphi^1,\varphi^2}(x_k)}\quad\forall x\in\mathbf{X}.\qquad\square$$

6. Proof of Theorem 4.7. Define the constants

$$\rho^l:=\sup_{\varphi^1\in\Phi^1}\inf_{\varphi^2\in\Phi^2}\rho(\varphi^1,\varphi^2)\quad\text{and}\quad\rho^u:=\inf_{\varphi^2\in\Phi^2}\sup_{\varphi^1\in\Phi^1}\rho(\varphi^1,\varphi^2).$$

We show in the next lemma that these constants are finite. Observe that this trivially holds if one assumes that the mean holding time function is bounded below by a positive constant.

LEMMA 6.1. *Suppose that Assumptions* 4.1, 4.2, 4.3, *and* 4.6 *hold. Then*

$$|\rho^l|<\infty\quad\textit{and}\quad|\rho^u|<\infty.$$

Proof of Lemma 6.1. Let φ^1 be a fixed but arbitrary stationary strategy for player 1 and consider the Markov (one player) model

$$\mathbf{M}=(\mathbf{X},K_B,\widetilde{Q},\widetilde{\tau}),$$

where $\mathbf{X}$ and K_B are as above, and the transition law and the "one-step cost" function are defined as

$$\widetilde{Q}(\cdot|x,b):=\int_{A(x)}Q(\cdot|x,a,b)\varphi^1(da|x),$$

$$\widetilde{\tau}(x,b):=\int_{A(x)}\tau(x,a,b)\varphi^1(da|x)$$

for all $(x,b)\in K_B$, respectively.

Thus following the notation in (2.4), for all $x\in\mathbf{X}$ and $\varphi^2\in\Phi^2$, define

$$\widetilde{Q}_{\varphi^2}(\cdot|x):=\int_{B(x)}\widetilde{Q}(\cdot|x,b)\varphi^2(db|x),$$

$$\widetilde{\tau}_{\varphi^2}(x):=\int_{B(x)}\widetilde{\tau}(x,b)\varphi^2(db|x).$$

Note that $\widetilde{Q}_{\varphi^2}(\cdot|\cdot)=Q_{\varphi^1,\varphi^2}(\cdot|\cdot)$ and $\widetilde{\tau}_{\varphi^2}(\cdot)=\tau_{\varphi^1,\varphi^2}(\cdot)$ for all $\varphi^2\in\Phi^2$.

The Markov model $\mathbf{M}$ satisfies all the conditions in [30, Thm. 3.6]; hence, in particular, there exists a stationary policy $\varphi^2_+\in\Phi^2$ such that

$$\mu_{\varphi^1,\varphi^2_+}(\tau_{\varphi^1,\varphi^2_+})=\mu_{\varphi^1,\varphi^2_+}(\widetilde{\tau}_{\varphi^2_+})=\inf_{\varphi^2\in\Phi^2}\mu_{\varphi^1,\varphi^2}(\widetilde{\tau}_{\varphi^2}).$$

Then, by Assumption 4.1, it holds that $\mu_{\varphi^1,\varphi^2_+}(\tau_{\varphi^1,\varphi^2_+})>0$. Next observe that

$$|\rho(\varphi^1,\varphi^2)|\le\frac{\mu_{\varphi^1,\varphi^2}(|R_{\varphi^1,\varphi^2}|)}{\mu_{\varphi^1,\varphi^2}(\tau_{\varphi^1,\varphi^2})}\le\frac{\mu_{\varphi^1,\varphi^2}(W)}{\mu_{\varphi^1,\varphi^2_+}(\tau_{\varphi^1,\varphi^2_+})}$$

$$\le\frac{k}{\mu_{\varphi^1,\varphi^2_+}(\tau_{\varphi^1,\varphi^2_+})},$$

where the last inequality follows from (4.5) with $k := \nu(W)[(1-\lambda)\nu(\mathbf{X})]^{-1}$. Hence

$$-\infty < -\frac{k}{\mu_{\varphi^1,\varphi^2_+}(\tau_{\varphi^1,\varphi^2_+})} \leq \inf_{\varphi^2\in\Phi^2} \rho(\varphi^1,\varphi^2) \leq \rho(\varphi^1,\varphi^2) \quad \forall \varphi^1 \in \Phi^1. \tag{6.1}$$

Now fix $\varphi^2 \in \Phi^2$ and proceed as above to get a stationary strategy $\varphi^1_+ \in \Phi$ such that

$$\mu_{\varphi^1_+,\varphi^2}(\tau_{\varphi^1_+,\varphi^2}) = \inf_{\varphi^1\in\Phi^1} \mu_{\varphi^1,\varphi^2}(\tau_{\varphi_1,\varphi^2}) > 0.$$

Then

$$\rho(\varphi^1,\varphi^2) \leq \frac{\mu_{\varphi^1,\varphi^2}(|R_{\varphi^1,\varphi^2}|)}{\mu_{\varphi^1,\varphi^2}(\tau_{\varphi^1,\varphi^2})} \leq \frac{k}{\mu_{\varphi^1_+,\varphi^2}(\tau_{\varphi^1_+,\varphi^2})} < +\infty.$$

Hence

$$\rho(\varphi^1,\varphi^2) \leq \sup_{\varphi^1\in\Phi^1} \rho(\varphi^1,\varphi^2) \leq \frac{k}{\mu_{\varphi^1_+,\varphi^2}(\tau_{\varphi^1_+,\varphi^2})}. \tag{6.2}$$

Therefore, by (6.1)–(6.2),

$$-\infty < \rho^l = \sup_{\varphi^1\in\Phi^1} \inf_{\varphi^2\in\Phi^2} \rho(\varphi^1,\varphi^2) \leq \rho^u = \inf_{\varphi^2\in\Phi^2} \sup_{\varphi^1\in\Phi^1} \rho(\varphi^1,\varphi^2) < +\infty,$$

which proves the desired result. □

For the proof of Theorem 4.7 introduce the following operators: for each $u \in B_W(\mathbf{X})$ define

$$L^l u(x,a,b) := R^l(x,a,b) + \int_{\mathbf{X}} u(y)\widehat{Q}(dy|x,a,b) \quad \forall (x,a,b) \in \mathbb{K}, \tag{6.3}$$

where

$$R^l(x,a,b) := R(x,a,b) - \rho^l \tau(x,a,b) \quad \forall (x,a,b) \in \mathbb{K}. \tag{6.4}$$

Thus, following the notation (2.4), for each strategy pair $(\varphi^1,\varphi^2) \in \Phi^1 \times \Phi^2$ define the operators

$$L^l_{\varphi^1,\varphi^2} u(\cdot) := R^l_{\varphi^1,\varphi^2}(\cdot) + \int_{\mathbf{X}} u(y)\widehat{Q}_{\varphi^1,\varphi^2}(dy|\cdot), \tag{6.5}$$

$$L^* u(\cdot) := \sup_{\varphi^1\in\mathbb{A}(x)} \inf_{\varphi^2\in\mathbb{B}(x)} L^l_{\varphi^1,\varphi^2} u(\cdot) \tag{6.6}$$

for each $u \in B_W(\mathbf{X})$.

The results in the next lemma are a combination of the well-known measurable selection theorem [24] and the Fan minimax theorem [4]. The proof is omitted since it is the same as the proof of Lemma 6.5 in [11] and Lemmas 2, 3, and 4 in [25].

LEMMA 6.2. *Suppose that Assumptions* 4.1, 4.2, 4.3, *and* 4.6 *hold and let* u *be a fixed function in* $B_W(\mathbf{X})$. *Then the following hold:*

(a) *For each* $x \in \mathbf{X}$, *the sets* $\mathbb{A}(x)$ *and* $\mathbb{B}(x)$ *are compact with respect to the weak convergence of measures.*

(b) *For each* $x \in \mathbf{X}, (\varphi^1, \varphi^2) \in \Phi^1 \times \Phi^2$, *and* $u \in B_W(\mathbf{X})$, *the mappings*

$$\varphi^1 \to L^l_{\varphi^1,\varphi^2} u(x),$$

$$\varphi^2 \to L^l_{\varphi^1,\varphi^2} u(x)$$

are upper semicontinuous and lower semicontinuous on $\mathbb{A}(x)$ *and* $\mathbb{B}(x)$, *respectively, with respect to the weak convergence of measures.*

(c) *Moreover, there exists a stationary strategy pair* $(\varphi^1_u, \varphi^2_u) \in \Phi^1 \times \Phi^2$ *such that*

$$L^* u(\cdot) = L^l_{\varphi^1_u,\varphi^2_u} u(\cdot)$$

$$= \max_{\varphi^1 \in \Phi^1} L^l_{\varphi^1,\varphi^2_u} u(\cdot) = \min_{\varphi^2 \in \Phi^2} L^l_{\varphi^1_u,\varphi^2} u(\cdot).$$

Hence $L^* u(\cdot)$ *is in* $B_W(\mathbf{X})$.

The proof of Theorem 4.7 follows the same scheme as that of Lemma 5.2, so we first show—in Lemma 6.3 below—that L^* is a contraction operator from $B_W(\mathbf{X})$ into itself with modulus λ; hence, by the Banach fixed-point theorem, there exists a unique function h^* in $B_W(\mathbf{X})$ such that

$$h^*(\cdot) = L^* h^*(\cdot) = \sup_{\varphi^1 \in \mathbb{A}(x)} \inf_{\varphi^2 \in \mathbb{B}(x)} L^l_{\varphi^1,\varphi^2} h^*(\cdot). \tag{6.7}$$

As a second step, in Lemma 6.4, we prove that

$$\rho^* := \rho^l = \rho^u \quad \text{and} \quad \nu(h^*) \le 0.$$

Once the latter is done, we show in Lemma 6.5 that

$$\nu(h^*) = 0.$$

Then (6.7) becomes

$$h^*(x) = \sup_{\varphi^1 \in \mathbb{A}(x)} \inf_{\varphi^2 \in \mathbb{B}(x)} \left[R_{\varphi^1,\varphi^2}(x) - \rho^* \tau_{\varphi^1,\varphi^2}(x) + \int_{\mathbf{X}} h^*(y) Q_{\varphi^1,\varphi^2}(dy|x) \right]$$

for all $x \in \mathbf{X}$. Hence Lemma 6.2 yields a stationary strategy pair $(\varphi^1_*, \varphi^2_*) \in \Phi^1 \times \Phi^2$ satisfying Theorem 4.7 (a).

LEMMA 6.3. *Suppose that the assumptions in Theorem* 4.7 *hold. Then,* L^* *in* (6.6) *is a contraction operator from* $B_W(\mathbf{X})$ *into itself with modulus* λ. *Thus, by the Banach fixed-point theorem and Lemma* 6.2, *there exist a unique function* h^* *in* $B_W(\mathbf{X})$ *and a stationary strategy pair* $(\varphi^1_*, \varphi^2_*) \in \Phi^1 \times \Phi^2$ *such that*

$$h^*(\cdot) = L^* h^*(\cdot) = L^l_{\varphi^1_*,\varphi^2_*} h^*(\cdot) \tag{6.8}$$

$$= \min_{\varphi^2 \in \mathbb{B}(x)} L^l_{\varphi^1_*,\varphi^2} h^*(\cdot) = \max_{\varphi^1 \in \mathbb{A}(x)} L^l_{\varphi^1,\varphi^2_*} h^*(\cdot). \tag{6.9}$$

Proof of Lemma 6.3. By Lemma 6.2 it remains only to prove that L^* is a contraction operator from $B_W(\mathbf{X})$ into itself with modulus λ. To prove this, consider

arbitrary functions u, v in $B_W(\mathbf{X})$ and observe, by Assumption 4.3 (b) and (4.2), that

$$\left|L^l_{\varphi^1,\varphi^2}u(\cdot) - L^l_{\varphi^1,\varphi^2}v(\cdot)\right| \le \|u-v\|_W \int_{\mathbf{X}} W(y)\widehat{Q}_{\varphi^1,\varphi^2}(dy|\cdot)$$

$$\le \|u-v\|_W \lambda W(\cdot)$$

for all $(\varphi^1,\varphi^2) \in \Phi^1 \times \Phi^2$. This implies that

$$L^l_{\varphi^1,\varphi^2}u(\cdot) \le L^l_{\varphi^1,\varphi^2}v(\cdot) + \|u-v\|_W \lambda W(\cdot) \quad \forall (\varphi^1,\varphi^2) \in \Phi^1 \times \Phi^2.$$

Thus the latter inequality together with Lemma 6.2 implies

$$\inf_{\varphi^2 \in \mathbb{B}(x)} L^l_{\varphi^1,\varphi^2}u(\cdot) \le \inf_{\varphi^2 \in \mathbb{B}(x)} L^l_{\varphi^1,\varphi^2}v(\cdot) + \|u-v\|_W \lambda W(\cdot) \quad \forall \varphi^1 \in \Phi^1,$$

which, using again Lemma 6.2, yields

$$L^*u(\cdot) \le L^*v(\cdot) + \|u-v\|_W \lambda W(\cdot).$$

Similarly, interchanging the role of u and v, it also holds that

$$L^*v(\cdot) \le L^*u(\cdot) + \|u-v\|_W \lambda W(\cdot).$$

Therefore,

$$\|L^*u - L^*v\|_W \le \lambda \|u-v\|_W .$$

That is, L^* is a contraction operator from $B_W(\mathbf{X})$ into itself with modulus λ. Now, the Banach fixed-point theorem together with Lemma 6.2 ensures the existence of a unique function $h^* \in B_W(\mathbf{X})$ and a stationary strategy pair $(\varphi^1_*, \varphi^2_*) \in \Phi^1 \times \Phi^2$ satisfying (6.8)–(6.9). □

LEMMA 6.4. *Suppose that the assumptions in Theorem* 4.7 *hold and let* h^* *be as in Lemma* 6.3. *Then*

$$\nu(h^*) \le 0 \quad \textit{and} \quad \rho^l = \rho^u.$$

Proof of Lemma 6.4. Let $(\varphi^1_*, \varphi^2_*)$ be as in Lemma 6.3. Then

$$h^*(x) = \min_{\varphi^2 \in \mathbb{B}(x)} \left[R^l_{\varphi^1_*,\varphi^2}(x) + \int_{\mathbf{X}} h^*(y)\widehat{Q}_{\varphi^1_*,\varphi^2}(dy|x) \right] \tag{6.10}$$

$$\le R^l_{\varphi^1_*,\varphi^2}(x) + \int_{\mathbf{X}} h^*(y)\widehat{Q}_{\varphi^1_*,\varphi^2}(dy|x)$$

$$= R^l_{\varphi^1_*,\varphi^2}(x) + \int_{\mathbf{X}} h^*(y)Q_{\varphi^1_*,\varphi^2}(dy|x) - \nu(h^*)S_{\varphi^1_*,\varphi^2}(x)$$

for all $x \in \mathbf{X}, \varphi^2 \in \Phi^2$. Then an integration with respect to the invariant probability measure $\mu_{\varphi^1_*,\varphi^2}$ yields

$$0 \le \mu_{\varphi^1_*,\varphi^2}(R^l_{\varphi^1_*,\varphi^2}) - \nu(h^*)\mu_{\varphi^1_*,\varphi^2}(S_{\varphi^1_*,\varphi^2}) \quad \forall \varphi^2 \in \Phi^2,$$

which implies that

$$\nu(h^*)\mu_{\varphi^1_*,\varphi^2}(S_{\varphi^1_*,\varphi^2}) \le \mu_{\varphi^1_*,\varphi^2}(R_{\varphi^1_*,\varphi^2}) - \rho^l \mu_{\varphi^1_*,\varphi^2}(\tau_{\varphi^1_*,\varphi^2})$$

$$= \mu_{\varphi^1_*,\varphi^2}(\tau_{\varphi^1_*,\varphi^2}) \left[\rho(\varphi^1_*,\varphi^2) - \rho^l\right]$$

for all $\varphi^2 \in \Phi^2$. Now, taking the infimum over Φ^2, we obtain

$$\inf_{\varphi^2\in\mathbb{B}(x)} \left[\frac{\nu(h^*)\mu_{\varphi^1_*,\varphi^2}(S_{\varphi^1_*,\varphi^2})}{\mu_{\varphi^1_*,\varphi^2}(\tau_{\varphi^1_*,\varphi^2})}\right] \le \inf_{\varphi^2\in\mathbb{B}(x)} \rho(\varphi^1_*,\varphi^2) - \rho^l \le 0,$$

which, by Assumption 4.1 and Lemma 5.1 (b), implies that

$$\nu(h^*) \le 0.$$

This inequality combined with (6.9) implies

$$h^*(x) = \max_{\varphi^1\in\mathbb{A}(x)} \left[R^l_{\varphi^1,\varphi^2_*}(x) + \int_{\mathbf{X}} h^*(y)\widehat{Q}_{\varphi^1,\varphi^2_*}(dy|x)\right]$$

$$\ge \max_{\varphi^1\in\mathbb{A}(x)} \left[R^l_{\varphi^1,\varphi^2_*}(x) + \int_{\mathbf{X}} h^*(y)Q_{\varphi^1,\varphi^2_*}(dy|x)\right]$$

$$\ge R^l_{\varphi^1,\varphi^2_*}(x) + \int_{\mathbf{X}} h^*(y)Q_{\varphi^1,\varphi^2_*}(dy|x)$$

for all $x \in \mathbf{X}, \varphi^1 \in \Phi^1$. Now, integrating both sides of the latter inequality with respect to the invariant probability measure $\mu_{\varphi^1,\varphi^2_*}$, we see that

$$0 \ge \mu_{\varphi^1,\varphi^2_*}(R^l_{\varphi^1,\varphi^2_*}) = \mu_{\varphi^1,\varphi^2_*}(R_{\varphi^1,\varphi^2_*}) - \rho^l \mu_{\varphi^1,\varphi^2_*}(\tau_{\varphi^1,\varphi^2_*}) \quad \forall \varphi^1 \in \Phi^1,$$

which implies that

$$\rho^l \ge \rho(\varphi^1,\varphi^2_*) = \frac{\mu_{\varphi^1,\varphi^2_*}(R_{\varphi^1,\varphi^2_*})}{\mu_{\varphi^1,\varphi^2_*}(\tau_{\varphi^1,\varphi^2_*})} \quad \forall \varphi^1 \in \Phi^1.$$

Hence

$$\rho^l \ge \sup_{\varphi^1\in\Phi^1} \rho(\varphi^1,\varphi^2_*) \ge \inf_{\varphi^2\in\Phi^2} \sup_{\varphi^1\in\Phi^1} \rho(\varphi^1,\varphi^2) = \rho^u.$$

Therefore, $\rho^l = \rho^u$. □

LEMMA 6.5. *Suppose that the assumptions in Theorem* 4.7 *hold and let* h^* *be as in Lemma* 6.3. *Then*

$$\nu(h^*) = 0.$$

Proof of Lemma 6.5. Let $(\varphi^1_*,\varphi^2_*)$ be as in Lemma 6.3 and put $\rho^* := \rho^l = \rho^u$. By (6.9), we have

$$h^*(x) = \max_{\varphi^1\in\mathbb{A}(x)} \left[R_{\varphi^1,\varphi^2_*}(x) - \rho^*\tau_{\varphi^1,\varphi^2_*}(x) + \int_{\mathbf{X}} h^*(y)\widehat{Q}_{\varphi^1,\varphi^2_*}(dy|x)\right]$$

$$\ge R_{\varphi^1,\varphi^2_*}(x) - \rho^*\tau_{\varphi^1,\varphi^2_*}(x) + \int_{\mathbf{X}} h^*(y)\widehat{Q}_{\varphi^1,\varphi^2_*}(dy|x)$$

for all $x \in \mathbf{X}, \varphi^1 \in \Phi^1$. As above, integrating with respect to the invariant probability measure $\mu_{\varphi^1,\varphi^2_*}$ in both sides of the latter inequality, we obtain

$$\begin{aligned}\nu(h^*)\mu_{\varphi^1,\varphi^2_*}(S_{\varphi^1,\varphi^2_*}) &\geq \mu_{\varphi^1,\varphi^2_*}(\tau_{\varphi^1,\varphi^2_*})\left[\rho(\varphi^1,\varphi^2_*) - \rho^*\right] \\ &= \mu_{\varphi^1,\varphi^2_*}(\tau_{\varphi^1,\varphi^2_*})\left[\rho(\varphi^1,\varphi^2_*) - \inf_{\varphi^2\in\Phi^2}\sup_{\varphi^1\in\Phi^1}\rho(\varphi^1,\varphi^2)\right] \\ &\geq \mu_{\varphi^1,\varphi^2_*}(\tau_{\varphi^1,\varphi^2_*})\left[\rho(\varphi^1,\varphi^2_*) - \sup_{\varphi^1\in\Phi^1}\rho(\varphi^1,\varphi^2_*)\right],\end{aligned}$$

which implies that

$$\frac{\nu(h^*)\mu_{\varphi^1,\varphi^2_*}(S_{\varphi^1,\varphi^2_*})}{\mu_{\varphi^1,\varphi^2_*}(\tau_{\varphi^1,\varphi^2_*})} \geq \rho(\varphi^1,\varphi^2_*) - \sup_{\varphi^1\in\Phi^1}\rho(\varphi^1,\varphi^2_*) \quad \forall\varphi^1 \in \Phi^1.$$

Then

$$\sup_{\varphi^1\in\Phi^1}\left[\frac{\nu(h^*)\mu_{\varphi^1,\varphi^2_*}(S_{\varphi^1,\varphi^2_*})}{\mu_{\varphi^1,\varphi^2_*}(\tau_{\varphi^1,\varphi^2_*})}\right] \geq 0.$$

This inequality implies that $\nu(h^*) \geq 0$. Hence, by Lemma 6.4, $\nu(h^*) = 0$. □

Finally, we are ready for the proof of Theorem 4.7.

Proof of Theorem 4.7. Let h^* and $(\varphi^1_*, \varphi^2_*)$ be as in Lemma 6.3. First note that the proof of part (a) is given throughout Lemmas 6.3, 6.4, and 6.5. Part (b) follows using standard dynamic programming arguments, while the first statement in part (c) is exactly Lemma 6.4. Thus it remains only to prove the equalities in (4.10). To do this first recall that $\mathbf{F}^i$ denotes the class of all stationary optimal strategies for player i, with $i = 1, 2$, which is nonempty because of part (b). Now, define the following operators on $B_W(\mathbf{X})$:

$$Mu(x) := \max_{\varphi^1\in\mathbb{A}(x)}\left[R_{\varphi^1_*,\varphi^2}(x) - \rho^*\tau_{\varphi^1_*,\varphi^2}(x) + \int_{\mathbf{X}} u(y)\widehat{Q}_{\varphi^1_*,\varphi^2}(dy|x)\right],$$

$$Nu(x) := \min_{\varphi^2\in\mathbb{B}(x)}\left[R_{\varphi^1_*,\varphi^2}(x) - \rho^*\tau_{\varphi^1_*,\varphi^2}(x) + \int_{\mathbf{X}} u(y)\widehat{Q}_{\varphi^1_*,\varphi^2}(dy|x)\right]$$

for all $x \in \mathbf{X}$. Proceeding as above, it is easy to check that M and N are well defined and that they are λ-contraction operators on $B_W(\mathbf{X})$ into itself. In fact, from part (a), h^* is the fixed point for both operators; that is,

$$h^*(\cdot) = Mh^*(\cdot) = Nh^*(\cdot).$$

Next choose an arbitrary strategy φ^1_0 in $\mathbf{F}^1$ and note that

$$\rho^* = \rho(\varphi^1_0, \varphi^2_*).$$

Then, by Theorem 4.5, there exists a unique function $h_{\varphi^1_0,\varphi^2_*}$ in $B_W(\mathbf{X})$, with $\nu(h_{\varphi^1_0,\varphi^2_*}) = 0$, which satisfies

$$h_{\varphi^1_0,\varphi^2_*}(x) = R_{\varphi^1_0,\varphi^2_*}(x) - \rho^*\tau_{\varphi^1_0,\varphi^2_*}(x) + \int_{\mathbf{X}} h_{\varphi^1_0,\varphi^2_*}(y)\widehat{Q}_{\varphi^1_0,\varphi^2_*}(dy|x) \quad \forall x \in \mathbf{X}.$$

Next, observe that

$$h_{\varphi_0^1,\varphi_*^2}(\cdot) \leq Mh_{\varphi_0^1,\varphi_*^2}(\cdot),$$

which implies that

$$h_{\varphi_0^1,\varphi_*^2}(\cdot) \leq M^n h_{\varphi_0^1,\varphi_*^2}(\cdot) \quad \forall n \in \mathbb{N}.$$

Now, since M is a contraction and h^* is its fixed point, we have

$$h_{\varphi_0^1,\varphi_*^2}(\cdot) \leq h^*(\cdot).$$

Hence, since $h^*(\cdot) = h_{\varphi_*^1,\varphi_*^2}(\cdot)$ and the policy φ_0^1 was chosen arbitrarily in $\mathbf{F}^1$, we have

$$\max_{\varphi^1 \in \mathbf{F}^1} h_{\varphi^1,\varphi_*^2}(\cdot) = h^*(\cdot).$$

Similar arguments, but using the operator N instead of M, show that

$$h^*(\cdot) = \min_{\varphi^2 \in \mathbf{F}^2} h_{\varphi_*^1,\varphi^2}(\cdot). \qquad \square$$

Acknowledgments. The author thanks Professor Onésimo Hernández-Lerma for his valuable comments on an early version of this work. He also thanks Professor Heinz-Uwe Küenle who kindly showed him a preliminary version of the paper [18].

REFERENCES

[1] E. Altman, A. Hordijk, and F. M. Spieksma, *Contraction conditions for average and α–discount optimality in countable state Markov games with unbounded rewards*, Math. Oper. Res., 22 (1997), pp. 588–618.

[2] S. Bhatnagar and V. S. Borkar, *A convex analytic framework for ergodic control of semi-Markov processes*, Math. Oper. Res., 20 (1995), pp. 923–936.

[3] V. S. Borkar and M. K. Ghosh, *Denumerable stochastic games with limiting average payoff*, J. Optim. Theory Appl., 76 (1993), pp. 539–560.

[4] K. Fan, *Minimax theorems*, Proc. Natl. Acad. Sci. USA, 39 (1953), pp. 42–47.

[5] A. Ferdegruen, P. J. Schweitzer, and H. C. Tijms, *Denumerable undiscounted semi-Markov decision processes with unbounded rewards*, Math. Oper. Res., 8 (1983), pp. 298–313.

[6] J. Filar and K. Vrieze, *Competitive Markov Decision Processes*, Springer-Verlag, New York, 1997.

[7] M. K. Ghosh and A. Bagchi, *Stochastic games with average payoff criterion*, Appl. Math. Optim., 38 (1998), pp. 283–301.

[8] J. I. González-Trejo, O. Hernández-Lerma, and L. F. Hoyos-Reyes, *Minimax control of discrete-time stochastic systems*, SIAM J. Control Optim., 41 (2003), pp. 1626–1659.

[9] E. Gordienko and O. Hernández-Lerma, *Average cost Markov control processes with weighted norms: Existence of canonical policies*, Appl. Math. (Warsaw), 23 (1995), pp. 199–218.

[10] O. Hernández-Lerma and J. B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.

[11] O. Hernández-Lerma and J. B. Lasserre, *Zero-sum stochastic games in Borel spaces: Average payoff criteria*, SIAM J. Control Optim., 39 (2001), pp. 1520–1539.

[12] O. Hernández-Lerma, R. Montes-de-Oca, and R. Cavazos-Cadena, *Recurrence conditions for MDPs with Borel state space*, Ann. Oper. Res., 28 (1991), pp. 29–46.

[13] O. Hernández-Lerma and O. Vega-Amaya, *Infinite-horizon Markov control processes with undiscounted cost criteria: From average to overtaking optimality*, Appl. Math. (Warsaw), 25 (1998), pp. 153–178.

[14] O. Hernández-Lerma, O. Vega-Amaya, and G. Carrasco, *Sample-path optimality and variance-minimization of average cost Markov control processes*, SIAM J. Control Optim., 38 (1999), pp. 79–93.

[15] A. Jaśkiewicz, *An approximation approach to ergodic semi-Markov control processes*, Math. Methods Oper. Res., 54 (2001), pp. 1–19.
[16] A. Jaśkiewicz and A. S. Nowak, *On the optimality equation for zero-sum ergodic stochastic games*, Math. Methods Oper. Res., 54 (2001), pp. 291–301.
[17] A. Jaśkiewicz, *Zero-sum semi-Markov games*, SIAM J. Control Optim., 41 (2002), pp. 723–739.
[18] H.-U. Küenle and R. Schurat, *The optimality equation and ε-optimal strategies in Markov games with average reward criterion*, Math. Methods Oper. Res., 56 (2003), pp. 439–449.
[19] M. Kurano, *Average optimal adaptive policies in semi-Markov decision processes including an unknown parameter*, J. Oper. Res. Soc. Japan, 28 (1985), pp. 252–266.
[20] A. K. Lal and S. Sinha, *Zero-sum two-person semi-Markov games*, J. Appl. Probab., 29 (1992), pp. 56–72.
[21] F. Luque-Vásquez, *Zero-sum semi-Markov games in Borel spaces: Discounted and average payoff*, Bol. Soc. Mat. Mexicana (3), 8 (2002), pp. 227–241.
[22] F. Luque-Vásquez and O. Hernández-Lerma, *Semi-Markov models with average costs*, Appl. Math. (Warsaw), 26 (1999), pp. 315–331.
[23] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.
[24] A. S. Nowak, *Measurable selection theorems for minimax stochastic optimization problems*, SIAM J. Control Optim., 23 (1985), pp. 466–476.
[25] A. S. Nowak, *Optimal strategies in a class of zero-sum ergodic stochastic games*, Math. Methods Oper. Res., 50 (1999), pp. 399–419.
[26] M. L. Puterman, *Markov Decision Processes. Discrete Stochastic Dynamic Programming*, Wiley, New York, 1994.
[27] U. Rieder, *Average optimality in Markov games with general state space*, in Proceedings of the 3rd Conference on Approximation Theory and Optimization, Puebla, México, 1995, http://www.emis.de/proceedings/.
[28] S. M. Ross, *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, 1970.
[29] P. J. Schweitzer, *Iterative solution of functional equations of undiscounted Markov renewal programming*, J. Math. Anal. Appl., 34 (1971), pp. 495–501.
[30] O. Vega-Amaya, *The average cost optimality equation: A fixed point approach*, Bol. Soc. Mat. Mexicana (3), 9 (2003), pp. 185–195.
[31] O. Vega-Amaya, *A Note on the Regularity Property of Semi-Markov Processes with Borel Stat Space*, Reporte de Investigación 18, Departamento de Matemáticas, Universidad de Sonora, México, 2002, http://fractus.mat.uson.mx/~tedi/reportes. Statist. Probab. Lett., to appear.
[32] O. Vega-Amaya and F. Luque-Vásquez, *Sample-path average cost optimality for semi-Markov control processes on Borel spaces: Unbounded costs and mean holding times*, Appl. Math. (Warsaw), 27 (2000), pp. 343–367.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1895–1904

A NEW APPROACH TO THE LUR'E PROBLEM IN THE THEORY OF ABSOLUTE STABILITY*

A. A. ZEVIN† AND M. A. PINSKY‡

Abstract. The classical Lur'e problem consists of finding conditions for absolute stability of a linear system with a nonlinear feedback contained within a prescribed sector. Most of the results obtained on this problem are based on the frequency domain or Lyapunov functions methods which are applied to systems with a time-invariant or periodic linear block. This paper develops a new approach providing a sufficient stability criterion for systems with time-variable coefficients, which is expressed in the transfer function of the linear block and the sector margins of the nonlinear block. The systems for which this criterion is precise are found. It is shown that stability of a system with a sign-constant transfer function is guaranteed by stability of the system with a limit linear feedback (so that, for such systems, the famous Aizerman conjecture is true). This, in particular, is the case for systems with a linear block consisting of an arbitrary number of first order time-dependent links. As an example, the stability criterion is applied to a second order system for which the obtained results are contrasted with ones delivered by the Popov criterion.

Key words. absolute stability, Lur'e problem, time-varying systems, exponential stability criteria, Aizerman class

AMS subject classification. 93D10

DOI. 10.1137/S0363012902409854

1. Introduction.

We consider a linear system controlled by a nonlinear feedback

$$\begin{aligned} \dot{x} &= A(t)x + b(t)\varphi(\sigma, t), \\ \sigma &= (c, x), \end{aligned} \tag{1.1}$$

where $x \in R^n$, $\sigma \in R^1$; the matrix $A(t)$ and the vector $b(t)$ are bounded and piece-wise continuous; (a, b) means the scalar product of vectors a and b.

It is assumed that the function $\varphi(\sigma, t)$ is contained within a prescribed sector, i.e.,

$$K_1\sigma^2 \le \varphi(\sigma, t)\sigma \le K_2\sigma^2. \tag{1.2}$$

Let $\Phi(K_1, K_2)$ be the set of functions $\varphi(\sigma, t)$ assuming (1.2). System (1.1), (1.2) is called absolutely stable in the class $\Phi(K_1, K_2)$ if for any $\varphi \in \Phi(K_1, K_2)$ and $x(0) \in R^n$, the corresponding solution $x(t) \to 0$ as $t \to \infty$. Finding conditions for absolute stability of system (1.1) in the class $\Phi(K_1, K_2)$ is the classical Lur'e problem [1], which has led to an extensive literature over the last few decades. Most of the known results on the problem are obtained by the frequency domain or Lyapunov functions methods and relate to systems with a time-invariant linear block (e.g., [2, 3, 4, 5, 6, 7, 8, 9]). Periodic systems ($A(t) = A(t + T)$) were considered by Yakubovich [10, 11, 12]; as mentioned in the most recent paper [12], a generalization of frequency domain methods on such systems faces significant complications. Stability analysis of more general nonperiodic systems confronts even more complex problems which, in principle, could

*Received by the editors June 17, 2002; accepted for publication (in revised form) May 31, 2003; published electronically December 17, 2003.

http://www.siam.org/journals/sicon/42-5/40985.html

†Transmag Research Institute, Academy of Sciences of Ukraine, 49005 Dnepropetrovsk, Piesarzhevsky 5, Ukraine (zevin@npkista.dp.ua).

‡Department of Mathematics, University of Nevada, Reno, NV 89511 (pinsky@unr.edu).

be tackled by the Lyapunov method, but there is no regular technique providing Lyapunov functions for such systems.

A characterization of a class of systems for which asymptotic stability of the corresponding linear system ($\varphi(\sigma) = k\sigma$) for any $k \in [K_1, K_2]$ implies absolute stability of system (1.1) in the class $\Phi(K_1, K_2)$ (the Aizerman problem [13]) is closely related to the absolute stability problem. Bergen and Williams [14] described the closed-loop systems of the third order belonging to this class. Trukhan extended this result on systems with a linear block consisting of up to five stable links (or any number of identical links) of the first order [15]. In [3] Pyatnitsky noted that the question of whether an analogous statement is true for closed-loop systems with arbitrary links of the first order is very interesting; to our knowledge, the answer still remains unknown after more than 30 years. Gil' proved [17] that a time-invariant ($A(t) \equiv A$) system (1.1), (1.2) belongs to the considered class, provided that the transfer function $w(t-s)$ of the linear part is nonnegative for $t \geq s$. Andrusevich showed [18] that if the off-diagonal elements of the matrix $A + \lambda bc'$ is nonnegative for $0 \leq \lambda \leq K$, then the system obeys the Aizerman conjecture (note that it can be shown that under this condition, the transfer function is nonnegative, and this result follows directly from the Gil' theorem).

This paper develops a new approach to stability analysis of system (1.1), (1.2) centered on studying the corresponding Volterra equation (2.4) for function $\sigma(t)$. A sufficient condition for absolute exponential stability of such systems is derived in section 2, Theorem 1. We indicate a class of systems for which this condition is precise (section 3, Theorem 3). We prove (section 3, Theorem 2) that systems with a sign-constant transfer function admit the Aizerman conjecture, which extends the Gil' theorem on time-varying systems. As a result, it is shown that a single-loop system with an arbitrary number of time-varying first order links belongs to the Aizerman class (section 5) yielding a positive answer to the long-standing question mentioned above.

In section 4 we consider stable system (1.1), (1.2) in the presence of an external perturbation. An upper bound for $|\sigma(t)|$ is found (Theorem 4); under a certain additional condition, it is proved that any two solutions $x_1(t)$ and $x_2(t)$ exponentially approach each other as $t \to \infty$ (Theorem 5).

The proposed approach is applied to a second order nonlinear system in section 5. For a time-invariant system, the absolute stability condition is derived in an explicit form and is compared with one implied by the known Popov criterion. It turns out that the developed criterion is superior for larger, and slightly more conservative for smaller, values of dissipation coefficient than the Popov condition.

2. Criterion for absolute exponential stability. Frequently, the class $\Phi(K_1, K_2)$ is reduced to $\Phi(0, K)$ ($K = K_2 - K_1$) by substitution $\varphi_1(\sigma, t) = \varphi(\sigma, t) - K_1\sigma$. In this paper, putting $\varphi_1(\sigma, t) = \varphi(\sigma, t) - K_1\sigma - 0.5K\sigma$ and retaining the previous notation we reduce it to $\Phi(-K/2, K/2)$; i.e., inequality (1.2) becomes

$$-0.5K\sigma^2 \leq \varphi(\sigma, t)\sigma \leq 0.5K\sigma^2. \tag{2.1}$$

Let $W(t, s)$ ($W(s, s) = I$, where I is the unit matrix) be the transition matrix of the linear equation

$$\dot{x} = A(t)x. \tag{2.2}$$

A solution $x(t)$ of (1.1) satisfies the equation

$$x(t) = W(t,0)x(0) + \int_0^t W(t,s)b(s)\varphi(\sigma(s),s)\mathrm{d}s. \tag{2.3}$$

Multiplying (2.3) by $c(t)$, we obtain the integral Volterra equation of the second kind about $\sigma(t)$:

$$\begin{gathered}\sigma(t) = f(t) + \int_0^t w(t,s)\varphi(\sigma(s),s)\mathrm{d}s,\\ f(t) = (c(t), W(t,0)x(0)), \quad w(t,s) = (c(t), W(t,s)b(s)),\end{gathered} \tag{2.4}$$

where $w(t,s)$ is called the transfer function.

Suppose that (2.2) is uniformly exponentially stable; i.e., there exist positive constants C and Δ such that for any solution $x(t)$,

$$\|x(t)\| \le C\exp[-\Delta(t-s)]\,\|x(s)\|, \quad t > s, \tag{2.5}$$

where $\|x\|$ is any norm of x. Note that for an asymptotically stable system with a constant matrix A, the largest real part of the corresponding eigenvalues can be taken as $-\Delta$.

In view of (2.5), system (1.1), (1.2) is absolutely stable in the class $\Phi(K_1, K_2)$ if for any $\varphi \in \Phi(K_1, K_2)$ and $x(0) \in R^n$, the solution of (2.4), $\sigma(t) \to 0$ as $t \to \infty$. Moreover, if there are positive constants N and β such that

$$|\sigma(t)| \le N\exp(-\beta t)\,|\sigma(0)|, \quad t > 0, \tag{2.6}$$

the stability of the system is exponential.

From (2.5) it follows that for some C,

$$\begin{gathered}\|W(t,s)\| \le C\exp[-\Delta(t-s)],\\ |w(t,s)| \le C\exp[-\Delta(t-s)], \quad |f(t)| \le C\exp(-\Delta t).\end{gathered} \tag{2.7}$$

In view of the second inequality (2.7),

$$M = \sup_t \int_0^t |w(t,s)|\mathrm{d}s < \infty. \tag{2.8}$$

Note that in the case of a time-invariant linear block, $w(t,s) = w(t-s)$. Thus, setting $t - s = z$, we get

$$M = \sup_t \int_0^t |w(z)|\mathrm{d}z = \int_0^\infty |w(z)|\mathrm{d}z. \tag{2.9}$$

The following theorem establishes a sufficient condition for absolute exponential stability of the system (1.1), (2.1).

THEOREM 1. *Under conditions* (2.5) *and*

$$KM < 2, \tag{2.10}$$

system (1.1) *is absolutely exponentially stable in the class* $\Phi(-K/2, K/2)$.

Proof. Let us put

$$M(\beta) = \sup_t \int_0^t \exp[\beta(t-s)]\,|w(t,s)|\,\mathrm{d}s. \tag{2.11}$$

As is seen from (2.11) and (2.8), $M(0) = M$. From the second inequality (2.7) it follows that $M(\beta)$ is bounded for $\beta < \Delta$. By (2.10), there exists β_* such that

$$M(\beta)K < 2 \text{ for } \beta \in [0, \beta_*). \tag{2.12}$$

Setting in (2.4)

$$\sigma(t) = \exp(-\beta t)y(t), \tag{2.13}$$

where $\beta \in (0, \beta_*)$, we obtain

$$y(t) = \exp(\beta t)f(t) + \exp(\beta t)\int_0^t w(t,s)\varphi[\exp(-\beta s)y(s), s]\mathrm{d}s. \tag{2.14}$$

Let us show that $|y(t)|$ is bounded for $t \in (0, \infty)$. In fact, otherwise a sequence $t_1, t_2, \ldots$ ($t_i \to \infty$ as $i \to \infty$) can be found such that $|y(t_i)| \geq |y(t)|$ for $t \in [0, t_i]$. Then from (2.1), (2.11), (2.14), and (2.13) one obtains

$$|y(t_i)| \leq \exp(\beta t_i)\,|f(t_i)| + 0.5M(\beta)K\,|y(t_i)|\,. \tag{2.15}$$

Since $\exp(\beta t)\,|f(t)| \to 0$ as $t \to \infty$ ($\beta < \Delta$) and $0.5M(\beta)K < 1$, inequality (2.15) fails for large i, proving that $|y(t)|$ is bounded for $t \in (0, \infty)$, which along with (2.13) proves inequality (2.6) and the theorem. □

Suppose that the nonlinearity bounds are time-dependent, i.e.,

$$-0.5K(t)\sigma^2 \leq \varphi(\sigma, t)\sigma \leq 0.5K(t)\sigma^2,\ K(t) \geq 0. \tag{2.16}$$

We denote the corresponding class $\Phi(-K(t)/2, K(t)/2)$. It is clear from the above proof that Theorem 1 holds true if KM in condition (2.10) is replaced by the value

$$\Phi = \sup_t \int_0^t |K(t)w(t,s)|\mathrm{d}s. \tag{2.17}$$

Note that the definition of absolute stability often assumes that $f(t)$ in (2.4) is any continuous function such that $|f(t)| \to 0$ as $t \to \infty$, which allows extending the admissible class of systems (1.1) to ones with an external perturbation $u(t)$ disappearing at infinity. Let λ be the Lyapunov exponent of the function $f(t)$, i.e.,

$$\lambda = \lim_{t\to\infty} \sup \frac{1}{t} \log |f(t)|\,.$$

Since $|f(t)| \to 0$ as $t \to \infty$, then $\lambda \leq 0$; to consider exponential stability, we assume $\lambda < 0$. Setting in the previous proof $\beta \in (0, -\lambda)$, we find that Theorem 1 holds true.

3. On the Aizerman problem. As is known (see, e.g., [5, 9]), the famous Aizerman conjecture that stability of system (1.1) with $\varphi(\sigma) = k\sigma$ for any $k \in [0, K]$ implies absolute stability in the class $\Phi(0, K)$ is, in general, false. So the Aizerman problem is to find the systems (1.1) for which such an assertion is actually true.

Suppose that the transfer function is sign-constant, i.e.,

$$w(t,s) \geq 0 \text{ or } w(t,s) \leq 0 \text{ for } t \geq s \geq 0. \tag{3.1}$$

Using the above stability definition, we assume that $f(t)$ in (2.4) is a continuous function such that $|f(t)| \to 0$ as $t \to \infty$ (functions with the Lyapunov exponent $\lambda = 0$ are also admitted; here nonexponential absolute stability is meant).

THEOREM 2. *System* (1.1), (2.1), (2.5), (3.1) *is absolutely stable in the class* $\Phi(-K/2, K/2)$ *if it is stable for* $\varphi = 0.5K\sigma$ $(w(t,s) \geq 0)$ *or for* $\varphi = -0.5K\sigma$ $(w(t,s) \leq 0)$.

Proof. Let $w(t,s) \geq 0$; suppose that for $\varphi = 0.5K\sigma$, the system is absolutely stable. Then for the solution of the equation

$$\sigma(t) = |f(t)| + 0.5\int_0^t Kw(t,s)\sigma(s)\mathrm{d}s, \tag{3.2}$$

$\sigma_1(t) \to 0$ as $t \to \infty$. Without loss of generality we assume that if $f(t) > 0$ for some $(0, t_0)$, then $\sigma_1(t) > 0$ for all $t > 0$ and the solution of (2.4), and $\sigma(t) > 0$ for $t \in (0, t_0)$. Since $\varphi(\sigma) \leq 0.5K\sigma$, then $\sigma_1(t) \geq \sigma(t) > 0$ for $t \in (0, t_0)$; this inequality holds for as long as $\sigma(t) > 0$. Let us show that $\sigma_1(t) \geq |\sigma(t)|$ for $t > 0$. Otherwise, $\sigma_1(t) \geq |\sigma(t)|$ for $t \in (0, t_1)$, and $\sigma_1(t_1) = -\sigma(t_1)$ for some t_1. Putting in (3.2) $\sigma = \sigma_1(t)$ and adding (2.4) and (3.2) for $t = t_1$, we obtain

$$0 = |f(t_1)| + f(t_1) + \int_0^{t_1} w(t,s)[0.5K\sigma_1(s) + \varphi(\sigma(s,s))]\mathrm{d}s. \tag{3.3}$$

Clearly, the integrand in (3.3) is nonnegative, so the right-hand side is positive. The obtained contradiction shows that $\sigma_1(t) \geq |\sigma(t)|$ for $t > 0$, and, therefore, $\sigma(t) \to 0$ as $t \to \infty$. In the case of $w(t,s) \leq 0$ for $t \geq s \geq 0$, the proof is quite analogous. ∎

Theorem 2 yields that for systems with a sign-constant transfer function, the Aizerman conjecture is true in the class $\Phi(-K/2, K/2)$; moreover, it is sufficient to check stability for $k = K/2$ or $k = -K/2$ only. As is clear from the proof, for the general class $\Phi(K_1, K_2)$, the theorem is true if the sign of the transfer function is identical to that of K_*, where K_* is the largest in modulus value of K_1, K_2. Then stability for $\varphi(\sigma) = K_*\sigma$ guarantees stability of the system in the class $\Phi(K_1, K_2)$ and even in the class $\Phi(-|K_*|, |K_*|)$. In particular, for the class $\Phi(0, K)$, the Aizerman conjecture is true if the transfer function is nonnegative (then it is also true for the class $\Phi(-K, K)$). This result significantly generalizes the Gil' theorem obtained for time-invariant systems only [17] (note that the technique of [17] is not extendible on time-varying systems).

Clearly, in the case of time-variable bounds, one has to replace K by $K(t)$ in Theorem 2.

Let K_+ correspond to the limit value of the sector angle in the class $\Phi(-K/2, K/2)$ so that the system is stable for $K < K_+$ and unstable for $K = K_+$. The following theorem gives an explicit expression for this value.

Suppose that the function $m(t) = \int_0^t |w(t,s)|\mathrm{d}s$ increases monotonically; then

$$M = \lim m(t) \text{ as } t \to \infty. \tag{3.4}$$

THEOREM 3. *If, in addition to the conditions of Theorem* 2, *condition* (3.4) *holds, then* $K_+ = 2/M$.

Proof. For $K < 2/M$, the stability follows from Theorem 1; let us show that for $K = 2/M$, the system is unstable.

Consider the equation

$$\begin{aligned} \dot{x} &= A(t)x + kb(t)\sigma + \delta(t), \\ \sigma &= (c, x). \end{aligned} \tag{3.5}$$

Setting in (3.5) $k = 0$, choose $x(0)$ and $\delta(t)$ so that for the corresponding solution $x_0(t)$,

$$\sigma_0(t) = (c, x_0(t)) = \sum_{i=1}^{n} c_i x_{0i}(t) = 1 - m(t)/M. \tag{3.6}$$

To this end, one can put $x_{0k}(t) \equiv \sigma_0(t)/c_k$ ($c_k \neq 0$) and $x_{0i}(t) \equiv 0$, $i = 1, \ldots, n; i \neq k$; then $\delta(t) = \dot{x}_0(t) - A(t)x_0(t)$. By (3.4) and (3.6), $\sigma_0(t) \to 0$, and, therefore,

$$\delta(t) \to 0 \text{ as } t \to \infty. \tag{3.7}$$

Suppose that (3.5) with $k = 1/M$ is exponentially stable. Then, in view of (3.7), for any solution

$$x(t) \to 0 \text{ as } t \to \infty. \tag{3.8}$$

Let $x(t)$ be the solution with $x(0) = x_0(0)$; then the corresponding function $\sigma(t)$ satisfies the equation

$$\sigma(t) = \sigma_0(t) + M^{-1} \int_0^t w(t, s)\sigma(s)\mathrm{d}s. \tag{3.9}$$

By a direct substitution we check that for the solution of (3.9), $\sigma(t) = (c, x(t)) \equiv 1$, which is impossible in view of (3.8). The contradiction obtained shows that for $K = 2/M$, (1.1) is not exponentially stable. □

Thus, for systems considered in Theorem 3, stability condition (2.10) is precise.

4. Systems with external perturbations. Consider now system (1.1), (2.1) in the presence of an external perturbation, i.e.,

$$\dot{x} = A(t)x + b\varphi(\sigma, t) + u(t). \tag{4.1}$$

We assume that if $u(t)$ is bounded on $(0, \infty)$, then so is the solution $f(t)$ of the corresponding linear system ($\varphi(\sigma, t) \equiv 0$). The following theorem gives an upper bound for $|\sigma(t)|$.

Let us put $\sigma^*(T) = \max_t |\sigma(t)|$ and $f^*(T) = \max_t |f(t)|$ for $t \in [0, T]$.

THEOREM 4. *Under condition* (2.10),

$$\sigma^*(T) \leq \frac{f^*(T)}{1 - 0.5KM}. \tag{4.2}$$

Proof. Really, let $\sigma(t^*) = \sigma^*(T)$ for $t^* \in [0, T]$. From (2.4) and (2.1) one gets

$$|\sigma(t^*)| = \sigma^*(T) \leq |f(t^*)| + 0.5KM\,|\sigma(t^*)| \leq f^*(T) + 0.5KM\sigma^*(T), \tag{4.3}$$

which implies required inequality (4.2). □

Let, in particular, $\sup_t |f(t)| \leq f^*$ for $t \in [0, \infty)$; then

$$\sup_t |\sigma(t)| \leq \frac{f^*}{1 - 0.5KM}. \tag{4.4}$$

Suppose that the function $\varphi(\sigma, t)$ satisfies the inequality

$$-\frac{K}{2} \le \frac{\varphi(\sigma_1, t) - \varphi(\sigma_2, t)}{\sigma_1 - \sigma_2} \le \frac{K}{2}. \tag{4.5}$$

As is known [19] for a time-invariant system (1.1), the Popov criterion, along with a condition analogous to (4.5), guarantees convergence of any two solutions $x_1(t)$ and $x_2(t)$ of (4.1) for $t \to \infty$. The following theorem establishes the solution convergence for a general time-variable case.

THEOREM 5. *There are positive constants C and β such that for any solutions $x_1(t)$ and $x_2(t)$ of system* (4.1), (2.1), (2.10), (4.5),

$$\|x_1(t) - x_2(t)\| \le C \exp[-\beta(t - t_0)] \, \|x_1(t_0) - x_2(t_0)\| . \tag{4.6}$$

Proof. The difference $y(t) = x_1(t) - x_2(t)$ satisfies the equation

$$\begin{aligned} &\dot{y} = A(t)y + b[\varphi(\sigma_1, t) - \varphi(\sigma_2, t)], \\ &\sigma_i = (c, x_i), \ i = 1, 2. \end{aligned} \tag{4.7}$$

Setting $\sigma = \sigma_1 - \sigma_2$ and $\phi(\sigma_1, \sigma_2, t) = \varphi(\sigma_1, t) - \varphi(\sigma_2, t)$, from (4.5) we find

$$-0.5K\sigma^2 \le \phi(\sigma_1, \sigma_2, t)\sigma \le 0.5K\sigma^2. \tag{4.8}$$

By Theorem 1, for system (4.7), (4.8), $\sigma(t)$ exponentially tends to zero, which guarantees an exponential decay of $y(t)$. □

5. Discussion. According to Theorem 2, a system with a sign-constant transfer function $w(t, s)$ satisfies the Aizerman conjecture in the class $\Phi(-K/2, K/2)$. Let us apply the obtained criteria to a single-loop system consisting of n links. Suppose that the individual transfer function $w_i(t, s)$ of each link is sign-constant. Since the output of such a link to sign-constant input is also sign-constant, the transfer function of the combined system $w(t, s)$ satisfies inequality (3.1), so the system satisfies Theorem 2.

Let, in particular, the links be of the first order; i.e., they are described by the equations

$$\dot{x}_i + a_i(t)x_i = 0, \ i = 1, \dots, n. \tag{5.1}$$

The functions $a_i(t)$ are not supposed to be necessarily positive, so some links may be unstable. For the individual transfer function of a link,

$$w_i(t, s) = \exp\left[-\int_s^t a_i(s)\mathrm{d}s\right] > 0 \quad \text{for } t > s, \tag{5.2}$$

so the transfer function $w(t, s)$ of the circuit is also positive. Therefore, for stability of the corresponding closed-loop system in the class $\Phi(-K/2, K/2)$, it is sufficient that the system be stable for $\varphi = 0.5K\sigma$ (as was mentioned earlier, the known results of this kind embrace, in contrast, only circuits with up to five stable time-invariant links [14], [15]).

Theorem 4 yields an upper bound for the value $|\sigma(t)|$ in the presence of an external perturbation. Theorem 5 provides exponential convergence of different solutions of a perturbed system. In contrast, the known results of this kind are mainly derived using the Popov approach developed only for time-invariant systems [19].

Note that for a time-variable matrix $A(t)$, finding the constant M requires, in general, numerical integration of the equation $\dot{x} = A(t)x$ to calculate the transfer

function $w(t,s)$. It is simplified by the fact that, as is seen from the second inequality (2.7), $w(t,s)$ exponentially tends to zero as $t-s$ increases.

In the case of a T-periodic matrix $A(t)$, the transition matrix satisfies the relations

$$W(t,s)=W(t,0)W^{-1}(s,0),\quad W(t+kT,0)=W(t,0)W(T,0)^k,$$

so it is sufficient to find $W(t,s)$ only for $0\le s\le t\le T$.

If A is time-invariant, an explicit formula for the function $w(t,s)=w(t-s)$ can be derived using A eigenvalues and eigenvectors.

Now we apply the developed approach to a second order system

$$\begin{aligned}&\dot x_1=x_2,\\ &\dot x_2=-a(t)x_1-2h(t)x_2+\varphi(x,t),\\ &-0,5K(t)x^2\le\varphi(x,t)x\le 0,5K(t)x^2.\end{aligned}\tag{5.3}$$

The corresponding transfer function $w(t,s)$ is a solution of the equation

$$\ddot x+2h(t)\dot x+a(t)x=0,\tag{5.4}$$

satisfying the conditions $w(t,t)=0$ and $\dot w(t,t)=1$.

Theorem 1 implies that for absolute exponential stability of system (5.3), it is sufficient that

$$\Phi=\sup_t\int_0^t|K(t)w(t,s)|\mathrm{d}s<2.\tag{5.5}$$

In fact, $w(t,s)>0$ for small $t-s>0$. If this inequality holds for all $t>s$, then, by Theorem 2, stability of (5.4) is provided by stability of the linear system with $\varphi=0.5K(t)x$. In this case (5.4) is nonoscillatory; i.e., any solution $x(t)$ has no more than one zero for $t\in[0,\infty)$. Note that there are numerous criteria ensuring this property directly through $h(t)$ and $a(t)$, e.g., [20]:

$$a(t)-h^2(t)-\dot h(t)\le 0.\tag{5.6}$$

If system (5.3) and bounds for the nonlinear term are time-invariant ($h(t)\equiv h,\ a(t)\equiv a,\ K(t)\equiv K$), then stability condition (5.5) can be explicitly expressed in these constants. Really, let $a>h^2$; then

$$w(t,s)=w(t-s)=\frac{\exp[-h(t-s)]}{\omega}\sin\omega(t-s),$$

where $\omega=\sqrt{a-h^2}$. Integrating (2.9), we obtain

$$M=\frac{1}{\omega^2}\left[1+2\sum_{k=1}^{\infty}\exp(k\beta)\right]=\frac{1+\exp\beta}{\omega^2(1-\exp\beta)},\quad \beta=-\frac{\pi h}{\omega}.\tag{5.7}$$

Thus, the system is absolutely stable if

$$K<K_0=2/M.\tag{5.8}$$

If $a\le h^2$, inequality (5.6) is true, so (5.4) is nonoscillatory and $w(t,s)>0$. Thus, in this case for exponential stability in the class $\Phi(-K/2,K/2)$, it is necessary and sufficient that the system with $\varphi(x)=0.5Kx$ be stable. Evidently, the last is true if $K<2a$ (one can directly check that for $K=2a$, system (5.3) admits the solution $x_1(t)=\text{const}$, $x_2(t)=0$, and thus is not absolutely stable).

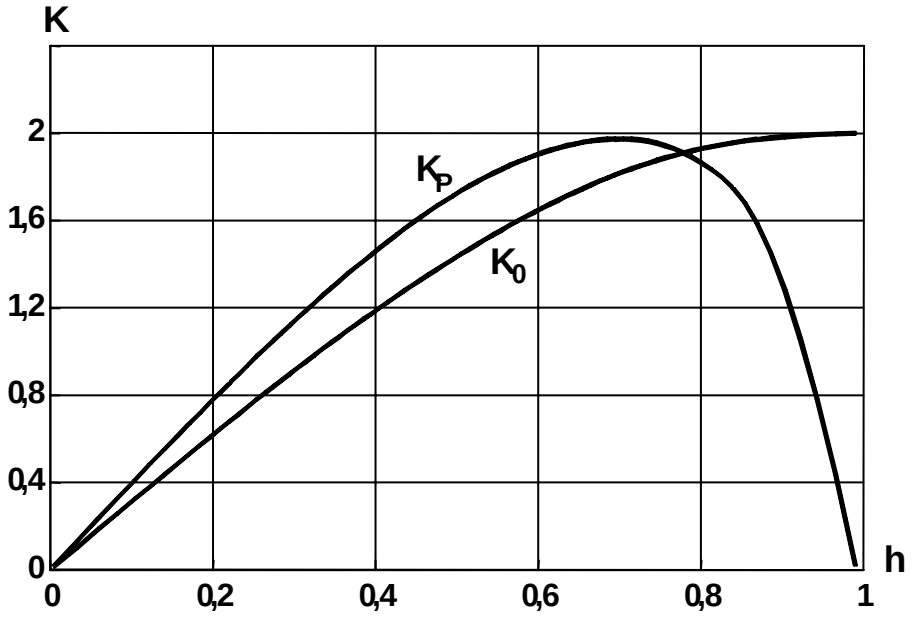


FIG. 1.

The corresponding time-invariant problem could be approached using known methods. For example, for the system

$$\begin{aligned} &\dot{x}_1 = x_2, \\ &\dot{x}_2 = -bx_1 - 2hx_2 - f(x,t), \\ &0 \le f(x,t)x \le Kx^2, \end{aligned} \tag{5.9}$$

the Popov criterion [16] guarantees absolute stability if for all $\omega \in [0, \infty)$,

$$\frac{1}{K} > \frac{\omega^2 - b}{(\omega^2 - b)^2 + 4\omega^2 h^2}. \tag{5.10}$$

The right-hand side of this inequality reaches its maximal value if $\omega^2 = b + 2h\sqrt{b}$, so condition (5.10) holds if

$$K < K_P = 4h\sqrt{b} + 4h^2.$$

Plugging in (5.9), $b = a - K/2$, $f(x,t) = \varphi(x,t) + 0.5Kx$, we reduce it to (5.3). Thus, by the Popov criterion, it is sufficient for absolute stability of system (5.3) that

$$K < K_P = 2h\sqrt{a - h^2}. \tag{5.11}$$

For $a = 1$, the limit values $K_0(h)$ and $K_P(h)$ are obtained using the developed and Popov criteria, and the obtained bounds (formulas (5.8) and (5.11)) are contrasted in Figure 1.

It is clear that for the considered system, the Popov condition is slightly less conservative for relatively small h, while for larger h, and especially close to 1, the presented approach is superior. In fact, for $h \to 1$, $K_0(h)$ approaches the precise bound $K_0(1) = 2$. Let us remind the reader that the described approach is equally applied to a wide class of time-varying systems where known techniques fail.

REFERENCES

[1] A.I. Lur'e, *Some Nonlinear Problems of Automatic Control Theory*, Gostehizdat, Moscow, Leningrad, 1951.
[2] R.E. Kalman, *Lyapunov functions for the problem of Lur'e in automatic control*, Proc. Natl. Acad. Sci. USA, 49 (1963), pp. 201–205.
[3] E.S. Pyatnitsky, *New studies in absolute stability of automatic control systems. Survey*, Avtomat. i Telemeh., 6 (1968), pp. 5–36 (English translation in Autom. Remote Control).
[4] J.C. Willems, *The Analysis of Feedback Systems*, MIT Press, Cambridge, MA, 1971.
[5] K.S. Narendra and J.F. Taylor, *Frequency Domain Criteria for Absolute Stability*, Academic Press, New York, 1973.
[6] M.G. Safonov, *Stability and Robustness of Multivariable Feedback Systems*, MIT Press, Cambridge, MA, 1980.
[7] M. Vidyasagar, *Nonlinear Systems Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[8] A. Megretski and A. Rantzer, *Systems analysis via integral quadratic constraints*, IEEE Trans. Automat. Control, 42 (1997), pp. 819–830.
[9] N.E. Barabanov, *The state space extension method in the theory of absolute stability*, IEEE Trans. Automat. Control, 45 (2000), pp. 2335–2339.
[10] V.A. Yakubovich, *Absolute stability of nonlinear systems with a periodically nonstationary linear part*, Dokl. Akad. Nauk SSSR, 298 (1988), pp. 299–303.
[11] V.A. Yakubovich, *Dichotomy and absolute stability of nonlinear systems with periodically nonstationary linear part*, Systems Control Lett., 11 (1988), pp. 221–228.
[12] V.A. Yakubovich, *Necessity for square-law criterion of an absolute stability of systems with a periodically non-stationary linear part*, Avtomat. i Telemeh., 12 (2000), pp. 62–74 (English translation in Autom. Remote Control).
[13] M.A. Aizerman, *On a conjecture from absolute stability theory*, Uspekhi Mat. Nauk, 4 (1949), pp. 25–49 (in Russian).
[14] A.R. Bergen and S. Williams, *Verification of Aizerman's conjecture*, IRE Trans. Automat. Control, AC-7 (1962), pp. 42–46.
[15] N.M. Trukhan, *On single-loop systems absolutely stable in the Hurwitzian angle*, Avtomat. i Telemeh, 11 (1966), pp. 5–8 (English translation in Autom. Remote Control).
[16] V.M. Popov, *Absolute stability of nonlinear systems of automatic control*, Autom. Remote Control, 22 (1962), pp. 857–875.
[17] M.I. Gil', *On one class of absolutely stable systems*, Soviet Phys. Dokl., 280 (1983), pp. 811–816.
[18] V.V. Andrusevich, *On one class of absolutely stable nonlinear nonstationary systems*, Avtomat. i Telemeh, 11 (1985), pp. 26–33 (English translation in Autom. Remote Control).
[19] V.A. Yakubovich, *The method of matrix inequalities in the theory of stability of nonlinear control systems. Absolute stability of forced oscillations*, Avtomat. i Telemeh, 25 (1964), pp. 1017–1029 (English translation in Autom. Remote Control).
[20] D. Willet, *Classification of second order differential equations with respect to oscillation*, Adv. Math., 3 (1969), pp. 594–623.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 5, pp. 1905–1910

ADDENDUM TO "GENERIC SIMPLICITY OF THE SPECTRUM AND STABILIZATION FOR A PLATE EQUATION"*

JAIME H. ORTEGA† AND ENRIQUE ZUAZUA‡

Abstract. In this addendum we clarify a technical point of the article [J. H. Ortega and E. Zuazua, *SIAM J. Control. Optim.*, 39 (2001), pp. 1585–1614], devoted to studying the multiplicity of the eigenvalues of the plate equation and some applications to the stabilization.

Key words. plate equation, stabilization, spectral theory, shape differentiation

AMS subject classifications. 35P05, 35J40, 93D15

DOI. 10.1137/S0363012902414032

In this addendum we clarify a technical point of the article [5], devoted to studying the multiplicity of the eigenvalues of the plate equation,

$$\text{(P)} \qquad \begin{cases} \triangle^2 y &= \lambda y & \text{in } \Omega, \\ y &= 0 & \text{on } \partial\Omega, \\ \dfrac{\partial y}{\partial n} &= 0 & \text{on } \partial\Omega, \end{cases}$$

where $\Omega \subseteq \mathbb{R}^d$ is a bounded domain with boundary of class C^4.

Problem (P) admits a sequence of positive eigenvalues

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \leq \cdots \longrightarrow \infty,$$

with finite multiplicity. The eigenfunctions $\{y_n\}_n \subset H_0^2(\Omega)$ of (P) can be chosen to form an orthonormal basis of $H_0^2(\Omega)$.

It is well known that, for some domains (some annular domains, for instance, [2]) the spectrum is not simple. In [5] we proved that, generically with respect to the domain, the spectrum is simple. The object of this addendum is to clarify a technical point of the proof.

To do that we recall the precise statement of the main result in [5, Theorem 1.1, p. 1586].

Given a bounded domain Ω of class C^4 of $\mathbb{R}^d$ and a deformation $u \in W^{5,\infty}(\Omega; \mathbb{R}^d)$, we introduce the following deformed domain:

$$\Omega + u = \left\{ z \in \mathbb{R}^d : z = x + u(x), x \in \Omega \right\}.$$

We then consider the plate equation in the deformed domain $\Omega + u$:

$$\text{(P}_\text{u}\text{)} \qquad \begin{cases} \triangle^2 y &= \lambda y & \text{in } \Omega + u, \\ y &= 0 & \text{on } \partial(\Omega + u), \\ \dfrac{\partial y}{\partial n} &= 0 & \text{on } \partial(\Omega + u). \end{cases}$$

*Received by the editors August 22, 2002; accepted for publication (in revised form) May 8, 2003; published electronically December 17, 2003.

http://www.siam.org/journals/sicon/42-5/41403.html

†Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, Departamento Ingeniería Matemática, Casilla 170-3, Correo 3, Santiago, Chile and Universidad del Bío-Bío, Departamento de Matemática, Casilla 5-C, Concepción, Chile (jortega@dim.uchile.cl). The research of this author was supported by grants FONDECYT 1000543 and 7000543.

‡Universidad Autónoma de Madrid, Departamento de Matemáticas, 28049 Madrid, Spain (enrique.zuazua@uam.es). The research of this author was supported by grant PB96-0663 of the DGES (Spain) and by the TMR Project of the EU "Homogenization and Multiple Scales."

The main result in [5] reads as follows.

THEOREM 1. *Let Ω be a bounded domain of $\mathbb{R}^d$ of class C^4. Let Γ_0 be an open nonempty subset of $\partial\Omega$.*

Then the set

$$A=\left\{u\in W^{5,\infty}(\Omega,\mathbb{R}^d) : u=0 \text{ on } \partial\Omega\setminus\Gamma_0 \text{ and the spectrum of } (\mathrm{P_u}) \text{ is simple}\right\}$$

is residual in

$$W_0=\left\{u\in W^{5,\infty}(\Omega,\mathbb{R}^d) : u=0 \text{ on } \partial\Omega\setminus\Gamma_0\right\}.$$

In other words, it is a countable intersection of dense open sets on W_0.

Theorem 1 guarantees the generic simplicity of the spectrum of the plate system with respect to perturbations of the domain Ω. Moreover, the theorem indicates that this property may be achieved by means of deformations that leave invariant most of the boundary of the domain Ω. The additional assumption $u\in W^{5,\infty}\left(\Omega;\mathbb{R}^d\right)$ is necessary for technical reasons. It guarantees the C^4 regularity of the perturbed domain.

In order to prove Theorem 1, in [5] we used classical tools on shape differentiation that we describe briefly. We denote by $\lambda(u)$ and $y(u)$ the eigenvalues and eigenfunctions of $(\mathrm{P_u})$. We perform the change of variables

$$Y(u)=y(u)\circ(I+u).$$

The following problem is equivalent to $(\mathrm{P_u})$: *to find $\lambda(u)\in\mathbb{R}$ and $Y(u)$ such that*

$$\left\{\begin{array}{ll} D_j^2(u)\left(Jac\,(I+u)\,D_i^2(u)\,Y(u)\right) & =\lambda(u)\,Y(u)\,Jac\,(I+u) \qquad \text{in } \Omega,\\ Y(u) & \in H_0^2(\Omega), \end{array}\right. \tag{1}$$

with

$$D_i(u)\,g=\left(\frac{\partial f}{\partial z_i}\right)\circ(I+u)=\sum_j M_{ij}(u)\,\frac{\partial g}{\partial x_j}$$

for $g=f\circ(I+u)$. Here and in what follows $M(u)=(M_{ij}(u))_{i,j=1}^d$ is defined by

$$M(u)=[M_{i,j}(u)]={}^t\left[\frac{\partial}{\partial x_j}(I+u)_i\right]^{-1}=\left[\frac{\partial}{\partial x_i}(I+u)_j\right]^{-1}.$$

In Theorem 3.5 (p. 1599) of [5] we claim that there exist h analytic families of eigenvalues and eigenfunctions of (1) (or, equivalently, $(\mathrm{P_u})$) with respect to the deformation $u\in W^{5,\infty}(\Omega;\mathbb{R}^d)$ of the domain. However, in general, this is not true. But this fact is not really needed in the proof of Theorem 1 above. Indeed, analyticity does hold with respect to scalar perturbations, and this suffices to develop the proof of the generic simplicity result in [5].

Consequently, the statement on the generic simplicity of the spectrum of Theorem 1 above holds. But in the proof given in [5] one has to replace Theorem 3.2 and Proposition 3.3 of [5] by Theorem 2 and Proposition 1 below.

More precisely, Theorem 3.2 in [5] should read as follows.

THEOREM 2. *Let E be a Hilbert space with inner product $\langle\cdot,\cdot\rangle$, and let Λ be a Banach space. Let $P : D(P)\subset E\to E$ be a self-adjoint operator densely defined*

in E. Assume that λ is an eigenvalue of multiplicity h of P, and let $\phi_1, \dots, \phi_h$ be the orthonormal eigenfunctions associated to λ. Moreover, assume that there exists a bounded linear operator $Q : E \to E$ such that $Q\Pi_N = 0$ and $Q(P+\lambda) = I - \Pi_N$, Π_N being the orthogonal projection in $N = \operatorname{Ker}(P+\lambda)$.

Let $R(u)$ be an analytic self-adjoint map in $B(E,F)$ for every u in a neighborhood of $u = 0$ in Λ such that $R(0) = 0$ and $P(u) = P + R(u)$.

Then there exist h continuous functions defined in a neighborhood of $u = 0$ in Λ with values in $\mathbb{R}$, $u \to \lambda_i(u)$, and h continuous functions $u \to \phi_i(u)$, with values in E, $i = 1, \dots, h$, defined in a neighborhood of $u = 0$ in Λ such that

1. $\lambda_j(0) = \lambda, \qquad j = 1, \dots, h.$
2. *For all u small enough, $(\lambda_j(u), \phi_j(u))$ is a solution of the eigenvalue problem $P(u)\phi_j(u) = \lambda_j(u)\phi_j(u)$.*
3. *For all u small enough the set $\{\phi_1(u), \dots, \phi_h(u)\}$ is orthonormal in E.*
4. *For each interval $I \subset \mathbb{R}$ such that $\overline{I}$ contains only the eigenvalue λ of P, there exists a neighborhood U of $u = 0$ such that there are exactly h eigenvalues (counting the multiplicity) $\lambda_1(u), \dots, \lambda_h(u)$ of $P(u)$ contained on I.*
5. *Moreover, for each u in a neighborhood U of $u = 0$ in Λ, the map $t \to (\lambda_j(tu), \phi_j(tu))$ is analytic in a neighborhood of $t = 0$.*

On the other hand, Proposition 3.3 in [5] should read as follows.

PROPOSITION 1. *Under the hypotheses of Theorem 2, if λ is an eigenvalue of multiplicity h of P and $\phi_1, \dots, \phi_h$ are orthonormal eigenfunctions associated to λ, then there exists at least a function $u \longrightarrow (\lambda(u), \phi(u)) \in \mathbb{R} \times E$ which is continuous in a neighborhood of $u = 0$ in Λ such that*

1. $\lambda(0) = \lambda$,
2. *$\phi(u)$ is an eigenfunction of $P(u)$ associated to the eigenvalue $\lambda(u)$.*
3. *Moreover, for each u in a neighborhood U of $u = 0$ in Λ, the map $t \to (\lambda(tu), \phi(tu))$ is analytic in a neighborhood of $t = 0$.*

Remark 1. In Theorem 3.2 and Proposition 3.3 (pp. 1593–1594) of [5], we claimed that the families of eigenvalues and eigenfunctions are analytic in a neighborhood of $u = 0$ in a Banach space Λ. But this statement is not correct. Indeed, in the case of nonscalar perturbations, there are well-known counterexamples (we refer, for instance, to [2]). This lack of analyticity is due to the fact that the perturbation parameter belongs to a multidimensional Banach space. In the case of scalar perturbations the analyticity result does hold ([4, Theorem 3.9, p. 392] and [7, Theorem 1, pp. 57–58]).

Remark 2. The proof of the results above is almost the same as in [5]. But, strictly speaking, that proof yields only continuity of the eigenvalues and eigenfunctions. This is due to the fact that when the problem is reduced to analyze a family of polynomials [5, equation (3.20), p. 1596]

$$p_u(\alpha) = \alpha^h + \sum_{k=0}^{h-1} a_k(u)\alpha^k,$$

with the coefficients depending regularly on u, and such that all its h zeros (counting multiplicity) are real, we can conclude that the branches of zeros of the polynomial depend continuously on u.

In fact, one may prove that the largest zero of p_u define a continuous function in a neighborhood $\mathcal{W}$ of $u = 0$ in Λ. One can then proceed by induction.

Indeed, let $\alpha_h(u)$ be the largest zero of p_u and consider a sequence $\{u_n\}_n \subset \mathcal{W}$, such that $u_n \longrightarrow u_0 \in \mathcal{W}$, and $|\alpha_h(u_n) - \alpha_h(u_0)| \geq \varepsilon > 0$ for all $n \in \mathbb{N}$.

Note that, since the coefficients are continuous in $\mathcal{W}$, they are bounded in $\mathcal{W}$. Therefore there exists $R > 0$ such that for all $u \in \mathcal{W}$

$$|p_u(\alpha)| > 1 \quad \text{if } |\alpha| > R.$$

This implies that the zeros of p_u belong to the interval $(-R, R)$, and therefore the sequence $\{\alpha_h(u_n)\}_n$ is bounded in $\mathbb{R}$. Thus, there exists a subsequence that, to simplify the notation, we will write as $\{\alpha_h(u_n)\}_n$, converging to a real value β. Thus we have that

$$0 = p_{u_n}(\alpha_h(u_n)) = \alpha_h^h(u_n) + \sum_{k=0}^{h-1} a_k(u_n)\alpha_h^k(u_n) \longrightarrow \beta^h + \sum_{k=0}^{h-1} a_k(u_0)\beta^k.$$

That is, β is a zero of p_{u_0}. Moreover, since

$$p_{u_n}(\alpha) > 0 \quad \text{and} \quad p'_{u_n}(\alpha) \geq 0 \quad \forall \alpha > \alpha_m(u_n),$$

we have that

$$p_{u_0}(\alpha) > 0 \quad \text{and} \quad p'_{u_0}(\alpha) \geq 0 \quad \forall \alpha > \beta,$$

which implies that $\beta = \alpha_h(u_0)$. This is in contradiction with our assumption and proves that the largest zero defines a continuous function.

We can then write

$$p_u(\alpha) = \alpha^h + \sum_{k=0}^{h-1} a_k(u)\alpha^k = (\alpha - \alpha_h(u)) \left(\alpha^{h-1} + \sum_{k=0}^{h-2} b_k(u)\alpha^k \right),$$

where the coefficients $b_j(u)$ are defined by

$$b_{h-2}(u) = \alpha_h(u) + a_{h-1}(u) \quad \text{and} \quad b_{k-1}(u) = b_k(u)\alpha_h(u) + a_j(u), \quad k = 1, \ldots, h-2,$$

which are continuous functions, and we may proceed by induction.

Moreover, using classical results (see [4, p. 117] and [7, p. 37]), we can obtain the analyticity of the eigenvalues with respect to perturbations of the form $u = tv$, v being fixed and $t \in \mathbb{R}$ being scalar.

These results may be applied to the map

$$P : W^{5,\infty}(\Omega, \mathbb{R}^d) \longrightarrow \mathcal{L}\left(H_0^2(\Omega); H^{-2}(\Omega)\right)$$

such that

$$P(u)\phi = \frac{1}{Jac\,(I+u)} D_j^2(u) \left(Jac\,(I+u)\, D_i^2(u)\, \phi \right). \tag{2}$$

Theorem 3.5 in [5] should reads as follows.

THEOREM 3. *Let $\Omega \subset \mathbb{R}^d$ be an open bounded domain of class C^4. Let λ be an eigenvalue of multiplicity h of the plate system* ($\mathrm{P_u}$) *for $u = 0$ with associated eigenfunctions $y_1, \ldots, y_h$.*

Then there exist h continuous functions with values in $\mathbb{R}$, $u \to \lambda_i(u)$, and h continuous functions $u \to y_i(u)$, with values in $H^4(\Omega+u) \cap H_0^2(\Omega+u)$, $i = 1, \ldots, h$, defined in a neighborhood of $u = 0$ in $W^{5,\infty}(\Omega, \mathbb{R}^d)$ such that

1. $\lambda_j(0) = \lambda, \qquad j = 1, \ldots, h.$

2. *For all u small enough, $(\lambda_j(u), y_j(u))$ is a solution of the plate system defined in the new domain $\Omega + u$.*
3. *For all u small enough the set $\{y(u), \dots, y(u)\}$ is orthonormal in $L^2(\Omega+u)$.*
4. *For each interval $I \subset \mathbb{R}$ such that $\overline{I}$ contains only the eigenvalue λ of* (P), *there exists a neighborhood U of $u = 0$ such that there are exactly h eigenvalues (counting the multiplicity) $\lambda_1(u), \dots, \lambda_h(u)$ of* ($\mathrm{P_u}$) *contained on I.*
5. *For each u in a neighborhood U of $u = 0$ in $W^{5,\infty}(\Omega, \mathbb{R}^d)$, the map $t \to (\lambda_j(tu), \phi_j(tu))$ is analytic in a neighborhood of $t = 0$.*

In the proof of Theorem 1 in [5], we also use the definitions of local derivative and the total derivative:

Let

$$\begin{array}{rccl} v: & W^{k,\infty}\left(\Omega, \mathbb{R}^d\right) & \to & W^{m,r}(\Omega + u), \\ & u & \to & v(u) \end{array}$$

be a function of the perturbation parameter u. Then if $k \geq m$, the *first local variation* can be defined as

$$v'(\Omega; u) = \lim_{t \to 0^+} \frac{v(tu)|_\omega - v(0)|_\omega}{t} \quad \text{in } \omega, \tag{3}$$

where $\omega \subset\subset \Omega$ and $v(tu)|_\omega$, $v(0)|_\omega$ are the restrictions of the functions $v(tu), v(0)$ to ω.

We define the *first order total variation* as

$$\dot{v}(\Omega; u) = \lim_{t \to 0^+} \frac{v(tu) \circ (I + tu) - v(0)}{t} \quad \text{in } \Omega. \tag{4}$$

It is important to observe that the total variation $\dot{v}(\Omega; u)$ is a function defined in the whole Ω while the local variation is a function defined "locally" in subsets $\omega \subset\subset \Omega$.

Moreover, we have that

$$v'(u) = \dot{v}(u) - u \cdot \nabla v(0).$$

We refer, for instance, to [6] or [8] for more details on this issue.

For each $u \in W^{5,\infty}(\Omega, \mathbb{R}^d)$, in a neighborhood of $u = 0$, we define the maps

$$t \to \lambda_j(tu), \qquad t \to \phi_j(tu),$$

where $\lambda_j(tu)$ and $\phi_j(tu)$ are eigenvalues and eigenfunctions of the Stokes system in the domain $\Omega + tu = \{x + tu(x) : x \in \Omega\}$.

According to Theorem 2 and Proposition 1, these functions are analytic in an interval $[0, T_u]$ (although analyticity cannot be guaranteed with respect to $u \in W^{5,\infty}(\Omega, \mathbb{R}^d)$).

Using the continuity of the eigenvalues with respect to $u \in W^{5,\infty}(\Omega, \mathbb{R}^d)$ and the analyticity with respect to the scalar perturbation, the proof of Theorem 1.1 in [5] applies without further changes.

REFERENCES

[1] D. Chenais and B. Rousselet, *Continuité et differentiabilité d'éléments propes: Application à l'optimisation de structures*, Appl. Math. Optim., 22 (1990), pp. 27–59.
[2] C. V. Coffman, R. J. Duffin, and D. H. Shaffer, *The fundamental mode of vibration of a clamped annular plate is not of one sign*, in Constructive Approaches to Mathematical Models, Academic Press, New York, 1979, pp. 267–277.
[3] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. I, Wiley-Insterscience, John Wiley and Sons, New York, 1989.
[4] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, Heidelberg, 1980.
[5] J. H. Ortega and E. Zuazua, *Generic simplicity of the spectrum and stabilization for a plate equation*, SIAM J. Control. Optim., 39 (2001), pp. 1585–1614.
[6] O. Pironneau, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, New York, 1983.
[7] F. Rellich, *Perturbation Theory of Eigenvalue Problems*, Gordon and Breach Science Publishers, New York, 1969.
[8] J. Simon, *Differentiation with respect to the domain in boundary value problems*, Numer. Funct. Anal. Optim., 2 (1980), pp. 649–687.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 1911–1933

NUMERICAL APPROXIMATIONS FOR STOCHASTIC DIFFERENTIAL GAMES: THE ERGODIC CASE*

HAROLD J. KUSHNER†

Abstract. The Markov chain approximation method is a widely used, relatively easy to use, and efficient family of methods for the bulk of stochastic control problems in continuous time for reflected-jump-diffusion-type models. It has been shown to converge under broad conditions, and there are good algorithms for solving the numerical problems if the dimension is not too high. We consider a class of stochastic differential games with a reflected diffusion system model and ergodic cost criterion and where the controls for the two players are separated in the dynamics and cost function. It is shown that the value of the game exists and that the numerical method converges to this value as the discretization parameter goes to zero. The actual numerical method solves a stochastic game for a finite state Markov chain and ergodic cost criterion. The essential conditions are nondegeneracy and that a weak local consistency condition hold "almost everywhere" for the numerical approximations, just as for the control problem.

Key words. stochastic differential games, numerical methods, Markov chain approximations

AMS subject classifications. 60F17, 65C30, 65C40, 91A15, 91A23, 93E25

DOI. 10.1137/S00363012901400342

1. Introduction.

The Markov chain approximation method of [18, 19, 22] is a widely used method for the numerical solution of virtually all of the standard forms of stochastic control problems with reflected-jump-diffusion models. It is robust and can be shown to converge under very broad conditions. Extensions to approximations for two-person differential games with discounted, finite time, stopping time, and pursuit-evasion games were given in [21] for reflected diffusion models where the controls for the two players are separated in the dynamics and cost rate functions. In this paper, the basic ideas will be extended to two-player stochastic dynamic games with the same systems model but where the cost function is ergodic. Such ergodic and "separated" models occur, for example, in risk-sensitive and robust control [2, 3, 7, 15]. In fact, the game formulation of risk sensitive control problems for queues in heavy traffic was our original motivation. See [15, section 7] for a formulation of risk sensitive control in terms of a stochastic differential game for a diffusion that fits our form once "numerical boundaries" are added, provided that the covariance does not depend on the control. While Markov chain approximation algorithms have been adapted to such problems, there were no proofs of convergence.

When the robust control is for controlled queues in heavy traffic, then the state is confined to some convex polyhedron by boundary reflection [20]. In many other applications, the state of the physical problem is confined to a bounded set. One example is the heavy traffic limit of controlled queueing networks with finite buffers [1, 20] or robust control of such systems as in [2, 3], where the set is a hyperectangle. Then robust control would lead to a game problem with a hyperrectangular state space. If the system state is not a priori confined to a bounded set, then for numerical purposes

*Received by the editors December 26, 2001; accepted for publication (in revised form) May 24, 2003; published electronically January 22, 2004. This work was partially supported by contract DAAD19-99-1-0223 from the Army Research Office and National Science Foundation grant ECS 0097447.

http://www.siam.org/journals/sicon/42-6/40034.html

†Applied Mathematics Department, Lefschetz Center for Dynamical Systems, Brown University, Providence, RI 02912 (hjk@dam.brown.edu).

it is commonly necessary to bound the state space artificially by adding a reflecting boundary and then experimenting with the bounds. Our systems model is confined to a state space G that is a convex polyhedron, and it is confined by a "reflection" on the boundary. More generally, the boundaries could be determined by a set of smooth curved surfaces as in [22], but we restrict attention to the polyhedral case, since that is the most common and it avoids minor details which can be distracting.

There are many results for various forms of the game problem, e.g., [4, 5, 6, 24, 28, 29]. However, there seems to be nothing available concerned with the ergodic problem for the reflected diffusion model. We will use purely probabilistic methods of proof. Such methods have the advantage of providing intuition concerning numerical approximations, they cover many of the problem formulations to date, and they converge under quite general conditions. The essential conditions are *weak-sense* existence and uniqueness of the solution to the controlled equations, "almost everywhere" continuity of the dynamical and cost rate terms, and a natural "local consistency" condition: The local consistency and continuity need hold only almost everywhere with respect to the measure of the basic model; hence discontinuities in the dynamics and cost function can be treated under appropriate conditions (see, in particular, the treatment of discontinuities and complex variational problems with singularities and Theorems 4.6 and 7.1 in [22]). Furthermore, the numerical approximations are represented as processes which are close to the original, which gives additional intuitive and practical meaning to the method.

Subsection 2.1 defines the basic systems model. First, the uncontrolled model is introduced. Then the control is added via the Girsanov transformation [17]. The dynamical model is the reflected stochastic differential equation (2.4), also called the *Skorohod problem* [12, 20, 22]. The conditions on the boundary of the state space are (A2.1)–(A2.2). Condition (A2.1) covers the great majority of cases of current interest, including those that arise from queueing and communications networks. The condition is obvious when the state space is a hyperrectangle with reflection directions being the interior normals. The strategies of the players are as follows. Player 1 wishes to minimize and player 2 wishes to maximize. For the infsup problem (the upper value), at the start of the game (i.e., at $t = 0$) player 1 selects a control. This can be either a pure (and time independent) feedback control or a relaxed feedback control (see subsection 2.1 for the definition). The selected feedback control will be used at all $t \geq 0$. Then player 2 selects its strategy. This can either be a relaxed feedback control or it can have the form of a classical relaxed control. Whatever it is, once selected, it cannot be changed.

The situation is analogous if player 2 selects first. Since the controls for the player who chooses first are time independent feedback and these are selected and fixed at the start of the game, and only the player choosing last can use time dependent controls, complications due to the notions of strategy in the time dependent case (e.g., concerning the definition of the value either via a limit of a discrete time game, or via the Elliott–Kalton definition) do not arise. In this sense the paper is simpler than [21]. On the other hand, the treatment of the ergodic cost criterion adds substantial new complications. Subsection 2.3 establishes the existence of the controls yielding the upper and lower values, using approximation methods from [20].

Suppose that, if feedback controls are used, then the asymptotic mean cost per unit time (for either the control or the game problem) does not depend on the initial state value. Then it is reasonable (and is the usual practice) to restrict the controls to being feedback. It is shown that the game has a value with this restriction. For

technical reasons, we allow player 2 to use arbitrary controls (not necessarily feedback). But, if player 1 uses a feedback control, then it is shown that player 2 cannot improve its cost by using a nonfeedback control, so the value is still determined within the space of feedback controls. We do not know whether it is possible for player 1 to achieve a better value by using some nonfeedback control strategy. But any approach that leads to a well posed Isaacs equation will yield feedback controls.

The methods to be used for the ergodic cost function are quite different than those used in [21]. They share the foundation in the theory of weak convergence [9, 13]; however, they depend heavily on the approximations to the ergodic cost control problem as developed in [20, Chapter 4]. The development depends on "continuity" properties of the invariant measures and ergodic costs in the control in an appropriate sense, and on the fact that we lose little in the costs if the controls are "smoothed." Similar results for the pure control problem are in [20], and the development of the paper has been structured to take advantage of the results in [20, 22], wherever possible. To facilitate the development, subsection 2.2 summarizes the results from [20] which will be needed here, with an occasional change of notation to suit that used here.

The Markov chain approximation numerical method is discussed in subsection 3.1. The methods for getting the approximating chain and cost function are the same as in [22] for the pure control problem, since it is the process for arbitrary controls that is approximated. The natural *local consistency condition* is stated. The proof of convergence of the numerical method is in subsection 3.2 and depends on the fact that the original game has a value. The numerical approximations are games for Markov chains. They might or might not have a value, depending on the form of the approximation. However, it is seen that the upper and lower values converge to the value of the original game as the approximation parameter goes to its limit. Finally, the proof that the original game has a value is given in section 4.

2. The dynamical model and background results.

2.1. Assumptions and the dynamical model. Assumptions. The first assumptions define the state space G.

A2.1. *The state space G is the intersection of a finite number of closed half spaces in Euclidean r-space $\mathbb{R}^r$ and is the closure of its interior (i.e., it is a closed convex polyhedron with an interior and planar sides). Let ∂G_i, $i = 1, \ldots,$ denote the faces of G, and n_i the interior normal to ∂G_i. Interior to ∂G_i, the reflection direction is denoted by the unit vector d_i, and $\langle d_i, n_i \rangle > 0$ for each i. The possible reflection directions at points on the intersections of the ∂G_i are in the convex hull of the directions on the adjoining faces. Let $d(x)$ denote the set of reflection directions at the point $x \in \partial G$, whether it is a singleton or not. No more than r constraints are active at any boundary point.*

A2.2. *For each $x \in \partial G$, define the index set $I(x) = \{i : x \in \partial G_i\}$. Suppose that $x \in \partial G$ lies in the intersection of more than one boundary; that is, $I(x)$ has the form $I(x) = \{i_1, \ldots, i_k\}$ for some $k > 1$. Let $N(x)$ denote the convex hull of the interior normals $n_{i_1}, \ldots, n_{i_k}$ to $\partial G_{i_1}, \ldots, \partial G_{i_k}$, resp., at x. Then there is some vector $v \in N(x)$ such that $\gamma' v > 0$ for all $\gamma \in d(x)$.*

There is a neighborhood $N(\partial G)$ and an extension of $d(\cdot)$ to $\overline{N(\partial G)}$ that is upper semicontinuous in the following sense: For each $\epsilon > 0$, there is $\rho > 0$ that goes to zero as $\epsilon \to 0$ and such that if $x \in N(\partial G) - \partial G$ and distance$(x, \partial G) \leq \rho$, then $d(x)$ is in the convex hull of the directions $\{d(v); v \in \partial G,$ distance$(x, v) \leq \epsilon\}$.

A2.3. *The $U_i, i = 1, 2$, are compact sets in some Euclidean space. The $(r \times r)$ matrix-valued function $\sigma(\cdot)$ on G is Hölder continuous, with $\sigma^{-1}(x)$ bounded, and the*

$\mathbb{R}^r$-valued functions $b_i(\cdot)$ on $G \times U_i$ are continuous.

Let $\alpha = (\alpha_1, \alpha_2), \alpha_1 \in U_1, \alpha_2 \in U_2$, denote the canonical control value, with α_i the canonical value for player i. The uncontrolled model is the solution to the Skorohod problem

$$dx(t) = \sigma(x(t))dw(t) + dz(t),\ x(t) \in G. \tag{2.1}$$

The controlled system will be defined via the Girsanov transformation, starting with (2.1). This transformation will add the control or "drift" term $b_1(x(t), \alpha_1)dt + b_2(x(t), \alpha_2)dt$.

The Skorokhod problem is a standard model for a reflected diffusion process. For a detailed discussion of the Skorohod problem and the assumptions (A2.1) and (A2.2), see [20, Chapter 3]. See also the brief comment below (A2.4). By a solution to (2.1) we mean the following. Let Ω denote the path space of $(x(\cdot), z(\cdot), w(\cdot))$, let $\{\mathcal{F}_t, t < \infty)\}$ denote the filtration on the space, and write $|z|(t)$ for the total variation of $z(\cdot)$ on $[0, t]$. Let $x(0) \in G, z(0) = 0$. The $x(\cdot)$ and $z(\cdot)$ are $\mathbb{R}^r$-valued, continuous and $\mathcal{F}_t$-adapted, and $w(\cdot)$ is an $\mathcal{F}_t$-standard $\mathbb{R}^r$-valued Wiener process. The $z(\cdot)$ is the reflection process and satisfies the following conditions: $|z|(t) < \infty$ with probability one (w.p.1) for all t, and there is a measurable function $\gamma(\cdot)$ with $\gamma(t) \in d(x(t))$ w.p.1 such that $z(t) = \int_0^t \gamma(s)d|z|(s)$.

Let Ω_T denote the restriction of Ω to functions defined on $[0, T]$. Define $\mathcal{F} = \lim_t \mathcal{F}_t$ and let P_x denote the measure when the initial condition is $x(0) = x$, with E_x the associated expectation. Let $P_{x,T}(\cdot)$ denote the probability measure, when we confine our interest to paths on the finite interval $[0, T]$.

We will also need the following condition.

A2.4. *There is a unique weak sense solution to* (2.1) *for each initial condition.*

Comments on (A2.1) and (A2.2). One can always construct the extension in (A2.2). Under (A2.1)–(A2.3), the choice of the reflection direction on the corners and edges of G has no effect on the process (w.p.1), since no matter what the choice of direction or of the controls, it spends zero "local time" at such points [20, Theorem 3.6, Chapter 4]. To see that (A2.1) is natural in application note the following. If the state space is being bounded for purely numerical reasons, then the reflections are introduced only to give a compact set G, which should be large enough so that the effects on the solution in the region of main interest are small. A common choice is a hyperrectangle with normal reflection directions, in which case the right side of (2.1) is zero. Next, consider a queueing network model in the heavy traffic limit [16, 20, 27] where the state space is the nonnegative orthant, and the probability that an output of the ith processor goes to the jth processor is q_{ij}. If the spectral radius of the routing matrix $Q = \{q_{ij}; i, j\}$ is less than unity, then all customers will eventually leave the system. The model is a special case of (2.4) with $z(t) = [I - Q']y(t)$, where $y_i(\cdot)$ is nondecreasing, continuous, and can increase only at t, where $x_i(t) = 0$. The condition (A2.1) implies (see [12, 20]), the so-called "completely-S" condition [16, 20, 26], which is used to ensure that $z(\cdot)$ has bounded variation w.p.1.

Classes of controls. A: Relaxed controls $r_i(\cdot)$. Suppose that for some filtration $\{\mathcal{F}_t, t < \infty\}$ and standard vector-valued $\mathcal{F}_t$-Wiener process $w(\cdot)$, each $r_i(\cdot), i = 1, 2$, is a measure on the Borel sets of $U_i \times [0, \infty)$ such that $r_i(U_i \times [0, t]) = t$ and $r_i(A \times [0, t])$ is $\mathcal{F}_t$-measurable for each Borel set $A \subset U_i$. Then $r_i(\cdot)$ is said to be an *admissible relaxed control* for player i, with respect to $w(\cdot)$. If the Wiener process and filtration have been given or are obvious or unimportant, then we simply say that $r_i(\cdot)$ is an admissible relaxed control for player i [14, 20, 22]. For Borel sets $A \subset U_i$,

we will write $r_i(A \times [0,t]) = r_i(A,t)$.

For almost all (ω,t) and each Borel $A \subset U_i$, one can define the derivative[1]

$$r_{i,t}(A,t) = \lim_{\delta \to 0} \frac{r_i(A,t) - r_i(A,t-\delta)}{\delta}.$$

Without loss of generality, we can suppose that the limit exists for each (ω,t). Then for all (ω,t), $r_{i,t}(\cdot,t)$ is a probability measure on the Borel sets of U_i and for any bounded Borel set B in $U_i \times [0,\infty)$,

$$r_i(B) = \int_0^\infty \int_{U_i} I_{\{(\alpha_i,t)\in B\}} r_{i,t}(d\alpha_i,t)dt.$$

An ordinary control $u_i(\cdot)$ can be represented in terms of the relaxed control $r_i(\cdot)$, defined by its derivative $r_{i,t}(A,t) = I_A(u_i(t))$, where $I_A(u_i)$ is unity if $u_i \in A$ and is zero otherwise. The weak topology [22] will be used on the space of admissible relaxed controls. Relaxed controls are commonly used in control theory to prove existence theorems, since any sequence of relaxed controls has a convergent subsequence.

Classes of controls. B: Relaxed feedback control $m_i(\cdot)$ [10, 20]. Suppose that $m_i(x,\cdot), i = 1,2$, is a probability measure on the Borel sets of U_i for each $x \in G$ and that $m_i(\cdot,A)$ is Borel measurable for each Borel set $A \subset U_i$. Then we say that $m_i(\cdot)$ is a relaxed feedback control. Define $U = U_1 \times U_2$. For relaxed feedback controls $m_i(\cdot)$, define the product measure $m(\cdot)$ by $m(x,d\alpha) = m_1(x,d\alpha_1)m_2(x,d\alpha_2)$. Then $m(\cdot)$ is also a relaxed feedback control but with control value space U. Unless explicitly noted otherwise, the symbol $m(\cdot)$ will always denote such a product form for some relaxed feedback controls $m_i(\cdot), i = 1,2$. If $x(\cdot)$ is a solution to (2.4), and $m(\cdot)$ a relaxed feedback control, then $m(\cdot)$ can be represented by a relaxed control $r(\cdot)$ with derivative having the product form $r_t(d\alpha,t) = r_{1,t}(d\alpha_1,t)r_{2,t}(d\alpha_2,t) = m(x(t),d\alpha)$.

The control for the player that chooses its control first will always be a relaxed feedback control, but the control for the player who chooses its control last might be either a relaxed feedback control or a relaxed control which is not representable in relaxed feedback form.

Defining the controlled dynamical system via the Girsanov transformation: Relaxed feedback controls. The controlled model will be defined via the Girsanov transformation [17]. Some of the well-known details will be described, since the equations will be needed for the approximations. This will be done first for the relaxed feedback controls. Let $m_i(\cdot), i = 1,2$, be relaxed feedback controls, and define the product $m(x,d\alpha) = m_1(x,d\alpha_1)m_2(x,d\alpha_2)$. Define

$$b_{i,m_i}(x) = \int_{U_i} b_i(x,\alpha_i)m_i(x,d\alpha_i), \quad b(x,\alpha) = b_1(x,\alpha_1) + b_2(x,\alpha_2),$$

and set $b_m(x) = \int_U b(x,\alpha)m(x,d\alpha) = b_{1,m_1}(x) + b_{2,m_2}(x)$. The $b(x,\alpha)$ will be the canonical drift term for the SDE and $b_m(x)$ the drift term when the control $m(\cdot)$ is used. For $T > 0$ and relaxed feedback control $m(\cdot)$, define

$$\zeta(T,m) = \int_0^T \left[\sigma^{-1}(x(s))b_m(x(s))\right]' dw(s) - \frac{1}{2}\int_0^T \left|\sigma^{-1}(x(s))b_m(x(s))\right|^2 ds,$$

and set

$$R(T,m) = e^{\zeta(T,m)}.$$

[1] In [20, 22], the derivative at t is written simply as $r_{i,t}(A)$.

For each $(x, T, m(\cdot))$, define the measure $P^m_{x,T}$ on $(\Omega_T, \mathcal{F}_T)$ via the Radon–Nikodym derivative $R(T, m)$:

$$dP^m_{x,T} = R(T, m)dP_{x,T}. \tag{2.2}$$

For each $(x, m(\cdot))$, the family $P^m_{x,T}$ of measures, indexed by T, is consistent and can be extended uniquely to a measure P^m_x on $(\Omega, \mathcal{F})$ that is consistent with the $P^m_{x,T}$. When there is no control (i.e., where the system is (2.1)), we omit the superscript m. The process $w_m(\cdot)$ defined by

$$dw_m(t) = dw(t) - \left[\sigma^{-1}(x(s))b_m(x(s))\right] dt \tag{2.3}$$

is an $\mathcal{F}_t$-standard Wiener process on $(\Omega, P^m_x, \mathcal{F})$ [17]. Now, rewrite the uncontrolled model (2.1) as

$$dx(t) = b_m(x(t))dt + \sigma(x(t))dw_m(t) + dz(t). \tag{2.4}$$

Under the measures $\{P^m_x, x \in G\}$, (2.4) is a Markov process, and we use $P^m(x, t, \cdot)$ for its transition function. Use $P(x, t, \cdot)$ for the transition function of the uncontrolled process (2.1). Strictly speaking, the process $w_m(\cdot)$ should be indexed also by the initial condition $x = x(0)$, but we omit it for notational simplicity.

The controlled dynamical system with relaxed controls. Let $r_i(\cdot)$ be a relaxed control for player i, with derivative $r_{i,t}(\cdot, t)$, and define $b_{i,r_i}(x, t) = \int_{U_i} b_i(x, \alpha)r_{i,t}(d\alpha_i, t)$. We will also have occasion to use relaxed (and not necessarily relaxed feedback) controls for one of the players. For specificity at this point, suppose that a relaxed control is used for player 1 and a relaxed feedback control is used for player 2. Write $b_{r_1,m_2}(x, t) = b_{1,r_1}(x, t) + b_{2,m_2}(x)$. Define $\xi(T, r_1, m_2)$ as $\xi(T, m)$ was defined but with $b_{r_1,m_2}(x(t), t)$ replacing $b_m(x(t))$. Using $\xi(T, r_1, m_2)$ in lieu of $\xi(T, m)$, define $P^{r_1,m_2}_{x,T}$, P^{r_1,m_2}_x, and $w_{r_1,m_2}(\cdot)$, analogously, and write the controlled equation with drift $b_{r_1,m_2}(x, t)$ as

$$dx(t) = b_{1,r_1}(x(t), t)dt + b_{2,m_2}(x)dt + \sigma(x(t))dw_{r_1,m_2}(t) + dz(t). \tag{2.5}$$

The measures P^{r_1,m_2}_x are used with (2.5). The development is analogous if player 1 uses the relaxed feedback control and if player 2 uses the relaxed control.

Representation of the reflection process $z(\cdot)$. Recall that d_i is the reflection direction on the interior of the ith face of G. For either the model (2.4) or (2.5), the process $z(\cdot)$ can be represented in terms of processes which increase only on the individual faces of G. In particular, we can write

$$z(t) = \sum_i y_i(t)d_i, \tag{2.6}$$

where the process $y_i(\cdot)$ is nondecreasing, right continuous, increases only at t, where $x(t)$ is on the ith face of G, and satisfies $y_i(0) = 0$. Under (A2.1), (A2.2), and (A2.4), the representation (2.6) is unique, and the contribution when $x(t)$ is on an edge or corner of G is zero, with probability one [20, Theorem 3.6, Chapter 4]. Let M_ϵ denote an ϵ-neighborhood of the boundary set where more than one constraint is active. Then the same theorem implies that, for $t > 0$, $\sup_{x,m} E^m_x |y(t)| I_{\{x(t) \in M_\epsilon\}} \to 0$ as $\epsilon \to 0$.

The representation (2.6) is useful since the individual $y_i(\cdot)$ often have the interpretation as "overflows" or "underflows," as, for example, in heavy traffic modeling of queueing or communications networks. There is often a cost associated with such overflows or underflows, with each component having its own weight.

2.2. Background results and the cost function. The development depends heavily on approximation, continuity, and limit results from [20, Chapter 4] for the control problem. The results carry over to the game problem, since they are concerned with arbitrary relaxed feedback and relaxed controls. To facilitate our development, several key results from [20] will be stated in the notation of this paper.

Illustration of the use of the Girsanov transformation: Mutual absolute continuity of the transition functions. The following theorem is [20, Theorem 3.1, Chapter 4]. We will outline the proof by copying some of the details from the reference, since similar "Girsanov transformation" methods underlie many of the results, there are some slight differences worth noting, and it gives a feeling for the approach. Unless otherwise noted, "almost all" refers to Lebesgue measure. The symbol $\Rightarrow$ denotes weak convergence.

THEOREM 2.1. *Assume* (A2.1)–(A2.4). *Let* $m^n(y,\cdot) \Rightarrow m(y,\cdot)$ *for almost all* $y \in G$ *as* $n \to \infty$, *where* $m(\cdot)$ *and* $m^n(\cdot)$ *are relaxed feedback controls. Then for any* $0 < t_0 < t_1 < \infty$ *and bounded and measurable real-valued function* $f(\cdot)$, *as* $n \to \infty$,

$$\int f(y)P^{m^n}(x,t,dy) \to \int f(y)P^m(x,t,dy) \tag{2.7}$$

uniformly for $(x,t) \in G \times [t_0,t_1]$. *For any* $t > 0$, $P^m(x,t,\cdot)$ *is absolutely continuous with respect to Lebesgue measure uniformly in* $m(\cdot)$ *and in* $(x,t) \in G \times [t_0,t_1]$. *For each relaxed feedback control* $m(\cdot)$, *the process defined by* (2.4) *is a strong Feller process, and it has a unique weak-sense solution for each initial condition* x.

Proof. We concentrate on the uniformity in x of the convergence (2.7). First note that, by the weak convergence and the product form of $m^n(\cdot)$, the limit $m(\cdot)$ can always be represented in the product form $m(x,d\alpha) = m_1(x,d\alpha_1)m_2(x,d\alpha_2)$ for some relaxed feedback controls $m_i(\cdot), i = 1,2$, for almost all x. The expression (2.7) can be written equivalently as

$$E_x f(x(t))R(t,m^n) - E_x f(x(t))R(t,m) \to 0 \ \text{ as } n \to \infty. \tag{2.8}$$

For notational simplicity, let $\sigma(x) = I$, the identity. We will use the following inequalities:

$$\left|e^a - e^b\right| \le |a-b| \left|e^a + e^b\right|, \tag{2.9a}$$

$$\begin{aligned} E_x \left| \int_0^t b'_m(x(s))dw(s) - \int_0^t b'_{m^n}(x(s))dw(s) \right|^2, \\ \le E_x \int_0^t \left|b_m(x(s)) - b_{m^n}(x(s))\right|^2 ds. \end{aligned} \tag{2.9b}$$

By the continuity and boundedness of $b(\cdot)$ and the weak convergence of the $m^n(y,\cdot)$ for almost all $y \in G$, we have, as $n \to \infty$,

$$b_{m^n}(y) = \int_U b(y,\alpha)m^n(y,d\alpha) \to b_m(y) = \int_U b(y,\alpha)m(y,d\alpha)$$

for almost all y. Define

$$\tilde{b}_n(y) = \left|b_m(y) - b_{m^n}(y)\right|^2.$$

Let $t \in [t_0, t_1]$, where $0 < t_0 < t_1 < \infty$. By Egoroff's theorem [11, Theorem 12, p. 149], for each $\epsilon > 0$, there is a measurable set A_ϵ with $l(A_\epsilon) \leq \epsilon$ such that $\tilde{b}_n(y) \to 0$ uniformly in $y \notin A_\epsilon$. Furthermore, $P(x,t,\cdot)$ is absolutely continuous with respect to Lebesgue measure for each x and $t > 0$ (and uniformly in $(x,t) \in G \times [t_0, t_1]$ for any $0 < t_0 < t_1 < \infty$). These facts imply that, as $n \to \infty$,

$$\int_0^t E_x \tilde{b}_n(x(s))ds \to 0$$

uniformly in $x \in G$. The last expression, together with the inequalities (2.9), implies (2.8) uniformly in $x \in G$. □

Additional background results. We will also need the results of Theorems 2.2–2.8, most of which are either taken from [20] or are minor adaptations of such results. Where an elaboration on a proof in [20] would be useful, additional comments will be made. Although the reference does not deal with games, the fact that the product $m(x, d\alpha) = m_1(x, d\alpha_1)m_2(x, d\alpha_2)$ is a relaxed feedback control allows the results to be carried over.

THEOREM 2.2 (from [20, Theorems 3.1–3.3, Chapter 4]). *Assume* (A2.1)–(A2.4). *The process $x(\cdot)$ defined by* (2.4) *has a unique invariant measure $\mu_m(\cdot)$ for each relaxed feedback control of the product form $m(x, d\alpha) = m_1(x, d\alpha_1)m_2(x, d\alpha_2)$. Furthermore, the transition function $P^m(x,t,\cdot)$ and Lebesgue measure are mutually absolutely continuous uniformly in all such $m(\cdot), x \in G$, and $t \in [t_0, t_1]$ for any $0 < t_0 < t_1 < \infty$.*

A smoothed control. Extend the definition of the relaxed feedback control $m_i(y, \cdot)$ so that it is defined as a relaxed feedback control for all $y \in \mathbb{R}^r$. For example, let it be concentrated on some fixed number in U for $y \notin G$. For small $\epsilon > 0$ and $x \in G$, define the smoothed control

$$m_{i,\epsilon}(x, \cdot) = \frac{1}{(2\pi\epsilon)^{r/2}} \int_{\mathbb{R}^r} e^{-|y-x|^2/2\epsilon} m_i(y, \cdot)dy, \quad x \in G.$$

Define $m_\epsilon(x, d\alpha) = m_{1,\epsilon}(x, d\alpha_1)m_{2,\epsilon}(x, d\alpha_2)$.

THEOREM 2.3 (this is [20, Theorem 3.4, Chapter 4]). *Assume* (A2.1)–(A2.4). *$m_\epsilon(\cdot)$ is a relaxed feedback control and $m_\epsilon(x, \cdot) \Rightarrow m(x, \cdot) = m_1(x, \cdot)m_2(x, \cdot)$ for almost all $x \in G$ as $\epsilon \to \infty$. The function $b_{m_\epsilon}(\cdot)$ is continuous for each ϵ, and, as $\epsilon \to 0$, $b_{m_\epsilon}(x) \to b_m(x)$ almost everywhere in G.*

THEOREM 2.4 (from [20, Theorem 4.2, Chapter 4]). *Assume* (A2.1)–(A2.4). *Then $\mu_m(\cdot)$ is continuous in the control in that if $m^n(x, \cdot) \Rightarrow m(x, \cdot)$ for almost all $x \in G$, as $n \to \infty$, then the associated invariant measures converge strongly in that for each Borel set $A \subset G$, as $n \to 0$,*

$$\mu_{m^n}(A) \to \mu_m(A).$$

The cost function. We will need the following assumption.

A2.5. *The real-valued functions $k_i(\cdot)$ on $G \times U_i, i = 1, 2$, are continuous. Also, c is a vector with nonnegative components.*

Define $k(x, \alpha) = k_1(x, \alpha_1) + k_2(x, \alpha_2)$. For a relaxed feedback control $m(x, d\alpha) = m_1(x, d\alpha_1)m_2(x, d\alpha_2)$, define $k_m(x) = \int_U k(x, \alpha)m(x, d\alpha)$ and

$$\gamma_T(x, m) = \frac{1}{T} E_x^m \int_0^T k_m(x(s))ds + \frac{1}{T} E_x^m c'y(T).$$

The $y_i(t)$, defined in (2.6), represents the total overflow or underflow from face i on $[0, t]$. Thus the cost has two parts: the average running cost per unit time and the average weighted overflow or underflow per unit time.

For relaxed feedback controls, the cost function of interest in this paper is

$$\gamma(m) = \lim_T \gamma_T(x, m). \tag{2.10}$$

Under our assumptions, $\gamma(m)$ will not depend on the initial condition (see Theorem 2.5). We will sometimes abuse terminology and write $\gamma(m) = \gamma(m_1, m_2)$.

It is convenient to use the same symbols γ_T and γ for the total mean cost on $[0, T]$ divided by T and its limit, resp., in other cases. We will abuse terminology and continue to use these symbols, with the arguments indicating the variables on which the quantities depend. In particular, if the initial condition x does not appear as an argument of a function, then the function does not depend on it. If player i uses a relaxed control $r_i(\cdot)$, then define

$$k_{r_i}(x, t) = \int_{U_i} k_i(x, \alpha_i) r_{i,t}(d\alpha_i, t).$$

If player 1 selects its control first and uses a relaxed feedback control and player 2 selects its control last and uses a relaxed control, then define (the use of lim inf is just a convention)

$$\gamma_T(x, m_1, r_2) = \frac{1}{T} E_x^{m_1, r_2} \int_0^T \left[k_{1,m_1}(x(s)) + k_{2,r_2}(x(s), s)\right] ds + \frac{1}{T} E_x^{m_1, r_2} c' y(T),$$

$$\gamma(x, m_1, r_2) = \liminf_T \gamma_T(x, m_1, r_2).$$

If player 2 selects its control first and uses a relaxed feedback control and player 1 uses a relaxed control, define (the use of lim sup is just a convention)

$$\gamma(x, r_1, m_2) = \limsup_T \gamma_T(x, r_1, m_2).$$

Representation of the cost in terms of a stationary system. Let $m(\cdot)$ be a relaxed feedback control. The system (2.4) starts with an arbitrary initial condition that does not necessarily have the stationary distribution. It turns out that the limit (2.10) is the same as if the initial condition were distributed as $\mu_m(\cdot)$. This is the assertion of the next theorem.

THEOREM 2.5 (this is [20, Theorem 4.1, Chapter 4]). *Assume* (A2.1)–(A2.5). *Let $m(\cdot)$ be a relaxed feedback control. Then the $E_x^m y_i(1)$ are continuous functions of x and*

$$\gamma(m) = \int k_m(x) \mu_m(dx) + \int E_x^m \left[c' y(1)\right] \mu_m(dx).$$

2.3. Existence of optimal controls for the upper and lower values. Define the upper and lower values, resp., for the game (fb denotes relaxed feedback, and rel denotes relaxed controls):

$$\bar{\gamma}^+ = \inf_{\text{relaxed fb } m_1} \sup_{\text{rel } r_2} \gamma(m_1, r_2), \tag{2.11a}$$

$$\bar{\gamma}^- = \sup_{\text{relaxed fb } m_2} \inf_{\text{rel } r_1} \gamma(r_1, m_2). \tag{2.11b}$$

It is shown below that the use of relaxed controls for the player selecting last offers no advantage over feedback controls. In section 4 it is shown that the game has a value in that $\bar{\gamma}^+ = \bar{\gamma}^- = \bar{\gamma}$. Then the numerical procedure converges to $\bar{\gamma}$ as the discretization level goes to zero (see section 3).

The definition (2.11a) is interpreted to mean that player 2 supposes that player 1 has selected a relaxed feedback control for itself, which will be fixed throughout the game (i.e., player 1 selects first). Given this presumed choice of player 1, player 2 can select any relaxed or relaxed feedback control and will choose so as to maximize. This maximizing control will exist and will actually be of the relaxed feedback control form (implied by Theorem 2.8). It will depend on the presumed choice of player 1. Given this relationship, player 1 will select a minimizing control. By Theorem 2.8, it will exist and be of the relaxed feedback form. The interpretation of (2.11b) is analogous.

THEOREM 2.6 (this is [20, Theorem 4.3, Chapter 4], adapted to the notation of the present case). *Assume* (A2.1)–(A2.5). *For a sequence* $\{m^n(\cdot)\}$ *of relaxed feedback controls, let* $m^n(x,\cdot)$ *converge weakly to* $m(x,\cdot)$ *for almost all* $x \in G$ *as* $n \to \infty$. *Then, as* $n \to \infty$, $\gamma(m^n) \to \gamma(m)$.

For fixed $m_1(\cdot)$, *maximize over* $m_2(\cdot)$, *and let* $\{m_2^n(\cdot)\}$ *be a maximizing sequence. Consider measures over the Borel sets of* $G \times U$ *which are defined by*

$$m^n(x, d\alpha)dx = m_1(x, d\alpha_1)m_2^n(x, d\alpha_2)dx \tag{2.12}$$

and take a weakly convergent subsequence. The limit can be factored into the form

$$m_1(x, d\alpha_1)\tilde{m}_2(x, d\alpha_2)dx, \tag{2.13}$$

where $\tilde{m}_2(\cdot)$ *is a relaxed feedback control for player* 2. *Since* $\tilde{m}_2(\cdot)$ *depends on* $m_1(\cdot)$, *write it as* $\tilde{m}_2(\cdot) = \overline{m}_2(\cdot; m_1)$. *Then, given* $m_1(\cdot)$, *the relaxed feedback control* $\overline{m}_2(\cdot; m_1)$ *is maximizing for player* 2 *in that*

$$\sup_{m_2} \gamma(m_1, m_2) = \gamma(m_1, \overline{m}_2(m_1)).$$

The analogous result holds in the other direction, where player 2 *chooses first.*

Remark on the proof. First, note that owing to the product form any weak sense limit of the sequence defined in (2.12) must be of the form (2.13), where $\tilde{m}_1(\cdot)$ is a relaxed feedback control. The reference [20, Theorem 4.3, Chapter 4] is concerned with a minimization problem. Changing minimization to maximization and adapting the notation to our case where there are two controls and one is fixed, it shows that the limit $m_1(x, d\alpha_1)\tilde{m}_2(x, d\alpha_2)$ is maximizing, which is the assertion of the second paragraph of the theorem. □

Relaxed controls for the player who chooses last. Suppose that with $m_1(\cdot)$ fixed, player 2 is allowed to use relaxed controls and not simply relaxed feedback controls. The following theorem says that the maximization over this larger class will not yield a better result for player 2. The analogue of the result for player 2 choosing first also holds.

THEOREM 2.7 (this is [20, Theorem 6.1, Chapter 4], adapted to the notation of the present case). *Assume* (A2.1)–(A2.5), *fix* $m_1(\cdot)$, *and let* $\overline{m}_2(\cdot; m_1)$ *be an optimal relaxed feedback control and* $r_2(\cdot)$ *an arbitrary relaxed control for player* 2. *Then for each* $x \in G$,

$$\gamma(x, m_1, r_2) \le \gamma(m_1, \overline{m}_2(m_1)).$$

THEOREM 2.8. *Assume* (A2.1)–(A2.5). *Let player* 1 *go first. Then it has an optimal control, denoted by* $\overline{m}_1^+(\cdot)$. *The analogous result holds if player* 2 *chooses first, and its optimal control is denoted by* $\overline{m}_2^-(\cdot)$.

Remark on the proof. The proof is essentially a consequence of [20, Theorem 4.3, Chapter 4], just as Theorem 2.6 was. Let player 1 go first and let $\{m_1^n(\cdot)\}$ be a minimizing sequence of relaxed feedback controls. By Theorem 2.6, if player 1 uses $m_1^n(\cdot)$, then player 2 would use the (maximizing) relaxed feedback control $\overline{m}_2(\cdot; m_1^n)$. Following the method of the reference that was used to prove Theorem 2.6, take a weakly convergent subsequence of the sequence of measures on the Borel sets of $G \times U$ that is defined by $m_1^n(x, d\alpha_1)\overline{m}_2(x, d\alpha_2; m_1^n)dx$ and denote the limit by $\overline{m}_1^+(x, d\alpha_1)\tilde{m}_2(x, d\alpha_2)dx$. Any weak sense limit must have this form, where the $\overline{m}_1^+(\cdot)$ and $\tilde{m}_2(\cdot)$ are relaxed feedback controls. For notational simplicity, let n index the weakly convergent subsequence. Then we must have $m_1^n(x, \cdot) \Rightarrow \overline{m}_1^+(x, \cdot)$ and $\overline{m}_2(x, \cdot; m_1^n) \Rightarrow \tilde{m}_2(x, \cdot)$ for almost all $x \in G$.

We need to show that $\overline{m}_1^+(\cdot)$ is optimal for player 1 if it chooses first and that it can be supposed that $\tilde{m}_2(\cdot) = \overline{m}_2(\cdot; \overline{m}_1^+)$. Since $\{m_1^n(\cdot)\}$ is minimizing for player 1 when it chooses first, $\gamma(m_1^n, \overline{m}_2(m_1^n)) \to \bar{\gamma}^+$. Suppose that $\bar{\gamma}^+ < \sup_{m_2} \gamma(\overline{m}_1^+, m_2)$. Then there is $\hat{m}_2(\cdot)$ such that $\bar{\gamma}^+ < \gamma(\overline{m}_1^+, \hat{m}_2)$. Now, let player 2 use $\hat{m}_2(\cdot)$ instead of $\overline{m}_2(\cdot; m_1^n)$ for large n. Since the sequence defined by $m_1^n(x, d\alpha_1)\hat{m}_2(x, d\alpha_2)dx$ converges weakly to the measure defined by $\overline{m}_1^+(x, d\alpha_1)\hat{m}_2(x, d\alpha_2)dx$, Theorem 2.6 implies that $\gamma(m_1^n, \hat{m}_2) \to \gamma(\overline{m}_1^+, \hat{m}_2) > \bar{\gamma}^+$. This contradicts the fact that $\{m_1^n(\cdot)\}$ is minimizing, since it implies that there is $\epsilon > 0$ such that $\gamma(m_1^n, \hat{m}_2) \geq \bar{\gamma}^+ + \epsilon$ for large n. Thus $\overline{m}_1^+(\cdot)$ is optimal for player 1 if it chooses first. Since $\bar{\gamma}^+ = \gamma(\overline{m}_1^+, \tilde{m}_2)$, without loss of generality we can suppose that $\tilde{m}_2(\cdot) = \overline{m}_2(\cdot; \overline{m}_1^+)$. □

Remark on smooth nearly optimal controls. In section 4 we will need the fact that the optimal relaxed feedback controls for either player can be smoothed with little loss. In particular, suppose that player 1 chooses first, let $\epsilon > 0$, and replace $\overline{m}_1^+(\cdot)$ by the smoothed $\overline{m}_{1,\epsilon}^+(\cdot)$ as defined above Theorem 2.3. It is true that

$$\lim_{\epsilon \to 0} \sup_{m_2} \gamma(\overline{m}_{1,\epsilon}^+, m_2) = \bar{\gamma}^+. \tag{2.14}$$

To prove (2.14), suppose that it does not hold in that there is $\delta > 0$ such that

$$\lim_{\epsilon \to 0} \sup_{m_2} \gamma(\overline{m}_{1,\epsilon}^+, m_2) \geq \bar{\gamma}^+ + \delta. \tag{2.15}$$

Then there are $m_{2,\epsilon}(\cdot)$ such that $\gamma(\overline{m}_{1,\epsilon}^+, m_{2,\epsilon}) \geq \bar{\gamma}^+ + \delta/2$ for all small $\epsilon > 0$. Let ϵ index a weakly convergent subsequence of $\overline{m}_{1,\epsilon}^+(x, d\alpha_1)m_{2,\epsilon}(x, d\alpha_2)dx$. The limit can be written as $\overline{m}_1^+(x, d\alpha_1)\tilde{m}_2(x, d\alpha_2)dx$ for some relaxed feedback control $\tilde{m}_2(\cdot)$. By Theorem 2.6, as $\epsilon \to 0$, $\gamma(\overline{m}_{1,\epsilon}^+, m_{2,\epsilon}) \to \gamma(\overline{m}_1^+, \tilde{m}_2) \geq \bar{\gamma}^+ + \delta/2$, a contradiction to the optimality of $\overline{m}_1^+(\cdot)$ for player 1 if it chooses first. Obviously, there is an analogue if player 2 chooses first.

3. Convergence of the numerical procedure.

3.1. The Markov chain approximation method. The numerical method to be employed is the Markov chain approximation method of [18, 19, 22]. The approximating processes are the same. But the numerical problem to be solved is an ergodic cost problem for a Markov chain. The method approximates the system process (2.4) by a discrete parameter finite state controlled Markov chain that is "locally consistent" with (2.4). The cost function is also approximated, and the

game problem is then solved. Some basic facts from [22] concerning the procedure will now be stated. Let h denote the approximation parameter. Many methods for getting suitable approximating chains are in the references (e.g., see [22, Chapter 5]). The approximating chain and local consistency conditions are the same for the game problems of this paper. In the present case, where $\sigma(x)\sigma'(x)$ is uniformly positive definite, for each small fixed value of h the constructed chains can be selected to be ergodic for each control [22, Chapter 7]), and this will be assumed to be the case. In fact, the chains can be chosen such that for each small h, the rate of convergence of the transition functions to the invariant measure (as time goes to infinity) will be uniform in the control. See [22, Chapter 7] for a discussion of the setup and convergence for the pure control problem.

To construct the approximation, one first defines S_h, a discretization of $\mathbb{R}^r$. For example, S_h might be a regular h-grid. The precise requirements are quite weak, and it is only the points in G and their immediate neighbors that are of interest. The state space for the chain is divided into two parts. The first part is $G_h = G \cap S_h$, on which the chain approximates the diffusion part of (2.4). If the chain tries to leave G_h, then it is returned immediately, consistently with the local reflection direction. Thus, define ∂G_h^+ to be the set of points not in G_h to which the chain might move in one step from some point in G_h. The set ∂G_h^+ is an approximation to the reflecting boundary. The use of ∂G_h^+ simplifies the analysis and allows us to get a reflection process $z^h(\cdot)$ (resp., $y^h(\cdot)$) that is analogous to $z(\cdot)$ (resp., to $y(\cdot)$).

Local consistency on G_h. Let $u_n^h = (u_{1,n}^h, u_{2,n}^h)$ denote the controls used at step n for the approximating chain ξ_n^h. Let $E_{x,n}^{h,\alpha}$ (resp., $\mathrm{Cov}_{x,n}^{h,\alpha}$) denote the expectation (resp., the covariance), given all of the data to step n, when $\xi_n^h = x, u_n^h = \alpha$. Then the chain satisfies the following consistency condition. There is $\Delta t^h(x,\alpha) = \Delta t^h$ (it does not depend on (x,α) for $x \in G$) such that $\Delta t^h \to 0$ as $h \to 0$ and

$$\begin{aligned} E_{x,n}^{h,\alpha}\left[\xi_{n+1}^h - x\right] &= b(x,\alpha)\Delta t^h + o(\Delta t^h), \\ \mathrm{Cov}_{x,n}^{h,\alpha}\left[\xi_{n+1}^h - x\right] &= a(x)\Delta t^h + o(\Delta t^h),\ a(x) = \sigma(x)\sigma'(x), \\ \|\xi_{n+1}^h - \xi_n^h\| &\le K_1 h \end{aligned} \tag{3.1}$$

for some real K_1. The $o(\Delta t^h)$ terms are uniform in (x,α). Let $P^h(x,y|\alpha_1,\alpha_2) = P^h(x,y|\alpha)$ denote the one-step transition probabilities. With the methods in [22], Δt^h is obtained automatically as a by-product of getting the $P^h(x,y|\alpha)$, and it is used as an interpolation interval. More generally, Δt^h can depend on x,α. But for theoretical purposes for the ergodic cost problem, the problem is rescaled to get constant intervals. See the discussion in [22, Chapter 7]. By (3.1), in G the conditional mean first two moments of $\xi_{n+1}^h - \xi_n^h$ are close to those of the differences of the solution to (2.4).

The first two lines of (3.1) give the conditional moments for any fixed control values $\alpha = (\alpha_1, \alpha_2)$. Suppose that the control for player $i = 1, 2$ is chosen at random, depending only on the current state (i.e., it is randomized feedback). Let $m_i^h(x, d\alpha_i)$ denote the associated probability, conditioned on the past and on the current state value x, and define $m^h(x, d\alpha) = m_1^h(x, d\alpha_1)m_2^h(x, d\alpha_2)$. Then the transition probability is

$$\int_U P^h(x,y|\alpha_1,\alpha_2)m_1^h(x,d\alpha_1)m_2^h(x,d\alpha_2).$$

The first two lines of (3.1) are now replaced by

$$\begin{aligned} E^{h,m^h}_{x,n}\left[\xi^h_{n+1}-x\right] &= b_{m^h}(x)\Delta t^h + o(\Delta t^h), \\ \mathrm{Cov}^{h,m^h}_{x,n}\left[\xi^h_{n+1}-x\right] &= a(x)\Delta t^h + o(\Delta t^h),\ a(x)=\sigma(x)\sigma'(x). \end{aligned} \tag{3.2}$$

Thus, the forms are the same as if relaxed feedback controls were used. Although the actual sample paths would differ, the transition probabilities are the same for the randomized and the relaxed feedback forms.

Local consistency on ∂G^+_h. From points in ∂G^+_h, the transitions of the chain are such that they move to G_h, with the conditional mean direction being a reflection direction at x. By (A2.2), it is always possible to construct such transitions. More precisely,

$$\lim_{h\to 0}\sup_{x\in\partial G^+_h}\mathrm{distance}(x,G_h)=0, \tag{3.3}$$

and there are $\theta_1>0$ and $\theta_2(h)\to 0$ as $h\to 0$ such that for all $x\in\partial G^+_h$,

$$\begin{aligned} E^{h,\alpha}_{x,n}\left[\xi^h_{n+1}-x\right] &\in \{a\gamma:\ \gamma\in d(x), \theta_2(h)\ge a\ge\theta_1 h\}, \\ \Delta t^h(x,\alpha) &= 0 \ \text{ for } x\in\partial G^+_h. \end{aligned} \tag{3.4}$$

The last line of (3.4) says that the reflection from states on ∂G^+_h is instantaneous. Without loss of generality, we can suppose that the transition probabilities are continuous in the control variables for each x (see [22, Chapter 5] for typical methods of construction).

Continuous time interpolation. Only the discrete time chain ξ^h_n is needed for the numerical computations. But, for the proofs of convergence, this chain must be interpolated into a continuous time process which approximates $x(\cdot)$. The interpolation intervals are suggested by the $\Delta t^h(\cdot)$ in (3.1) and (3.4). We will use a Markovian interpolation, called $\psi^h(\cdot)$. Recall that the intervals between jumps of a continuous time Markov chain on a finite state space are (conditionally, given the current state value) exponentially distributed. In our interpolation, the (conditional) mean values will just be the interpolation intervals. This interpolation is used for analytical purposes only. Let $\{\Delta\tau^h_n, n<\infty\}$ be conditionally mutually independent and "exponential" random variables in that

$$P^{h,\alpha}_{x,n}\left\{\Delta\tau^h_n\ge t\right\}=e^{-t/\Delta t^h(x,\alpha)}.$$

Note that $\Delta\tau^h_n=0$ if ξ^h_n is on the reflecting boundary ∂G^+_h. Define $\tau^h_0=0$, and for $n>0$, set $\tau^h_n=\sum_{i=0}^{n-1}\Delta\tau^h_i$. The τ^h_n will be the jump times of $\psi^h(\cdot)$. Let E^h_i denote the expectation conditioned on the data to step i. Now define $\psi^h(\cdot)$ and the interpolated reflection processes by

$$\psi^h(t)=x(0)+\sum_{\tau^h_{i+1}\le t}[\xi^h_{i+1}-\xi^h_i],$$

$$Z^h(t)=\sum_{\tau^h_{i+1}\le t}[\xi^h_{i+1}-\xi^h_i]I_{\{\xi^h_i\in\partial G^+_h\}},$$

$$z^h(t) = \sum_{\tau^h_{i+1} \le t} E^h_i[\xi^h_{i+1} - \xi^h_i] I_{\{\xi^h_i \in \partial G^+_h\}}.$$

Define the continuous time interpolations $u^h_i(\cdot)$ of the controls analogously. Let $r^h_i(\cdot)$ denote the relaxed control representation of $u^h_i(\cdot)$. The process $\psi^h(\cdot)$ is a continuous time Markov chain. When the state is x and control pair is α, the jump rate out of $x \in G_h$ is $1/\Delta t^h(x,\alpha)$. So the conditional mean interpolation interval is $\Delta t^h(x,\alpha)$; i.e., $E^{h,\alpha}_{x,n}[\tau^h_{n+1} - \tau^h_n] = \Delta t^h(x,\alpha)$.

Define $\tilde{z}^h(\cdot)$ by $Z^h(t) = z^h(t) + \tilde{z}^h(t)$. This representation splits the effects of the reflection into two parts. The first is composed of the "conditional mean" parts $E^h_i[\xi^h_{i+1} - \xi^h_i] I_{\{\xi^h_i \in \partial G^+_h\}}$, and the second is composed of the perturbations about these conditional means [22, section 5.7.3]. Both components can change only at t, where $\psi^h(t)$ can leave G_h. Suppose that at some time t, $Z^h(t) - Z^h(t-) \neq 0$, with $\psi^h(t-) = x \in G_h$. Then by (3.4), $z^h(t) - z^h(t-)$ points in a direction in $d(N_h(x))$, where $N_h(x)$ is a neighborhood with radius that goes to zero as $h \to 0$. The process $\tilde{z}^h(\cdot)$ is the "error" due to the centering of the increments of the reflection term about their conditional means and has bounded (uniformly in x, h) second moments, and it converges to zero, as will be seen in Theorem 3.1. It is convenient to represent $z^h(\cdot)$ in a form analogous to (2.6). By (A2.1), (A2.2), and the local consistency condition (3.4), we can write (modulo an asymptotically negligible term)

$$z^h(t) = \sum_i d_i y^h_i(t),$$

where $y^h_i(0) = 0$, and $y^h_i(\cdot)$ is nondecreasing and can increase only when $\psi^h(t)$ is arbitrarily close (as $h \to 0$) to the ith face of ∂G.

A representation for $\psi^h(\cdot)$. The process $\psi^h(\cdot)$ has a representation which resembles (2.4) and is useful in the convergence proofs. Let $\xi^h_0 = x$. By [22, Sections 5.7.3 and 10.4.1], we can write

$$\begin{aligned} \psi^h(t) = x \; &+ \int_0^t b(\psi^h(s), u^h(s)) ds \\ &+ \int_0^t \sigma(\psi^h(s)) dw^h(s) + Z^h(s) + \epsilon^h(s), \end{aligned} \tag{3.5}$$

where $\psi^h(t) \in G$. The process $\epsilon^h(\cdot)$ is due to the $o(\cdot)$ terms in (3.1) and is asymptotically unimportant in that, for any T, $\lim_h \sup_{x,u^h} \sup_{s \le T} E^{h,u^h}_x |\epsilon^h(s)|^2 = 0$. The process $w^h(\cdot)$ is a martingale with respect to the filtration induced by $(\psi^h(\cdot), u^h(\cdot), w^h(\cdot))$ and converges weakly to a standard (vector-valued) Wiener process. The $w^h(t)$ is obtained from $\{\psi^h(s), s \le t\}$. All of the processes in (3.5) are constant on the intervals $[\tau^h_n, \tau^h_{n+1})$.

Let $|z^h|(T)$ denote the variation of the process $z^h(\cdot)$ on the time interval $[0,T]$. Then we have the following theorem from [22]. Recall that $z^h(0) = 0$.

THEOREM 3.1 (Theorem 11.1.3 and equation (5.7.5) [22]). *Assume* (A2.1), (A2.2), *the local consistency conditions, and let $b(\cdot)$ and $\sigma(\cdot)$ be bounded and measurable. Then for any $T < \infty$, there are $K_2 < \infty$ and δ_h, where $\delta_h \to 0$ as $h \to 0$, and which do not depend on the controls or initial condition, such that*

$$E\left|z^h\right|^2(T) \le K_2, \tag{3.6}$$

$$E \sup_{s \le T} \left|\tilde{z}^h(s)\right|^2 = \delta_h E \left|z^h\right|(T). \tag{3.7}$$

Owing to the fact that the reflection directions at any corner or edge are linearly independent, the inequalities hold for $y^h(\cdot)$ replacing $z^h(\cdot)$.

The cost function and upper and lower values for the discrete game. Relaxed feedback controls, when applied to the Markov chain, are equivalent to randomized controls. Let $u^h(\cdot) = (u_1^h(\cdot), u_2^h(\cdot))$ be feedback controls for the approximating chain. Then the cost is

$$\begin{gathered}\gamma_T^h(x, u^h) = \gamma_T^h(x, u_1^h, u_2^h) = \frac{1}{T} E_x^{h,u^h} \int_0^T k_{u^h}(\psi^h(s))ds + E_x^{h,u^h} \frac{c' y^h(T)}{T}, \\ \gamma^h(u^h) = \lim_T \gamma_T^h(x, u^h).\end{gathered} \tag{3.8}$$

Now suppose that $m^h(\cdot)$ represents a randomized control (as discussed above (3.2)). Then the cost function can be written as

$$\begin{gathered}\gamma_T^h(x, m^h) = \gamma_T^h(x, m_1^h, m_2^h) = \frac{1}{T} E_x^{h,m^h} \int_0^T k_{m^h}(\psi^h(s))ds + E_x^{h,m^h} \frac{c' y^h(T)}{T}, \\ \gamma^h(m^h) = \lim_T \gamma_T^h(x, m^h).\end{gathered} \tag{3.9}$$

With the relaxed feedback control representation of an ordinary feedback control, (3.8) is a special case of (3.9). Also, we can always take the controls in (3.9) to be randomized feedback.

Suppose that player 1 chooses its control first and uses the relaxed feedback (or randomized feedback) control $m_1^h(\cdot)$. Then player 2 has a maximization problem for a finite state Markov chain. The approximating chain is ergodic for any feedback control, whether randomized or not. Then, since the transition probabilities and cost rates are continuous in the control of the second player, the optimal control of the second player exists and is a pure feedback control (not randomized) [8, volume 2], [25]. The cost does not depend on the initial condition. The analogous situation holds if player 2 chooses its control first. These facts will be used in the next theorem. We use $m_i^h(\cdot)$ to denote either a randomized feedback, relaxed feedback, or the relaxed feedback representation of an ordinary feedback control. Define the upper and lower values, resp.:

$$\bar{\gamma}^{+,h} = \inf_{m_1^h} \sup_{m_2^h} \gamma^h(m_1^h, m_2^h),$$

$$\bar{\gamma}^{-,h} = \sup_{m_2^h} \inf_{m_1^h} \gamma^h(m_1^h, m_2^h).$$

Under our hypotheses, the upper and lower values might be different, although Theorem 3.2 says that they converge to the same value asymptotically. If the dynamics are separated in the sense that $P^h(x, y|\alpha)$ can be written as a function of (x, y, α_1) plus a function of (x, y, α_2), then $\bar{\gamma}^{+,h} = \bar{\gamma}^{-,h}$. (The proof is similar to that giving the analogous result in section 4, except that the state space is discrete here.) One can choose the transition probability so that it is separated, if desired.

3.2. Convergence of the numerical procedure.

THEOREM 3.2. *Assume* (A2.1)–(A2.5) *and suppose that* [2]

$$\bar{\gamma}^+ = \bar{\gamma}^- = \bar{\gamma}. \tag{3.10}$$

Then

$$\bar{\gamma}^- \le \liminf_h \bar{\gamma}^{-,h} \le \limsup_h \bar{\gamma}^{+,h} \le \bar{\gamma}^+. \tag{3.11}$$

Hence

$$\lim_h \bar{\gamma}^{+,h} = \lim_h \bar{\gamma}^{-,h} = \bar{\gamma}, \tag{3.12}$$

and both the upper and lower values for the numerical approximation converge to the value for the original game.

Proof. Let player 1 choose its control first and let $\epsilon > 0$. Let $\overline{m}^+_{\epsilon,1}(\cdot)$ be an ϵ-smoothing of the optimal control $\overline{m}^+_1(\cdot)$ for player 1, when it chooses first, as discussed at the end of section 2. That discussion implies that, given $\delta > 0$, there is $\epsilon > 0$ such that $\overline{m}^+_{1,\epsilon}(\cdot)$ is δ-optimal for player 1 for the original problem. Now, let player 1 use $\overline{m}^+_{1,\epsilon}(\cdot)$ on the approximating chain, either as a randomized feedback or a relaxed feedback control. Given that player 1 chooses first and uses $\overline{m}^+_{1,\epsilon}(\cdot)$, we have a simple control problem for player 2. As noted above, the optimal control for player 2 exists and is pure feedback, and we denote it by $\tilde{u}^h_2(\cdot)$, with relaxed feedback control representation $\tilde{m}^h_2(\cdot)$.

By the definition of the upper value,

$$\bar{\gamma}^{+,h} \le \sup_{u^h_2} \gamma^h(\overline{m}^+_{1,\epsilon}, u^h_2) = \sup_{m^h_2} \gamma^h(\overline{m}^+_{1,\epsilon}, m^h_2) = \gamma^h(\overline{m}^+_{1,\epsilon}, \tilde{u}^h_2), \tag{3.13}$$

where $u^h_2(\cdot)$ denotes an arbitrary ordinary feedback control, and $m^h_2(\cdot)$ an arbitrary randomized feedback control. The maximum value $\gamma^h(\overline{m}^+_{1,\epsilon}, \tilde{u}^h_2)$ of the control problem for player 2 with player 1's control fixed at $\overline{m}^+_{1,\epsilon}(\cdot)$ does not depend on the initial condition. Hence, without loss of generality, the corresponding continuous time interpolation $\psi^h(\cdot)$ can be considered to be stationary. Then, using the continuity in (x, α_2) of $\int_{U_1} b(x,\alpha)\overline{m}^+_{1,\epsilon}(x, d\alpha_1)$ and of $\int_{U_1} k(x,\alpha)\overline{m}^+_{1,\epsilon}(x, d\alpha_1)$ (and replacing the minimization problem by a maximization problem) yields [22, Theorem 3.1, Chapter 11] that there is a relaxed control $\tilde{r}_2(\cdot)$ for the original problem such that[3]

$$\limsup_h \bar{\gamma}^{+,h} \le \limsup_h \gamma^h(\overline{m}^+_{1,\epsilon}, \tilde{u}^h_2) = \gamma(\overline{m}^+_{1,\epsilon}, \tilde{r}_2) \le \bar{\gamma}^+ + \delta. \tag{3.14}$$

The last inequality of (3.14) follows from Theorem 2.7 and the δ-optimality of $\overline{m}^+_{1,\epsilon}(\cdot)$ in the class of relaxed feedback controls for player 1 if it chooses first.

Now, let player 2 choose first. Then there is an analogous result with analogous notation: In particular, given $\delta > 0$, there is an $\epsilon > 0$ and an ϵ-smoothing $\overline{m}^-_{2,\epsilon}(\cdot)$ of the optimal control and a relaxed control $\tilde{r}_1(\cdot)$ for the original problem (2.4) such that

$$\liminf_h \bar{\gamma}^{-,h} \ge \liminf_h \gamma^h(\tilde{u}^h_1, \overline{m}^-_{2,\epsilon}) = \gamma(\tilde{r}_2, \overline{m}^-_{2,\epsilon}) \ge \bar{\gamma}^- - \delta. \tag{3.15}$$

Hence, since δ is arbitrary, (3.11) holds. This, with (3.10), yields the theorem. □

[2] Equation (3.10) will be proved in the next section.

[3] In [22, Theorem 3.1, Chapter 11], the symbol $m(\cdot)$ is used for a relaxed control and not a relaxed feedback control. That reference does not use relaxed feedback controls.

4. Existence of the value of the game. An approach to the proof. The existence of the value, namely (3.10), will be proved in this section. Before proceeding with the proof, we will motivate what will be needed by outlining a tentative approach. The outline is purely formal. But, later, it will be seen that the method can be carried out.

Suppose for the moment that the game for the numerical approximation has a value in that $\bar{\gamma}^{+,h} = \bar{\gamma}^{-,h}$, and let there be controls $\overline{m}_1^h(\cdot), \overline{m}_2^h(\cdot)$ for the numerical method (written in relaxed feedback form) which ttain the value, no matter who chooses first; i.e., $\overline{m}_i^h(\cdot)$ is optimal for player i whether it chooses its control first or last. Thus,

$$\bar{\gamma}^{+,h} = \bar{\gamma}^{-,h} = \bar{\gamma}^h = \gamma^h(\overline{m}_1^h, \overline{m}_2^h). \tag{4.1}$$

Suppose also that there are relaxed feedback controls $\tilde{m}_i(\cdot)$ such that, for some subsequence of $h \to 0$,

$$\overline{m}_1^h(x, d\alpha_1)\overline{m}_2^h(x, d\alpha_2)dx \Rightarrow \tilde{m}_1(x, d\alpha_1)\tilde{m}_2(x, d\alpha_2)dx. \tag{4.2}$$

Finally, suppose that for any sequence (indexed by $h \to 0$) of relaxed feedback controls $\{m_i^h(\cdot)\}, i = 1, 2$, for which $m_1^h(x, d\alpha_1)m_2^h(x, d\alpha_2)dx$ converges weakly to, say, $m_1(x, d\alpha_1)m_2(x, d\alpha_2)dx$, we have the convergence of the costs

$$\gamma^h(m_1^h, m_2^h) \to \gamma(m_1, m_2). \tag{4.3}$$

Then by (3.11) it follows that

$$\bar{\gamma}^- \leq \gamma(\tilde{m}_1, \tilde{m}_2) \leq \bar{\gamma}^+.$$

We claim that, under the above hypotheses, the limit control $\tilde{m}_i(\cdot)$ is optimal for player i if it chooses first. To prove this claim one can proceed as follows. Suppose that $\tilde{m}_1(\cdot)$ is not optimal for player 1 if it chooses first, in that $\sup_{m_2} \gamma(\tilde{m}_1, m_2) > \bar{\gamma}^+$. Then there are $\delta > 0$ and $\hat{m}_2(\cdot)$ such that $\gamma(\tilde{m}_1, \hat{m}_2) \geq \bar{\gamma}^+ + 2\delta$. Following the approach in Theorem 3.2, for $\epsilon > 0$ let $\hat{m}_{2,\epsilon}(\cdot)$ be an ϵ-smoothing of $\hat{m}_2(\cdot)$. Then, for small $\epsilon > 0$, $\gamma(\tilde{m}_1, \hat{m}_{2,\epsilon}) \geq \bar{\gamma}^+ + \delta$. Then apply $\hat{m}_{2,\epsilon}(\cdot)$ to the approximating controlled process $\psi^h(\cdot)$ to get a contradiction to the optimality of $(\overline{m}_1^h(\cdot), \overline{m}_2^h(\cdot))$ for small h. Such a contradiction implies that $\sup_{m_2} \gamma(\tilde{m}_1, m_2) \leq \bar{\gamma}^+$. But, the strict inequality $<$ is impossible due to the definition of the upper value. Hence $\sup_{m_2} \gamma(\tilde{m}_1, m_2) = \bar{\gamma}^+$, as desired.

To get the desired contradiction to the optimality of $(\overline{m}_1^h(\cdot), \overline{m}_2^h(\cdot))$ for small h, let h index a weakly convergence subsequence of the measures defined in the left side of (4.2). The limit must be of the form on the right side of (4.2) for some $\tilde{m}_i(\cdot)$, $i = 1, 2$, where $\overline{m}_i^h(x, \cdot) \Rightarrow \tilde{m}_i(x, \cdot)$ for almost all $x \in G, i = 1, 2$. Apply the control pair $(\overline{m}_1^h(\cdot), \hat{m}_{2,\epsilon}(\cdot))$ to $\psi^h(\cdot)$. Then (along the chosen subsequence of h)

$$\overline{m}_1^h(x, d\alpha_1)\hat{m}_{2,\epsilon}(x, d\alpha_2)dx \Rightarrow \tilde{m}_1(x, d\alpha_1)\hat{m}_{2,\epsilon}(x, d\alpha_2)dx.$$

Since (4.3) implies that $\gamma^h(\overline{m}_1^h, \hat{m}_{2,\epsilon}) \to \gamma(\tilde{m}_1, \hat{m}_{2,\epsilon})$, for small enough ϵ and h, we must have $\gamma^h(\overline{m}_1^h, \hat{m}_{2,\epsilon}) \geq \bar{\gamma}^{+,h} + \delta/2$, which is a contradiction to the optimality of $\overline{m}_1^h(\cdot)$. We can now conclude that

$$\sup_{m_2} \gamma(\tilde{m}_1, m_2) = \bar{\gamma}^+ = \gamma(\tilde{m}_1, \tilde{m}_2). \tag{4.4}$$

Thus, if player 1 chooses its control first and uses its optimal control $\tilde{m}_1(\cdot)$, then $\tilde{m}_2(\cdot)$ is optimal for player 2. By repeating the procedure with the order of the players reversed, we can finally conclude that, if (4.1)–(4.3) hold (at least for some subsequence of h), then (3.10) holds.

The approach outlined above for proving (3.10) is attractive. But it *cannot* work for the class of processes $\psi^h(\cdot)$ which are used for the actual Markov chain approximation numerical method in section 3, since for each h, the state space is only some finite set. Hence, the controls are not defined for all $x \in G$, and the transition function is not mutually absolutely continuous with respect to Lebesgue measure. However, in this section we are concerned only with proving (3.10), and not with the numerical procedure. Thus, we can use the approach which was outlined above for an appropriately chosen alternative approximating process for which (3.11) also holds. A discrete time process will be constructed for which (3.11) and (4.1)–(4.3) hold. This process is to be used solely to prove (3.10). It is not suitable for numerical solution. For future use, note that if the $\overline{m}_i^h(\cdot), i = 1, 2$, are relaxed feedback controls for each h and the $\overline{m}_i^h(x, \cdot)$ are defined for almost all x, then there is always a subsequence and relaxed feedback controls $\tilde{m}_i(\cdot), i = 1, 2$, for which (4.2) holds.

An alternative approximating process. To get the approximating process, time will be discretized but not space. Let $\Delta > 0$ denote the time discretization interval. We need to construct a process whose n-step transition functions $P^\Delta(x, n\Delta, \cdot|\alpha)$ have densities that are mutually absolutely continuous with respect to Lebesgue measure uniformly in (Δ, control, $t_0 \le n\Delta \le t_1$) for any $0 < t_0 < t_1 < \infty$.

Consider the following procedure. Start with the process (2.4) but with the controls held constant on the intervals $[l\Delta, l\Delta + \Delta)$, $l = 0, 1, \ldots$. The discrete approximation will be the samples at times $l\Delta, l = 0, 1, \ldots$. The controls are chosen at $t = 0$, with one of the players selected to choose first, just as for the original game. Let $u_i^\Delta(\cdot), i = 1, 2$, denote the controls if they are in pure feedback (not relaxed or randomized) form. In relaxed control notation write the controls as $m_i^\Delta(\cdot), i = 1, 2$. These controls are used henceforth, whenever control is applied. The chosen controls are applied at random as follows. At each time, only one of the players will use its control. At each time $l\Delta$, $l = 0, 1, \ldots$, flip a fair coin. With probability 1/2, player 1 will use its control during the interval $[l\Delta, l\Delta + \Delta)$ and player 2 will not. Otherwise, player 2 will use its control, and player 1 will not. The values of the controls during the interval will depend on the state at its start. The optimal controls will be feedback. Define $x^\Delta(t) = x(l\Delta)$ on $[l\Delta, l\Delta + \Delta)$. For pure (not randomized or relaxed) feedback controls $u_i^\Delta(\cdot), i = 1, 2$, the system can be written as

$$dx = b^\Delta(x, u^\Delta(x^\Delta))dt + \sigma(x)dw + dz, \tag{4.5a}$$

where the value of $b^\Delta(\cdot)$ is determined by the coin tossing randomization procedure at the times $l\Delta, l = 0, 1 \ldots$. In particular, at $t \in [l\Delta, l\Delta + \Delta)$, $b^\Delta(x, u^\Delta(x^\Delta))$ is defined by $2b_i(x(t), u_i^\Delta(x^\Delta(t)))$, for either $i = 1$ or $i = 2$ according to the random choice made at $l\Delta$. If the controls $u_i^\Delta(x)$ are replaced by relaxed feedback controls $m_i^\Delta(x, \cdot)$, then write the model as

$$dx = b^\Delta(x, m^\Delta(x^\Delta))dt + \sigma(x)dw + dz, \tag{4.5b}$$

where at $t \in [l\Delta, l\Delta+\Delta)$, $b^\Delta(x, m^\Delta(x^\Delta))$ is $2\int_{U_i} b_i(x(t), \alpha_i)m_i^\Delta(x(l\Delta), d\alpha_i)$, for either $i = 1$ or $i = 2$ according to the random choice made at $l\Delta$. Following the Girsanov transformation based usage in (2.4), the Wiener process $w(\cdot)$ should be indexed by the controls $u^\Delta(\cdot)$ or $m^\Delta(\cdot)$, but we omit it for notational simplicity. In the rest of

this section, we consider $m^\Delta(\cdot) = (m_1^\Delta(\cdot), m_2^\Delta(\cdot))$ to be a vector and not a product measure.

Let $E^{\Delta,i,\alpha_i}_{x(l\Delta)}$ denote the expectation of functionals on $[l\Delta, l\Delta+\Delta)$ when player i acts on that interval and uses control action α_i. Let $P_i^\Delta(x,\cdot|\alpha_i)$ denote the measure of $x(\Delta)$, given that the initial condition is x, player i acts and uses control action α_i. The conditional mean increment in the total cost function on the time interval $[l\Delta, l\Delta+\Delta)$ is, for $u_i^\Delta(x(l\Delta)) = \alpha_i, i = 1,2$,

$$C^\Delta(x(l\Delta),\alpha) = \frac{1}{2}\sum_{i=1,2} E^{\Delta,i,\alpha_i}_{x(l\Delta)}\left[\int_{l\Delta}^{l\Delta+\Delta} 2k_i(x(s),\alpha_i))ds + c'\left(y(l\Delta+\Delta) - y(l\Delta)\right)\right]. \tag{4.6}$$

Note that $C^\Delta(x,\alpha)$ is the sum of two terms, one depending on (x,α_1) and the other on (x,α_2). The weak sense uniqueness of the solution to (2.4) for any control and initial condition implies the following result. Define $C^\Delta(x, m^\Delta(x))$ analogously, and we will sometimes write it as $C^\Delta(x, m_1^\Delta(x), m_2^\Delta(x))$.

THEOREM 4.1. *Assume* (A2.1)–(A2.5). *Then for each* $\Delta > 0$, $C^\Delta(\cdot)$ *is continuous, and the measures* $P_i^\Delta(\cdot)$ *are weakly continuous in that for any bounded and continuous real-valued function* $f(\cdot)$, $\int f(y)P_i^\Delta(x,dy|\alpha)$ *and* $C^\Delta(x,\alpha)$ *are continuous in* (x,α).

The reason for choosing the acting controls at random at each time $l\Delta, l = 0,1,\ldots,$ is that the randomization "separates" the cost rates and dynamics in the controls for the two players. By separation, we mean that both the cost function and transition function are the sum of two terms, one depending on (x,α_1) and the other on (x,α_2). This separation is important since it gives the "Isaacs condition" which is needed to assure the existence of a value for the game for the discrete time process, as seen in Theorem 4.2. Proceeding formally at this point, let $\mu^\Delta_{m^\Delta}(\cdot)$ denote the invariant measure under the control $m^\Delta(\cdot)$. Define the stationary cost increment

$$\lambda^\Delta(m^\Delta) = \int_G \mu^\Delta_{m^\Delta}(dx)\left[\int_U C(x,\alpha)m_1^\Delta(x,d\alpha_1)m_2^\Delta(x,d\alpha_2)\right].$$

Note that, due to the scaling, $\lambda^\Delta(m^\Delta)$ is an average over an interval of length Δ: hence $\lambda^\Delta(m^\Delta) = \Delta\gamma^\Delta(m^\Delta)$. Suppose for the moment that there is an optimal control $\overline{m}_i^\Delta(\cdot), i = 1,2$, for each $\Delta > 0$, and define $\overline{\lambda}^\Delta = \lambda^\Delta(\overline{m}^\Delta)$. The "separation" is easily seen from the formal Isaacs equation for the value of the discrete time problem, namely,

$$\overline{\lambda}^\Delta + \overline{g}^\Delta(x) = \inf_{\alpha_1}\sup_{\alpha_2}\left[\frac{1}{2}\int \overline{g}^\Delta(x+y)P_1^\Delta(x,dy|\alpha_1) + \frac{1}{2}\int \overline{g}^\Delta(x+y)P_2^\Delta(x,dy|\alpha_2) + C^\Delta(x,\alpha)\right], \tag{4.7}$$

where $\overline{g}^\Delta(\cdot)$ is the relative value or potential function.

THEOREM 4.2. *Assume* (A2.1)–(A2.5). *Then* (3.10) *holds.*

Proof. We will work with the approximating process $x(l\Delta), l = 0,1,\ldots,$ just described, where $x(\cdot)$ is defined by (4.5) with the piecewise constant control, and verify the conditions imposed in the formal discussion at the beginning of the section. Results from [20] will be exploited whenever possible. The result (3.11) holds (with Δ replacing h) for the same reasons that it holds for the numerical approximating process

of the last section. For any sequence of relaxed feedback controls $m_i^\Delta(\cdot), i = 1, 2$, there is a subsequence (indexed by $\Delta \to 0$) and relaxed feedback controls $\tilde{m}_i^\Delta(\cdot), i = 1, 2$, such that

$$m_1^\Delta(x, d\alpha_1)m_2^\Delta(x, d\alpha_2)dx \Rightarrow \tilde{m}_1(x, d\alpha_1)\tilde{m}_2(x, d\alpha_2)dx.$$

One needs to show the analogue of (4.3), namely (along the same subsequence, indexed by Δ)

$$\gamma^\Delta(m^\Delta) \to \gamma(\tilde{m}). \tag{4.8}$$

The process $\{x(l\Delta)\}$ based on (4.5) inherits the crucial properties of (2.4), as developed in [20, Chapter 4] and summarized in subsection 2.2. In particular, for each positive Δ and n the n-step transition probability $P^\Delta(x, n\Delta, \cdot|m^\Delta)$ is mutually absolutely continuous with respect to Lebesgue measure uniformly in the control and in $x \in G$, $n\Delta \in [t_0, t_1]$, for any $0 < t_0 < t_1 < \infty$, and it is a strong Feller process. The invariant measures are mutually absolutely continuous with respect to Lebesgue measure, again uniformly in the control. Then the proof of (4.8) is very similar to the corresponding proof for (2.4) given in [20, Theorem 4.3, Chapter 4], and the details are omitted. There are relaxed feedback controls $\overline{m}_i^{\Delta,+}(\cdot), i = 1, 2$, which are optimal if player 1 chooses its control first (i.e., for the upper value), and $\overline{m}_i^{\Delta,-}(\cdot), i = 1, 2$, which are optimal if player 2 chooses its control first (i.e., for the lower value).

We will concentrate on showing the analogue of (4.1), namely,

$$\bar{\gamma}^{+,\Delta} = \bar{\gamma}^{-,\Delta}. \tag{4.9}$$

By the (uniform in the controls) mutual absolute continuity (with respect to Lebesgue measure) of the one-step transition probabilities for each $\Delta > 0$, the process satisfies a Doeblin condition uniformly in the control. Hence it is uniformly ergodic (uniformly in the control) [23, Theorems 16.2.1 and 16.2.3]. In particular it follows that there are constants K_Δ and ρ_Δ, with $\rho_\Delta < 1$ such that

$$\sup_{x,m^\Delta} \left| E_x^{\Delta,m^\Delta} \int_U C(x(n\Delta), \alpha)m^\Delta(x(n\Delta), d\alpha) - \lambda^\Delta(m^\Delta) \right| \le K_\Delta \left[\rho_\Delta\right]^n,$$

where $\lambda^\Delta(m^\Delta)$ is defined above (4.7).

Define the relative value function

$$g^\Delta(x, m^\Delta) = \sum_{l=0}^{\infty} \left[E_x^{\Delta,m^\Delta} C(x(l\Delta), m^\Delta(x(n\Delta)) - \lambda^\Delta(m^\Delta) \right].$$

The summands converge to zero exponentially, uniformly in $(x, m^\Delta(\cdot))$. Also, by the strong Feller property the summands (for $l > 0$) are continuous. Define $g^{\Delta,+}(x) = g^\Delta(x, \overline{m}^{\Delta,+})$ and $g^{\Delta,-}(x) = g^\Delta(x, \overline{m}^{\Delta,-})$. Then a direct evaluation yields

$$\bar{\lambda}^{\Delta,+} + g^{\Delta,+}(x) = E_x^{\Delta,\overline{m}^{\Delta,+}} \left[g^{\Delta,+}(x(\Delta)) + C^\Delta(x, \overline{m}^{\Delta,+}(x)) \right]. \tag{4.10}$$

Next we show that under $\overline{m}_1^{\Delta,+}(\cdot)$ for player 1 (and for almost all x)

$$\bar{\lambda}^{\Delta,+} + g^{\Delta,+}(x) = \sup_{\alpha_2} \left[E_x^{\Delta,\overline{m}_1^{\Delta,+},\alpha_2} g^{\Delta,+}(x(\Delta)) + C^\Delta(x, \overline{m}_1^{\Delta,+}(x), \alpha_2) \right]. \tag{4.11}$$

By (4.10), (4.11) holds for almost all x with the equality replaced by the inequality $\leq$. The function in brackets in (4.11) is continuous in α_2 uniformly in $x \in G$. Suppose that (4.11) does not hold on a set $A \subset G$ of Lebesgue measure $l(A) > 0$. Let $\tilde{m}_2^\Delta(\cdot)$ denote the (relaxed feedback control representation of the) maximizing control in (4.11). Then

$$\bar{\lambda}^{\Delta,+} + g^{\Delta,+}(x) \leq \left[E_x^{\Delta,\overline{m}_1^{\Delta,+},\tilde{m}_2^\Delta} g^{\Delta,+}(x(\Delta)) + C^\Delta(x, \overline{m}_1^{\Delta,+}(x), \tilde{m}_2^\Delta(x)) \right], \tag{4.12}$$

with strict inequality for $x \in A$. Now, integrate both sides of (4.12) with respect to the invariant measure $\mu^\Delta_{\{\overline{m}_1^{\Delta,+},\tilde{m}_2^\Delta\}}(\cdot)$ corresponding to the control $(\overline{m}_1^\Delta(\cdot), \tilde{m}_2^\Delta(\cdot))$ and note that

$$\int g^{\Delta,+}(x)\mu^\Delta_{\{\overline{m}_1^{\Delta,+},\tilde{m}_2^\Delta\}}(dx) = \int \left[E_x^{\Delta,\overline{m}_1^{\Delta,+},\tilde{m}_2^\Delta} g^{\Delta,+}(x(\Delta)) \right] \mu^\Delta_{\{\overline{m}_1^{\Delta,+},\tilde{m}_2^\Delta\}}(dx). \tag{4.13}$$

Also, by definition,

$$\lambda^\Delta(\overline{m}_1^{\Delta,+}, \tilde{m}_2^\Delta) = \int C^\Delta(x, \overline{m}_1^{\Delta,+}(x), \tilde{m}_2^\Delta(x))\mu^\Delta_{\{\overline{m}_1^{\Delta,+},\tilde{m}_2^\Delta\}}(dx).$$

Then, canceling the terms in (4.13) from the integrated inequality and using the fact that the invariant measure is mutually absolutely continuous with respect to Lebesgue measure yields $\bar{\lambda}^{\Delta,+} < \lambda^\Delta(\overline{m}_1^{\Delta,+}, \tilde{m}_2^\Delta)$, which contradicts the optimality of $\overline{m}_2^{\Delta,+}(\cdot)$ for player 2 if player 1 selects its control first. Thus, (4.11) holds.

Next, given that (4.11) holds, let us show that for almost all x

$$\bar{\lambda}^{\Delta,+} + g^{\Delta,+}(x) = \inf_{\alpha_1} \sup_{\alpha_2} E_x^{\Delta,\alpha_1,\alpha_2} \left[g^{\Delta,+}(x(\Delta)) + C^\Delta(x, \alpha_1, \alpha_2) \right]. \tag{4.14}$$

By (4.11), this last equation holds if $\overline{m}_1^{\Delta,+}(\cdot)$ replaces α_1 and the inf is dropped. Suppose that (4.14) is false. Then there are $A \in G$ with $l(A) > 0$ and $\epsilon > 0$ such that for $x \in A$ the equality is replaced by the inequality $\geq$ plus ϵ, with the inequality $\geq$ holding for almost all other $x \in G$. More particularly, let $\hat{m}_1^{\Delta,+}(\cdot)$ denote the minimizing control for player 1 in (4.14). Then we have, for almost all x and any $m_2^\Delta(\cdot)$,

$$\bar{\lambda}^{\Delta,+} + g^{\Delta,+}(x) \geq E_x^{\Delta,\hat{m}_1^\Delta,m_2^\Delta} \left[g^{\Delta,+}(x(\Delta)) + C^\Delta(x, \hat{m}_1^\Delta(x), m_2^\Delta(x)) \right] + \epsilon I_{\{x\in A\}}. \tag{4.15}$$

Now, repeating the procedure used to prove (4.11), integrate both sides of (4.15) with respect to the invariant measure associated with $(\hat{m}_1^\Delta(\cdot), m_2^\Delta(\cdot))$, use the fact that the invariant measure is mutually absolutely continuous with respect to Lebesgue measure, uniformly in the controls, and cancel the terms which are analogous to those in (4.13) to get that

$$\bar{\lambda}^{\Delta,+} > \sup_{m_2^\Delta} \lambda^\Delta(\hat{m}_1^\Delta, m_2^\Delta).$$

This implies that $\overline{m}_1^{\Delta,+}(\cdot)$ is not optimal for player 1 if it selects its control first, a contradiction. Thus, (4.14) holds. The analogous procedure can be carried out for the lower value where player 2 selects its control first.

Now the fact that the dynamics and cost rate are separated in the control implies that $\inf_{\alpha_1}\sup_{\alpha_2} = \sup_{\alpha_2}\inf_{\alpha_1}$ in (4.14). Thus, (4.14) holds with the order of the sup

and inf inverted. By working with (4.14) with the sup and inf inverted and following an argument similar to that used to prove (4.14), one can show that $\bar{\lambda}^{\Delta,+} = \bar{\lambda}^{\Delta,-}$ and that $\overline{m}_i^{\Delta}(\cdot)$ is optimal for player i whether it selects first or last. The rest of the details are left to the reader. □

REFERENCES

[1] E. Altman and H.J. Kushner, *Admission control for combined guaranteed performance and best effort communications systems under heavy traffic*, SIAM J. Control Optim., 37 (1999), pp. 1780–1807.

[2] J.A. Ball, M. Day, and P. Kachroo, *Robust feedback control of a single server queueing system*, Math. Control Signals Systems, 12 (1999), pp. 307–345.

[3] J.A. Ball, M. Day, P. Kachroo, and T. Yu, *Robust L_2-gain for nonlinear systems with projection dynamics and input constraints: An example from traffic control*, Automatica J. IFAC, 35 (1999), pp. 429–444.

[4] M. Bardi, S. Bottacin, and M. Falcone, *Convergence of discrete schemes for discontinuous value functions of pursuit-evasion games*, in New Trends in Dynamic Games and Applications, G.J. Olsder, ed., Birkhäuser Boston, Cambridge, MA, 1995, pp. 273–304.

[5] M. Bardi, M. Falcone, and P. Soravia, *Fully discrete schemes for the value function of pursuit-evasion games*, in Advances in Stochastic Games and Applications, T. Basar and A. Haurie, eds., Birkhäuser Boston, Cambridge, MA, 1994.

[6] M. Bardi, M. Falcone, and P. Soravia, *Numerical methods for pursuit-evasion games via viscosity solutions*, in Stochastic and Differential Games: Theory and Numerical Methods, M. Bardi, T.E.S. Raghavan, and T. Parthasarathy, eds., Birkhäuser Boston, Cambridge, MA, 1998, pp. 105–175.

[7] T. Basar and P. Bernhard, *H_∞-Optimal Control and Related Minimax Problems*, Birkhäuser Boston, Cambridge, MA, 1991.

[8] D.M. Bertsekas, *Dynamic Programming and Optimal Control*, Athena-Scientific, Belmont, MA, 1995.

[9] P. Billingsley. *Convergence of Probability Measures*, 2nd ed., Wiley, New York, 1999.

[10] V.S. Borkar, *Optimal Control of Diffusion Processes*, Longman Scientific and Technical, Harlow, Essex, UK, 1989.

[11] N. Dunford and J.T. Schwartz, *Linear Operators, Part* 1: *General Theory*, Wiley-Interscience, New York, 1966.

[12] P. Dupuis and H. Ishii, *On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications*, Stochastics Stochastics Rep., 35 (1991), pp. 31–62.

[13] S.N. Ethier and T.G. Kurtz, *Markov Processes: Characterization and Convergence*, Wiley, New York, 1986.

[14] W.F. Fleming, *Generalized solutions in optimal stochastic control*, in Differential Games and Control Theory: III, P.T. Liu, E. Roxin, and R. Sternberg, eds., Marcel Dekker, New York, 1977, pp. 147–165.

[15] W.H. Fleming and W. McEneaney, *Risk-sensitive control on an infinite time horizon*, SIAM J. Control Optim., 33 (1995), pp. 1881–1915.

[16] J.M. Harrison and R.J. Williams, *Brownian models of open queueing networks with homogeneous customer populations*, Stochastics Stochastics Rep., 22 (1987), pp. 77–115.

[17] I. Karatzas and S.E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.

[18] H.J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.

[19] H.J. Kushner, *Numerical methods for stochastic control problems in continuous time*, SIAM J. Control Optim., 28 (1990), pp. 999–1048.

[20] H.J. Kushner, *Heavy Traffic Analysis of Controlled Queueing and Communication Networks*, Springer-Verlag, Berlin, New York, 2001.

[21] H.J. Kushner, *Numerical methods for stochastic differential games*, SIAM J. Control Optim., 41 (2003), pp. 457–486.

[22] H.J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, Berlin, New York, 1992. Second edition, 2001.

[23] S.P. Meyn and R.I. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, Berlin, New York, 1994.

[24] O. Pourtallier and M. Tidball, *Approximation of the Value Function for a Class of Dif-*

ferential Games with Target, Research report 2942, INRIA, France, 1996.
[25] M.L. Puterman, *Markov Decision Processes*, Wiley, New York, 1994.
[26] M.I. Reiman and R.J. Williams, *A boundary property of semimartingale reflecting Brownian motions*, Prob. Theory Rel. Fields, 77 (1988), pp. 87–97.
[27] M.I. Reiman, *Open queueing networks in heavy traffic*, Math. Oper. Res., 9 (1984), pp. 441–458.
[28] M. Tidball, *Undiscounted zero-sum differential games with stopping times*, in New Trends in Dynamic Games and Applications, G.J. Olsder, ed., Birkhäuser Boston, Cambridge, MA, 1995, pp. 305–322.
[29] M. Tidball and R.L.V. González, *Zero-sum differential games with stopping times: Some results and about its numerical resolution*, in Advances in Dynamic Games and Applications, T. Basar and A. Haurie, eds., Birkhäuser Boston, Cambridge, MA, 1994, pp. 106–124.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 1934–1959

WEAK CONVERSE LYAPUNOV THEOREMS AND CONTROL-LYAPUNOV FUNCTIONS*

CHRISTOPHER M. KELLETT† AND ANDREW R. TEEL‡

Abstract. Given a *weakly uniformly globally asymptotically stable* closed (not necessarily compact) set $\mathcal{A}$ for a differential inclusion that is defined on $\mathbb{R}^n$, is locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$, and satisfies other basic conditions, we construct a weak Lyapunov function that is locally Lipschitz on $\mathbb{R}^n$. Using this result, we show that uniform global asymptotic controllability to a closed (not necessarily compact) set for a locally Lipschitz nonlinear control system implies the existence of a locally Lipschitz control-Lyapunov function, and from this control-Lyapunov function we construct a feedback that is robust to measurement noise.

Key words. converse Lyapunov theorem, weak set stability, differential inclusions, control-Lyapunov function, asymptotic controllability

AMS subject classifications. 34A60, 93B05, 93D15, 93D20, 93C57

DOI. 10.1137/S0363012901398186

1. Introduction.

Herein we consider the related questions of the existence of a weak Lyapunov function for differential inclusions, and the existence of a control-Lyapunov function for controlled differential equations, under the assumption of weak asymptotic stability (respectively, asymptotic controllability) of (to) a closed, not necessarily compact, set $\mathcal{A}$.

The converse question in Lyapunov theory has received a great deal of attention. In the case of the differential inclusion

$$\dot{x} \in F(x),\ x \in \mathbb{R}^n, \tag{1.1}$$

we ask the following question: Given the weak (or strong) asymptotic stability of an attractor $\mathcal{A}$, does there exist a positive definite, proper function $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ satisfying a specific decrease condition along at least one trajectory (respectively, all trajectories)?

Numerous results on the existence of smooth, strong Lyapunov functions for *strong asymptotic stability* have been established, that is, when *all* solutions of (1.1) satisfy stability and uniform attractivity properties with respect to the set $\mathcal{A}$. Among these results is the pioneering work of Kurzweil [17] and Wilson [31] with recent results by Lin, Sontag, and Wang [18], Clarke, Ledyaev, and Stern [7], and Teel and Praly [30]. A more complete summary of these results can be found in Teel and Praly [30].

All of the above references for strong asymptotic stability generate *smooth* Lyapunov functions. Lower semicontinuous Lyapunov functions for weak stability (not

*Received by the editors November 13, 2001; accepted for publication (in revised form) May 31, 2003; published electronically January 22, 2004. Portions of this article appeared in preliminary form in *Proceedings of Mathematical Theory of Networks and Systems*, Perpignan, France, June 2000 and *Proceedings of the 39th IEEE Conference on Decision and Control*, Sydney, Australia, December 2000. This research was supported in part by the NSF under grant ECS-9896140 and the AFOSR under grant F49620-00-1-0106.

http://www.siam.org/journals/sicon/42-6/39818.html

†Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106. Current address: Department of Electrical and Electronic Engineering, University of Melbourne, VIC 3052, Australia (ckellett@ee.mu.oz.au).

‡Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 (teel@ece.ucsb.edu).

asymptotic) were given by Roxin [21, Theorem 9.5] for arbitrary closed sets $\mathcal{A}$ and by Deimling [10, Proposition 14.2] for $\mathcal{A} = \{0\}$. Clarke, Ledyaev, and Stern [7, Remark 1.5 and Theorem 6.1] noted that the existence of a continuously differentiable Lyapunov function for weak asymptotic stability of $\mathcal{A} = \{0\}$ implies the existence of a smooth Lyapunov function. However, they further demonstrated that the set-valued map $F(\cdot)$ must satisfy a nongeneric covering condition to admit a continuously differentiable weak Lyapunov function. Since we would like to consider inclusions which do not satisfy this covering condition, we require a decrease condition for nondifferentiable functions. We will make use of the Dini subderivate.

DEFINITION 1. *The Dini subderivate of a function $V : \mathbb{R}^n \to \mathbb{R}$ at a point $x \in \mathbb{R}^n$ in the direction $v \in \mathbb{R}^n$ is defined as*

$$DV(x; v) := \liminf_{w \to v, \varepsilon \to 0^+} \frac{V(x + \varepsilon w) - V(x)}{\varepsilon} .$$

For a locally Lipschitz function $V : \mathcal{O} \to \mathbb{R}$ ($\mathcal{O}$ open) at $x \in \mathcal{O}$ in the direction $v \in \mathbb{R}^n$ this simplifies to (see [8, p. 136])

$$DV(x; v) = \liminf_{\varepsilon \to 0^+} \frac{V(x + \varepsilon v) - V(x)}{\varepsilon} .$$

Therefore, in order to consider a more general class of inclusions (i.e., the class that includes inclusions which do not satisfy the previously mentioned covering condition), we search for a locally Lipschitz function $V(\cdot)$ and specify the decrease condition as

$$\min_{w \in F(x)} DV(x; w) \leq -V(x) \ \forall x \in \mathbb{R}^n.$$

Under appropriate conditions on the map $F : \mathbb{R}^n \to$ subsets of $\mathbb{R}^n$, the existence of a locally Lipschitz weak Lyapunov function for (possibly) noncompact attractors is asserted in Theorem 2.1. This result first appeared in Kellett and Teel [14]. We note that, while their results were stated for controlled differential equations, Clarke et al. [5] and Rifford [20] used differential inclusions as an intermediary. Implicitly, Clarke et al. [5] constructed a weak Lyapunov function, locally Lipschitz on compact sets disjoint from the origin, given, essentially, weak asymptotic stability of the origin. Rifford [20] then combined these functions in a clever way to obtain a weak Lyapunov function that is locally Lipschitz on $\mathbb{R}^n$, given weak asymptotic stability to the origin.

Now consider the controlled differential equation

$$\dot{x} = f(x, u), \ x \in \mathbb{R}^n, \ u \in \mathcal{U} , \tag{1.2}$$

where $\mathcal{U}$ is the set of all possible controls, and, rather than a weak Lyapunov function, we are interested in a control-Lyapunov function. An early result related to the existence of control-Lyapunov functions came from Roxin [22]. Results on the existence of control-Lyapunov functions, given asymptotic controllability to the origin, came from Sontag [24] (cf. Sontag and Sussman [29]), Clarke et al. [6], Sontag [28], and Clarke et al. [5]. In [20], Rifford generated a locally Lipschitz control-Lyapunov function, given asymptotic controllability to the origin, answering a longstanding question on the existence of same. A continuous control-Lyapunov function was generated by Albertini and Sontag [1] under the assumption of asymptotic controllability to a closed (not necessarily compact) set.

It is well known that continuously differentiable control-Lyapunov functions fail to exist for generic systems; i.e., systems which fail to satisfy Brockett's covering condition (see Brockett [4] or Ryan [23] for a definition). Again, we would like to consider a broader class of systems, including those that do not satisfy Brockett's condition. Toward this end, we examine the differential inclusion obtained by allowing the controls to range over the admissable control set; i.e.,

$$F(x) := \{z \in \mathbb{R}^n : z = f(x,u), u \in \mathcal{U}\} \ .$$

Intuitively, then, weak asymptotic stability of $\mathcal{A}$ for the differential inclusion thus defined is equivalent to the asymptotic controllability of the differential equation to $\mathcal{A}$. That is, the trajectory of $\dot{x} \in F(x)$ which does not wander too far from and is attracted to the set $\mathcal{A}$ is generated by a particular control selection. The notion of controllability used herein will be made precise in section 3. Following the arguments presented for inclusions, the corresponding decrease condition of the control-Lyapunov function then becomes

$$\min_{w \in f(x,\mathcal{U})} DV(x;w) \leq -V(x) \ \forall x \in \mathbb{R}^n \ .$$

That is, there exists a control selection such that the direction of the vector field defining the system causes the Dini subderivate of $V(\cdot)$ to decrease. Note that this is an intuitive discussion and we have avoided concerning ourselves with specifics such as a necessary "small-control property," precise regularity conditions on $f(\cdot,\cdot)$, and other technical details. These specifics are addressed in the following sections.

Remark 1. The above decrease condition involving the Dini subderivate was also used in Clarke et al. [5], [6]. In both references, use is also made of an equivalent formulation in terms of the proximal subgradient; i.e.,

$$\min_{u \in \mathcal{U}} \langle f(x,u), \zeta \rangle \leq -V(x) \ \forall x \in \mathbb{R}^n \ \forall \zeta \in \partial_P V(x) \ . \qquad \square$$

The significance of this novel result stems from the important role that control-Lyapunov functions have played in the development of stabilizing state feedbacks over the years. As examples, we refer the reader to Artstein [2], Sontag [25], Freeman and Kokotović [12], and Krstić, Kanellakopoulos, and Kokotović [16] for the case of continuously differentiable control-Lyapunov functions and to Clarke et al. [6], Sontag [28], and Clarke et al. [5] for locally Lipschitz control-Lyapunov functions. Similar to the latter articles, in section 4 we present the design of a (discontinuous) stabilizing state feedback that is robust to small additive disturbances and measurement noise using our derived control-Lyapunov function.

Our approach is to convert the control system into a differential inclusion (which is the approach also taken in Clarke et al. [5] and Rifford [20]) and then use the result on the existence of a Lyapunov function for the differential inclusion to get the promised control-Lyapunov function. The novelty of the current control-Lyapunov function is that the result is derived for closed (possibly noncompact) attractors. The proof technique is also novel in that it follows directly from a comparison function formulation of the controllability or stability property. This result first appeared in Kellett and Teel [15].

Our paper is organized as follows. Section 2 contains the precise statement of our weak converse Lyapunov theorem for differential inclusions with the associated proof in section 5. Section 3 contains our control-Lyapunov function result. A stabilizing

feedback construction for use with locally Lipschitz control-Lyapunov functions, such as the one presented in section 3, is given in section 4, with a robustness result for this feedback given in section 4.4. Section 6 contains necessary technical proofs.

2. A weak converse Lyapunov theorem. Having given some insight for the results which follow, we begin to make these ideas more precise. In what follows we let $|\cdot|$ denote the Euclidean norm on $\mathbb{R}^n$; i.e., $|x| = \sqrt{\langle x, x\rangle}$. For a closed set $\mathcal{A} \subset \mathbb{R}^n$ we write the distance from a point $x \in \mathbb{R}^n$ to the set $\mathcal{A}$ as $|x|_{\mathcal{A}} := \inf_{a\in\mathcal{A}} |x - a|$. We let $\overline{\mathcal{B}}_n(x, r)$ denote the closed ball in $\mathbb{R}^n$ of radius r centered at x; i.e., $\overline{\mathcal{B}}_n(x, r) := \{\xi \in \mathbb{R}^n : |\xi - x| \le r\}$. We define $\overline{\mathcal{B}}_n := \overline{\mathcal{B}}_n(0, 1)$, where 0 denotes the origin in $\mathbb{R}^n$. Recall that a function $\alpha : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ belongs to class-$\mathcal{K}_\infty$ ($\alpha \in \mathcal{K}_\infty$) if it is continuous, zero at zero, strictly increasing, and unbounded. A function $\beta : \mathbb{R}_{\ge 0} \times \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ belongs to class-$\mathcal{KL}$ if, for each $t \ge 0$, $\beta(\cdot, t)$ is nondecreasing and $\lim_{s\to 0^+} \beta(s, t) = 0$, and, for each $s \ge 0$, $\beta(s, \cdot)$ is nonincreasing and $\lim_{t\to\infty} \beta(s, t) = 0$.

A function $x : [0, T] \to \mathbb{R}^n$ ($T > 0$) is said to be a *solution* of the differential inclusion $\dot{x} \in F(x)$ if it is absolutely continuous and satisfies, for almost all $t \in [0, T]$, $\dot{x}(t) \in F(x(t))$. A function $x : [0, T) \to \mathbb{R}^n$ ($0 < T \le +\infty$) is said to be a *maximal solution* of the differential inclusion if it does not have an extension which is a solution belonging to $\mathbb{R}^n$; i.e., either $T = \infty$ or there does not exist a solution $y : [0, T_+] \to \mathbb{R}^n$ with $T_+ > T$ such that $y(t) = x(t)$ for all $t \in [0, T)$. We use $\phi(\cdot, x)$ to denote a solution of $\dot{x} \in F(x)$ starting at x. We denote by $\mathcal{S}[0, T](x)$, or $\mathcal{S}[0, T)(x)$, the set of maximal solutions starting at x that are defined on the time interval $[0, T]$, or $[0, T)$.

The following basic conditions guarantee existence of solutions for differential inclusions.

DEFINITION 2. *The set-valued map $F(\cdot)$ is said to satisfy the* basic conditions on $\mathbb{R}^n$ *if, for each $x \in \mathbb{R}^n$, $F(x)$ is nonempty, compact, and convex and if $F(\cdot)$ is upper semicontinuous on $\mathbb{R}^n$; i.e., for each $x \in \mathbb{R}^n$ and $\varepsilon > 0$ there exists $\delta > 0$ such that, for all $\xi \in \mathbb{R}^n$ satisfying $|x - \xi| < \delta$, we have $F(\xi) \subseteq F(x) + \varepsilon\mathcal{B}_n$.*

Previous results which obtained lower semicontinuous Lyapunov functions (see [21], [10]) assumed the set-valued $F(\cdot)$ was merely upper semicontinuous. Our stronger result (i.e., existence of a locally Lipschitz Lyapunov function) will require a correspondingly stronger regularity property.

DEFINITION 3. *A set-valued map $F : \mathbb{R}^n \to$ subsets of $\mathbb{R}^n$ is* locally Lipschitz on $\mathcal{O} \subseteq \mathbb{R}^n$ *if for all $x \in \mathcal{O}$ there exists a neighborhood $\mathcal{U} \subseteq \mathcal{O}$ of x and $L > 0$ such that $x_1, x_2 \in \mathcal{U}$ implies $F(x_1) \subseteq F(x_2) + L|x_1 - x_2|\overline{\mathcal{B}}_n$. We say that this property is* uniform in distance to the closed set $\mathcal{A}$ *if for any $\ell > 0$ the above neighborhood can be defined as $\mathcal{U} := \{x \in \mathbb{R}^n : |x|_{\mathcal{A}} \le \ell\}$.*

DEFINITION 4. *Analogous to our terminology for set-valued maps, we say that a function $g : \mathbb{R}^n \to \mathbb{R}^n$ is locally Lipschitz, uniformly in distance to the closed set $\mathcal{A}$, if for any $\ell > 0$ and a closed set $\mathcal{U} := \{x \in \mathbb{R}^n : |x|_{\mathcal{A}} \le \ell\}$ there exists $L > 0$ such that for every $x_1, x_2 \in \mathcal{U}$ we have $|g(x_1) - g(x_2)| \le L|x_1 - x_2|$.*

Our stability condition is phrased in the language of comparison functions (as is our controllability condition in the next section). This presents a simple and concise way to summarize both uniform boundedness and attractivity.

DEFINITION 5. *For the differential inclusion $\dot{x} \in F(x)$, the closed set $\mathcal{A} \subset \mathbb{R}^n$ is said to be* weakly uniformly globally asymptotically stable (weakly UGAS) *if there exists $\beta \in \mathcal{KL}$ such that, for each $x \in \mathbb{R}^n$, there exists a solution $\phi \in \mathcal{S}[0, \infty)(x)$ satisfying $|\phi(t, x)|_{\mathcal{A}} \le \beta(|x|_{\mathcal{A}}, t)$ for all $t \ge 0$.*

We will make an assumption that follows [1, Definition 1.5]. This rules out finite time trajectory escape unobservable through distance to the set $\mathcal{A}$. Or, intuitively, it prevents the trajectory escaping to infinity in a direction parallel to the set $\mathcal{A}$.

Assumption 1. For each $r > 0$ there exists $M_r > 0$ such that $|x|_{\mathcal{A}} \leq r$ implies $\sup_{w \in F(x)} |w| \leq M_r$.

We are now in a position to assert the existence of a locally Lipschitz weak Lyapunov function.

THEOREM 2.1. *Suppose $F(\cdot)$ satisfies the basic conditions on $\mathbb{R}^n$, is locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$, satisfies Assumption* 1*, and, for $\dot{x} \in F(x)$, the closed set $\mathcal{A}$ is weakly UGAS. Then there exists a (weak Lyapunov) function $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ that is locally Lipschitz on $\mathbb{R}^n$ and $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ such that for all $x \in \mathbb{R}^n$*

$$\alpha_1(|x|_{\mathcal{A}}) \leq V(x) \leq \alpha_2(|x|_{\mathcal{A}}) \quad \text{and} \tag{2.1}$$

$$\min_{w \in F(x)} DV(x; w) \leq -V(x) \ . \tag{2.2}$$

Furthermore, if $F(\cdot)$ is locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$, uniformly in distance to the set $\mathcal{A}$, then there exists a locally Lipschitz weak Lyapunov function for $\dot{x} \in F(x)$ with respect to $\mathcal{A}$ where the local Lipschitz property is uniform in distance to the set $\mathcal{A}$.

The theorem is proved in section 5.

Remark 2. The use of the minimum is justified here, and throughout the paper, in place of an infimum by virtue of the fact that the set over which the infimum is taken is compact and the function is continuous; i.e., $DV(x; \cdot)$ is locally Lipschitz for all $x \in \mathbb{R}^n$ when $V(\cdot)$ is locally Lipschitz (see [8, Exercise 3.4.1a]). ∎

3. A control-Lyapunov function. In this section we state our result that uniform asymptotic controllability to a set implies the existence of a locally Lipschitz control-Lyapunov function. In what follows, we take $\mathcal{U}$ to be a locally compact metric space with a unique zero element, "0," and, by abuse of notation, $|u| := d(u, 0)$. We define the closed unit ball in the metric space $\mathcal{U}$ as $\overline{\mathcal{B}}_{\mathcal{U}} := \{\xi \in \mathcal{U} : d(\xi, 0) \leq 1\}$.

DEFINITION 6. *Let $\mathcal{A} \subset \mathbb{R}^n$ be a closed, nonempty set, and let $\sigma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be nondecreasing. We say that* (1.2) *is* uniformly globally asymptotically controllable (UGAC) to $\mathcal{A}$ with $\mathcal{U} \cap \sigma$ controls *if there exists a function $\beta \in \mathcal{KL}$ such that for each $x \in \mathbb{R}^n$ there exist a measurable, essentially bounded function $u : [0, \infty) \to \mathcal{U}$ and a solution $\phi(\cdot, x, u)$ of $\dot{x} = f(x, u(t))$ satisfying*

$$\begin{aligned} |\phi(t, x, u)|_{\mathcal{A}} &\leq \beta(|x|_{\mathcal{A}}, t), \\ |u(t)| &\leq \sigma(|\phi(t, x, u)|_{\mathcal{A}}) \quad \text{for almost all } t \geq 0. \end{aligned} \tag{3.1}$$

Remark 3. We note that $\mathcal{U} \cap \sigma$ is an abuse of notation. It is shorthand for allowing controls from $\mathcal{U} \cap \sigma(|x|_{\mathcal{A}})\overline{\mathcal{B}}_{\mathcal{U}} = \{u \in \mathcal{U} : |u| \leq \sigma(|x|_{\mathcal{A}})\}$ for each $x \in \mathbb{R}^n$. ∎

Note that the usual definition of UGAC (such as in [24, Definition 2.2] or [27]) limits the control based on the size of the initial condition of the state, whereas for UGAC with $\mathcal{U} \cap \sigma$ controls we limit the control through the size of the trajectory. The following lemma is proved in [13, section 5.3.6].

LEMMA 3.1. *The system* (1.2) *is UGAC to $\mathcal{A}$ with $\mathcal{U} \cap \sigma$ controls if and only if there exist $\beta_c \in \mathcal{KL}$ and $\sigma_c : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ nondecreasing such that for each $x \in \mathbb{R}^n$ there exist a measurable, essentially bounded function $u : [0, \infty) \to \mathcal{U}$ and a maximal solution $\phi(t, x, u(t))$ of* (1.2) *such that $\|u\|_\infty \leq \sigma_c(|x|_{\mathcal{A}})$ and $|\phi(t, x, u(t))|_{\mathcal{A}} \leq \beta_c(|x|_{\mathcal{A}}, t)$.*

DEFINITION 7. *Let $\sigma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be nondecreasing. We say a locally Lipschitz function $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is a* control-Lyapunov function with $\mathcal{U} \cap \sigma$ controls *for the*

system (1.2) *if there exist* $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ *such that* $\alpha_1(|x|_{\mathcal{A}}) \leq V(x) \leq \alpha_2(|x|_{\mathcal{A}})$, *and* $V(\cdot)$ *satisfies the weak infinitesimal decrease property*

$$\min_{w \in \overline{\text{co}} f(x, \mathcal{U} \cap \sigma(|x|_{\mathcal{A}})\overline{\mathcal{B}}_{\mathcal{U}})} DV(x; w) \leq -V(x) \ \forall x \in \mathbb{R}^n \ .$$

Our result will require the following technical assumption which parallels Assumption 1 and which is essentially [1, Definition 1.5]. Again, this rules out finite time escape of trajectories which is not observable through distance to the set $\mathcal{A}$.

Assumption 2. For all $r_1, r_2 \in \mathbb{R}_{>0}$, there exists $M_{r_1,r_2} > 0$ such that

$$\sup_{\{|x|_{\mathcal{A}} \leq r_1, |u| \leq \sigma(r_2)\}} |f(x,u)| \leq M_{r_1,r_2}.$$

With all the necessary definitions in hand, we state the following theorem.

THEOREM 3.2. *Suppose* (1.2) *satisfies Assumption* 2 *and is UGAC to the set* $\mathcal{A}$ *with* $\mathcal{U} \cap \sigma$ *controls. Furthermore, assume that the set-valued map*

$$F(x) := \overline{\text{co}} f\left(x, \mathcal{U} \cap \sigma(|x|_{\mathcal{A}})\overline{\mathcal{B}}_{\mathcal{U}}\right)$$

satisfies the basic conditions on $\mathbb{R}^n$ *and is locally Lipschitz on* $\mathbb{R}^n \backslash \mathcal{A}$. *Then there exists a locally Lipschitz control-Lyapunov function with* $\mathcal{U} \cap \sigma$ *controls for* (1.2).

Furthermore, if $F(\cdot)$ *is locally Lipschitz on* $\mathbb{R}^n \backslash \mathcal{A}$, *uniformly in distance to the set* $\mathcal{A}$, *then there exists a locally Lipschitz control-Lyapunov function with* $\mathcal{U} \cap \sigma$ *controls which is locally Lipschitz, uniformly in distance to the set* $\mathcal{A}$.

Remark 4. Two examples of regularity conditions on $f(\cdot, \cdot)$ which would give rise to a locally Lipschitz $F(\cdot)$ are as follows.

1. Let $c > 0$ be constant (possibly $+\infty$) and $\sigma(\cdot) \equiv c$. Furthermore, let $f(x, \cdot)$ be measurable for each $x \in \mathbb{R}^n$ and $f(\cdot, u)$ be locally Lipschitz uniformly in $u \in \mathcal{U} \cap c\overline{\mathcal{B}}_{\mathcal{U}}$. Then $F(\cdot)$ is locally Lipschitz.
2. Consider $\mathcal{U} = \mathbb{R}^m$, and let $\sigma(\cdot)$ be locally Lipschitz (and nondecreasing) and $f(\cdot, \cdot)$ be locally Lipschitz. Then $F(\cdot)$ is locally Lipschitz.

Note that we have not assumed that the set-valued map $f\left(x, \mathcal{U} \cap \sigma(|x|_{\mathcal{A}})\overline{\mathcal{B}}_{\mathcal{U}}\right)$ is convex. However, if $f\left(x, \mathcal{U} \cap \sigma(|x|_{\mathcal{A}})\overline{\mathcal{B}}_{\mathcal{U}}\right)$ is Lipschitz, then $\overline{\text{co}} f\left(x, \mathcal{U} \cap \sigma(|x|_{\mathcal{A}})\overline{\mathcal{B}}_{\mathcal{U}}\right)$ is also Lipschitz (see [3, section 1.1, Proposition 6]).

These examples extend easily to the case of generating a set-valued map $F(\cdot)$ which is locally Lipschitz, uniformly in distance to the set $\mathcal{A}$. This is done by requiring the corresponding Lipschitz property on f to be uniform in distance to the set $\mathcal{A}$. □

The proof follows by noting that the set $\mathcal{A}$ will be weakly UGAS for the set-valued map $F(\cdot)$, allowing the application of Theorem 2.1. See [15] and [13] for details.

4. Control construction. By making use of the control-Lyapunov function of Theorem 3.2, we can construct a (discontinuous) time-invariant feedback stabilizer that, when implemented with a sample-and-hold strategy, guarantees semiglobal practical asymptotic stability of the set $\mathcal{A}$ and robustness to small additive disturbances and measurement noise. By sample and hold we mean that the system state is "sampled," a control action is computed, and then it is implemented (or "held") for a fixed holding period. The procedure is then repeated.

Results of this type were presented by Clarke et al. [6] and Sontag [28] for the case where $\mathcal{A}$ is compact, and a construction is given by Clarke et al. [5] that applies to the case of noncompact sets $\mathcal{A}$. Our construction resembles the constructions used

in both of these references. In comparison to the construction in [5], we use proximal aiming to a point that minimizes the control-Lyapunov function in a ball around the current point rather than proximal aiming to a sublevel set of the control-Lyapunov function. This permits a very concise statement of the control synthesis algorithm.

The following assumptions, under which we construct our feedback law, all follow directly from Theorem 3.2. However, these assumptions are somewhat weaker as they simplify the exposition. Specifically, as we use a sample-and-hold strategy to implement our feedback control, we are concerned with only the semiglobal practical qualities of the control-Lyapunov function.

4.1. Assumptions. For $V : \mathbb{R}^n \to \mathbb{R}$ and $\ell_1, \ell_2 \in \{-\infty\} \bigcup \mathbb{R}$ such that $\ell_1 < \ell_2$, we define $\mathcal{V}(\ell_1, \ell_2) := \{x \in \mathbb{R}^n : \ell_1 \leq V(x) \leq \ell_2\}$. We denote $\mathcal{V}(-\infty, \ell_2)$ by $\mathcal{V}(\ell_2)$. Suppose $\sigma(\cdot)$ is nondecreasing and we are given $\ell_1 < \ell_2$, $\varepsilon_2 > 0$, $\varepsilon_3 > 0$, $\varepsilon_4 > 0$, $c > 0$, $L_V > 0$, $L_f > 0$, $\widetilde{M} > 0$ such that

1. for $x_1, x_2 \in \mathcal{V}(\ell_1, \ell_2 + \varepsilon_2) + \varepsilon_3 \overline{\mathcal{B}}_n$ and $u \in \mathcal{U} \cap \sigma(\max\{|x_1|_{\mathcal{A}}, |x_2|_{\mathcal{A}}\} + \varepsilon_4)\overline{\mathcal{B}}_{\mathcal{U}}$
   (a) $|V(x_1) - V(x_2)| \leq L_V |x_1 - x_2|$,
   (b) $|f(x_1, u) - f(x_2, u)| \leq L_f |x_1 - x_2|$,
   (c) $\min_{w \in \overline{\text{co}} f(x, \mathcal{U} \cap \sigma(|x|_{\mathcal{A}})\overline{\mathcal{B}}_{\mathcal{U}})} DV(x; w) \leq -2c$;
2. $f(\cdot, u)$ is continuous and bounded in norm by $\widetilde{M}$ on $\mathcal{V}(\ell_2 + \varepsilon_2) + \varepsilon_3 \overline{\mathcal{B}}_n$ for all $u \in \mathcal{U} \cap \sigma(|\cdot|_{\mathcal{A}} + \varepsilon_4)\overline{\mathcal{B}}_{\mathcal{U}}$.

Note again that, with appropriate values for the constants, these assumptions are all satisfied by the control-Lyapunov function of Theorem 3.2.

4.2. Control design. For the control system

$$\dot{x} = f(x, u) + d, \quad u \in \mathcal{U} \cap \sigma(|x|_{\mathcal{A}} + \varepsilon_4)\overline{\mathcal{B}}_{\mathcal{U}}, \tag{4.1}$$

we define a (discontinuous) control law as follows.

1. Let $r \in (0, \min\{\frac{\varepsilon_2}{L_V}, \varepsilon_3, \varepsilon_4, \frac{c}{L_f L_V}\}]$.
2. For each $x \in \mathcal{V}(\ell_1, \ell_2 + \varepsilon_2)$,
   (a) let $s \in \overline{\mathcal{B}}_n(x, r)$ be such that $V(s) \leq V(\xi)$ for all $\xi \in \overline{\mathcal{B}}_n(x, r)$;
   (b) let $\alpha \in \mathcal{U} \cap \sigma(|x|_{\mathcal{A}} + r)\overline{\mathcal{B}}_{\mathcal{U}}$ be such that $\langle x - s, f(x, \alpha)\rangle \leq -\frac{c}{L_V}|x - s|$.
3. For any $x \notin \mathcal{V}(\ell_1, \ell_2 + \varepsilon_2)$ let $\alpha \in \mathcal{U} \cap \sigma(|x|_{\mathcal{A}})\overline{\mathcal{B}}_{\mathcal{U}}$ be arbitrary.
4. Take $u = \alpha(x)$.

4.3. Closed-loop results. We let $T_1 > 0$ be such that $\frac{cT_1}{16L_V} \leq r - \sqrt{r^2 - \frac{rcT_1}{4L_V}}$. Such a value exists since the derivative with respect to T_1 of the function on the right-hand side evaluated at $T_1 = 0$ is equal to $\frac{c}{8L_V}$. We define $M := \widetilde{M} + \frac{c}{2L_V}$, $a_1 := M^2 L_f$, $a_2 := M(M + rL_f)$, $a_3 := \frac{cr}{4L_V}$, and

$$T^* := \min\left\{T_1, \ \frac{\ell_2 - \ell_1}{L_V M}, \ \frac{\varepsilon_3}{M}, \frac{\sqrt{a_2^2 + 4a_1 a_3} - a_2}{2a_1}\right\}.$$

We note that $T^* > 0$ and $T^* \to 0$ as $r \to 0$. This is evident from the last term that defines T^*.

THEOREM 4.1. *Suppose $u = \alpha(x)$ is implemented by sampling and holding with holding period $T \in (0, T^*]$. Then for every $x_0 \in \mathcal{V}(\ell_2)$, for all $d(\cdot)$ such that $||d||_\infty \leq \frac{c}{2L_V}$, and for all $t \geq 0$, the resulting solutions satisfy*

$$V(x(t)) \leq \max\left\{V(x_0) - \frac{c^2 \max\{t - T, 0\}}{8L_V M}, \ell_1 + L_V M T\right\} + L_V r.$$

An outline of the proof of this theorem may be found in [15] with full details to be found in [13].

4.4. Robustness to measurement noise. In this section we will demonstrate that our control design is robust with respect to small measurement errors. That is, if we implement our control using a corrupted measurement $x + n$ rather than with the true state x, the trajectory of the controlled system will still approach the set $\mathcal{A}$. Similar results are established by Sontag [28] and Clarke et al. [5]. The important observation here is that, while the measurement noise may be persistent, nondifferentiable, and unknown, since we implement the control via a sample-and-hold procedure, it is only the noise values at the sampling instants that are important.

Consider the system

$$\dot{x} = f(x, \alpha(x_i + n_i)),$$

where n_i represents samples of a bounded noise function $n : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$. We construct a fake noise function $n_L(\cdot)$ that is globally Lipschitz and matches $n(\cdot)$ at sampling instances. If N is a bound for $|n(\cdot)|$ and T is the sampling period, then $n_L(\cdot)$ can be constructed so that it is bounded by N and its Lipschitz constant is $2N/T$. Note that such an approximation always exists. For instance, taking a linear interpolation between each noise sample yields such a function. Also note that, in what follows, precise knowledge of this signal is unnecessary. That is, we have not assumed that we know this function, only that it is bounded and we know its Lipschitz constant. We perform a coordinate change, $z = x + n_L$, in order to write

$$\dot{z} = f(z - n_L, \alpha(z_i)) + \dot{n}_L \ .$$

We rewrite the system as

$$\dot{z} = f(z - n_L, \alpha(z_i)) + f(z, \alpha(z_i)) - f(z, \alpha(z_i)) + \dot{n}_L = f(z, \alpha(z_i)) + d \ ,$$

where we have defined $d := f(z - n_L, \alpha(z_i)) - f(z, \alpha(z_i)) + \dot{n}_L$. Utilizing the Lipschitz constant for $f(\cdot, u)$ and the bound on $|\dot{n}_L|$ we have that $|d| \leq N(L_f + 2/T)$. Therefore the result of Theorem 4.1 applies if we insist that $N \leq \frac{Tc}{2L_V(2+L_fT)}$. That is, selecting the noise bound N appropriately yields $|d| \leq c/2L_V$, which allows us to appeal to the result of the theorem.

Remark 5. It is well known that fast sampling is advantageous for stability properties. Specifically, fast sampling is needed for large states to guarantee stability and is desirable for small states to decrease the size of the "practical stability" region. However, as shown by the bound required on the noise, as the sampling time T becomes small, the allowable noise also becomes small. Therefore, one wants to sample fast but not too fast. This observation was also made by Sontag [28, Theorem 1]. However, the observation is worth repeating as the result is made transparent via the above coordinate change. □

5. Proof of weak converse result. We turn now to the proof of Theorem 2.1. We prove the result without the assumption that $F(\cdot)$ is locally Lipschitz, uniform in distance to the set $\mathcal{A}$. It is easy to see that the Lyapunov function then inherits this property.

5.1. Technical preliminaries. In what follows we will appeal to the following lemma to demonstrate the local Lipschitz property of our Lyapunov function by demonstrating an appropriate bound on the Dini subderivate (see [9, Corollary 3.7]).

LEMMA 5.1. *Let the function $V : \mathcal{O} \to (-\infty, \infty]$ be lower semicontinuous. Let $\mathcal{U} \subset \mathcal{O}$ be open and convex. The function $V(\cdot)$ is Lipschitz with Lipschitz constant M on $\mathcal{U}$ if and only if $DV(x; v) \leq M|v|$, for all $x \in \mathcal{U}$, and all $v \in \mathbb{R}^n$.*

The following lemma is [26, Corollary 10].

LEMMA 5.2. *For each $\omega_0 \in \mathcal{K}_\infty$ there exist $\omega_1, \omega_2 \in \mathcal{K}_\infty$ such that $\omega_0(rs) \leq \omega_1(r)\omega_2(s)$ for all $r, s \geq 0$.*

Next, we state some lemmas that are proved in section 6. We start with a slight refinement of Sontag's lemma on $\mathcal{KL}$-estimates wherein we specify the required regularity property of one of the $\mathcal{K}_\infty$ functions [26, Proposition 7].

LEMMA 5.3. *For each $\beta \in \mathcal{KL}$ and $\lambda > 0$, there exist $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ such that $\alpha_1(\cdot)$ is Lipschitz on its domain, continuously differentiable (C^1) on $(0, \infty)$, $\alpha_1(s) \leq s\alpha_1'(s)$ for all $s > 0$, and $\alpha_1(\beta(s, t)) \leq \alpha_2(s)e^{-\lambda t}$ for all $(s, t) \in \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$.*

The next lemma comes from [19, Lemmas 11 and 12].

LEMMA 5.4. *For each continuous, positive definite function $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ there exists $\rho \in \mathcal{K}_\infty$ such that $\rho(\cdot)$ is locally Lipschitz on its domain, continuously differentiable on $(0, \infty)$, and $\rho(s) \leq \alpha(s)\rho'(s)$ for all $s > 0$.*

LEMMA 5.5. *For each $\omega_2 \in \mathcal{K}_\infty$ there exist a locally Lipschitz, strictly increasing, unbounded function $\kappa : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 1}$ and a continuous, nonincreasing function $\vartheta : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ such that, with $s_0 := \omega_2^{-1}(0.5)$,*

$$\kappa(0) = 1, \tag{5.1}$$

$$\kappa(t)\kappa(T) \geq \kappa(t + T) \qquad \forall t, T \geq 0\ , \tag{5.2}$$

$$\frac{\kappa(t)}{\kappa(T)} \leq e^{2(t-T)} \qquad \forall t \geq T \geq 0\ , \tag{5.3}$$

$$\kappa(t) \leq \min\left\{e^t, \frac{1}{\omega_2(s_0 e^{-t})}\right\}\ \forall t \geq 0, \tag{5.4}$$

$$\max_{s\in[0,T]} \frac{\kappa(s)}{\kappa(s + t)} \leq 1 - \vartheta(T)t \quad \forall t \in [0, 1]\ . \tag{5.5}$$

The next fact follows [18, Lemma 4.3].

LEMMA 5.6. *Let $\mathcal{A} \subset \mathbb{R}^n$ be a closed set. Suppose $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$, continuous on $\mathbb{R}^n$, $V(x) = 0$ for all $x \in \mathcal{A}$, and $V(x) > 0$ for all $x \in \mathbb{R}^n \backslash \mathcal{A}$. Then there exists a function $\rho \in \mathcal{K}_\infty$ that is C^1 on $(0, \infty)$, satisfies $\rho(s) \leq s\rho'(s)$ for all $s > 0$, and is such that $V_L := \rho \circ V$ is locally Lipschitz on $\mathbb{R}^n$ and $DV_L(x; v) = 0$ for all $x \in \mathcal{A}$ and all $v \in \mathbb{R}^n$.*

The next lemma is used twice in the proof of our main result.

LEMMA 5.7. *Suppose we are given the following:*

1. *A function $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ defined as*

$$V(x) := \inf_{\phi\in\mathcal{S}[0,\infty)(x)} \sup_{t\geq 0} g(\phi(t, x))\kappa(t),$$

where $g(\cdot)$ is locally Lipschitz on $\mathbb{R}^n$ and $\kappa : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 1}$ is locally Lipschitz, strictly increasing, unbounded, $\kappa(t)\kappa(T) \geq \kappa(t + T)$ for all $t, T \geq 0$, and $\kappa(0) = 1$.

2. *$\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ such that $\alpha_1(|x|_\mathcal{A}) \leq g(x)$, $V(x) \leq \alpha_2(|x|_\mathcal{A})$.*

3. *A sequence $\{\ell_j\}_{j=1}^\infty$, where $\ell_j \to 0$ as $j \to \infty$ such that $|x|_\mathcal{A} = \ell_j$, implies $V(x) = g(x)$. Then V is continuous on $\mathbb{R}^n$, locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$, and there exists a function $T : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ that is continuous on $\mathbb{R}_{>0}$ and, for each $x \in \mathbb{R}^n$,*

there exists $\phi^* \in \mathcal{S}[0,\infty)(x)$ *such that*

$$V(x) \;=\; \sup_{t\geq 0} g(\phi^*(t,x))\kappa(t) \;=\; \max_{t\in[0,T(|x|_{\mathcal{A}})]} g(\phi^*(t,x))\kappa(t)\ .$$

5.2. Construction and Lipschitz property of $V(x)$. The following proposition is the key to constructing our Lyapunov function and is proved in section 5.5.

PROPOSITION 5.8. *Under the assumptions of Theorem* 2.1*, there exists a family of locally Lipschitz functions* $W_i : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$*,* $i \in \mathbb{Z}_{>0}$*, that have the following properties.*

1. *There exist* $\tilde{\alpha}_1, \alpha \in \mathcal{K}_\infty$ *such that, for each* $x \in \mathbb{R}^n$*,* $i \in \mathbb{Z}_{>0}$*,*

$$\tilde{\alpha}_1(|x|_{\mathcal{A}}) \leq W_i(x) \leq \alpha(|x|_{\mathcal{A}}). \tag{5.6}$$

2. *There exists* $\tilde{\alpha}_2 \in \mathcal{K}_\infty$ *such that, for every* $x \in \mathbb{R}^n$*, there exists* $\hat{\phi} \in \mathcal{S}[0,\infty)(x)$ *such that, for all* $i \in \mathbb{Z}_{>0}$*,*

$$W_i(\hat{\phi}(t,x)) \leq \tilde{\alpha}_2(|x|_{\mathcal{A}})e^{-2t}\ \forall t \geq 0. \tag{5.7}$$

3. *For every* $i \in \mathbb{Z}_{>0}$ *and* $|x|_{\mathcal{A}} \geq \frac{1}{i}$*, there exists* $\phi_i \in \mathcal{S}[0,\infty)(x)$ *such that the set* $\left\{t : |\phi_i(t,x)|_{\mathcal{A}} = \frac{1}{i}\right\}$ *is nonempty and, with* $T := \inf\left\{t : |\phi_i(t,x)|_{\mathcal{A}} = \frac{1}{i}\right\}$*,*

$$W_i(\phi_i(t,x)) \leq W_i(x)e^{-2t} \quad \forall t \in [0,T]\ . \tag{5.8}$$

Remark 6. Note that the functions $W_i(\cdot)$ are locally Lipschitz weak Lyapunov functions on sets defined by $|x|_{\mathcal{A}} \geq \frac{1}{i}$. This follows from items 1 and 3 above. ∎

Let $\tilde{\alpha}_1, \tilde{\alpha}_2 \in \mathcal{K}_\infty$ come from Proposition 5.8, and let $\omega_1, \omega_2 \in \mathcal{K}_\infty$ come from Lemma 5.2 satisfying

$$\tilde{\alpha}_2 \circ \tilde{\alpha}_1^{-1}(rs) \leq \omega_1(r)\omega_2(s) \quad \forall r,s \geq 0\ . \tag{5.9}$$

Define $s_0 := \omega_2^{-1}(0.5)$, and let $\kappa(\cdot)$ and $\vartheta(\cdot)$ come from Lemma 5.5 satisfying (5.1)–(5.5). Let $\widetilde{\omega} \in \mathcal{K}_\infty$ satisfy

$$\widetilde{\omega}(s) \leq s \qquad \forall s \geq 0 \text{ and} \tag{5.10}$$

$$\widetilde{\omega}(s) \leq s_0^2 \frac{\left(\omega_1^{-1}(s)\right)^2}{s} \quad \forall s \in (0,1]\ . \tag{5.11}$$

Let $\alpha \in \mathcal{K}_\infty$ also come from Proposition 5.8, and define

$$\alpha_f := \tilde{\alpha}_2^{-1} \circ \widetilde{\omega} \circ \tilde{\alpha}_1 \circ \alpha^{-1} \circ \tilde{\alpha}_1\ . \tag{5.12}$$

The function $\alpha_f(\cdot)$ belongs to class-$\mathcal{K}_\infty$. Also

$$\alpha_f(s) \leq s \qquad \forall s \geq 0 \tag{5.13}$$

since, from (5.6), $\alpha^{-1} \circ \tilde{\alpha}_1(s) \leq s$, and from (5.6), (5.7) with $t = 0$, and (5.10), $\tilde{\alpha}_2^{-1} \circ \widetilde{\omega} \circ \tilde{\alpha}_1(s) \leq s$. Now choose a function $q : \mathbb{Z}_{\geq 0} \to \mathbb{Z}_{>0}$ satisfying

$$q(0) = 1\ , \quad q(j+1) \geq \frac{1}{\alpha_f(\frac{1}{q(j)+1})}\ ,\ j \in \mathbb{Z}_{\geq 0}\ . \tag{5.14}$$

It follows from (5.13) that $\frac{1}{q(j)+1} \geq \frac{1}{q(j+1)}$; i.e., $q(j)+1 \leq q(j+1)$. Therefore, the sequence of integers $q(j)$ is strictly increasing with j. Let the locally Lipschitz functions $\lambda_i : \mathbb{R}^n \to [0,1]$ $(i \in \mathbb{Z}_{>0})$ be such that

$$\frac{1}{q(j+1)} \leq |x|_{\mathcal{A}} \leq \frac{1}{q(j)+1} \Longrightarrow \lambda_{q(j+1)}(x) = 1 \ , \tag{5.15}$$

and for each $x \in \mathbb{R}^n$ there exists a finite index set $\mathcal{I}_x$ such that $\sum_i \lambda_i(x) = \sum_{i \in \mathcal{I}_x} \lambda_i(x) = 1$. Let the functions $W_i(\cdot)$ come from Proposition 5.8, and define

$$f(x) = \sum_i \lambda_i(x) W_i(x) \ . \tag{5.16}$$

Note that, on the strips defined by (5.15), this corresponds to $f(x) = W_{q(j+1)}(x)$ and a convex combination of the $W_i(\cdot)$ functions between these strips. It follows from the properties of $W_i(\cdot)$, given in Proposition 5.8, and $\lambda_i(\cdot)$ that $f(\cdot)$ is locally Lipschitz,

$$\tilde{\alpha}_1(|x|_{\mathcal{A}}) \leq f(x) \leq \alpha(|x|_{\mathcal{A}}) \quad \forall x \in \mathbb{R}^n, \tag{5.17}$$

and, for each $x \in \mathbb{R}^n$, there exists $\hat{\phi} \in \mathcal{S}[0,\infty)(x)$ such that

$$f(\hat{\phi}(t,x)) \leq \tilde{\alpha}_2(|x|_{\mathcal{A}}) e^{-2t} \quad \forall t \geq 0 \ . \tag{5.18}$$

Next define

$$V_1(x) := \inf_{\phi \in \mathcal{S}[0,\infty)(x)} \sup_{t \geq 0} f(\phi(t,x)) \kappa(t) \tag{5.19}$$

and note that, using (5.1) and (5.17), we have

$$V_1(x) \geq f(x)\kappa(0) \geq \tilde{\alpha}_1(|x|_{\mathcal{A}}) \tag{5.20}$$

and, using (5.4) and (5.18), we have

$$V_1(x) \leq \sup_{t \geq 0} f(\hat{\phi}(t,x))\kappa(t) \leq \tilde{\alpha}_2(|x|_{\mathcal{A}}) \ . \tag{5.21}$$

The following claim is proved at the end of this section.

Claim 1. There exists a decreasing sequence $\{\ell_j\}_{j=1}^{\infty}$ such that $\ell_j \to 0$ as $j \to \infty$ and $|x|_{\mathcal{A}} = \ell_j$ implies $V_1(x) = f(x)$.

Since (5.1) and (5.2) hold, we can apply Lemma 5.7 with $V(x) := V_1(x)$, $g(x) := f(x)$, $\alpha_1(s) := \tilde{\alpha}_1(s)$, $\alpha_2(s) := \tilde{\alpha}_2(s)$, and the sequence $\{\ell_j\}_{j=1}^{\infty}$ constructed in Claim 1. So $V_1(\cdot)$ is continuous on $\mathbb{R}^n$, locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$, and there exists a function $T : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ that is continuous on $\mathbb{R}_{>0}$ and for each $x \in \mathbb{R}^n$ there exists $\phi^* \in \mathcal{S}[0,\infty)(x)$ such that

$$V_1(x) = \max_{t \in [0, T(|x|_{\mathcal{A}})]} f(\phi^*(t,x))\kappa(t) \ . \tag{5.22}$$

We define $\alpha_3 : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ by $\alpha_3(0) = 0$ and

$$\alpha_3(s) := \min_{r \in [\tilde{\alpha}_2^{-1}(s), \tilde{\alpha}_1^{-1}(s)]} \vartheta(T(r)+1)s \ . \tag{5.23}$$

Since $\vartheta(\cdot)$ is continuous and positive, $T(\cdot)$ is continuous, and $\tilde{\alpha}_1^{-1}(\cdot)$ and $\tilde{\alpha}_2^{-1}(\cdot)$ are continuous, $\alpha_3(\cdot)$ is continuous. It is also positive definite. We apply Lemma 5.4 with $\alpha_3(\cdot)$ to get $\rho_1 \in \mathcal{K}_\infty$, locally Lipschitz on its domain, and C^1 on $(0,\infty)$ such that $\rho_1(s) \leq \alpha_3(s)\rho_1'(s)$ for all $s > 0$. We define $V_2 := \rho_1 \circ V_1$, and we note that $V_2(\cdot)$ is continuous on $\mathbb{R}^n$, locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$, and from (5.20) and (5.21) we have

$$\rho_1 \circ \tilde{\alpha}_1(|x|_{\mathcal{A}}) \leq V_2(x) \leq \rho_1 \circ \tilde{\alpha}_2(|x|_{\mathcal{A}}) \ . \tag{5.24}$$

It follows that we can apply Lemma 5.6 with $V_2(\cdot)$ to get $\rho_2 \in \mathcal{K}_\infty$ that is C^1 on $(0,\infty)$ such that $\rho(s) \leq s\rho'(s)$ for all $s > 0$ and $V := \rho_2 \circ V_2$ is locally Lipschitz on $\mathbb{R}^n$ and satisfies $DV(x;v) = 0$ for all $x \in \mathcal{A}$ and all $v \in \mathbb{R}^n$. Moreover, it follows from (5.24) that (2.1) holds with the class-$\mathcal{K}_\infty$ functions $\alpha_1 := \rho_2 \circ \rho_1 \circ \tilde{\alpha}_1$ and $\alpha_2 := \rho_2 \circ \rho_1 \circ \tilde{\alpha}_2$.

5.3. Infinitesimal decrease. The infinitesimal decrease condition holds for each $x \in \mathcal{A}$ since $DV(x;v) = 0 = -V(x)$, for any $x \in \mathcal{A}$, and $v \in \mathbb{R}^n$. Let $x \in \mathbb{R}^n \backslash \mathcal{A}$, and let $\phi^* \in \mathcal{S}[0,\infty)(x)$ satisfy (5.22). Define $T := T(|x|_{\mathcal{A}})$, and let $t^* \in (0,1]$ be such that $T(|\phi^*(t,x)|_{\mathcal{A}}) \leq T+1$ for all $t \in [0,t^*]$. With (5.22), for each $t \in [0,t^*]$ let $\psi^* \in \mathcal{S}[0,\infty)(\phi^*(t,x))$ satisfy

$$V_1(\phi^*(t,x)) = \max_{s \in [0, T(|\phi^*(t,x)|_{\mathcal{A}})]} f(\psi^*(s,\phi^*(t,x)))\kappa(s) \ .$$

Then, for each $t \in [0,t^*]$, we have

$$\begin{aligned} V_1(\phi^*(t,x)) &= \max_{s \in [0, T(|\phi^*(t,x)|_{\mathcal{A}})]} f(\psi^*(s,\phi^*(t,x)))\kappa(s) \leq \max_{s \in [0,T+1]} f(\phi^*(s+t,x))\kappa(s) \\ &= \max_{s \in [0,T+1]} f(\phi^*(s+t,x))\kappa(s+t)\frac{\kappa(s)}{\kappa(s+t)} \\ &\leq \sup_{\tau \geq 0} f(\phi^*(\tau,x))\kappa(\tau)\,[1-\vartheta(T+1)t] = V_1(x)\,[1-\vartheta(T+1)t]\,. \end{aligned} \tag{5.25}$$

For almost all s, let $g(s)$ be the unique closest point in $F(x)$ to $\dot{\phi^*(s,x)}$. Such a unique closest point exists since $F(x)$ is convex (see [11, section 5, Lemma 2]). Since $F(\cdot)$ is locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$, $x \in \mathbb{R}^n \backslash \mathcal{A}$, $\phi^*(\cdot,x)$ is (absolutely) continuous, $\dot{\phi^*(s,x)} \in F(\phi^*(s,x))$ for almost all s, and $F(\cdot)$ is locally bounded, there exist $\bar{s}, M, L > 0$ such that, for almost all $s \in [0,\bar{s}]$,

$$|g(s) - \dot{\phi^*(s,x)}| \leq L|x - \phi^*(s,x)| \leq LMs \ . \tag{5.26}$$

We can also write

$$\frac{\phi^*(t,x)-x}{t} = t^{-1}\int_0^t \dot{\phi^*(s,x)}ds = t^{-1}\int_0^t g(s)ds + t^{-1}\int_0^t [\dot{\phi^*(s,x)} - g(s)]ds \ . \tag{5.27}$$

Then note that the first term on the right-hand side belongs to $F(x)$ for all t since $F(x)$ is convex (see [11, section 5, Lemma 12]) and the second term on the right-hand side converges to zero as t converges to zero (using (5.26)).

Call the first term on the last line of (5.27) $v(t)$ and the sum $w(t)$. Since $F(x)$ is compact, there exists an accumulation point $v \in F(x)$ for $v(t)$, i.e., a sequence of

times t_j converging to zero so that $v(t_j) \to v$. So, we can write $\phi^*(t_j, x) = x + t_j w(t_j)$, where $w(t_j) \to v$ as $t_j \to 0$. Then, using (5.23) and (5.25),

$$\begin{aligned} DV_1(x;v) &\leq \liminf_{j\to\infty} [V_1(x + t_j w(t_j)) - V_1(x)]/t_j = \liminf_{j\to\infty} [V_1(\phi^*(t_j, x)) - V_1(x)]/t_j \\ &\leq -V_1(x)\vartheta(T(|x|_{\mathcal{A}}) + 1) \leq -\alpha_3(V_1(x)). \end{aligned}$$

Since $V_2 = \rho_1 \circ V_1$ and $\rho_1(s) \leq \alpha_3(s)\rho_1'(s)$ for all $s > 0$, we have, for all $x \in \mathbb{R}^n \backslash \mathcal{A}$,

$$\min_{w\in F(x)} DV_2(x;w) \leq -\rho_1'(V_1(x))\alpha_3(V_1(x)) \leq -\rho_1(V_1(x)) = -V_2(x).$$

Similarly, since $V = \rho_2 \circ V_2$ and $\rho_2(s) \leq s\rho_2'(s)$ for all $s > 0$, we have, for all $x \in \mathbb{R}^n \backslash \mathcal{A}$,

$$\min_{w\in F(x)} DV(x;w) \leq -\rho_2'(V_2(x))V_2(x) \leq -\rho_2(V_2(x)) = -V(x).$$

5.4. Proof of Claim 1. For each $j \in \mathbb{Z}_{\geq 0}$ define

$$\ell_j := \tilde{\alpha}_1^{-1} \circ \widetilde{\omega}^{-1} \circ \tilde{\alpha}_2 \left(\frac{1}{q(j+1)} \right) . \tag{5.28}$$

Since $q(j)$ is strictly increasing and unbounded, it follows that ℓ_j decreases monotonically to zero. We now restrict our attention to integers j sufficiently large so that $\alpha(\ell_j) \leq 1$. We consider $x \in \mathbb{R}^n \backslash \mathcal{A}$ such that $|x|_{\mathcal{A}} = \ell_j$, and we will show that $V_1(x) = f(x)$. Clearly, $V_1(x) \geq f(x)$. It remains to show that $V_1(x) \leq f(x)$. First, it follows from (5.17) that $0 < f(x) \leq 1$. Second, it follows from (5.17), (5.18) with $t = 0$, and (5.28) that

$$|z|_{\mathcal{A}} = \frac{1}{q(j+1)} \quad \Longrightarrow \quad f(z) \leq \widetilde{\omega}(f(x)) . \tag{5.29}$$

Third, it follows from (5.11) and $0 < f(x) \leq 1$ that

$$\widetilde{\omega}(f(x)) \leq s_0^2 \frac{\omega_1^{-1}(f(x))^2}{f(x)} . \tag{5.30}$$

Fourth, since $\tilde{\alpha}_1^{-1} \circ \widetilde{\omega}^{-1} \circ \tilde{\alpha}_2(s) \geq s$ for all $s \geq 0$, it follows from (5.28) that $|x|_{\mathcal{A}} \geq \frac{1}{q(j+1)}$. From property 3 of Proposition 5.8, there exists a trajectory $\phi_{q(j+1)} \in \mathcal{S}[0,\infty)(x)$ such that the set $\{t : |\phi_{q(j+1)}(t,x)|_{\mathcal{A}} = \frac{1}{q(j+1)}\}$ is nonempty and for all $t \in [0,T]$, where $T := \inf\{t : |\phi_{q(j+1)}(t,x)|_{\mathcal{A}} = \frac{1}{q(j+1)}\}$, we have

$$W_{q(j+1)}(\phi_{q(j+1)}(t,x)) \leq W_{q(j+1)}(x)e^{-2t}. \tag{5.31}$$

It follows from the definition of T that, for all $t \in [0,T]$,

$$|\phi_{q(j+1)}(t,x))|_{\mathcal{A}} \geq \frac{1}{q(j+1)} . \tag{5.32}$$

With (5.6), $|x|_{\mathcal{A}} = \ell_j$, (5.28), (5.12), and (5.14), one can show that (5.31) implies that, for all $t \in [0,T]$, $|\phi_{q(j+1)}(t,x)|_{\mathcal{A}} \leq \frac{1}{q(j)+1}$. Combining this with (5.32), (5.31), (5.15), and (5.16), we have

$$f(\phi_{q(j+1)}(t,x)) \leq f(x)e^{-2t} \quad \forall t \in [0,T] \tag{5.33}$$

and thus, using (5.4), we have for all $t \in [0,T]$,

$$f(\phi_{q(j+1)}(t,x))\kappa(t) \le f(x)e^{-2t}\kappa(t) \le f(x)\ . \tag{5.34}$$

Now define $z := \phi_{q(j+1)}(T,x)$ and note that

$$|z|_{\mathcal{A}} = \frac{1}{q(j+1)}\ . \tag{5.35}$$

Let $\tilde{\phi} \in \mathcal{S}[0,\infty)(z)$ satisfy (see (5.18))

$$f(\tilde{\phi}(t,z)) \le \tilde{\alpha}_2(|z|_{\mathcal{A}})e^{-2t} \qquad \forall t \ge 0\ . \tag{5.36}$$

Define $\tilde{\psi} \in \mathcal{S}[0,\infty)(x)$ as

$$\tilde{\psi}(t,x) := \begin{cases} \phi_{q(j+1)}(t,x), & t \in [0,T], \\ \tilde{\phi}(t-T,z), & t \ge T. \end{cases}$$

Now, using (5.36), (5.3), (5.17), (5.33), (5.9), (5.4), (5.35), (5.29), and (5.30) one may show that, for $t \ge T$, $f(\tilde{\phi}(t-T,z))\kappa(t) \le f(x)$. Combining this with (5.34), we have $f(\tilde{\psi}(t,x))\kappa(t) \le f(x)$ for all $t \ge 0$, which implies $V_1(x) \le f(x)$. ∎

5.5. Proof of Proposition 5.8. Given $\dot{x} \in F(x)$, where $F(\cdot)$ satisfies the basic conditions and is locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$, let M_1 come from Assumption 1. We define a modified inclusion

$$\dot{x} \in F_i(x) := F(x) + \gamma_i(|x|_{\mathcal{A}})\,[M_1+1]\,\overline{\mathcal{B}}_n, \tag{5.37}$$

where $\gamma_i : [0,\infty) \to [0,1]$ is locally Lipschitz and $\gamma(s) = 0$ for $s \ge \frac{1}{i}$ and $\gamma(s) = 1$ for $s \le \frac{1}{i+1}$. It follows from the properties of $F(\cdot)$ and $\gamma_i(\cdot)$ that $F_i(\cdot)$ satisfies the basic conditions on $\mathbb{R}^n$ and is locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$.

Note that $F(x) \subseteq F_i(x)$ for all $x \in \mathbb{R}^n$ by taking the origin in $\overline{\mathcal{B}}_n$. Furthermore, note that for all

$$|x|_{\mathcal{A}} \le \frac{1}{i+1} \quad \Longrightarrow \quad \overline{\mathcal{B}}_n \subseteq F_i(x). \tag{5.38}$$

To see this, let $x_1 \in \overline{\mathcal{B}}_n$ and $x_2 \in F(x)$, where $|x|_{\mathcal{A}} \le \frac{1}{i+1}$. Therefore, $-x_2 + x_1 \in \left[\max_{z\in F(x)} |z| + 1\right]\overline{\mathcal{B}}_n \subseteq [M_1+1]\,\overline{\mathcal{B}}_n$, which implies that $x_1 = x_2 + (-x_2 + x_1) \in F(x) + [M_1+1]\,\overline{\mathcal{B}}_n = F_i(x)$. We will denote the set of maximal solutions of (5.37) starting at the point x as $\mathcal{S}_i(x)$.

Since $F(x) \subseteq F_i(x)$, the assumption of weak-UGAS of $\mathcal{A}$ for $\dot{x} \in F(x)$ gives $\hat{\phi} \in \mathcal{S}_i[0,\infty)(x)$ such that

$$|\hat{\phi}(t,x)|_{\mathcal{A}} \le \beta(|x|_{\mathcal{A}}, t)\ \forall t \ge 0\ . \tag{5.39}$$

Let $\lambda = 4$, and then let $\widehat{\alpha}_1 \in \mathcal{K}_\infty$ and $\widehat{\alpha} \in \mathcal{K}_\infty$ come from Lemma 5.3 so that $\widehat{\alpha}_1(\cdot)$ is Lipschitz on its domain, C^1 on $(0,\infty)$, $\widehat{\alpha}_1(s) \le s\widehat{\alpha}_1'(s)$ for all $s \in (0,\infty)$, and

$$\widehat{\alpha}_1(\beta(s,t)) \le \widehat{\alpha}(s)e^{-4t} \quad \forall s \ge 0,\ t \ge 0\ . \tag{5.40}$$

We make the following definitions:

- $\widetilde{W}_i(x) := \inf_{\phi \in \mathcal{S}_i[0,\infty)(x)} \sup_{t\geq 0} \widehat{\alpha}_1(|\phi(t,x)|_{\mathcal{A}})e^{2t}$;
- $\beta_2(s,t) := \widehat{\alpha} \circ \widehat{\alpha}_1^{-1}\left(\widehat{\alpha}(s)e^{-4t}\right)$ (note that $\beta_2 \in \mathcal{KL}$);
- applying Lemma 5.3 with $\beta_2 \in \mathcal{KL}$ and $\lambda = 2$, we get $\alpha_m \in \mathcal{K}_\infty$ and $\tilde{\alpha}_2 \in \mathcal{K}_\infty$ such that $\alpha_m(\cdot)$ is Lipschitz on its domain, C^1 on $(0,\infty)$, $\alpha_m(s) \leq s\alpha_m'(s)$ for all $s \in (0,\infty)$, and

$$\alpha_m(\beta_2(s,t)) \leq \tilde{\alpha}_2(s)e^{-2t} \quad \forall s \geq 0,\ t \geq 0; \tag{5.41}$$

- $W_i := \alpha_m \circ \widetilde{W}_i$, $\alpha := \alpha_m \circ \widehat{\alpha}$, $\tilde{\alpha}_1 := \alpha_m \circ \widehat{\alpha}_1$.

5.5.1. Proof of (5.6). Given the bound for $\hat{\phi} \in \mathcal{S}_i[0,\infty)(x)$ in (5.39) and the $\mathcal{KL}$-estimate given by (5.40), we have

$$W_i(x) \leq \alpha_m\left(\sup_{t\geq 0} \widehat{\alpha}_1(\beta(|x|_{\mathcal{A}},t))e^{2t}\right) \leq \alpha_m\left(\sup_{t\geq 0} \widehat{\alpha}(|x|_{\mathcal{A}})e^{-2t}\right) \leq \alpha(|x|_{\mathcal{A}}), \tag{5.42}$$

while the lower bound follows from

$$W_i(x) \geq \alpha_m\left(\inf_{\phi\in\mathcal{S}_i[0,\infty)(x)} \widehat{\alpha}_1(|\phi(t,x)|_{\mathcal{A}})e^{2t}\Big|_{t=0}\right) = \alpha_m(\widehat{\alpha}_1(|x|_{\mathcal{A}})) = \tilde{\alpha}_1(|x|_{\mathcal{A}}). \tag{5.43}$$

5.5.2. Proof of (5.7). From the upper bound on $W_i(x)$, the $\mathcal{KL}$-bound in (5.39), and the $\mathcal{KL}$-estimates in (5.40) and (5.41) we can write, for all $x \in \mathbb{R}^n$ and all $t \geq 0$,

$$\begin{aligned} W_i(\hat{\phi}(t,x)) &\leq \alpha(|\hat{\phi}(t,x)|_{\mathcal{A}}) = \alpha_m \circ \widehat{\alpha}(|\hat{\phi}(t,x)|_{\mathcal{A}}) \leq \alpha_m \circ \widehat{\alpha}(\beta(|x|_{\mathcal{A}},t)) \\ &\leq \alpha_m \circ \widehat{\alpha} \circ \widehat{\alpha}_1^{-1}\left(\widehat{\alpha}(|x|_{\mathcal{A}})e^{-4t}\right) = \alpha_m\left(\beta_2(|x|_{\mathcal{A}},t)\right) \leq \tilde{\alpha}_2(|x|_{\mathcal{A}})e^{-2t}\ . \end{aligned}$$

5.5.3. Proof of (5.8). We make the following claim and defer its proof to the end of this section.

Claim 2. For $|x|_{\mathcal{A}} \leq \frac{1}{i+1}$, we have $\widetilde{W}_i(x) = \widehat{\alpha}_1(|x|_{\mathcal{A}})$.

Construct any decreasing sequence $\{\ell_j\}_{j=1}^\infty$ such that $\ell_j \to 0$ as $j \to \infty$ and $\ell_1 = \frac{1}{i+1}$. Apply Lemma 5.7 with $V(x) := \widetilde{W}_i(x)$, $\alpha_1(s) := \widehat{\alpha}_1(s)$, $\alpha_2(s) := \widehat{\alpha}(s)$, $\kappa(t) = e^{2t}$, and the sequence $\{\ell_j\}_{j=1}^\infty$ constructed above. Therefore, for each $x \in \mathbb{R}^n$, there exists $\tilde{\phi}_i \in \mathcal{S}_i[0,\infty)(x)$ such that $\widetilde{W}_i(x) = \sup_{t\geq 0} \widehat{\alpha}_1(|\tilde{\phi}_i(t,x)|_{\mathcal{A}})e^{2t}$. Then

$$\begin{aligned} \widetilde{W}_i(x)e^{-2t} &= \sup_{\tau\geq 0} \widehat{\alpha}_1(|\tilde{\phi}_i(\tau,x)|_{\mathcal{A}})e^{2\tau}e^{-2t} \geq \sup_{\tau\geq t} \widehat{\alpha}_1(|\tilde{\phi}_i(\tau,x)|_{\mathcal{A}})e^{2(\tau-t)} \\ &= \sup_{s\geq 0} \widehat{\alpha}_1(|\tilde{\phi}_i(s+t,x)|_{\mathcal{A}})e^{2s} \geq \inf_{\phi\in\mathcal{S}_i[0,\infty)(\tilde{\phi}_i(t,x))} \sup_{s\geq 0} \widehat{\alpha}_1(|\phi(s,\tilde{\phi}_i(t,x))|_{\mathcal{A}})e^{2s} \\ &= \widetilde{W}_i(\tilde{\phi}_i(t,x)). \end{aligned} \tag{5.44}$$

If we define $U(t) := \alpha_m(\widetilde{W}_i(x)e^{-2t})$, then $U(\cdot)$ is C^1 on $(0,\infty)$ since $\alpha_m(\cdot)$ is C^1 on $(0,\infty)$, and we can write $\dot{U}(t) = (-2\widetilde{W}_i(x)e^{-2t})\alpha_m'(\widetilde{W}_i(x)e^{-2t}) \leq -2U(t)$, where the inequality follows from the property $\alpha_m(s) \leq s\alpha_m'(s)$ for all $s \in (0,\infty)$. By a standard comparison lemma we obtain $U(t) \leq U(0)e^{-2t}$; i.e., $\alpha_m(\widetilde{W}_i(x)e^{-2t}) \leq \alpha_m(\widetilde{W}_i(x))e^{-2t}$. Combining this result with (5.44) we get

$$\begin{aligned} W_i(\tilde{\phi}_i(t,x)) &= \alpha_m(\widetilde{W}_i(\tilde{\phi}_i(t,x))) \leq \alpha_m\left(\widetilde{W}_i(x)e^{-2t}\right) \leq \alpha_m\left(\widetilde{W}_i(x)\right)e^{-2t} \\ &= W_i(x)e^{-2t}. \end{aligned} \tag{5.45}$$

Fix $i \in \mathbb{Z}_{>0}$, and let $|x|_{\mathcal{A}} \geq \frac{1}{i}$. It follows from (5.45) with (5.6) that the set $\{t : |\tilde{\phi}_i(t,x)|_{\mathcal{A}} = \frac{1}{i}\}$ is nonempty. Moreover, since $F_i(x) = F(x)$ on the set $|x|_{\mathcal{A}} \geq \frac{1}{i}$, it follows, with $T := \inf\{t : |\tilde{\phi}_i(t,x)|_{\mathcal{A}} = \frac{1}{i}\}$, that $\tilde{\phi} \in \mathcal{S}[0,T](x)$. Let $\psi \in \mathcal{S}[0,\infty)(\tilde{\phi}_i(T,x))$ be arbitrary. Define

$$\phi_i(t,x) := \begin{cases} \tilde{\phi}_i(t,x), & t \in [0,T], \\ \psi(t-T, \tilde{\phi}_i(T,x)), & t \geq T\ . \end{cases}$$

Therefore, $\phi_i \in \mathcal{S}[0,\infty)(x)$. Furthermore, for all $t \in [0,T]$, $W_i(\phi_i(t,x)) \leq W_i(x)e^{-2t}$.

5.5.4. Local Lipschitz property of $W_i(x)$. From Lemma 5.7 we know that $\widetilde{W}_i(\cdot)$ is locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$. Consequently, since $\alpha_m(\cdot)$ is locally Lipschitz, $W_i(\cdot)$ is locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$. This, coupled with Claim 2 and the local Lipschitz property of $\widehat{\alpha}_1(\cdot)$ and $\alpha_m(\cdot)$, implies that $W_i(\cdot)$ is locally Lipschitz on $\mathbb{R}^n$.

5.5.5. Proof of Claim 2. Via the same argument as in (5.43), we observe that $\widetilde{W}_i(x) \geq \widehat{\alpha}_1(|x|_{\mathcal{A}})$. So, we just need to show that $\widetilde{W}_i(x) \leq \widehat{\alpha}_1(|x|_{\mathcal{A}})$ for all $|x|_{\mathcal{A}} \leq \frac{1}{i+1}$.

Let $|x|_{\mathcal{A}} \leq \frac{1}{i+1}$, $a \in \mathcal{A}$, be a closest point to x, and define

$$\psi(t,x) := \begin{cases} x + \frac{a-x}{|a-x|}t & \forall t \in [0, |a-x|], \\ a & \forall t \geq |a-x|\ . \end{cases} \tag{5.46}$$

We see that $\psi(\cdot,x)$ is absolutely continuous and

$$\left|\frac{d}{dt}\psi(t,x)\right| = \begin{cases} \left|\frac{a-x}{|a-x|}\right| = 1 \in \overline{\mathcal{B}}_n & \forall t \in (0,|a-x|), \\ 0 \in \overline{\mathcal{B}}_n & \forall t > |a-x|, \end{cases}$$

which, with (5.38), implies that (5.46) is a solution to (5.37); i.e., $\psi \in \mathcal{S}_i[0,\infty)(x)$. For $t \in [0,|a-x|]$ we can write

$$\begin{aligned} |x - \psi(t,x)| &= \left|\tfrac{a-x}{|a-x|}t\right| = t \ \text{ and} \\ |\psi(t,x) - a| &= \left|(x-a)\left(1 - \tfrac{1}{|a-x|}t\right)\right| = |x-a| - t \end{aligned} \tag{5.47}$$

so that $|a-x| = |x-\psi(t,x)| + |\psi(t,x)-a|$. Furthermore, $|x-a| = |x|_{\mathcal{A}}$, and $|x|_{\mathcal{A}} \leq \frac{1}{i+1}$. These facts, together with (5.47), imply that $|\psi(t,x)|_{\mathcal{A}} \leq \frac{1}{i+1}$ for all $t \geq 0$.

Now, for some $t \in (0,|a-x|]$, assume there exists $b \in \mathcal{A}$ such that $|b-\psi(t,x)| < |a-\psi(t,x)|$; i.e., b is closer than a to $\psi(t,x)$. The triangle inequality yields

$$|x-b| \leq |x-\psi(t,x)| + |b-\psi(t,x)| < |x-\psi(t,x)| + |a-\psi(t,x)| = |x-a|\ ,$$

which contradicts that $a \in \mathcal{A}$ is a closest point to x. Therefore, $a \in \mathcal{A}$ is also a closest point to $\psi(t,x)$ for $t \in [0,|a-x|]$. More explicitly, $|\psi(t,x)|_{\mathcal{A}} = |\psi(t,x)-a|$. Combining this with (5.47) yields, for almost all $t \in [0,|a-x|]$,

$$\frac{d}{dt}|\psi(t,x)|_{\mathcal{A}} = \frac{d}{dt}\left(|x-a| - t\right) = -1\ . \tag{5.48}$$

Using (5.47) and (5.48), and since $\widehat{\alpha}_1(\cdot)$ is C^1 on $(0,\infty)$ and $\widehat{\alpha}_1(s) \leq s\widehat{\alpha}_1'(s)$, we can write, for almost all $t \in [0, |a-x|]$,

$$\frac{d}{dt}\left(\widehat{\alpha}_1(|\psi(t,x)|_{\mathcal{A}})e^{2t}\right) = \left(2\widehat{\alpha}_1(|\psi(t,x)|_{\mathcal{A}}) + \widehat{\alpha}_1'(|\psi(t,x)|_{\mathcal{A}})(\dot{\overbrace{|\psi(t,x)|_{\mathcal{A}}}})\right)e^{2t}$$
$$\leq \widehat{\alpha}_1'(|\psi(t,x)|_{\mathcal{A}})\left(2(|x|_{\mathcal{A}} - t) - 1\right)e^{2t} \quad \leq \quad 0,$$

where the final step follows from $|x|_{\mathcal{A}} \leq \frac{1}{i+1} \leq \frac{1}{2}$. Furthermore, for $t > |a-x|$ we have

$$\frac{d}{dt}\left(\widehat{\alpha}_1(|\psi(t,x)|_{\mathcal{A}})e^{2t}\right) = \frac{d}{dt}\left(\widehat{\alpha}_1(0)e^{2t}\right) = 0\ .$$

This gives $\frac{d}{dt}\left(\widehat{\alpha}_1(|\psi(t,x)|_{\mathcal{A}})e^{2t}\right) \leq 0$ for almost all $t \geq 0$. It follows by integrating that $\sup_{t\geq 0}\widehat{\alpha}_1(|\psi(t,x)|_{\mathcal{A}})e^{2t} = \widehat{\alpha}_1(|x|_{\mathcal{A}})$. So $\widetilde{W}_i(x) \leq \widehat{\alpha}_1(|x|_{\mathcal{A}})$ for all $|x|_{\mathcal{A}} \leq \frac{1}{i+1}$. □

6. Proofs of Lemmas 5.3–5.7. Collected in this section are the proofs of lemmas utilized in proving Theorem 2.1.

6.1. Proof of Lemma 5.3. The proof relies on the proof of [30, Lemma 3]. First we pick $\rho \in \mathcal{K}_\infty$ and a function $\theta : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ continuous and strictly decreasing with $\lim_{t\to\infty}\theta(t) = 0$ such that, for all $t \geq 0$, $\beta(\rho(t), t) \leq \theta(t)$. (The proof that such functions exist can be found at the beginning of [30, Lemma 3].) Next, let $\theta^{-1}(\cdot)$ be the inverse of $\theta(\cdot)$, which is defined and continuous on $(0, \theta(0)]$. It is also strictly decreasing with $\lim_{s\to 0}\theta^{-1}(s) = +\infty$. It follows that the function $e^{-2\lambda\theta^{-1}(s)}$ is well defined, continuous, positive and strictly increasing on $(0, \theta(0)]$. Then we define a continuous, positive definite, nondecreasing function $\pi(\cdot)$ as

$$\pi(s) := \begin{cases} 0, & s = 0, \\ \frac{1}{\theta(0)}e^{-2\lambda\theta^{-1}(s)}, & s \in (0, \theta(0)], \\ \frac{1}{\theta(0)}, & s \geq \theta(0). \end{cases}$$

We then define $\alpha_1(s) := \int_0^s \pi(\tau)d\tau$. It follows that $\alpha_1(\cdot)$ is Lipschitz on its domain (with global Lipschitz constant $\frac{1}{\theta(0)}$), continuously differentiable on $(0,\infty)$, and of class-$\mathcal{K}_\infty$. From the definition of $\alpha_1(\cdot)$, $\alpha_1(s) \leq s\pi(s) = s\alpha_1'(s)$, for all $s > 0$, and $\alpha_1(\cdot)$ satisfies, for any $s \in (0, \theta(0)]$,

$$\alpha_1(s) \leq s\pi(s) \leq \theta(0)\pi(s) \leq e^{-2\lambda\theta^{-1}(s)}\ . \tag{6.1}$$

One may now follow [30, Lemma 3].

6.2. Proof of Lemma 5.4. We follow the proofs of [19, Lemmas 11 and 12]. Without loss of generality, we assume that $\alpha(\cdot)$ is C^1 at the origin, $\alpha(s) \leq s$, and $\alpha'(0) = 0$. Now define $\rho(0) = 0$ and

$$\rho(s) := \exp\left(2\int_1^s \frac{1}{\alpha(\tau)}d\tau\right) \qquad \forall s > 0\ . \tag{6.2}$$

Since $\alpha(s) \leq s$, for $s \geq 1$ we have $\rho(s) \geq \exp\left(2\int_1^s \frac{d\tau}{\tau}\right) = s^2$ and for $s \leq 1$ we have $\rho(s) \leq \exp(-2\int_s^1 \frac{d\tau}{\tau}) = s^2$. It follows with the fact that $\rho(\cdot)$ is C^1 on $(0,\infty)$ that $\rho \in \mathcal{K}_\infty$. By differentiating (6.2), we have, for all $s > 0$, $\rho'(s) = \frac{2\rho(s)}{\alpha(s)}$ so that $\rho(s) = \frac{1}{2}\alpha(s)\rho'(s) \leq \alpha(s)\rho'(s)$. The Lipschitz property for $\rho(\cdot)$ on its domain, demonstrated by showing the boundedness of $\rho'(s)$ for small s, follows from the proof of [19, Lemma 12].

6.3. Proof of Lemma 5.5. Pick $\kappa(t) = 1 + \int_0^t q(s)ds$, where $q : \mathbb{R}_{\geq 0} \to (0, 1]$ is piecewise continuous, nonintegrable (so that $\kappa(\cdot)$ is unbounded), nonincreasing, and such that

$$1 + \int_0^t q(s)ds \leq \frac{1}{\omega_2(s_0 e^{-t})}. \tag{6.3}$$

We can always pick a nonintegrable, nonincreasing, piecewise continuous function $q(\cdot)$ taking values in $(0, 1]$ satisfying (6.3) as follows: We take $q(\cdot)$ to be piecewise constant with $q(s) = q_j$ for $s \in [j-1, j)$, where $\{q_j\}_{j=1}^{\infty}$ is a sequence of nonincreasing strictly positive numbers chosen inductively as follows: Define $q_0 = 1$. Let $\kappa_i = \sum_{j=0}^{i} q_j$. Note that $1 = \kappa_0 \leq 1/\omega_2(s_0 e^{-0}) = 2$. Define

$$m_i := \frac{1}{\omega_2(s_0 e^{-i})} - \kappa_i \tag{6.4}$$

and pick $q_{i+1} = \min\{m_i, q_i\}$. It follows that q_i is nonincreasing. It also follows that

$$\kappa_{i+1} = \kappa_i + q_{i+1} \leq \frac{1}{\omega_2(s_0 e^{-i})} < \frac{1}{\omega_2(s_0 e^{-(i+1)})},$$

from which it follows that $m_i > 0$ and so $q_i > 0$. Let $t \geq 0$, and let i be such that $t \in [i, i+1)$. Then

$$\kappa_i \leq 1 + \int_0^t q(s)ds \leq \kappa_{i+1} \leq \frac{1}{\omega_2(s_0 e^{-i})} \leq \frac{1}{\omega_2(s_0 e^{-t})}. \tag{6.5}$$

Now, suppose that $q(\cdot)$ is integrable. Since the function $\frac{1}{\omega_2(s_0 e^{-t})}$ is unbounded, (6.4) and (6.5) imply that $\lim_{i\to\infty} m_i = \infty$, which implies that there exists $j > 0$ such that $q_k = q_j > 0$ for all $k \geq j$. This contradicts $q(\cdot)$ being integrable.

That $\kappa(\cdot)$ is strictly increasing and (5.1) holds (i.e., $\kappa(0) = 0$) is obvious from the definition. We observe that

$$\frac{\kappa(t)}{\kappa(T)} = \frac{\kappa(T) + \int_T^t q(s)ds}{\kappa(T)} \leq 1 + \frac{t-T}{\kappa(T)} \leq 1 + (t-T) \leq e^{t-T}; \tag{6.6}$$

that is, (5.3) holds. Furthermore, by considering $T = 0$ in (6.6) and the constraint (6.3) we see that (5.4) holds.

Since $q(s) > 0$ for all $s \in \mathbb{R}_{\geq 0}$ and $q(\cdot)$ is nonincreasing, we have

$$\begin{aligned}\kappa(t)\kappa(T) &\geq 1 + \int_0^t q(s)ds + \int_0^T q(s)ds \geq 1 + \int_0^t q(s)ds + \int_t^{t+T} q(s)ds \\ &= 1 + \int_0^{t+T} q(s)ds = \kappa(t+T);\end{aligned}$$

that is, (5.2) holds.

To see that we can simultaneously satisfy (5.5), let $q_s : \mathbb{R}_{\geq 0} \to (0, 1]$ be continuous, nonincreasing, and such that $q_s(t) \leq q(t)$. Then, using the mean value theorem for locally Lipschitz functions, we have that for all $t \in [0, 1]$,

$$\max_{s\in[0,T]} \frac{\kappa(s)}{\kappa(s+t)} = \max_{s\in[0,T]} 1 - \frac{\kappa(s+t) - \kappa(s)}{\kappa(s+t)} \leq 1 - \frac{q_s(T+1)}{\kappa(T+1)} t =: 1 - \vartheta(T)t.$$

Clearly, $\vartheta(\cdot)$ is continuous, nonincreasing, and takes values in $\mathbb{R}_{>0}$. ∎

6.4. Proof of Lemma 5.6. We require some preliminary definitions. Let $\Omega \subset \mathbb{R}^n \backslash \mathcal{A}$ denote the set (of measure zero) where the gradient of $V(\cdot)$ is not defined in $\mathbb{R}^n \backslash \mathcal{A}$. For $(c,d) \in \mathbb{R}_{\geq 0} \times (0,1]$ define

$$\eta(c,d) := \sup_{\{|x| \leq c, V(x) \in [d,1], x \notin \Omega\}} |\nabla V(x)| \ .$$

For each $(c,d) \in \mathbb{R}_{\geq 0} \times (0,1]$, $\eta(c,d)$ is finite since for these (c,d) the set

$$\{x \in \mathbb{R}^n : |x| \leq c, \ V(x) \in [d,1]\}$$

is a compact subset of $\mathbb{R}^n \backslash \mathcal{A}$, the latter being a set on which $V(\cdot)$ is locally Lipschitz. Also $\eta(\cdot,d)$ is nondecreasing and $\eta(c,\cdot)$ is nonincreasing so that $\sigma : \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ defined as

$$\sigma(s) := \begin{cases} \eta\left(\frac{1}{s}, s\right) & \forall s \in (0,1), \\ \eta(1,1) & \forall s \geq 1 \end{cases}$$

is nonincreasing. We claim that

$$x \in \left\{\xi \in (\mathbb{R}^n \backslash \mathcal{A}) \backslash \Omega : \max\{1, |\xi|\} \leq \frac{1}{V(\xi)}\right\} \quad \Longrightarrow \quad |\nabla V(x)| \leq \sigma(V(x)) \tag{6.7}$$

since, in the indicated set, $V(x) \in (0,1]$, $|x| \leq \frac{1}{V(x)}$ so that $\sigma(V(x)) \geq \eta(|x|, V(x)) \geq |\nabla V(x)|$.

Let $\varphi : \mathbb{R}_{\geq 0} \to [0,1]$ be continuous, zero at zero, positive definite, nondecreasing, and let $\pi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be continuous, positive definite, nondecreasing and such that

$$\pi(s)\sigma(s) \leq \varphi(s) \leq 1 \qquad \forall s > 0 \ . \tag{6.8}$$

Define $\rho(r) := \int_0^r \pi(s) ds$. Since $\pi(\cdot)$ is continuous, positive definite, and nondecreasing, we have that $\rho \in \mathcal{K}_\infty$ and $\rho(r) \leq r\pi(r) = r\rho'(r)$ for all $r > 0$.

We now demonstrate the local Lipschitz property for $V_L := \rho \circ V$ on $\mathbb{R}^n$. Since both $\rho(\cdot)$ and $V(\cdot)$ are continuous on $\mathbb{R}^n$, $V_L := \rho \circ V$ is continuous on $\mathbb{R}^n$. Since $\rho(\cdot)$ is locally Lipschitz and $V(\cdot)$ is locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$ and on the interior of $\mathcal{A}$, $V_L(\cdot)$ is locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$ and on the interior of $\mathcal{A}$. To show that $V_L(\cdot)$ is locally Lipschitz on $\mathbb{R}^n$ it is sufficient to establish that, for each $x \in \text{bdry}\mathcal{A}$ (the boundary of $\mathcal{A}$), there exists $M > 0$ and an open, convex neighborhood $\mathcal{U}$ of x such that $DV_L(\xi; v) \leq M|v|$ for all $\xi \in \mathcal{U}, v \in \mathbb{R}^n$, and then use Lemma 5.1. Using $\text{int}(\mathcal{A})$ for the interior of $\mathcal{A}$,

$$\xi \in \text{int}(\mathcal{A}) \quad \Longrightarrow \quad DV_L(\xi; v) = 0 \ \forall v \in \mathbb{R}^n \ . \tag{6.9}$$

Also, it follows from (6.7), (6.8), and $\rho'(s) = \pi(s)$ that

$$\xi \in \left\{\zeta \in (\mathbb{R}^n \backslash \mathcal{A}) \backslash \Omega \ , \ \max\{1, |\zeta|\} \leq \frac{1}{V(\zeta)}\right\} \quad \Longrightarrow \quad |\nabla V_L(\xi)| \leq \varphi(V(\xi)) \leq 1.$$

That $DV_L(\xi; v) \leq \limsup_{y \to \xi} \{\langle \nabla V_L(y), v\rangle : y \notin \Omega\}$ is a combination of [8, Corollary 2.8.2] and [8, Exercise 3.4.1]. Consequently, we have

$$\xi \in \left\{\zeta \in \mathbb{R}^n \backslash \mathcal{A} : \max\{1, 2|\zeta|\} \leq \frac{1}{V(\zeta)}\right\} \quad \Longrightarrow \quad DV_L(\xi; v) \leq |v| \ \forall v \in \mathbb{R}^n. \tag{6.10}$$

With (6.9) and (6.10), we will be done if we can show that

$$x \in \text{bdry}\mathcal{A} \quad \Longrightarrow \quad DV_L(x;v) = 0 \ \forall v \in \mathbb{R}^n \ . \tag{6.11}$$

Since $V(x) = 0$ and $V(\xi) \geq 0$ for all $\xi \in \mathbb{R}^n$, we just need to show $DV_L(x;v) \leq 0$. Let $x \in \text{bdry}\mathcal{A}$. Since $\varphi \circ V(\cdot)$ is continuous and $\varphi(V(x)) = 0$, for each $\varepsilon > 0$ there exists a convex neighborhood $\mathcal{U}$ of x such that

$$\xi \in (\mathbb{R}^n \backslash \mathcal{A}) \cap \mathcal{U} \quad \Longrightarrow \quad \begin{cases} \max\{1, |\xi|\} \leq \frac{1}{V(\xi)} \ , \\ \varphi(V(\xi)) \leq \varepsilon. \end{cases}$$

It follows from (6.7), (6.8), and $\rho'(s) = \pi(s)$ that

$$\xi \in ((\mathbb{R}^n \backslash \mathcal{A}) \backslash \Omega) \cap \mathcal{U} \quad \Longrightarrow \quad |\nabla V_L(\xi)| \leq \varepsilon \ . \tag{6.12}$$

It follows from continuity of $V_L(\cdot)$ that for each $\xi \in (\mathbb{R}^n \backslash \mathcal{A}) \bigcap \mathcal{U}$ there exists $s^*(\xi) \in (0,1)$ such that

$$\begin{aligned} V_L(x + s(\xi - x)) &\geq \frac{V_L(\xi)}{2} \quad \forall s \in [s^*(\xi), 1] \ , \\ V_L(x + s^*(\xi)(\xi - x)) &= \frac{V_L(\xi))}{2} \ . \end{aligned}$$

It follows from the convexity of $\mathcal{U}$ that $\xi \in (\mathbb{R}^n \backslash \mathcal{A}) \cap \mathcal{U}$, and $s \in [s^*(\xi), 1]$ imply $x + s(\xi - x) \in (\mathbb{R}^n \backslash \mathcal{A}) \cap \mathcal{U}$. So, with (6.12) and the mean value theorem for locally Lipschitz functions (see [8, Theorem 2.2.4]) together with [8, Theorem 2.8.1], for all $\xi \in (\mathbb{R}^n \backslash \mathcal{A}) \bigcap \mathcal{U}$,

$$\begin{aligned} 0 \leq V_L(\xi) - V_L(x) &= 2\left(V_L(\xi) - V_L(x + s^*(\xi)(\xi - x))\right) \\ &\leq 2\varepsilon(1 - s^*(\xi))|\xi - x| \leq 2\varepsilon|\xi - x|. \end{aligned}$$

Since, for every $\xi \in \mathcal{A} \bigcap \mathcal{U}$, we have $V_L(\xi) - V_L(x) = 0$, it follows that, for all $\xi \in \mathcal{U}$, we have $0 \leq V_L(\xi) - V_L(x) \leq 2\varepsilon|\xi - x|$. Since $\varepsilon > 0$ can be taken arbitrarily small, (6.11) holds. □

6.5. Proof of Lemma 5.7. For $\dot{x} \in F(x)$ we denote by $\mathcal{R}_{\leq T}(\mathcal{C})$ the set of points reachable from a compact set $\mathcal{C} \subset \mathbb{R}^n$ in time $T > 0$; i.e., $\mathcal{R}_{\leq T}(\mathcal{C}) := \{\xi \in \mathbb{R}^n : \xi = \phi(t,x), t \in [0,T], x \in \mathcal{C}, \phi \in \mathcal{S}(x)\}$.

DEFINITION 8. *The differential inclusion is said to be forward complete if, for every $x \in \mathbb{R}^n$, all $\phi \in \mathcal{S}(x)$ are defined on $[0,\infty)$.*

We require several preliminary lemmas prior to proving Lemma 5.7. The first is also [11, section 7, Theorem 3].

LEMMA 6.1. *Let $F(\cdot)$ satisfy the basic conditions on $\mathbb{R}^n$, and suppose $\dot{x} \in F(x)$ is forward complete. For each compact set $\mathcal{C} \subset \mathbb{R}^n$ and $T \in \mathbb{R}_{>0}$, $\mathcal{R}_{\leq T}(\mathcal{C})$ is a compact subset of $\mathbb{R}^n$ and $\mathcal{S}[0,T](\mathcal{C})$ is a compact set in the metric of uniform convergence.*

An easy corollary of the above appeared as [30, Lemma 5].

LEMMA 6.2. *Let $F(\cdot)$ satisfy the basic conditions on $\mathbb{R}^n$, and let $\dot{x} \in F(x)$ be forward complete. Then each sequence $\{\phi_n\}_{n=1}^{\infty}$ of solutions satisfying $\phi_n \in \mathcal{S}[0,\infty)(x_n)$, where $x_n \to x$, has a subsequence converging to a function $\phi \in \mathcal{S}[0,\infty)(x)$, and the convergence is uniform on each compact time interval.*

The following appeared as [30, Lemma 10] (which derives from [10, Lemma 8.3] and [8, Theorem 4.3.11]) and describes how solutions to locally Lipschitz inclusions depend on initial conditions in a locally Lipschitz manner.

LEMMA 6.3. *Let $F(\cdot)$ satisfy the basic conditions on $\mathbb{R}^n$ and be locally Lipschitz on the open set $\mathcal{O} \subseteq \mathbb{R}^n$. For each $T > 0$ and each compact set $\mathcal{C} \subset \mathcal{O}$, there exist L and $\delta > 0$ such that, for each $\xi \in \mathcal{C}$, each $\phi \in \mathcal{S}(\xi)$, and each $\bar{v}$ satisfying $|\bar{v}| \leq \delta$, there exists $\psi \in \mathcal{S}(\xi + \bar{v})$ with the property $|\phi(t,\xi) - \psi(t,\xi+\bar{v})| \leq L|\bar{v}|$ for all $t \in [0, T_\phi]$, where $T_\phi \in [0,T]$ is such that $\phi(t,\xi) \in \mathcal{C}$ for all $t \in [0,T_\phi]$.*

6.5.1. Preliminaries. We define, for all $(s,\mu) \in \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$,

$$\gamma(s,\mu) := \begin{cases} 1 & \text{if} \quad s \leq 2+\mu, \\ 3+\mu-s & \text{if} \quad s \in [2+\mu, 3+\mu], \\ 0 & \text{if} \quad s \geq 3+\mu. \end{cases}$$

It follows from this definition that $\gamma(\cdot,\mu)$ is Lipschitz of rank one (i.e., Lipschitz with a Lipschitz constant of one). We define the set-valued map

$$F_\mu(x) := \gamma\left(|x|_{\mathcal{A}}, \mu\right) F(x) \ .$$

We note that, for each μ, $F_\mu(\cdot)$ satisfies the basic conditions and is locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{A}$. Also, using Assumption 1 and the definition of $F_\mu(x)$, we have $\max_{w \in F_\mu(x)} |w| \leq M_{3+\mu}$ for all $x \in \mathbb{R}^n$. Therefore, for each $\mu \geq 0$, the inclusion $\dot{x} \in F_\mu(x)$ is forward complete. We define $\mathcal{S}_\mu(x)$ to be the set of maximal solutions for $\dot{x} \in F_\mu(x)$. We let $\mathcal{Q}_\mu(x)$ be the subset of $\mathcal{S}_\mu(x)$ whose elements $\phi(\cdot,x)$ satisfy

$$|\phi(t,x)|_{\mathcal{A}} \leq \alpha_1^{-1}\left(\alpha_1(1+\mu)/\kappa(t)\right) \leq 1+\mu \quad \forall t \geq 0 \ . \tag{6.13}$$

Notice that $\mathcal{Q}_\mu(x)$ may be empty. It follows from the definitions of $\gamma(\cdot,\mu)$ and $F_\mu(\cdot)$ and (6.13) that when $\mathcal{Q}_\mu(x)$ is nonempty we have $\mathcal{Q}_\mu(x) \subseteq \mathcal{S}[0,\infty)(x)$.

6.5.2. An optimal trajectory. We require the following claim.

Claim 3. There exists $\phi^* \in \mathcal{S}[0,\infty)(x)$ such that $V(x) = \sup_{t \geq 0} g(\phi^*(t,x))\kappa(t)$ and, for every $\mu \geq \alpha_1^{-1} \circ \alpha_2(|x|_{\mathcal{A}})$, $\phi^* \in \mathcal{Q}_\mu(x)$.

Proof. Let $\mu \geq \alpha_1^{-1} \circ \alpha_2(|x|_{\mathcal{A}})$, which, since $\alpha_1 \in \mathcal{K}_\infty$, implies $\alpha_1(1+\mu) - \alpha_2(|x|_{\mathcal{A}}) > 0$, and let $\varepsilon > 0$ be such that $\varepsilon \leq \alpha_1(1+\mu) - \alpha_2(|x|_{\mathcal{A}})$. Let $\{\phi_k\}_{k=1}^\infty$ be a sequence within $\mathcal{S}[0,\infty)(x)$, and let $\{\varepsilon_k\}_{k=1}^\infty$ be a sequence of positive real numbers satisfying $\varepsilon_k \leq \varepsilon$ and monotonically decreasing to zero such that

$$\sup_{t \geq 0} g(\phi_k(t,x))\kappa(t) \leq V(x) + \varepsilon_k \ . \tag{6.14}$$

Since $V(x) + \varepsilon_k \leq \alpha_2(|x|_{\mathcal{A}}) + \varepsilon \leq \alpha_1(1+\mu)$ and $g(x) \geq \alpha_1(|x|_{\mathcal{A}})$, it follows that $|\phi_k(t,x)|_{\mathcal{A}} \leq \alpha_1^{-1}(\alpha_1(1+\mu)/\kappa(t)) \leq 1+\mu$ for all $t \geq 0$; i.e., $\phi_k \in \mathcal{Q}_\mu(x) \subseteq \mathcal{S}_\mu(x)$. Since $\dot{x} \in F_\mu(x)$ is forward complete and $F_\mu(\cdot)$ satisfies the basic conditions, it follows from Lemma 6.2 that the sequence ϕ_k has a converging subsequence converging to an element $\phi^* \in \mathcal{S}_\mu(x)$ and that the convergence is uniform on each compact time interval. (We now use ϕ_k to refer to this converging subsequence and ε_k to refer to the corresponding values in (6.14).)

We claim that $\phi^* \in \mathcal{Q}_\mu(x)$. If not, then there exist $\widetilde{t} \geq 0$ and $\widetilde{\rho} > 0$ such that

$$|\phi^*(\widetilde{t},x)|_{\mathcal{A}} \geq \alpha_1^{-1}\left(\alpha_1(1+\mu)/\kappa(t)\right) + \widetilde{\rho} \ . \tag{6.15}$$

From the uniform convergence of ϕ_k to ϕ^* on the interval $[0,\widetilde{t}]$ and since $|\cdot|_{\mathcal{A}}$ is Lipschitz of rank one, there exists $\widetilde{k}$ such that

$$|\phi_{\widetilde{k}}(\widetilde{t},x)|_{\mathcal{A}} \geq |\phi^*(\widetilde{t},x)|_{\mathcal{A}} - \frac{\widetilde{\rho}}{2}\ . \tag{6.16}$$

Combining (6.15) and (6.16), we get that $\phi_{\widetilde{k}} \notin \mathcal{Q}_\mu(x)$, which is a contradiction.

Next we claim that

$$V(x) = V^*(x) := \sup_{t\geq 0} g(\phi^*(t,x))\kappa(t)\ . \tag{6.17}$$

Since $\mathcal{Q}_\mu(x) \subseteq \mathcal{S}[0,\infty)(x)$, we have that $V(x) \leq V^*(x)$. So if (6.17) does not hold, then there exists $\rho > 0$ and $\bar{t} \geq 0$ such that

$$\sup_{t\in[0,\bar{t}]} g(\phi^*(t,x))\kappa(t) \geq V(x) + \rho. \tag{6.18}$$

Now, from the uniform convergence of ϕ_k to ϕ^* on the interval $[0,\bar{t}]$, the compactness of $\mathcal{R}^{\mu}_{\leq\bar{t}}(\{x\})$ (the set of points reachable from x in time $\bar{t}$ for the modified inclusion $\dot{x} \in \overline{F}_\mu(x)$) from Lemma 6.1, and the continuity of $g(\cdot)$, there exists k such that $\varepsilon_k < \rho/2$ and

$$\sup_{t\in[0,\bar{t}]} g(\phi_k(t,x))\kappa(t) \geq \sup_{t\in[0,\bar{t}]} g(\phi^*(t,x))\kappa(t) - \frac{\rho}{2}\ . \tag{6.19}$$

Combining (6.19) with (6.18) and $\varepsilon_k < \rho/2$ we get

$$\sup_{t\in[0,\bar{t}]} g(\phi_k(t,x))\kappa(t) \geq V(x) + \frac{\rho}{2} > V(x) + \varepsilon_k,$$

which contradicts (6.14). This establishes (6.17). □

6.5.3. The local Lipschitz property. We make the following claim.

Claim 4. $V(\cdot)$ is lower semicontinuous on $\mathbb{R}^n$; i.e., for each $x \in \mathbb{R}^n$ and any sequence $x_k \to x$, $\liminf_{k\to\infty} V(x_k) \geq V(x)$.

Proof. Fix $x \in \mathbb{R}^n$, and let $\mu \geq \alpha_1^{-1} \circ \alpha_2\left(|x|_{\mathcal{A}} + 1\right)$. Let x_k be a sequence converging to x and, without loss of generality, assume that $|x_k - x| \leq 1$ for all k so that $\mu \geq \alpha_1^{-1} \circ \alpha_2\left(|x_k|_{\mathcal{A}}\right)$ for all k. Define $\underline{V}(x) := \liminf_{k\to\infty} V(x_k)$. By extracting a suitable subsequence from x_k (and using x_k to denote the subsequence), we can construct a sequence ε_k monotonically decreasing to zero so that

$$V(x_k) \leq \underline{V}(x) + \varepsilon_k\ . \tag{6.20}$$

Let $\phi_k^* \in \mathcal{S}[0,\infty)(x_k)$ come from Claim 3 so that $\phi_k^* \in \mathcal{Q}_\mu(x_k)$. With Lemma 6.2 this sequence has a converging subsequence converging to an element $\psi \in \mathcal{S}_\mu(x)$, and the convergence is uniform on each compact time interval. (We now use ϕ_k and x_k to refer to this convergent subsequence and ε_k to refer to the corresponding values in (6.20).) We can use the same argument as in the proof of Claim 3 to establish that $\psi \in \mathcal{Q}_\mu(x)$. We claim

$$\underline{V}(x) \geq \sup_{t\geq 0} g(\psi(t,x))\kappa(t)\ . \tag{6.21}$$

If this is not the case, then there exist $\rho > 0$ and $\bar{t} \geq 0$ such that

$$\sup_{t\in[0,\bar{t}]} g(\psi(t,x))\kappa(t) \geq \underline{V}(x) + \rho \ . \tag{6.22}$$

Now, from the uniform convergence of $\phi_k^*(\cdot, x_k)$ to $\psi(\cdot, x)$ on the interval $[0, \bar{t}]$, the compactness of $\mathcal{R}^{\mu}_{\leq \bar{t}}(\{x\} + \overline{\mathcal{B}}_n)$ from Lemma 6.1, and the continuity of $g(\cdot)$, there exists k such that $\varepsilon_k < \rho/2$ and

$$\sup_{t\in[0,\bar{t}]} g(\phi_k^*(t,x_k))\kappa(t) \geq \sup_{t\in[0,\bar{t}]} g(\psi(t,x))\kappa(t) - \frac{\rho}{2} \ . \tag{6.23}$$

Combining (6.23) with (6.22) and $\varepsilon_k < \rho/2$ we get

$$V(x_k) \geq \sup_{t\in[0,\bar{t}]} g(\phi_k^*(t,x_k))\kappa(t) \geq \underline{V}(x) + \frac{\rho}{2} > \underline{V}(x) + \varepsilon_k,$$

which contradicts (6.20); i.e., (6.21) holds. Using (6.21) and $\psi \in \mathcal{Q}_\mu(x) \subseteq \mathcal{S}[0,\infty)(x)$, it follows that $\underline{V}(x) \geq V(x)$. □

We now proceed to set ourselves up to apply Lemma 5.1 in order to prove that $V(\cdot)$ is locally Lipschitz. Without loss of generality, we can assume that the sequence ℓ_j is strictly decreasing. For notational convenience, we set $\ell_0 = +\infty$. For $s > 0$, define $\underline{j}(s) := \inf\{j : s \geq \ell_j\}$ and then let $\Upsilon : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ be a continuous, nondecreasing function satisfying $\Upsilon(s) \leq \ell_{\underline{j}(s)+3}$. To see that such a function exists, we note that $\underline{j}(s)$ is nonincreasing and $\underline{j}(s) \to \infty$ as $s \to 0$. Consequently, $\ell_{\underline{j}(s)+3}$ is nondecreasing and $\ell_{\underline{j}(s)+3} \to 0$ as $s \to 0$.

We note that $s \geq \ell_{\underline{j}(s)} \geq \ell_{\underline{j}(s)+3} \geq \Upsilon(s)$. With the fact that $s \leq \alpha_1^{-1} \circ \alpha_2(s)$ for all $s \geq 0$ and that $\kappa(\cdot)$ has a continuous inverse on $[1, \infty)$, this allows us to define $T : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ by $T(0) = 0$ and

$$T(s) := \kappa^{-1}\left(\frac{\alpha_1(1 + \alpha_1^{-1} \circ \alpha_2(s+2))}{\alpha_1(\Upsilon(s))}\right) \quad \forall s > 0 \ .$$

It follows that $T(\cdot)$ is continuous on $\mathbb{R}_{>0}$.

Let $\zeta \in \mathbb{R}^n \backslash \mathcal{A}$, $j := \underline{j}(|\zeta|_{\mathcal{A}})$, $\mu := \alpha_1^{-1} \circ \alpha_2(|\zeta|_{\mathcal{A}} + 1)$ and

$$\mathcal{U} := \{\zeta\} + \frac{1}{2}\min\{1, (\ell_{j-2} - \ell_{j-1}), (\ell_j - \ell_{j+1})\}\,\mathcal{B}_n.$$

Notice that $\mathcal{U}$ is open and convex and that $\ell_j \leq |\zeta|_{\mathcal{A}} < \ell_{j-1}$. Since $\mu > \ell_j > \ell_{j+2}$, we can define $\overline{T} := \kappa^{-1}(\frac{\alpha_1(1+\mu)}{\alpha_1(\ell_{j+2})})$. It follows from the definition of $\mathcal{U}$ and since $|\cdot|_{\mathcal{A}}$ is globally Lipschitz of rank one that, for all $\xi \in \overline{\mathcal{U}}$,

$$|\zeta|_{\mathcal{A}} - \frac{1}{2}\min\{1, (\ell_j - \ell_{j+1})\} \leq |\xi|_{\mathcal{A}} \leq |\zeta|_{\mathcal{A}} + \frac{1}{2}\min\{1, \ell_{j-2} - \ell_{j-1}\} \ . \tag{6.24}$$

Therefore for all $\xi \in \overline{\mathcal{U}}$ we have that

$$\mu \geq \alpha_1^{-1} \circ \alpha_2(|\xi|_{\mathcal{A}}) \ , \tag{6.25}$$

$$|\xi|_{\mathcal{A}} + 2 \geq |\zeta|_{\mathcal{A}} + 1 \ , \text{ and} \tag{6.26}$$

$$|\xi|_{\mathcal{A}} \geq \ell_{j+1} \ . \tag{6.27}$$

Finally, using the upper bound on $|\xi|_{\mathcal{A}}$ and $|\zeta|_{\mathcal{A}} < \ell_{j-1}$ one may show that $|\xi|_{\mathcal{A}} \leq \ell_{j-2}$. We note that from the definition of $\underline{j}$ we have that $\underline{j}(s) \geq i+1$ if $s < \ell_i$. Consequently, $\underline{j}(|\xi|_{\mathcal{A}}) \geq j-1$, which implies

$$\Upsilon(|\xi|_{\mathcal{A}}) \leq \ell_{\underline{j}(|\xi|_{\mathcal{A}})+3} \leq \ell_{j+2} \ . \tag{6.28}$$

Inequalities (6.28) and (6.26) imply that $\overline{T} \leq \min_{\xi \in \overline{\mathcal{U}}} T(|\xi|_{\mathcal{A}})$.

Define $\mathcal{C} := \mathcal{R}^{\mu}_{\leq \overline{T}}(\overline{\mathcal{U}}) \bigcap \{z : |z|_{\mathcal{A}} \geq \ell_{j+2}\}$. It follows that $\mathcal{C}$ is a compact subset of $\mathbb{R}^n \backslash \mathcal{A}$, the latter being an open set where $F_\mu(\cdot)$ is locally Lipschitz. We apply Lemma 6.3 to generate $L > 0$ and $\delta > 0$.

Let $\xi \in \mathcal{U}$. Using (6.25), let $\phi^* \in \mathcal{Q}_\mu(\xi)$ come from Claim 3 so that

$$V(\xi) = \sup_{t \geq 0} g(\phi^*(t,\xi))\kappa(t) \text{ and} \tag{6.29}$$

$$|\phi^*(t,\xi)|_{\mathcal{A}} \leq \alpha_1^{-1}\left(\alpha_1(1+\mu)/\kappa(t)\right) \ \forall t \geq 0 \ . \tag{6.30}$$

It follows from combining (6.30) and (6.27) that the set $\{t : |\phi^*(t,\xi)|_{\mathcal{A}} = \ell_{j+2}\}$ is nonempty and $T_\xi := \inf\{t : |\phi^*(t,\xi)|_{\mathcal{A}} = \ell_{j+2}\}$ is well defined. Since $\phi^*(\cdot,\xi)$ is continuous, we see that

$$\ell_{j+2} = |\phi^*(T_\xi,\xi)|_{\mathcal{A}} \leq \alpha_1^{-1}\left(\alpha_1(1+\mu)/\kappa(T_\xi)\right), \tag{6.31}$$

which leads to $\kappa(T_\xi) \leq \alpha_1(1+\mu)/\alpha_1(\ell_{j+2}) = \kappa(\overline{T})$, which, since $\kappa(\cdot)$ is strictly increasing, implies that $T_\xi \leq \overline{T}$.

Again since $\phi^*(\cdot,\xi)$ is continuous, it follows that $|\phi^*(t,\xi)|_{\mathcal{A}} \geq \ell_{j+2}$ for all $t \in [0,T_\xi]$; i.e., $\phi^*(t,\xi) \in \mathcal{C}$ for all $t \in [0,T_\xi]$. According to the result of Lemma 6.3, for all $\bar{v} \in \mathbb{R}^n$ such that $|\bar{v}| \leq \delta$, there exists $\psi \in \mathcal{S}_\mu(\xi+\bar{v})$ such that

$$|\phi^*(t,\xi) - \psi(t,\xi+\bar{v})| \leq L|\bar{v}| \ \forall t \in [0,T_\xi] \ . \tag{6.32}$$

We define $\delta_1 := \min\left\{\delta, \frac{1}{L}, \frac{1}{L}\left(\ell_{j+1}-\ell_{j+2}\right), \frac{1}{2}\left(\ell_j - \ell_{j+1}\right)\right\}$. Henceforth we assume $|\bar{v}| \leq \delta_1$. It follows with (6.24) that

$$|\xi+\bar{v}|_{\mathcal{A}} \geq |\xi|_{\mathcal{A}} - \frac{1}{2}\left(\ell_j - \ell_{j+1}\right) \geq |\zeta|_{\mathcal{A}} - \left(\ell_j - \ell_{j+1}\right) \geq \ell_{j+1} \tag{6.33}$$

and with (6.32) that

$$|\phi^*(t,\xi) - \psi(t,\xi+\bar{v})| \leq L|\bar{v}| \leq \min\left\{1, (\ell_{j+1}-\ell_{j+2})\right\} \quad \forall t \in [0,T_\xi]. \tag{6.34}$$

The relations (6.34), (6.30), and (6.31) imply that

$$|\psi(t,\xi+\bar{v})|_{\mathcal{A}} \leq |\phi^*(t,\xi)|_{\mathcal{A}} + 1 \leq 2+\mu \ \ \forall t \in [0,T_\xi] \text{ and} \tag{6.35}$$

$$|\psi(T_\xi,\xi+\bar{v})|_{\mathcal{A}} \leq |\phi^*(T_\xi,\xi)|_{\mathcal{A}} + \ell_{j+1} - \ell_{j+2} = \ell_{j+1} \ . \tag{6.36}$$

It follows from (6.35) and the definition of $F_\mu(x)$ that

$$\psi \in \mathcal{S}[0,T_\xi](\xi+\bar{v}) \ . \tag{6.37}$$

The inequalities (6.33) and (6.36) imply that the set $\{t : |\psi(t,\xi+\bar{v})|_{\mathcal{A}} = \ell_{j+1}\}$ is nonempty and we can define $T_{\psi,\bar{v}} := \inf\{t : |\psi(t,\xi+\bar{v})|_{\mathcal{A}} = \ell_{j+1}\}$. We see that $T_{\psi,\bar{v}} \leq T_\xi \leq \overline{T}$, and, since $\psi(\cdot,\xi+\bar{v})$ is continuous, $|\psi(T_{\psi,\bar{v}},\xi+\bar{v})|_{\mathcal{A}} = \ell_{j+1}$. For clarity, define

$z := \psi(T_{\psi,\bar{v}}, \xi + \bar{v})$. From the third assumption of the lemma $g(z) = V(z)$; so, with (6.37) and $\kappa(\tau)\kappa(T_{\psi,\bar{v}}) \geq \kappa(\tau + T_{\psi,\bar{v}})$ for all $\tau \geq 0$, $T_{\psi,\bar{v}} \geq 0$, we have

$$\begin{aligned}\max_{t\in[0,T_{\psi,\bar{v}}]} g(\psi(t,\xi+\bar{v}))\kappa(t) &= \max\left\{\max_{t\in[0,T_{\psi,\bar{v}}]} g(\psi(t,\xi+\bar{v}))\kappa(t), V(z)\kappa(T_{\psi,\bar{v}})\right\}\\ &= \max\left\{\max_{t\in[0,T_{\psi,\bar{v}}]} g(\psi(t,\xi+\bar{v}))\kappa(t), \inf_{\phi\in\mathcal{S}[0,\infty)(z)} \sup_{\tau\geq 0} g(\phi(\tau,z))\kappa(\tau)\kappa(T_{\psi,\bar{v}})\right\}\\ &\geq \inf_{\phi\in\mathcal{S}[0,\infty)(\xi+\bar{v})} \sup_{t\geq 0} g(\phi(t,\xi+\bar{v}))\kappa(t) \;=\; V(\xi+\bar{v}). \qquad (6.38)\end{aligned}$$

One consequence of this calculation, which comes by taking $\bar{v} = 0$ and $\psi = \phi^*$ so that (6.32) is satisfied, is that for each $\xi \in \mathcal{U}$,

$$V(\xi) \geq \max_{t\in[0,T(|\xi|_{\mathcal{A}})]} g(\phi^*(t,\xi))\kappa(t) \geq \max_{t\in[0,T_{\phi^*,0}]} g(\phi^*(t,\xi))\kappa(t) \geq V(\xi)\ ;$$

i.e., $\max_{t\in[0,T(|\xi|_{\mathcal{A}})]} g(\phi^*(t,\xi))\kappa(t) = V(\xi)$. Let $L_{\mathcal{C}} > 0$ be such that $|g(x_1) - g(x_2)| \leq L_{\mathcal{C}}|x_1 - x_2|$ for all $x_1, x_2 \in \mathcal{C} + \overline{\mathcal{B}}_n$. From (6.29), (6.32), and (6.38) it follows that

$$\begin{aligned}V(\xi) &\geq \max_{t\in[0,T_{\psi,\bar{v}}]} g(\phi^*(t,\xi))\kappa(t) \geq \max_{t\in[0,T_{\psi,\bar{v}}]} g(\psi(t,\xi+\bar{v}))\kappa(t) - LL_{\mathcal{C}}|\bar{v}|\kappa(\overline{T})\\ &\geq V(\xi+\bar{v}) - LL_{\mathcal{C}}|\bar{v}|\kappa(\overline{T})\ .\end{aligned}$$

Therefore, with the definition $M := LL_{\mathcal{C}}\kappa(\overline{T})$ for all $\xi \in \mathcal{U}$ and all $v \in \mathbb{R}^n$, we have

$$DV(\xi;v) = \liminf_{w\to v,\varepsilon\to 0^+} \frac{V(\xi+\varepsilon w) - V(\xi)}{\varepsilon} \leq \liminf_{w\to v,\varepsilon\to 0^+} LL_{\mathcal{C}}|w|\kappa(\overline{T}) = M|v|\ .$$

Having already shown that $V(\cdot)$ is lower semicontinuous on $\mathbb{R}^n$ (Claim 4), it follows from Lemma 5.1 that $V(\cdot)$ is locally Lipschitz on the open, convex set $\mathcal{U}$ with Lipschitz constant M. Since $\zeta \in \mathbb{R}^n\backslash\mathcal{A}$ was arbitrary, $V(\cdot)$ is locally Lipschitz on $\mathbb{R}^n\backslash\mathcal{A}$. With this Lipschitz property and the relation $0 \leq V(x) \leq \alpha_2(|x|_{\mathcal{A}})$, it follows that $V(\cdot)$ is continuous on $\mathbb{R}^n$. □

REFERENCES

[1] F. Albertini and E. D. Sontag, *Continuous control-Lyapunov functions for asymptotically controllable time-varying systems*, Internat. J. Control, 72 (1999), pp. 1630–1641.

[2] Z. Artstein, *Stabilization with relaxed controls*, Nonlinear Anal., 7 (1983), pp. 1163–1173.

[3] J. P. Aubin and A. Cellina, *Differential Inclusions: Set-Valued Maps and Viability Theory*, Springer-Verlag, Berlin, 1984.

[4] R. W. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, Birkhäuser Boston, Cambridge, MA, 1982, pp. 181–191.

[5] F. H. Clarke, Yu. S. Ledyaev, L. Rifford, and R. J. Stern, *Feedback stabilization and Lyapunov functions*, SIAM J. Control Optim., 39 (2000), pp. 25–48.

[6] F. H. Clarke, Y. S. Ledyaev, E. D. Sontag, and A. I. Subbotin, *Asymptotic controllability implies feedback stabilization*, IEEE Trans. Automat. Control, 42 (1997), pp. 1394–1407.

[7] F. H. Clarke, Y. S. Ledyaev, and R. J. Stern, *Asymptotic stability and smooth Lyapunov functions*, J. Differential Equations, 149 (1998), pp. 69–114.

[8] F. H. Clarke, Y. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.

[9] F. H. Clarke, R. J. Stern, and P. R. Wolenski, *Subgradient criteria for monotonicity, the Lipschitz condition, and convexity*, Canad. J. Math., 45 (1993), pp. 1167–1183.

[10] K. Deimling, *Multivalued Differential Equations*, Walter de Gruyter, Berlin, 1992.

[11] A. F. Filippov, *Differential Equations with Discontinuous Righthand Sides*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.

[12] R. Freeman and P. Kokotović, *Robust Nonlinear Control Design: State-Space and Lyapunov Techniques*, Birkhäuser Boston, Cambridge, MA, 1996.

[13] C. M. Kellett, *Advances in Converse and Control Lyapunov Functions*, Ph.D. thesis, University of California, Santa Barbara, CA, 2002.

[14] C. M. Kellett and A. R. Teel, *A converse Lyapunov theorem for weak uniform asymptotic stability of sets*, in Proceedings of Mathematical Theory of Networks and Systems, Perpignan, France, 2000.

[15] C. M. Kellett and A. R. Teel, *Uniform asymptotic controllability to a set implies locally Lipschitz control-Lyapunov function*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia, 2000.

[16] M. Krstić, I. Kanellakopoulos, and P. Kokotović, *Nonlinear and Adaptive Control Design*, John Wiley and Sons, New York, 1995.

[17] J. Kurzweil, *On the inversion of Ljapunov's second theorem on stability of motion*, Amer. Math. Soc. Transl. Ser. 2, 24 (1956), pp. 19–77.

[18] Y. Lin, E. D. Sontag, and Y. Wang, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.

[19] L. Praly and Y. Wang, *Stabilization in spite of matched unmodeled dynamics and an equivalent definition of input-to-state stability*, Math. Control Signals Systems, 9 (1996), pp. 1–33.

[20] L. Rifford, *Existence of Lipschitz and semiconcave control-Lyapunov functions*, SIAM J. Control Optim., 39 (2000), pp. 1043–1064.

[21] E. Roxin, *Stability in general control systems*, J. Differential Equations, 1 (1965), pp. 115–150.

[22] E. Roxin, *On asymptotic stability in control systems*, Rend. Circ. Mat. Palermo (2), 15 (1966), pp. 193–208.

[23] E. P. Ryan, *On Brockett's condition for smooth stabilizability and its necessity in a context of nonsmooth feedback*, SIAM J. Control Optim., 32 (1994), pp. 1597–1604.

[24] E. D. Sontag, *A Lyapunov-like characterization of asymptotic controllability*, SIAM J. Control Optim., 21 (1983), pp. 462–471.

[25] E. D. Sontag, *A "universal" construction of Artstein's theorem on nonlinear stabilization*, Systems Control Lett., 13 (1989), pp. 117–123.

[26] E. D. Sontag, *Comments on integral variants of ISS*, Systems Control Lett., 34 (1998), pp. 93–100.

[27] E. D. Sontag, *Stability and stabilization: Discontinuities and the effect of disturbances*, in Nonlinear Analysis, Differential Equations, and Control, NATO Sci. Ser. C. Math. Phys. Sci. 528, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 551–598.

[28] E. D. Sontag, *Clocks and insensitivity to small measurement errors*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 537–557.

[29] E. D. Sontag and H. J. Sussman, *Nonsmooth control-Lyapunov functions*, in Proceedings of the 34th Conference on Decision and Control, New Orleans, LA, 1995, pp. 2799–2805.

[30] A. R. Teel and L. Praly, *A smooth Lyapunov function from a class-$\mathcal{KL}$ estimate involving two positive semidefinite functions*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 313–367.

[31] F. W. Wilson, *Smoothing derivatives of functions and applications*, Trans. Amer. Math. Soc., 139 (1969), pp. 413–428.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 1960–1971

FAST CONSTRUCTION OF ROBUSTNESS DEGRADATION FUNCTION*

XINJIA CHEN†, KEMIN ZHOU†, AND JORGE L. ARAVENA†

Abstract. We develop a fast algorithm to construct the robustness degradation function, which describes quantitatively the relationship between the proportion of systems guaranteeing the robustness requirement and the radius of the uncertainty set. This function can be applied to predict whether a controller design based on an inexact mathematical model will perform satisfactorily when implemented on the true system.

Key words. robustness, uncertainty, sample reuse algorithm

AMS subject classifications. 93D09, 93D15, 68W20, 68W40

DOI. 10.1137/S0363012902409234

1. Introduction.

In recent years, there has been growing interest in the development of probabilistic methods for robustness analysis and design problems aimed at overcoming the computational complexity and the conservatism issue of the deterministic worst-case framework [16, 17, 14, 12, 19, 2, 5, 4, 18, 8, 9, 6, 7, 21, 22, 15, 3]. In the deterministic worst-case framework, one is interested in knowing if the robustness requirement is guaranteed for every value of the uncertainty. However, it should be borne in mind that the uncertainty set may include worst cases which never happen in reality. Instead of seeking the worst-case guarantee, it is sometimes "acceptable" that the robustness requirement is satisfied for most of the cases. It has been demonstrated that the proportion of systems guaranteeing the robustness requirement can be close to 1 even if the radii of the uncertainty set are much larger than the worst-case deterministic robustness margin [2, 13, 4, 8, 18]. Therefore, it is of practical importance to construct a function which describes quantitatively the relationship between the proportion of systems guaranteeing the robustness requirement and the radius of the uncertainty set. Such a function can serve as a guide for control engineers in evaluating the robustness of a control system once a controller design is completed. Such a function, referred to as a *robustness degradation function*, has been proposed by a number of researchers [2, 8]. For example, Barmish, Lagoa, and Tempo [2] have constructed a curve of the robustness margin amplification versus risk in a probabilistic setting. In a similar spirit, Calafiore, Dabbene, and Tempo [8, 9] have constructed a probability degradation function in the context of real and complex parametric uncertainty.

In this paper, allowing the robustness analysis to be performed in a distribution-free manner, we introduce the concept of *proportion* and adopt the assumption from the classical robust control framework that uncertainty is deterministic and bounded. It follows naturally that the robustness of a system can be reasonably measured by the ratio of the volume (Lebesgue measure) of the set of uncertainty guaranteeing the robustness requirement to the overall set of uncertainty [19]. Evaluation of such a measure of robustness requires generating samples with uniform distribution over un-

*Received by the editors June 5, 2002; accepted for publication (in revised form) August 6, 2003; published electronically January 22, 2004. This research was supported in part by grants from NASA (NCC5-573) and LEQSF (NASA/LEQSF(2001-04)-01).

http://www.siam.org/journals/sicon/42-6/40923.html

†Department of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA 70803 (chan@ece.lsu.edu, kemin@ece.lsu.edu, aravena@ece.lsu.edu).

certainty sets such as a spectral normal ball or an l_p ball. The difficulty of generating such samples has been successfully resolved in [8, 9].

The conventional method for constructing the robustness function is to perform, independently, a certain number of simulations for each value of the uncertainty radius and then plot the function. Although such a curve can be applied to evaluate the robustness of the control system, it may be computationally expensive. This is especially true when many cycles of controller synthesis and robustness analysis are needed in the development of a high performance control system. Motivated by this situation, we focus on the machinery that can make the construction of such a function efficient. We have developed a sample reuse algorithm that allows the simulations to be conducted in an iterative manner. The idea is to start simulation from the larger uncertainty set and save appropriate evaluations of the robust requirement for the use of later simulations on the smaller uncertainty set. In this way the total number of simulations can be reduced significantly as compared to the conventional method.

In addition to deriving our sample reuse algorithm from the worst-case deterministic framework, we show that the technique is also applicable when considering the random nature of the uncertainty. In such cases, the worst-case properties of uniform distribution given in the pioneering work [5, 2, 1] allow our algorithm to be applied to efficiently solve a wide variety of robustness analysis problems. In particular, the radial truncation theory [2] can be applied to robustness analysis problems with uncertainty bounding sets defined as spectral norm balls and l_p balls.

The organization of the paper is as follows. Section 2 gives the problem formulation. Section 3 presents our sample reuse algorithm. Section 4 is the performance analysis of the algorithm. Section 5 applies the algorithm to examples. Section 6 shows the justification of the algorithm for the case of random uncertainties. Section 7 is the conclusion. The proofs of the theorems are included as an appendix.

2. Problem formulation. We adopt the assumption, from the classical robust control framework, that the uncertainty is deterministic and bounded. We formulate a general robustness analysis problem as follows.

Let $\mathbf{P}$ denote a robustness requirement. The definition of $\mathbf{P}$ can be a fairly complicated combination of the following:

- stability or $\mathcal{D}$-stability;
- H_∞ norm of the closed-loop transfer function;
- time specifications such as overshoot, rise time, settling time, and steady state error.

Let $\mathcal{B}(r)$ denote the set of uncertainties with size smaller than r. In applications, we are usually dealing with uncertainty sets such as the following:

- l_p ball $\mathcal{B}_p(r) := \{\Delta \in \mathbf{R}^n : \ ||\Delta||_p \le r\}$, where $||.|_p$ denotes the l_p norm and $p = 1, 2, \ldots, \infty$. In particular, $\mathcal{B}_\infty(r)$ denotes a box.
- Spectral norm ball $\mathcal{B}_\sigma(r) := \{\Delta \in \mathbf{\Delta} : \bar{\sigma}(\Delta) \le r\}$, where $\bar{\sigma}(\Delta)$ denotes the largest singular value of Δ. The class of allowable perturbations is

  $$\mathbf{\Delta} := \{\text{blockdiag}[q_1 I_{r_1}, \ldots, q_s I_{r_s}, \Delta_1, \ldots, \Delta_c]\}, \tag{2.1}$$

  where $q_i \in \mathbb{F}$, $i = 1, \ldots, s$ are scalar parameters with multiplicity $r_1, \ldots, r_s$ and $\Delta_i \in \mathbb{F}^{n_i \times m_i}$, $i = 1, \ldots, c$ are possibly repeated full blocks. Here $\mathbb{F}$ is either the complex field $\mathbf{C}$ or the real field $\mathbf{R}$.
- Homogeneous star-shaped bounding set $\mathcal{B}_H(r) := \{r(\Delta - \Delta_0) + \Delta_0 : \Delta \in Q\}$, where $Q \subset \mathbf{R}^n$ and $\Delta_0 \in Q$ (see [2] for a detailed illustration).

Throughout this paper, $\mathcal{B}(r)$ refers to any type of uncertainty set described above.

Define a function $\ell(.)$ such that, for any X,

$$\ell(X) := \min\{r : X \in \mathcal{B}(r)\},$$

i.e., $\mathcal{B}(\ell(X))$ includes X exactly in the boundary. By such definition,

$$\ell(X) = \min\left\{r : \frac{X - \Delta_0}{r} + \Delta_0 \in Q\right\},$$

$$\ell(X) = \bar{\sigma}(X),$$

and

$$\ell(X) = ||X||_p$$

in the context of a homogeneous star-shaped bounding set, spectral norm ball, and l_p ball, respectively.

To allow the robustness analysis to be performed in a distribution-free manner, we introduce the notion of *proportion* as follows. For any $\Delta \in \mathcal{B}(r)$ there is an associated system $G(\Delta)$. We define *proportion* as follows:

$$\mathbb{P}(r) := \frac{\text{vol}(\{\Delta \in \mathcal{B}(r) : \text{The associated system } G(\Delta) \text{ guarantees } \mathbf{P}\})}{\text{vol}(\mathcal{B}(r))}$$

with

$$\text{vol}(S) := \int_{q \in S} dq,$$

where the notion of dq is illustrated as follows:

- (I): If $q = [x_{rs}]_{n\times m}$ is a real matrix in $\mathbf{R}^{n\times m}$, then $dq = \prod_{r=1}^{n}\prod_{s=1}^{m} dx_{rs}$.
- (II): If $q = [x_{rs} + jy_{rs}]_{n\times m}$ is a complex matrix in $\mathbf{C}^{n\times m}$, then $dq = \prod_{r=1}^{n}\prod_{s=1}^{m}(dx_{rs}dy_{rs})$.
- (III): If $q \in \mathbf{\Delta}$, i.e., q possesses a block structure defined by (2.1), then $dq = (\prod_{i=1}^{s} dq_i)(\prod_{i=1}^{c} d\Delta_i)$, where the notion of dq_i and $d\Delta_i$ is defined by (I) and (II).

It follows that $\mathbb{P}(r)$ is a reasonable measure of the robustness of the system [8, 20]. In the worst-case deterministic framework, we are interested only in knowing if $\mathbf{P}$ is guaranteed for every Δ. However, one should bear in mind that the uncertainty set in our model may include worst cases which never happen in reality. Thus, it would be "acceptable" in many applications if the robustness requirement $\mathbf{P}$ is satisfied for most of the cases. Hence, due to the inaccuracy of the model, we should also obtain the value of $\mathbb{P}(r)$ for uncertainty radius r which exceeds the deterministic robustness margin.

Clearly, $\mathbb{P}(r)$ is deterministic in nature. However, we can resort to a probabilistic approach to evaluate $\mathbb{P}(r)$. To see this, one needs to observe that a random variable with *uniform distribution* over $\mathcal{B}(r)$, denoted by $\mathbf{\Delta}^u$, guarantees that

$$\Pr\{\mathbf{\Delta}^u \in S\} = \frac{\text{vol}(S \bigcap \mathcal{B}(r))}{\text{vol}(\mathcal{B}(r))}$$

for any S, and thus

$$\mathbb{P}(r) = \Pr\{\text{The associated system } G(\mathbf{\Delta}^u) \text{ guarantees } \mathbf{P}\}.$$

It follows that a Monte Carlo method can be employed to estimate $\mathbb{P}(r)$ based on independently and identically distributed (i.i.d.) observations of $\mathbf{\Delta}^u$.

It is interesting to know how the function $\mathbb{P}(r)$ degrades with respect to r when r increases from a to b, where $b > a \geq 0$. In a similar spirit, such a function has been proposed as a *confidence degradation function* in [2] and as a *probability degradation*

function in [8, 9]. In this paper, we refer to the function $\mathbb{P}(.)$ as a *robustness degradation function* for the following reasons. First, we introduce the confidence interval for assessing the accuracy of the estimate of $\mathbb{P}(r)$. To be useful, every numerical method should be associated with an assessment for the accuracy of the estimate. Monte Carlo simulation is no exception. To avoid confusion, we reserve the notion of "confidence" for the purpose of interval estimation. Second, we introduce the concept of *proportion* for measuring robustness, which has no probabilistic content. Third, $\mathbb{P}(r)$ is a robustness measure and is usually decreasing with respect to r when $\mathbb{P}(r)$ is close to 1.

To construct such a function of practical importance, the conventional way is to grid the interval $[a, b]$ as $a = \rho_1 < \rho_2 < \cdots < \rho_l = b$ and estimate $\mathbb{P}(\rho_i)$ by conducting N i.i.d. sampling experiments for each ρ_i. In total, we need Nl samples. In the next section we show that the number of experiments can be significantly reduced.

3. Sample reuse algorithm. To improve efficiency, we shall make use of the following simple yet important observation.

Let q^* be an observation of a random variable with uniform distribution over $\mathcal{B}(r^*) \supseteq \mathcal{B}(r)$ such that $q^* \in \mathcal{B}(r)$. Then q^* can also be viewed as an observation of a random variable with uniform distribution over $\mathcal{B}(r)$.

In our algorithm, we flip the order of ρ_i by defining

$$r_i = \rho_{l+1-i}$$

for $i = 1, 2, \ldots, l$. Thus, the direction of simulation is backward. Our algorithm is described as follows.

Sample Reuse Algorithm

- Input: Sample size N, confidence parameter $\delta \in (0, 1)$ and uncertainty radii r_i, $i = 1, 2, \ldots, l$.
- Output: Proportion estimate $\widehat{\mathbb{P}}_i$ and the related confidence interval for $i = 1, \ldots, l$. In the following, m_{i1} denotes the number of sampling experiments conducted at r_i, and m_{i2} denotes the number of observations guaranteeing $\mathbf{P}$ during the m_{i1} sampling experiments.
- *Step* 1 (initialization). Let $M = [m_{ij}]_{l\times 2}$ be a zero matrix.
- *Step* 2 (backward iteration). For $i = 1$ to $i = l$ do the following:
  - Let $r \leftarrow r_i$.
  - While $m_{i1} < N$ do the following:
    - Generate uniform sample q from $\mathcal{B}(r)$. Evaluate the robustness requirement $\mathbf{P}$ for q.
    - Let $m_{s1} \leftarrow m_{s1} + 1$ for any s such that $r \geq r_s \geq \ell(q)$.
    - If robustness requirement $\mathbf{P}$ is satisfied for q, then let $m_{s2} \leftarrow m_{s2}+1$ for any s such that $r \geq r_s \geq \ell(q)$.
  - Let $\widehat{\mathbb{P}}_i \leftarrow \frac{m_{i2}}{N}$ and construct the confidence interval of confidence level $100(1-\delta)\%$.

It follows that q can be viewed as an observation of a random variable with uniform distribution over $\mathcal{B}(r_j)$ if and only if $r \geq r_j \geq \ell(q)$. Hence, if the robustness requirement $\mathbf{P}$ has been evaluated for $\mathcal{B}(r_i)$ at sample q, the result can be accepted without repeated evaluation of $\mathbf{P}$ for all $\mathcal{B}(r_j)$ such that $r \geq r_j \geq \ell(q)$. *Thus, sample reuse allows us to save both the sample generation and the evaluation of $\mathbf{P}$ for the sample. It is also interesting to point out that the samples collected for each r_i are i.i.d. and thus the confidence interval can be rigorously constructed based on the evaluation of $\mathbf{P}$ for the samples.*

4. Sample reuse factor. Let $\mathbf{n}_i$ be the number of simulations required at r_i. Define *sample reuse factor* as follows:

$$\mathcal{F}_{reuse} := \frac{Nl}{\mathcal{E}[\sum_{i=1}^{l} \mathbf{n}_i]},$$

where $\mathcal{E}(X)$ denotes the expectation of random variable X. Obviously, $\mathcal{F}_{reuse}$ measures the improvement of efficiency upon the conventional method. We demonstrate that the improvement can be significant in most applications.

THEOREM 1. *The sample reuse factor* $\mathcal{F}_{reuse} = l/l - \sum_{i=2}^{l} \left(\frac{r_i}{r_{i-1}}\right)^d$, *where* $d = n$ *for* l_p *ball* $\mathcal{B}_p(r)$ *and homogeneous star-shaped bounding set* $\mathcal{B}_H(r)$*; and*

$$d = \sum_{i=1}^{s} \kappa(q_i) + \sum_{j=1}^{c} \kappa(\Delta_j)$$

for spectral norm ball $\mathcal{B}_\sigma(r)$ *with* $\kappa(.)$ *defined as*

$$\kappa(X) := \begin{cases} 2mn & \text{if } X \text{ is a variable in } \mathbf{C}^{n\times m} \\ mn & \text{if } X \text{ is a variable in } \mathbf{R}^{n\times m}. \end{cases}$$

See the appendix for proof. For illustration purposes, we choose $r_i = b - \frac{(b-a)(i-1)}{l-1}$ for $i = 1, 2, \ldots, l$. By Theorem 1, $\mathcal{F}_{reuse} = l/l - \sum_{i=2}^{l} \left(1 - \frac{1}{\frac{l-1}{1-\frac{a}{b}} - i + 2}\right)^d$. Figures 1 and 2 show that the improvement over the conventional approach is significant when d is not large. These figures also reveal that the sample reuse factor does not scale well with the uncertainty dimension. For example, when $d > 160$, the efficiency gained from sample reuse techniques may not be attractive.

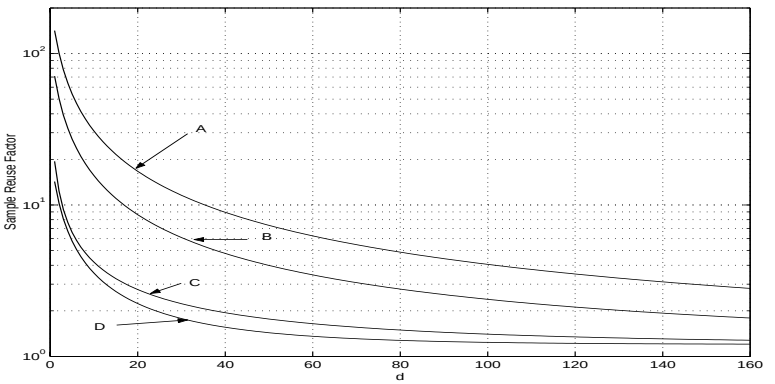


FIG. 1. *Performance improvement (*A : $l = 200,\ b = 2a$; B : $l = 100,\ b = 2a$; C : $l = 100,\ a = 0$; D : $l = 20,\ b = 2a$*).*

5. Illustrative examples. In this section we demonstrate through examples the power of the sample reuse algorithm in solving a wide variety of complicated robustness analysis problems which are intractable in the classical deterministic framework.

First, we consider an example which has been studied in [11] by a deterministic approach. The system is as shown in Figure 3.

The compensator is $C(s) = \frac{s+2}{s+10}$ and the plant is $P(s) = \frac{800(1+0.1\delta_1)}{s(s+4+0.2\delta_2)(s+6+0.3\delta_3)}$ with parametric uncertainty $\Delta = [\delta_1, \delta_2, \delta_3]^{\mathrm{T}}$. The nominal system is stable. The closed-loop roots of the nominal system are

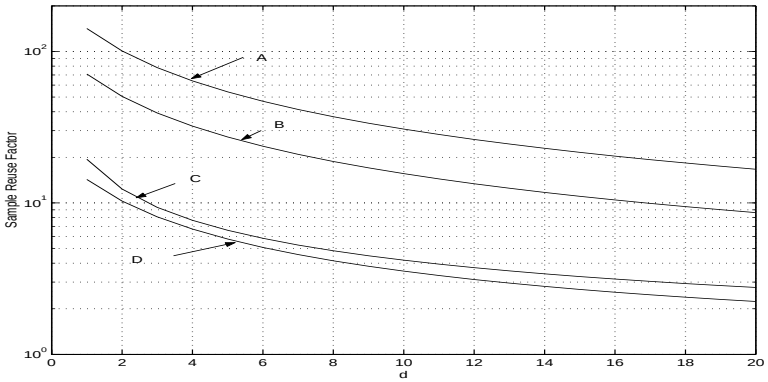


FIG. 2. *Performance improvement (*A : $l = 200,\ b = 2a$; B : $l = 100,\ b = 2a$; C : $l = 100,\ a = 0$; D : $l = 20,\ b = 2a$*).*

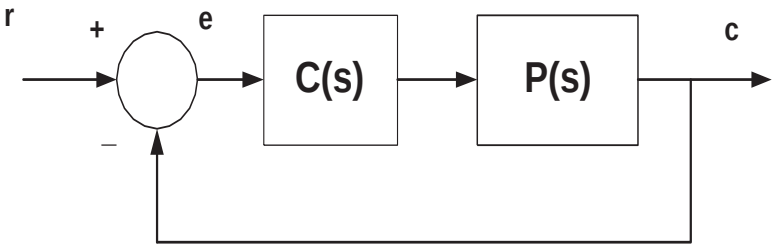


FIG. 3. *Uncertain system.*

$$z_1 = -15.9178,\ \ z_2 = -1.8309,\ \ z_3 = -1.1256 + 7.3234i,\ \ z_4 = -1.1256 - 7.3234i.$$

The H_∞ norm of the nominal closed-loop transfer function is $||T^0||_\infty = 2.78$. The peak value, rise time, and settling time of step response of the nominal system are, respectively, $P^0_{peak} = 1.47$, $t^0_r = 0.185$, and $t^0_s = 3.175$. In all of the following examples, we take $l = 100$. To guarantee that the absolute error of the estimate for the proportion is less than 0.01 with confidence level 99%, we choose $N = 26,492$ based on the well-known Chernoff bound (see [12, 19] for "sharper" bounds). Since the Chernoff bound is conservative, we also performed a post-experimental evaluation of the estimates by constructing confidence intervals with confidence level 99% based on Clopper–Pearson's method [10].

Figure 4 is the robustness degradation curve for robust stability over uncertainty set $\mathcal{B}_\infty(r) := \{\Delta :\ ||\Delta||_\infty \le r\}$. It demonstrates that a significant enhancement of the robustness margin can be achieved at the price of a small risk.

Figure 5 is the robustness degradation curve, with the robustness requirement $\mathbf{P}$ defined as stability and H_∞ norm $< 170\%\ ||T^0||_\infty$, and the uncertainty set defined as the ellipsoid $\mathcal{B}_2(r) := \{\Delta :\ ||\Delta||_2 \le r\}$.

Figure 6 is the robustness degradation curve with the robustness requirement $\mathbf{P}$ defined as $\mathcal{D}$-stability with the domain of poles defined as: real part < -1.5, or it falls within one of the two disks centered at z_3 and z_4 with radius 0.3. The uncertainty set is defined as the polytope

$$\mathcal{B}_H(r) := \left\{ r\Delta + (1-r)\,\frac{\sum_{i=1}^4 \Delta^i}{4} : \Delta \in \mathrm{conv}\{\Delta^1, \Delta^2, \Delta^3, \Delta^4\} \right\},$$

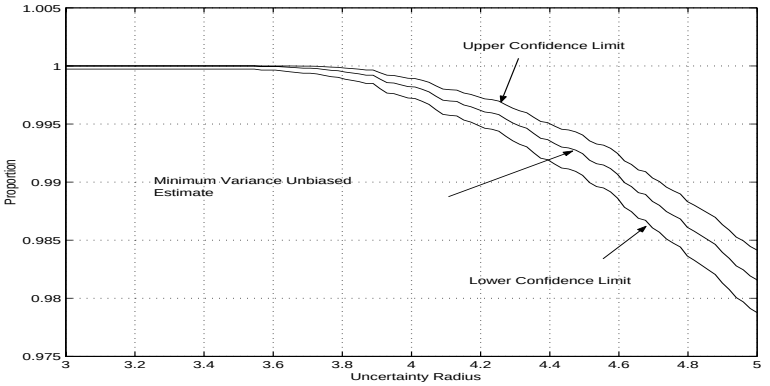


FIG. 4. *Robustness degradation curve (reuse factor = 41).*

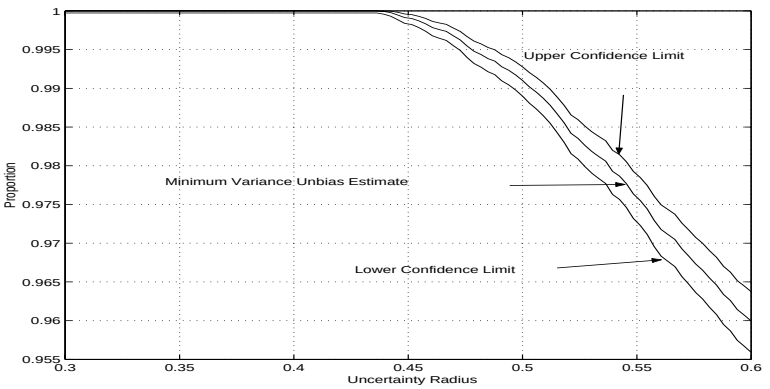


FIG. 5. *Robustness degradation curve (reuse factor = 43).*

where "conv" denotes the convex hull of $\Delta^i = [\frac{1}{2}\sin(\frac{2i-1}{3}\pi),\ \frac{1}{2}\cos(\frac{2i-1}{3}\pi),\ -\frac{\sqrt{3}}{2}]^{\mathrm{T}}$ for $i = 1, 2, 3$ and $\Delta^4 = [0,\ 0,\ 1]^{\mathrm{T}}$.

Figure 7 is the robustness degradation curve for the case where the uncertainty set is $\mathcal{B}_\infty(r) := \{\Delta :\ ||\Delta||_\infty \le r\}$, the robustness requirement $\mathbf{P}$ is: stability, and rise time $t_r < 135\%\ t_r^0 = 0.25$, settling time $t_s < 110\%\ t_s^0 = 3.5$, and overshoot $P_{peak} < 116\%\ P_{peak}^0 = 1.7$.

Finally, we consider the same example in [8] where the class of uncertainty is defined as

$$\mathbf{\Delta} := \{\text{blockdiag}[q_1 I_5,\ q_2 I_5,\ \Delta_1]\},$$

where $\Delta_1 \in \mathbf{C}^{4\times 4}$, and I_5 denotes the identity matrix of 5×5. By Theorem 1, we have $d = 34$. Figure 8 shows the robustness degradation curve. An improvement (of efficiency) about fivefold is achieved by our algorithm.

6. A probabilistic perspective. In sections 2 and 3, we have derived our sample reuse algorithm from the worst-case deterministic framework. In this section, we show that the proposed algorithm is also applicable from the perspective of the random nature of uncertainty. In situations where we need to take into account the random nature of uncertainty, the pioneering work of Barmish, Lagoa, Tempo, Bai, and Fu [2, 1] allows our sample reuse algorithm to be applied to solve efficiently a

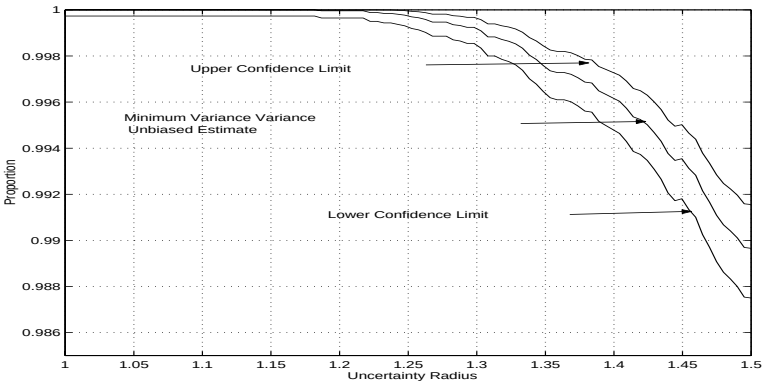


FIG. 6. *Robustness degradation curve (reuse factor = 49).*

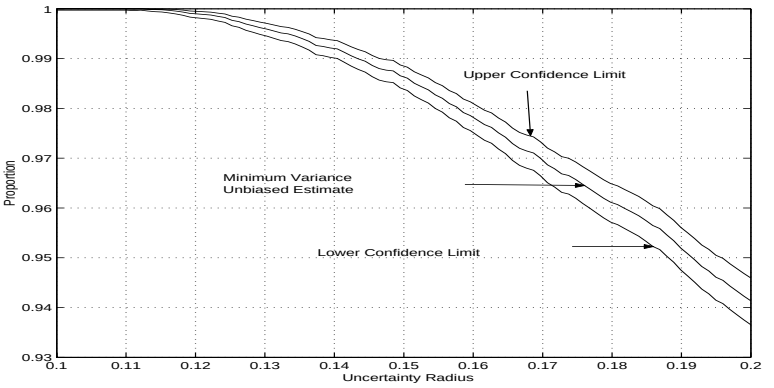


FIG. 7. *Robustness degradation curve (reuse factor = 38).*

wide variety of robustness problems. The following theorem plays an important role.

THEOREM 2 (see [2]). *Suppose that uncertainty* $\mathbf{\Delta}$ *is a random variable with a density function* $f(\Delta)$ *which depends only on* $\ell(\Delta)$ *and is nonincreasing with respect to* $\ell(\Delta)$. *Then*

$$\Pr\{\text{The associated system } G(\mathbf{\Delta}) \text{ guarantees } \mathbf{P} \mid \mathbf{\Delta} \in \mathcal{B}(r)\} \geq \inf_{0 \leq \rho \leq r} \mathbb{P}(\rho).$$

Remark 1. A remarkable fact of Theorem 2 is that no assumption needs to be imposed on the robustness requirement $\mathbf{P}$. The assumption in Theorem 2 is roughly interpreted to mean that the probability measure of the uncertainty is radially symmetrical with respect to the nominal value. In many applications, small perturbations are more likely than large perturbations, and the uncertainty is sufficiently unstructured so as to be treated equally likely in the surface of $\mathcal{B}(r)$ [2].

Remark 2. It should be noted that Theorem 2 applies to a homogeneous star-shaped bounding set, l_p ball, and spectral norm ball. We introduce in Theorem 2 a conditional probability based on the following reason: It does not seem logical to treat the uncertainty as different bounded random variables. For example, if the uncertainty possesses a certain distribution over $\mathcal{B}(r_1)$, it would be a contradiction that the uncertainty possesses another distribution over $\mathcal{B}(r_2)$ for $r_2 > r_1$. In fact, if the uncertainty is of random nature, then the associated distribution is unique.

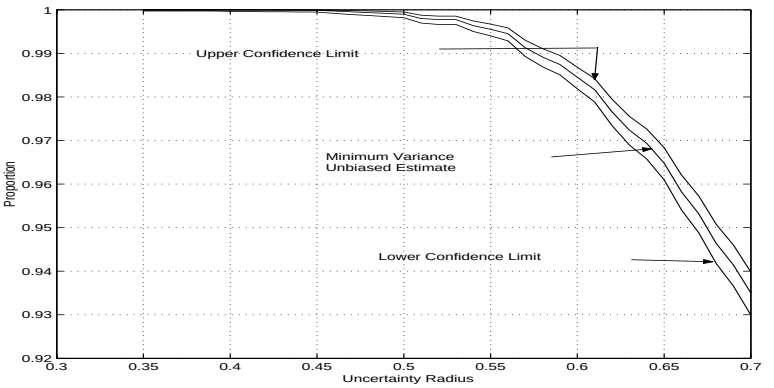


FIG. 8. *Robustness degradation curve (reuse factor = 5).*

Based on Theorem 2, we can apply the sample reuse algorithm to estimate $\mathbb{P}(r)$ for $r \in [0, b]$, from which we can construct the lower bounds for $\Pr\{G(\mathbf{\Delta})$ guarantees $\mathbf{P} \mid \mathbf{\Delta} \in \mathcal{B}(r)\}$.

7. Conclusion. We develop a fast algorithm for computing the robustness degradation function which overcomes the computational complexity and conservatism issue of the deterministic worst-case methods. We also demonstrate that our algorithm can provide efficient solutions for a wide variety of robustness analysis problems which are intractable by the deterministic worst-case methods. We derive our algorithm from the worst-case deterministic framework and also show that the algorithm is applicable from a probabilistic perspective.

Appendix. Proof of Theorem 1.

The following lemma follows essentially from the definition of volume function vol(.).

LEMMA 3. *Let* $X = \{\Delta \in \mathbf{C}^{n\times m} : \bar{\sigma}(\Delta) \le r\}$ *and* $Y = \{\Delta \in \mathbf{C}^{m\times n} : \bar{\sigma}(\Delta) \le r\}$. *Let* $Z = \{\Delta \in \mathbf{R}^{n\times m} : \bar{\sigma}(\Delta) \le r\}$ *and* $W = \{\Delta \in \mathbf{R}^{m\times n} : \bar{\sigma}(\Delta) \le r\}$. *Then* $\mathrm{vol}(X) = \mathrm{vol}(Y)$ *and* $\mathrm{vol}(Z) = \mathrm{vol}(W)$.

LEMMA 4. *Let* $m \ge n$. *Define spectral norm ball* $\mathcal{B}_\sigma^C(r) = \{\Delta \in \mathbf{C}^{n\times m} : \bar{\sigma}(\Delta) \le r\}$ *and spectral norm ball* $\mathcal{B}_\sigma^R(r) = \{\Delta \in \mathbf{R}^{n\times m} : \bar{\sigma}(\Delta) \le r\}$. *Then* $\mathrm{vol}(\mathcal{B}_\sigma^C(r)) = \mathrm{vol}(\mathcal{B}_\sigma^C(1))r^d$ *with* $d = 2mn$ *and* $\mathrm{vol}(\mathcal{B}_\sigma^R(r)) = \mathrm{vol}(\mathcal{B}_\sigma^R(1))r^d$ *with* $d = mn$.

Proof. By Theorem 1 of [8], we have

$$\frac{\Upsilon_C}{\mathrm{vol}(\mathcal{B}_\sigma^C(r))} \int_{r\ge\sigma_1>\sigma_2>\cdots>\sigma_n>0} \prod_{i=1}^{n} \sigma_i^{2(m-n)+1} \times \prod_{1\le i<k\le n} (\sigma_i^2 - \sigma_k^2)^2 \, d\sigma_1 d\sigma_2 \cdots d\sigma_n = 1$$

with $\Upsilon_C = \frac{2^n \pi^{mn}}{\prod_{k=1}^{n}(n-k)!(m-k)!}$. Performing a change of variables with $x_i = \frac{\sigma_i}{r}$ for $i = 1, \ldots, n$, we have

$$\frac{r^{2mn}\Upsilon_C}{\mathrm{vol}(\mathcal{B}_\sigma^C(r))} \int_{1\ge x_1>x_2>\cdots>x_n>0} \prod_{i=1}^{n} x_i^{2(m-n)+1} \times \prod_{1\le i<k\le n} (x_i^2 - x_k^2)^2 \, dx_1 dx_2 \cdots dx_n = 1.$$

Thus

$$\mathrm{vol}(\mathcal{B}_\sigma^C(1)) = \Upsilon_C \int_{1\geq x_1>x_2>\cdots>x_n>0} \prod_{i=1}^{n} x_i^{2(m-n)+1} \times \prod_{1\leq i<k\leq n} (x_i^2 - x_k^2)^2 \, dx_1 dx_2 \cdots dx_n,$$

and

$$\mathrm{vol}(\mathcal{B}_\sigma^C(r)) = \mathrm{vol}(\mathcal{B}(1))\, r^{2mn}.$$

Similarly, by Theorem 2 of [8], we can show that $\mathrm{vol}(\mathcal{B}_\sigma^R(r)) = \mathrm{vol}(\mathcal{B}_\sigma^R(1))r^d$ with $d = mn$. □

LEMMA 5. *$\mathrm{vol}(\mathcal{B}(r)) = \mathrm{vol}(\mathcal{B}(1))r^d$ where $d = n$ for l_p ball $\mathcal{B}_p(r)$ and homogeneous star-shaped bounding set $\mathcal{B}_H(r)$; and $d = \sum_{i=1}^{s} \kappa(q_i) + \sum_{j=1}^{c} \kappa(\Delta_j)$ for spectral norm ball $\mathcal{B}_\sigma(r)$.*

Proof. The truth is obvious for cases of an l_p ball and homogeneous star-shaped bounding set. To prove the lemma for the case of a spectral norm ball, we need to apply Lemmas 3 and 4. □

LEMMA 6. *For $i = 1, \ldots, l-1$,*

$$\mathcal{E}\left[\mathbf{n}_{i+1}\right] = N - \sum_{j=1}^{i} \left(\frac{r_{i+1}}{r_j}\right)^d \mathcal{E}\left[\mathbf{n}_i\right].$$

Proof. Let $q^1, q^2, \ldots, q^{\mathbf{n}_j}$ be the samples generated on r_j. For $l = 1, \ldots, \mathbf{n}_j$, define binomial random variable $X_{j,i+1}^l$ such that

$$X_{j,i+1}^l := \begin{cases} 1 & \text{if } q^l \text{ fall in } \mathcal{B}(r_{i+1}), \\ 0 & \text{otherwise.} \end{cases}$$

By the rule of the sample reuse algorithm,

$$N = \mathbf{n}_{i+1} + \sum_{j=1}^{i} \sum_{l=1}^{\mathbf{n}_j} X_{j,i+1}^l.$$

Thus for $i = 1, \ldots, l-1$,

$$\begin{aligned} \mathcal{E}\left[\mathbf{n}_{i+1}\right] &= N - \sum_{j=1}^{i} \mathcal{E}\left[\sum_{l=1}^{\mathbf{n}_j} X_{j,i+1}^l\right] \\ &= N - \sum_{j=1}^{i} \sum_{n\in\Omega_{\mathbf{n}_j}} \sum_{l=1}^{n} \mathcal{E}\left[X_{j,i+1}^l \mid \mathbf{n}_j = n\right] \Pr\{\mathbf{n}_j = n\}, \end{aligned}$$

where $\Omega_{\mathbf{n}_j}$ denotes the sample space of $\mathbf{n}_j$. Since q^l is a random variable with uniform distribution over $\mathcal{B}(r_i)$, it follows from Lemma 5 that

$$\mathcal{E}[X_{j,i+1}^l \mid \mathbf{n}_j = n] = \frac{\mathrm{vol}(\mathcal{B}(r_{i+1}))}{\mathrm{vol}(\mathcal{B}(r_j))} = \left(\frac{r_{i+1}}{r_j}\right)^d.$$

Therefore,

$$\begin{aligned} \mathcal{E}\left[\mathbf{n}_{i+1}\right] &= N - \sum_{j=1}^{i} \sum_{n\in\Omega_{\mathbf{n}_j}} n \left(\frac{r_{i+1}}{r_j}\right)^d \Pr\{\mathbf{n}_j = n\} \\ &= N - \sum_{j=1}^{i} \left(\frac{r_{i+1}}{r_j}\right)^d \sum_{n\in\Omega_{\mathbf{n}_j}} n \Pr\{\mathbf{n}_j = n\} \end{aligned}$$

$$= N - \sum_{j=1}^{i} \left(\frac{r_{i+1}}{r_j}\right)^d \mathcal{E}[\mathbf{n}_i]. \qquad \square$$

LEMMA 7. *For* $i = 2, \cdots, l$,

$$\mathcal{E}[\mathbf{n}_i] = N - N\left(\frac{r_i}{r_{i-1}}\right)^d.$$

Proof. We use induction. Obviously,

$$\mathcal{E}[\mathbf{n}_1] = N.$$

By Lemma 6, we get

$$\mathcal{E}[\mathbf{n}_2] = N - N\left(\frac{r_2}{r_1}\right)^d.$$

Suppose it is true that

$$\mathcal{E}[\mathbf{n}_i] = N - N\left(\frac{r_i}{r_{i-1}}\right)^d.$$

Then

$$\begin{aligned}
\sum_{j=1}^{i} \left(\frac{r_{i+1}}{r_j}\right)^d \mathcal{E}[\mathbf{n}_j] &= \sum_{j=1}^{i} \left(\frac{r_{i+1}}{r_j}\right)^d \left[N - N\left(\frac{r_j}{r_{j-1}}\right)^d\right] \\
&= \sum_{j=1}^{i} \left[N\left(\frac{r_{i+1}}{r_j}\right)^d - N\left(\frac{r_{i+1}}{r_{j-1}}\right)^d\right] \\
&= N\left(\frac{r_{i+1}}{r_i}\right)^d.
\end{aligned}$$

It follows from Lemma 6 that

$$\begin{aligned}
\mathcal{E}[\mathbf{n}_{i+1}] &= N - \sum_{j=1}^{i} \left(\frac{r_{i+1}}{r_j}\right)^d \mathcal{E}[\mathbf{n}_j] \\
&= N - N\left(\frac{r_{i+1}}{r_i}\right)^d.
\end{aligned}$$

The proof of Lemma 7 is thus completed by induction. $\square$

Now we are in the position to prove Theorem 1. By Lemmas 6 and 7, we have

$$\mathcal{E}\left[\sum_{i=1}^{l} \mathbf{n}_i\right] = N + \sum_{i=2}^{l} \left[N - N\left(\frac{r_i}{r_{i-1}}\right)^d\right] = Nl - N \sum_{i=2}^{l} \left[1 - \left(\frac{r_i}{r_{i-1}}\right)^d\right].$$

Therefore,

$$\begin{aligned}
\mathcal{F}_{reuse} &= \frac{Nl}{\mathcal{E}[\sum_{i=1}^{l} \mathbf{n}_i]} \\
&= \frac{l}{l - \sum_{i=2}^{l} \left(\frac{r_i}{r_{i-1}}\right)^d}
\end{aligned}$$

and thus the proof of Theorem 1 is completed. $\square$

REFERENCES

[1] E. W. Bai, R. Tempo, and M. Fu, *Worst-case properties of the uniform distribution and randomized algorithms for robustness analysis*, Math. Control Signals Systems, 11 (1998), pp. 183–196.

[2] B. R. Barmish, C. M. Lagoa, and R. Tempo, *Radially truncated uniform distributions for probabilistic robustness of control systems*, in Proceedings of American Control Conference, Albuquerque, NM, American Automatic Control Council, 1997, pp. 853–857.

[3] B. R. Barmish, *A probabilistic result for a multilinearly parameterized H_∞ norm*, in Proceedings of American Control Conference, Chicago, IL, American Automatic Control Council, 2000, pp. 3309–3310.

[4] B. R. Barmish and B. T. Polyak, *A new approach to open robustness problems based on probabilistic predication formulae*, in Proceedings of the IFAC World Congress, Vol. H, San Francisco, CA, International Federation of Automatic Control, 1996, pp. 1–6.

[5] B. R. Barmish and C. M. Lagoa, *The uniform distribution: A rigorous justification for its use in robustness analysis*, Math. Control Signals Systems, 10 (1997), pp. 203–222.

[6] X. Chen and K. Zhou, *Order statistics and probabilistic robust control*, Systems Control Lett., 35 (1998), pp. 175–182.

[7] X. Chen and K. Zhou, *Constrained robustness analysis and synthesis by randomized algorithms*, IEEE Trans. Automat. Control, 45 (2000), pp. 1180–1186.

[8] G. Calafiore, F. Dabbene, and R. Tempo, *Randomized algorithms for probabilistic robustness with real and complex structured uncertainty*, IEEE Trans. Automat. Control, 45 (2000), pp. 2218–2235.

[9] G. Calafiore, F. Dabbene, and R. Tempo, *Radial and uniform distributions in vector and matrix spaces for probabilistic robustness*, in Topics in Control and Its Applications, D. Miller and L. Qiu, eds., Springer-Verlag, London, 1999, pp. 17–31.

[10] C. J. Clopper and E. S. Pearson, *The use of confidence or fiducial limits illustrated in the case of the binomial*, Biometrika, 26 (1934), pp. 404–413.

[11] R. R. De Gaston and M. G. Safonov, *Exact calculation of the multiloop stability margin*, IEEE Trans. Automat. Control, 33 (1988), pp. 156–171.

[12] P. P. Khargonekar and A. Tikku, *Randomized algorithms for robust control analysis and synthesis have polynomial complexity*, in Proceedings of the 35th Conference on Decision and Control, Kobe, Japan, IEEE Control Systems Society, 1996, pp. 3470–3475.

[13] C. M. Lagoa, *Probabilistic enhancement of classic robustness margins: A class of nonsymmetric distributions*, in Proceedings of American Control Conference, Chicago, IL, American Automatic Control Council, 2000, pp. 3802–3806.

[14] C. Marrison and R. F. Stengel, *Robust control system design using random search and genetic algorithms*, IEEE Trans. Automat. Control, 42 (1997), pp. 835–839.

[15] S. R. Ross and B. R. Barmish, *Distributionally robust gain analysis for systems containing complexity*, in Proceedings of the 40th Conference on Decision and Control, Orlando, FL, IEEE Control Systems Society, 2001, pp. 5020–5025.

[16] L. R. Ray and R. F. Stengel, *A Monte Carlo approach to the analysis of control system robustness*, Automatica J. IFAC, 3 (1993), pp. 229–236.

[17] R. F. Stengel and L. R. Ray, *Stochastic robustness of linear time-invariant systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 82–87.

[18] B. T. Polyak and P. S. Scherabakov, *Random spherical uncertainty in estimation and robustness*, IEEE Trans. Automat. Control, 45 (2000), pp. 2145–2150.

[19] R. Tempo, E. W. Bai, and F. Dabbene, *Probabilistic robustness analysis: Explicit bounds for the minimum number of samples*, Systems Control Lett., 30 (1997), pp. 237–242.

[20] R. Tempo and F. Dabbene, *Probabilistic robustness analysis and design of uncertain systems*, in Dynamical Systems, Control, Coding, Computer Vision, Progr. Systems Control Theory 25, G. Picci, ed., Birkhäuser, Basel, 1999, pp. 263–282.

[21] M. Vidyasagar and V. D. Blondel, *Probabilistic solutions to NP-hard matrix problems*, Automatica J. IFAC, 37 (2001), pp. 1597–1405.

[22] M. Vidyasagar, *Randomized algorithms for robust controller synthesis using statistical learning theory*, Automatica J. IFAC, 37 (2001), pp. 1515–1528.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 1972–1993

STOCHASTIC FOREST STAND VALUE AND OPTIMAL TIMBER HARVESTING*

LUIS H. R. ALVAREZ†

Abstract. We consider a Faustmann timber harvesting problem arising in the literature on rational forest management by modeling the value of the harvested resource as a time homogeneous, regular, and linear diffusion. We state a set of easily verifiable general conditions under which the existence and uniqueness of an optimal cutting value and, consequently, an optimal impulse control, are guaranteed. We also present a set of conditions under which increased volatility increases both the value and the optimal harvesting threshold at which the irreversible harvesting strategy is exercised.

Key words. optimal rotation, impulse control, recursive optimal stopping, diffusions, minimal r-excessive mappings

AMS subject classifications. 93E20, 60G40, 60J65, 49J40, 35R35

DOI. 10.1137/S0363012901393456

1. Introduction. Assume that a forest stand value is subject to a given potentially state-dependent and random growth rate (i.e., the growth dynamics are stochastic). Assume also that the discount rate (i.e., the rate of intertemporal preference) is a known exogenously determined constant. Given these assumptions, consider now the following problem: When should the forest stand be optimally harvested? Put somewhat differently, what is the optimal rotation period of the forest stand? This problem, which is known as the optimal rotation problem or the so-called *Faustmann problem* (or the *ongoing rotation problem*; cf. [27]) is probably the best known impulse control problem in natural resource economics (see, for example, [9], [14], [15], [17], [19], [23], [26], [27], [31], [32], [33], [34], and [36]). The Faustmannian tradition relies on the net present value (NPV) of all future harvests from the present up to a potentially infinite future, the relevance of which has been postulated on the seminal study [28], where the economic foundations of different approaches to the optimal rotation problem were also discussed.

While attempts to both characterize and solve this general stochastic rotation problem have been made (cf. [15], [27], [32], [34], and [35]), nobody has yet established easily verifiable sufficient conditions under which the general optimal rotation problem is solvable. This is somewhat surprising especially in light of the vast literature on the classical theory of linear diffusions and its representation theorems of r-excessive mappings and Markovian functionals in general (see, for example, [10], [18], and [20]). Moreover, given the recent studies emphasizing the mean reverting nature of commodity prices (cf. [16], [17], [30], and [34]) it is clear that a general analysis valid for a broad class of stochastic models of the value of a forest stand could provide us with valuable information on the general properties of the optimal rotation policy and its value.

Motivated by these arguments, we plan to consider in this study the general

*Received by the editors August 8, 2001; accepted for publication (in revised form) July 2, 2003; published electronically January 22, 2004. This research was supported by the Foundation for the Promotion of the Actuarial Profession, the Finnish Insurance Society, and the Yrjö Jahnsson Foundation.

http://www.siam.org/journals/sicon/42-6/39345.html

†Department of Economics, Quantitative Methods in Management, Turku School of Economics and Business Administration, FIN-20500 Turku, Finland (luis.alvarez@tukkk.fi).

stochastic optimal ongoing rotation problem when the value of the forest stand is assumed to evolve according to a general time homogeneous, regular, and linear diffusion. By relying on a combination of the classical theory of diffusions, stochastic calculus, and ordinary nonlinear programming, *we first state, in terms of the growth rate of the value of the forest stand, the reforestation costs, and the discount rate, a set of easily verifiable conditions under which the considered stochastic impulse control problem is solvable* (see [1], [2], [4], and [5] for a similar approach to singular control and [7] and [8] for a similar approach to optimal stopping). In economic terms, our condition essentially requires that the appreciation rate of the NPV of the forest stand has to be declining and negative for sufficiently high values of the forest stand. Since this condition is satisfied by all mean reverting diffusion models for the value of a forest stand, we find that our conditions imply that the ongoing rotation problem is always solvable in such a case. *We then present an algebraic equation from which the unique optimal cutting value can be determined.* Interestingly, and in line with previous studies of optimal stopping and singular stochastic control, a direct implication of this equation is that the smooth-fit principle can be viewed as an ordinary first order necessary condition for an optimum (cf. [1], [2], [5], [8], and [11]). Given the optimal cutting value, we then derive the expected cumulative present value of the future harvests from the present up to an arbitrarily distant future in terms of the increasing minimal r-excessive mapping for the controlled diffusion. In this way, we generalize previous results obtained in studies relying on explicit diffusion models for the value growth of the forest stand (cf. [15], [27], [32], [34], and [35]). Given our general results, we are able to present *a representation for the Faustmann formula in terms of the underlying diffusion process characterizing the value growth of the forest stand.* In line with previous studies of the optimal rotation problem (see, for example, [9], [32], and [34]), we also establish a close connection between the optimal ongoing rotation problem and an associated optimal stopping problem, which is also known as the *Wicksellian rotation problem* (i.e., the *single rotation problem* where the cutting decision is viewed as a once-and-for-all strategy; cf. [27]). In line with the results obtained in explicit models, we are able to prove that *the optimal cutting value in the single rotation case dominates the optimal cutting value in the infinite rotation case.* Consequently, we are able to confirm that *the required exercise premium is typically higher in the single rotation model than in the ongoing rotation model.* Moreover, we also consider the impact of increased volatility on the value of the optimal rotation problem and show that typically *increased volatility increases the expected cumulative present value of the future harvests and expands the set where harvesting is suboptimal and thus is postponed into the future.* Consequently, we are able to confirm that increased volatility increases the required exercise premium of a rational harvester and therefore postpones the optimal exercise of the irreversible policy.

It is at this point worth mentioning that there are many different approaches to the optimal rotation problem in the presence of uncertainty (for a survey of deterministic models, see [28]; see also [14] and [19]). In the studies [9] and [31], the stumpage price process is assumed to be stochastically fluctuating but differentiable with respect to time. Consequently, the nondifferentiability of diffusion processes is avoided and both the optimal policy and its value are derived by relying on standard optimization techniques. One of the major advantages of that approach is that it can be applied to establishing that the relationship between increased uncertainty and the optimal rotation period is positive (by relying on arguments first derived by Sandmo in [29]). However, it also simultaneously limits the class of stochastic processes which

can actually be applied for such a modeling effort, since one is left with the problem of finding a stochastic model satisfying the required differentiability properties. The analysis of the above studies was later extended to the diffusion case in, among others, [33] and [36]. However, in those studies the rotation problem is considered by relying on a discrete approximation of the original continuous time diffusion model (the authors rely on the Cox–Ross–Rubinstein-type binomial approximation of a diffusion). In this way the optimal policy and its value can be derived numerically by relying on standard recursive dynamic programming techniques. Another approach to the rotation problem relies on option valuation techniques and models the value of the optimal policy as a European forward contract written on a dividend paying asset and with known maturity (the rotation length). For studies relying on this approach and its variants, see, for example, [15], [17], [26], and [27]. Finally, the authors of [35] consider the optimal rotation problem under the assumption that both the timber volume and the timber price evolve according to two potentially correlated geometric Brownian motions. They consider the rotation strategies under different criteria and present a comparison of the various policies.

The contents of this study are as follows. In section 2 we present the considered optimal rotation problem and prove a set of useful results needed for the verification of optimality and the analysis of the comparative static properties of the value. In section 3 we then solve the considered optimal impulse control problem and present an associated optimal stopping problem when the increasing minimal r-excessive mapping for the controlled diffusion is convex. Section 4 extends our results to the case where the increasing minimal r-excessive mapping for the controlled diffusion is not necessarily globally convex. Our results are then explicitly illustrated in section 5 in a model based on logistic growth subject to a stochastic intrinsic growth rate. Finally, some concluding comments are presented in section 6.

2. The optimal rotation problem. Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ be a filtered probability space, and assume that the dynamics of the controlled diffusion process (i.e., the *stochastic forest stand value growth*) are described up to extinction by the generalized Itô equation (cf. [34])

$$X_t^\nu = x + \int_0^t \mu(X_s^\nu)ds + \int_0^t \sigma(X_s^\nu)dW_s - \sum_{\tau_k \leq t} \zeta_k, \quad 0 \leq t \leq \tau^\nu(0), \tag{2.1}$$

where $\tau^\nu(0) = \inf\{t \geq 0 : X_t^\nu \leq 0\} \leq \infty$ denotes the possibly finite first exit time from the state-space $(0, \infty)$ (i.e., *the extinction date*) and $\mu : \mathbb{R}_+ \mapsto \mathbb{R}$ and $\sigma : \mathbb{R}_+ \mapsto \mathbb{R}_+$ are known sufficiently smooth (at least continuous) mappings for guaranteeing the existence of a solution for (2.1). In order to avoid interior singularities, we also assume that $\sigma(x) > 0$ for all $x \in (0, \infty)$. As in [34], a *cutting strategy* for the system (2.1) is a possibly finite sequence

$$\nu = (\tau_1, \tau_2, \ldots, \tau_k, \ldots; \zeta_1, \zeta_2, \ldots, \zeta_k, \ldots)_{k\leq N} \quad (N \leq \infty),$$

where $\{\tau_k\}_{k\leq N}$ is an increasing sequence of $\mathcal{F}_t$-stopping times (known as the *cutting times*) for which $\tau_1 \geq 0$, and $\{\zeta_k\}_{k\leq N}$ denote a sequence of nonnegative impulses (i.e., $\zeta_k \geq 0$ for all $k \leq N$, which can be interpreted as the *cutting values*) exerted at the corresponding intervention times $\{\tau_k\}_{k\leq N}$, respectively. In line with previous studies of the optimal rotation problem, we assume that whenever the cutting decision is made, the system is instantaneously driven to a known state $x_0 \in \mathbb{R}_+$. More precisely, if the forest stand is cut when the system is in the state $y \in \mathbb{R}_+$, it is instantaneously

driven to the new lower state $y-(y-x_0)=x_0$ (i.e., $y-x_0$ is the size of the impulse). Consequently, *given the optimal threshold value $y^*_{x_0}$ at which the irreversible cutting decision is made, the subsequent value of the implemented impulse control is always a known constant.* That is, if the initial state of the system is above the optimal cutting value $y^*_{x_0}$, then the value of the first impulse is $(x-x_0)^+$. The size of the subsequent impulses is then always a constant $y^*_{x_0}-x_0$ (i.e., $\zeta_k=y^*_{x_0}-x_0$ for $k\geq 2$). We denote as $\mathcal{V}$ the class of *admissible impulse controls* $\nu=(\tau_1,\tau_2,\ldots,\ldots;\zeta_1,\zeta_2,\ldots)_{k\leq N}$ and assume that $\tau_k\to\tau^\nu(0)$ almost surely for all $\nu\in\mathcal{V}$ and $x\in\mathbb{R}_+$. In accordance with the economic and biologic literature considering impulse control models, we assume that the upper boundary ∞ is *natural for the controlled diffusion in the absence of regulation.* Since the unregulated diffusion dominates the controlled diffusion almost surely, we find that the controlled diffusion is never expected to become arbitrarily great in finite time. If the lower boundary is regular for the unregulated diffusion, we assume that it is killing (in line with the concept of *extinction*). As usual, we denote as

$$\mathcal{A}=\frac{1}{2}\sigma^2(x)\frac{d^2}{dx^2}+\mu(x)\frac{d}{dx}$$

the differential operator representing the infinitesimal generator of the controlled diffusion. Given the stochastic forest stand value growth described in (2.1) and our assumptions, define *the expected cumulative NPV of all future harvests from the present up to a potentially infinite future* as

$$J^\nu(x)=\mathbf{E}_x\left[\sum_{k=1}^{N}e^{-r\tau_k}(X^\nu_{\tau_k}-c)\right], \tag{2.2}$$

where $c>0$ denotes the reforestation cost of the forest stand. Given the definition of $J^\nu(x)$, we consider the optimal rotation problem

$$V(x)=\sup_{\nu\in\mathcal{V}}J^\nu(x),\quad x\in\mathbb{R}_+, \tag{2.3}$$

and to determine an admissible cutting policy $\nu^*\in\mathcal{V}$ for which

$$J^{\nu^*}(x)=V(x),\quad x\in\mathbb{R}_+.$$

That is, we plan to consider and determine the rotation policy maximizing the expected cumulative NPV of the future harvests from the present up to a potentially arbitrarily distant future.

Denote now as X_t the controlled diffusion in the absence of interventions. For the sake of boundedness, we will assume throughout this study that

$$\mathbf{E}_x[e^{-rt}(X_t-c);t<\tau(0)]<\infty$$

for all $(t,x)\in\mathbb{R}^2_+$. Given our assumptions above, we now define the value of the associated single rotation (Wicksellian) problem as

$$\bar{V}(x)=\sup_{\tau}\mathbf{E}_x\left[e^{-r\tau}(X_\tau-c)\right], \tag{2.4}$$

where τ is an arbitrary $\mathcal{F}_t$-stopping time.

In line with the standard literature on diffusion processes and their resolvent operators, we denote as $\mathcal{L}^1(\mathbb{R}_+)$ the class of measurable mappings $f : \mathbb{R}_+ \mapsto \mathbb{R}$ satisfying the uniform integrability condition

$$\mathbf{E}_x \int_0^{\tau(0)} e^{-rs} |f(X_s)| ds < \infty,$$

where $\tau(0) = \inf\{t \geq 0 : X_t \leq 0\}$. Given the class $\mathcal{L}^1(\mathbb{R}_+)$, we define for $f \in \mathcal{L}^1(\mathbb{R}_+)$ the (resolvent) functional $(R_r f) : \mathbb{R}_+ \mapsto \mathbb{R}$ as

$$(R_r f)(x) = \mathbf{E}_x \int_0^{\tau(0)} e^{-rs} f(X_s) ds.$$

As is well known from the literature on linear diffusions, if $f \in \mathcal{L}^1(\mathbb{R}_+)$, then

$$(R_r f)(x) = B^{-1}\varphi(x) \int_0^x \psi(y) f(y) m'(y) dy + B^{-1}\psi(x) \int_x^\infty \varphi(y) f(y) m'(y) dy,$$

where $\psi(x)$ denotes the increasing and $\varphi(x)$ the decreasing fundamental solution of the ordinary second order differential equation $(\mathcal{A}u)(x) = ru(x)$ (defined on the domain of the operator of the diffusion $\{X_t; t \in [0, \tau(0))\}$; see [10, pp. 18–20] for a thorough characterization of the fundamental solutions and the Green function of a linear diffusion), $B = \frac{\psi'(x)}{S'(x)}\varphi(x) - \frac{\varphi'(x)}{S'(x)}\psi(x) > 0$ denotes the constant (with respect to the scale) Wronskian determinant,

$$S'(x) = \exp\left(-\int^x \frac{2\mu(y)}{\sigma^2(y)} dy\right)$$

denotes the density of the scale function S of X, and

$$m'(x) = \frac{2}{\sigma^2(x) S'(x)}$$

denotes the density of the speed measure m of X. It is worth pointing out that since ∞ was assumed to be natural, we know that $\lim_{x\to\infty} \psi(x) = \infty$, $\lim_{x\to\infty} \psi'(x)/S'(x) = \infty$, $\lim_{x\to\infty} \varphi(x) = 0$, and $\lim_{x\to\infty} \varphi'(x)/S'(x) = 0$. It is also worth pointing out that the fundamental solutions $\psi(x)$ and $\varphi(x)$ constitute the minimal r-excessive mappings for the underlying diffusion $\{X_t; t \in [0, \tau(0))\}$ in the sense that any nontrivial r-excessive mapping for $\{X_t; t \in [0, \tau(0))\}$ can be expressed in terms of these mappings (an integral representation of r-excessive mappings; cf. [10, p. 32]). Before proceeding any further in our analysis, we first prove the following.

LEMMA 2.1. *Assume that $\theta \in \mathcal{L}^1(\mathbb{R}_+)$, where $\theta : \mathbb{R}_+ \mapsto \mathbb{R}$ is defined as $\theta(x) = \mu(x) - rx$ (the net value appreciation rate of X). Then, for all $x \in \mathbb{R}_+$ we have that*

$$\psi''(x) = \frac{2S'(x)}{\sigma^2(x)} \left[r \int_0^x \psi(y)\theta(y) m'(y) dy - \theta(x) \frac{\psi'(x)}{S'(x)} \right]. \tag{2.5}$$

In particular, if 0 is unattainable for X, then

$$\psi''(x) = \frac{2rS'(x)}{\sigma^2(x)} \int_0^x \psi(y)(\theta(y) - \theta(x)) m'(y) dy. \tag{2.6}$$

Proof. As in [3] and [6], we find by applying Dynkin's theorem to the identity mapping $x \mapsto x$ that for all $x \in (a,b) \subseteq (0,\infty)$ we have that

$$\mathbf{E}_x\left[e^{-r\tau^*}X_{\tau^*}\right] = x + \mathbf{E}_x\int_0^{\tau^*} e^{-rs}\theta(X_s)ds, \tag{2.7}$$

where $\tau^* = \inf\{t \geq 0 : X_t \notin (a,b)\}$ denotes the first exit time of the diffusion X from the set (a,b). It is well known from the classical theory of diffusions that (2.7) can be rewritten as

$$a\frac{\hat{\varphi}(x;b)}{\hat{\varphi}(a;b)} + b\frac{\hat{\psi}(x;a)}{\hat{\psi}(b;a)} = x + \int_a^b \tilde{G}(x,y)\theta(y)m'(y)dy, \tag{2.8}$$

where

$$\tilde{G}(x,y) = \begin{cases} \tilde{B}^{-1}\hat{\varphi}(x;b)\hat{\psi}(y;a), & x \geq y, \\ \tilde{B}^{-1}\hat{\varphi}(y;b)\hat{\psi}(x;a), & x \leq y, \end{cases}$$

denotes the Green function of the diffusion X killed at a and b,

$$\hat{\psi}(x;a) = \psi(x) - \frac{\psi(a)}{\varphi(a)}\varphi(x),$$

$$\hat{\varphi}(x;b) = \varphi(x) + \frac{\varphi(b)}{\psi(b)}\psi(x),$$

and $\tilde{B} = B\varphi(a)/\hat{\varphi}(a;b) = B\psi(b)/\hat{\psi}(b;a)$ denotes the Wronskian of the fundamental solutions $\hat{\psi}(x;a)$ and $\hat{\varphi}(x;b)$. Standard differentiation of (2.8) now yields that

$$\begin{aligned} a\frac{\hat{\varphi}'(x;b)}{\hat{\varphi}(a,;b)} + b\frac{\hat{\psi}'(x;a)}{\hat{\psi}(b;a)} &= 1 + \tilde{B}^{-1}\hat{\varphi}'(x;b)\int_a^x \hat{\psi}(y;a)\theta(y)m'(y)dy \\ &\quad + \tilde{B}^{-1}\hat{\psi}'(x;a)\int_x^b \hat{\varphi}(y;b)\theta(y)m'(y)dy. \end{aligned}$$

Dividing this equation first with the term $\hat{\psi}'(x;a)$ and reordering terms then yield

$$\begin{aligned} \frac{1}{\hat{\psi}'(x;a)} &= \frac{a\hat{\varphi}'(x;b)}{\hat{\psi}'(x;a)\hat{\varphi}(a;b)} + \frac{b}{\hat{\psi}(b;a)} - \tilde{B}^{-1}\frac{\hat{\varphi}'(x;b)}{\hat{\psi}'(x;a)}\int_a^x \hat{\psi}(y;a)\theta(y)m'(y)dy \\ &\quad - \tilde{B}^{-1}\int_x^b \hat{\varphi}(y;b)\theta(y)m'(y)dy. \end{aligned}$$

Differentiating this equation with respect to x and observing that

$$\frac{d}{dx}\left[\frac{\hat{\varphi}'(x;b)}{\hat{\psi}'(x;a)}\right] = \frac{2r\tilde{B}S'(x)}{\sigma^2(x)\hat{\psi}'^2(x;a)}$$

then yield

$$\frac{\hat{\psi}''(x;a)}{\hat{\psi}'^2(x;a)} = \frac{2S'(x)}{\sigma^2(x)\hat{\psi}'^2(x;a)}\left[r\int_a^x \hat{\psi}(y;a)\theta(y)m'(y)dy - \theta(x)\frac{\hat{\psi}'(x;a)}{S'(x)} - \frac{arB}{\varphi(a)}\right]. \tag{2.9}$$

Equation (2.5) then follows from (2.9) after we notice that $\lim_{a\downarrow 0}\hat{\psi}(x;a) = \psi(x)$, $\lim_{a\downarrow 0}\hat{\psi}''(x;a) = \psi''(x)$, and $\lim_{a\downarrow 0} a/\varphi(a) = 0$ and invoke the assumption $\theta \in \mathcal{L}^1(\mathbb{R}_+)$. Equation (2.6) then follows from (2.5) after we notice that

$$\frac{\psi'(x)}{S'(x)} = r\int_0^x \psi(y)m'(y)dy$$

whenever 0 is unattainable for X. □

Lemma 2.1 presents an integral representation of the second derivative of the increasing fundamental solution $\psi(x)$. Two key consequences of this representation are now summarized in the following corollary.

COROLLARY 2.2. (A) *Assume that* 0 *is unattainable (i.e., either natural or entrance) for* X, *that* $\theta \in \mathcal{L}^1(\mathbb{R}_+)$, *and that there is a threshold* $\tilde{x} \in [0,\infty]$ *such that* $\mu(x) \le 0$ *for* $x \ge \tilde{x}$ *and* $\theta(x)$ *is nonincreasing on* $(0,\tilde{x})$. *Then* $\psi(x)$ *is strictly convex on* $\mathbb{R}_+$.

(B) *Assume that* 0 *is attainable (i.e., either regular or exit) for* X, *that* $\theta \in \mathcal{L}^1(\mathbb{R}_+)$, *that* $\lim_{x\downarrow 0}\mu(x) \le 0$, *and that there is a threshold* $\tilde{x} \in [0,\infty]$ *such that* $\mu(x) \le 0$ *for* $x \ge \tilde{x}$ *and* $\theta(x)$ *is nonincreasing on* $(0,\tilde{x})$. *Then* $\psi(x)$ *is strictly convex on* $\mathbb{R}_+$.

Proof. (A) Assume now that 0 is unattainable for X and that there is a threshold $\tilde{x} \in [0,\infty]$ such that $\mu(x) \le 0$ for $x \ge \tilde{x}$ and $\theta(x)$ is nonincreasing on $(0,\tilde{x})$. It is clear that the r-harmonicity of $\psi(x)$ implies that for all $x \in [\tilde{x},\infty)$ we have

$$\frac{1}{2}\sigma^2(x)\psi''(x) = r\psi(x) - \mu(x)\psi'(x) > 0,$$

since $\psi(x) > 0$, $\psi'(x) > 0$, and $\mu(x) \le 0$ on $[\tilde{x},\infty)$. On the other hand, the assumed monotonicity of $\theta(x)$ and (2.6) implies that $\psi''(x) > 0$ on $\mathbb{R}_+$, that is, that $\psi(x)$ is strictly convex on $\mathbb{R}_+$.

(B) As in the case of part (A), $\psi(x)$ is strictly convex on $[\tilde{x},\infty)$. Define the functional $I : \mathbb{R}_+ \mapsto \mathbb{R}$ as

$$I(x) = r\int_0^x \psi(y)\theta(y)m'(y)dy - \theta(x)\frac{\psi'(x)}{S'(x)}.$$

It is clear that

$$\lim_{x\downarrow 0} I(x) = -\lim_{x\downarrow 0}\mu(x)\frac{\psi'(x)}{S'(x)} \ge 0,$$

since $\psi'(x)/S'(x) > 0$ and $\lim_{x\downarrow 0}\mu(x) \le 0$. Assume now that $0 < x_1 < x_2 < \tilde{x}$. Then

$$\begin{aligned} I(x_2) - I(x_1) &= r\int_{x_1}^{x_2} \psi(y)\theta(y)m'(y)dy + \theta(x_1)\frac{\psi'(x_1)}{S'(x_1)} - \theta(x_2)\frac{\psi'(x_2)}{S'(x_2)} \\ &\ge r\theta(x_2)\int_{x_1}^{x_2} \psi(y)m'(y)dy + \theta(x_1)\frac{\psi'(x_1)}{S'(x_1)} - \theta(x_2)\frac{\psi'(x_2)}{S'(x_2)} \\ &= (\theta(x_1) - \theta(x_2))\frac{\psi'(x_1)}{S'(x_1)} \ge 0, \end{aligned}$$

proving that $I(x)$ is nondecreasing on the set where $\theta(x)$ is nonincreasing. Consequently, we find that $\psi''(x) > 0$ on $\mathbb{R}_+$, that is, that $\psi(x)$ is strictly convex on $\mathbb{R}_+$. □

Corollary 2.2 states a set of conditions under which the increasing minimal r-excessive mapping for the controlled diffusion $\{X_t; t \in [0, \tau(0))\}$ is strictly convex on $\mathbb{R}_+$. As we will later observe, the strict convexity of the increasing fundamental solution is useful when proving the existence and uniqueness of the optimal cutting policy and its value. Another auxiliary result which is helpful in proving the existence of an optimal exercise threshold in both the single rotation (Wicksellian) and the ongoing rotation (Faustmannian) case is now summarized in the following (extending the results obtained in Proposition 1 in [8]).

LEMMA 2.3. *Assume that there is a threshold* $\check{x} \in \mathbb{R}_+$ *such that* $\theta(x) + rc \gtreqless 0$ *when* $x \lesseqgtr \check{x}$. *Then there is a threshold* $\bar{y} = \operatorname{argmax}\{\frac{(x-c)}{\psi(x)}\} > \max(\check{x}, c)$ *satisfying the ordinary first order condition*

$$\psi(\bar{y}) = \psi'(\bar{y})(\bar{y} - c). \tag{2.10}$$

Moreover,

$$\frac{d}{dx}\left[\frac{x-c}{\psi(x)}\right] \gtreqless 0, \quad x \lesseqgtr \bar{y}.$$

In particular, $\bar{\tau} = \inf\{t \geq 0 : X_t \notin (0, \bar{y})\}$ *is the optimal stopping time and*

$$\bar{V}(x) = \psi(x) \sup_{y \geq x}\left[\frac{y-c}{\psi(y)}\right] = \begin{cases} x - c, & x \geq \bar{y}, \\ \frac{\psi(x)}{\psi'(\bar{y})}, & x < \bar{y}. \end{cases} \tag{2.11}$$

Proof. Since the mapping $x - c$ is bounded on any bounded and open interval $(0, y) \subset \mathbb{R}_+$, Dynkin's theorem implies that for all $x \in \mathbb{R}_+$ we have that

$$\mathbf{E}_x\left[e^{-r\tau(0,y)}(X_{\tau(0,y)} - c)\right] = x - c + \mathbf{E}_x \int_0^{\tau(0,y)} e^{-rs}(\theta(X_s) + rc)ds, \tag{2.12}$$

where $\tau(0, y) = \inf\{t \geq 0 : X_t \notin (0, y)\}$ and $y \in (0, \infty)$. It is well known from the classical theory of diffusions that for $x \in (0, y)$ (2.12) can be rewritten as

$$\begin{aligned}(y-c)\frac{\psi(x)}{\psi(y)} - c\frac{\bar{\varphi}(x)}{\bar{\varphi}(0)} &= x - c + B^{-1}\bar{\varphi}(x)\int_0^x \psi(t)(rc + \theta(t))m'(t)dt \\ &\quad + B^{-1}\psi(x)\int_x^y \bar{\varphi}(t)(rc + \theta(t))m'(t)dt,\end{aligned} \tag{2.13}$$

where $\bar{\varphi}(x) = \varphi(x) - \varphi(y)\psi(x)/\psi(y)$. Dividing (2.13) now with $\psi(x)$ and differentiating then yield

$$\frac{d}{dx}\left[\frac{x-c}{\psi(x)}\right] = \frac{S'(x)}{\psi^2(x)}\left[\frac{cB}{\bar{\varphi}(0)} + \int_0^x \psi(y)(rc + \theta(y))m'(y)dy\right].$$

Letting y tend to infinity in $\bar{\varphi}(x)$ and simplifying then yield

$$\psi(x) - \psi'(x)(x-c) = S'(x)\left[\frac{cB}{\varphi(0)} + \int_0^x \psi(y)(rc + \theta(y))m'(y)dy\right].$$

Since $S'(x) > 0$, it is sufficient to consider the behavior of the functional

$$I(x) = \frac{cB}{\varphi(0)} + \int_0^x \psi(y)(rc + \theta(y))m'(y)dy.$$

It is clear that $I(x) > 0$ as long as $x \leq \check{x}$. If $x > K > \check{x}$, then the mean value theorem implies that

$$I(x) = I(K) + \int_K^x \psi(y)(rc + \theta(y))m'(y)dy = I(K) + \frac{rc + \theta(\xi)}{r}\left[\frac{\psi'(x)}{S'(x)} - \frac{\psi'(K)}{S'(K)}\right],$$

where $\xi \in (K, x)$. Since $rc + \theta(\xi) < 0$ and $\lim_{x\to\infty} \psi'(x)/S'(x) = \infty$, we find that $\lim_{x\to\infty} I(x) = -\infty$. Since $I(x)$ is continuous on $\mathbb{R}_+$ and monotonically decreasing on $(\check{x}, \infty)$, we find that there is a unique threshold $\bar{y}$ for which $I(\bar{y}) = 0$ and, consequently, for which $\psi(\bar{y}) = \psi'(\bar{y})(\bar{y} - c)$.

Denote now the proposed value function as $\bar{V}_p(x)$. Since $\bar{V}_p(x) = \mathbf{E}_x[e^{-r\bar{\tau}}(X(\bar{\tau}) - c)]$ and the stopping time in (2.4) is arbitrary, we immediately find that $\bar{V}(x) \geq \bar{V}_p(x)$. In order to prove the opposite inequality, we first observe that the proposed value function is continuously differentiable on $\mathbb{R}_+$ and twice continuously differentiable on $\mathbb{R}_+\backslash\{\bar{y}\}$ and dominates the exercise payoff $(x - c)$ for all $x \in \mathbb{R}_+$. Moreover, since $(\mathcal{A}\bar{V}_p)(x) = r\bar{V}_p(x)$ on $(0, \bar{y})$ and

$$(\mathcal{A}\bar{V}_p)(x) - r\bar{V}_p(x) = \theta(x) + rc < 0$$

on $(\bar{y}, \infty)$, we find that $\bar{V}_p(x)$ is an r-excessive majorant of the exercise payoff $x - c$. However, since $V(x)$ is the least of such majorants, we find that $\bar{V}(x) \leq \bar{V}_p(x)$, completing the proof of our theorem. □

Lemma 2.3 states a set of weak conditions under which the single rotation problem (2.4) is solvable and under which the mapping $\psi(x) - \psi'(x)(x - c)$ has a unique root on $(0, \infty)$. In economic terms, Lemma 2.3 essentially states that if there is a threshold above which the expected present value of the NPV of the project is decreasing (that is, if there is a threshold above which the futures price effect dominates the option effect; cf. [22]), then there is a unique threshold at which the harvesting opportunity should be irreversibly exercised. As we will later observe, this finding plays an important role in the determination of the optimal threshold in the ongoing rotation case as well.

3. The nonlinear programming approach. Having presented the considered optimal rotation problem and a set of auxiliary results, it is our purpose to now solve the stochastic impulse control problem (2.3) explicitly. It is now clear that if there is a state $y \in \mathbb{R}_+$ at which the diffusion is instantaneously driven to the lower state $x_0 \in (0, y)$ (a *suboptimal impulse control*), then for all $x < y$ we have that the value of such a cutting strategy reads as

$$\begin{aligned} F_y(x) &= \mathbf{E}_x\left[e^{-r\tau(0,y)}\left(X_{\tau(0,y)} - c + F_y(x_0)\right)\right] \\ &= (y - c + F_y(x_0))\,\frac{\psi(x)}{\psi(y)}, \end{aligned} \tag{3.1}$$

where $\tau(0, y) = \inf\{t \geq 0 : X_t \geq y\}$ and $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ denotes the increasing fundamental solution of the ordinary second order differential equation $(\mathcal{A}u)(x) = ru(x)$ subject to the boundary condition $\psi(0) = 0$ whenever 0 is a regular boundary for X_t. It is now clear from (3.1) that letting x tend to y yields the *value-matching condition*

$$F_y(y) = y - c + F_y(x_0). \tag{3.2}$$

Similarly, letting $x \downarrow x_0$ in (3.1) then yields that

$$F_y(x_0) = (y - c + F_y(x_0))\,\frac{\psi(x_0)}{\psi(y)},$$

implying that

$$F_y(x_0) = \frac{\psi(x_0)(y-c)}{\psi(y)-\psi(x_0)}. \tag{3.3}$$

Plugging (3.3) into (3.1) then yields that for all $x \in (0,y)$ we have that

$$F_y(x) = \frac{\psi(x)(y-c)}{\psi(y)-\psi(x_0)}, \tag{3.4}$$

which can be rewritten as

$$F_y(x) = \frac{\mathbf{E}_x[e^{-r\tau(y)}(X_{\tau(y)}-c)]}{1-\mathbf{E}_{x_0}[e^{-r\tau(y)}]},$$

where $\tau(y) = \inf\{t \geq 0 : X_t = y\}$. Consequently, we find that for all potentially suboptimal rotation policies described in the beginning of this section the expected cumulative present value of the future harvests from the present up to a potentially infinite future reads as

$$F_y(x) = \begin{cases} x - c + \psi(x_0)g(y), & x \geq y, \\ \psi(x)g(y), & x < y, \end{cases} \tag{3.5}$$

where

$$g(y) = \frac{(y-c)}{\psi(y)-\psi(x_0)}.$$

Standard differentiation now yields that

$$g'(y) = \frac{[\psi(y)-\psi(x_0)-\psi'(y)(y-c)]}{(\psi(y)-\psi(x_0))^2}. \tag{3.6}$$

Therefore, if a threshold $y^*_{x_0}$ maximizing the mapping $g(y)$ on the set (x_0,∞) exists, it has to satisfy the ordinary first order condition

$$\psi(y^*_{x_0}) = \psi(x_0) + \psi'(y^*_{x_0})(y^*_{x_0}-c). \tag{3.7}$$

Since

$$g''(y^*_{x_0}) = -\frac{\psi''(y^*_{x_0})}{\psi'(y^*_{x_0})(\psi(y^*_{x_0})-\psi(x_0))},$$

we find that $y^*_{x_0}$ can be a maximum of $g(y)$ on the set (x_0,∞) provided that the condition $\psi''(y^*_{x_0}) > 0$ is satisfied (i.e., the increasing fundamental solution has to be locally convex at the optimal threshold). Moreover, if a threshold $y^*_{x_0} \in (x_0,\infty)$ satisfying the necessary condition (3.7) exists, then $g(y^*_{x_0}) = 1/\psi'(y^*_{x_0})$ and, therefore,

$$F_{y^*_{x_0}}(x) = \begin{cases} x - c + \frac{\psi(x_0)}{\psi'(y^*_{x_0})}, & x \geq y^*_{x_0}, \\ \frac{\psi(x)}{\psi'(y^*_{x_0})}, & x < y^*_{x_0}. \end{cases} \tag{3.8}$$

It is at this point worth observing that dividing the necessary condition (3.7) with the term $\psi(y^*_{x_0})$ and reordering terms then yield

$$1 - \frac{\psi(x_0)}{\psi(y^*_{x_0})} = \frac{\psi'(y^*_{x_0})}{\psi(y^*_{x_0})}(y^*_{x_0}-c).$$

Since $x_0 < y^*_{x_0}$ and, consequently, $\mathbf{E}_{x_0}[e^{-r\tau(y^*_{x_0})}] = \psi(x_0)/\psi(y^*_{x_0})$, we observe that the equation above can be rewritten as

$$1 - \mathbf{E}_{x_0}[e^{-r\tau(y^*_{x_0})}] = \frac{\psi'(y^*_{x_0})}{\psi(x_0)}\mathbf{E}_{x_0}[e^{-r\tau(y^*_{x_0})}(X_{\tau(y^*_{x_0})} - c)].$$

A simple algebraic manipulation of this equation then yields that

$$\frac{\mathbf{E}_{x_0}[e^{-r\tau(y^*_{x_0})}]}{\mathbf{E}_{x_0}[e^{-r\tau(y^*_{x_0})}(X_{\tau(y^*_{x_0})} - c)]}\frac{r\psi(y^*_{x_0})}{\psi'(y^*_{x_0})} = \frac{r}{1 - \mathbf{E}_{x_0}[e^{-r\tau(y^*_{x_0})}]}. \tag{3.9}$$

Invoking now the r-harmonicity of the fundamental solution $\psi(x)$ then yields that $r\psi(y^*_{x_0}) = \mu(y^*_{x_0})\psi'(y^*_{x_0}) + \frac{1}{2}\sigma^2(y^*_{x_0})\psi''(y^*_{x_0})$ and, therefore, that (3.9) can be rewritten as

$$\frac{\mathbf{E}_{x_0}[e^{-r\tau(y^*_{x_0})}B(X_{\tau(y^*_{x_0})})]}{\mathbf{E}_{x_0}[e^{-r\tau(y^*_{x_0})}(X_{\tau(y^*_{x_0})} - c)]} = \frac{r}{1 - \mathbf{E}_{x_0}[e^{-r\tau(y^*_{x_0})}]}, \tag{3.10}$$

where $B(x) = \mu(x) + \frac{1}{2}\sigma^2(x)\psi''(x)/\psi'(x) = \mu(x) + \frac{1}{2}\sigma^2(x)\frac{d}{dx}\ln\psi'(x)$. *Equation* (3.10) *is the stochastic version of the well-known Faustmann formula for the optimal rotation period* (cf. [34]). As is clear from (3.10), the major difference between the stochastic and the deterministic version of the Faustmann formula is the term $\frac{1}{2}\sigma^2(x)\frac{\psi''(x)}{\psi'(x)}$ measuring the *required risk premium of the decision maker*. It is also clear that the deterministic Faustmann formula for the optimal rotation period can be obtained from (3.10) simply by letting $\sigma(x) \downarrow 0$. Given the results of Corollary 2.2, we can now prove the following.

COROLLARY 3.1. *Assume that the conditions of Corollary* 2.2 *are met. Then*

$$\frac{\mathbf{E}_{x_0}[e^{-r\tau(y^*_{x_0})}\mu(X_{\tau(y^*_{x_0})})]}{\mathbf{E}_{x_0}[e^{-r\tau(y^*_{x_0})}(X_{\tau(y^*_{x_0})} - c)]} \leq \frac{r}{1 - \mathbf{E}_{x_0}[e^{-r\tau(y^*_{x_0})}]}. \tag{3.11}$$

Proof. Given the conditions of Corollary 2.2, $\psi(x)$ is strictly convex and, therefore, $B(x) \geq \mu(x)$ for all $x \in \mathbb{R}_+$. □

Although our analysis above demonstrates what an optimal policy has to satisfy whenever it exists, it is not clear whether an *optimal exercise threshold* (i.e., *the optimal cutting value*) exists. Fortunately, we can now prove the following.

LEMMA 3.2. *Assume that $x_0 < c$ and that the conditions of Corollary* 2.2 *and Lemma* 2.3 *are met. Then there is a unique threshold $y^*_{x_0} \in (c, \bar{y})$ satisfying* (3.7).

Proof. Consider now the mapping $f : \mathbb{R}_+ \mapsto \mathbb{R}$ defined as

$$f(y) = \psi(y) - \psi(x_0) - \psi'(y)(y - c).$$

Since $f'(y) = -\psi''(y)(y-c)$ and $\psi(x)$ is strictly convex given the conditions of Corollary 2.2, we find that $f'(y) \gtreqless 0$ when $y \lesseqgtr c$, implying that $c = \text{argmax}\{f(y)\}$, that is, that the mapping $f(y)$ attains its global maximum at c. Since $\psi(x)$ is increasing, we find that $f(c) = \psi(c) - \psi(x_0) > 0$. However, since

$$f(\bar{y}) = \psi(\bar{y}) - \psi(x_0) - \psi'(\bar{y})(\bar{y} - c) = -\psi(x_0) < 0$$

and $\psi(y) < \psi'(y)(y - c)$ for all $y > \bar{y}$ by Lemma 2.3, we find that the mapping $f(y)$ has a unique root $y^*_{x_0} \in (c, \bar{y})$. Combining this finding with (3.6) then implies that $y^*_{x_0} = \text{argmax}\{g(y)\}$. □

Lemma 3.2 essentially establishes that if the increasing fundamental solution is strictly convex, the conditions of Lemma 2.3 are met, and $x_0 < c$, then a unique threshold $y^*_{x_0}$ maximizing $g(y)$ always exists. It is worth noticing that if $x_0 > c$, then $f(c) = \psi(c) - \psi(x_0) < 0$ and, therefore, in that case the necessary condition (3.7) can never be satisfied. If $c = x_0$, then the equation $f(y) = 0$ has a unique root at c (as was noted in [34]). However, since $f(y)$ does not change sign in that case, we find that c constitutes an inflection point for $g(y)$ and, therefore, that no interior root maximizing $g(y)$ exists. It is worth pointing out that the results of Lemma 3.2 extend the results obtained in [34], since they show not only that an optimal threshold $y^*_{x_0}$ exists but that the threshold $y^*_{x_0}$ also maximizes the representation $F_y(x)$ on the *continuation set* $x \in (0, y)$ (i.e., in the *do-nothing region*). As we will observe in the subsequent analysis, the findings of Lemma 3.2 play a key role in the verification of the existence and uniqueness of an optimal cutting strategy. Before stating the main result of our study, we assume from now on that $x_0 < c$. Given this assumption, consider the *recursive optimal stopping problem*

$$H(x) = \sup_{\tau < \tau(0)} \mathbf{E}_x \left[e^{-r\tau} (X_\tau - c + H(x_0)) \right], \tag{3.12}$$

where τ is an arbitrary $\mathcal{F}_t$-stopping time satisfying the constraint $\tau < \tau(0)$. Since $H(x) \geq x - c + H(x_0)$ for all $x \in \mathbb{R}_+$, we find by letting $x \to x_0$ that especially $H(x_0) \geq x_0 - c + H(x_0)$, that is, that $x_0 \leq c$. Consequently, we find that the assumption $x_0 < c$ is actually a *consistency condition needed for the existence of a solution for the recursive stopping problem* (3.12). We can now prove the following.

LEMMA 3.3. *Assume that $x_0 < c$ and that the mapping $M : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is r-excessive for the diffusion $\{X_t; t \in [0, \tau(0))\}$ and satisfies the recursive condition $M(x) \geq x - c + M(x_0)$ for all $x \in \mathbb{R}_+$. Then*

$$M(x) \geq \sup_{\tau < \tau(0)} \mathbf{E}_x \left[e^{-r\tau} (X_\tau - c + M(x_0)) \right]$$

for all $x \in \mathbb{R}_+$.

Proof. Our assumptions imply that for all $x \in \mathbb{R}_+$ we have that

$$M(x) \geq \mathbf{E}_x \left[e^{-r\tau_n} M(X_{\tau_n}) \right] \geq \mathbf{E}_x \left[e^{-r\tau_n} (X_{\tau_n} - c + M(x_0)) \right],$$

where $\tau_n = n \wedge \inf\{t \geq 0 : X_t \notin (n^{-1}, n)\} \wedge \tau$ is an almost surely finite $\mathcal{F}_t$-stopping time. Letting $n \to \infty$ and invoking Fatou's lemma then yield

$$M(x) \geq \mathbf{E}_x \left[e^{-r(\tau(0) \wedge \tau)} (X_{\tau(0) \wedge \tau} - c + M(x_0)) \right].$$

Since this inequality is valid for an arbitrary stopping time, it must be valid for the optimal one as well and, therefore,

$$M(x) \geq \sup_{\tau < \tau(0)} \mathbf{E}_x \left[e^{-r\tau} (X_\tau - c + M(x_0)) \right]$$

for all $x \in \mathbb{R}_+$. □

Having stated the auxiliary Lemma 3.3, we are now in a position to state our main results on the optimal cutting strategy and its value. These results are now summarized in the following.

THEOREM 3.4. *Assume that $x_0 < c$ and that the conditions of Lemma 3.2 are met. Then the optimal stopping time is $\tau(0, y^*_{x_0}) = \inf\{t \geq 0 : X_t \geq y^*_{x_0}\}$ and the value of*

the optimal stopping problem (3.12) *reads as* $H(x) = F_{y^*_{x_0}}(x)$, *where* $F_{y^*_{x_0}}(x)$ *is defined as in* (3.8) *and* $y^*_{x_0}$ *is the optimal exercise threshold satisfying* (3.7) *and defined as in Lemma* 3.2. *Moreover, the optimal cutting value is* $y^*_{x_0}$ *and* $V(x) = H(x) = F_{y^*_{x_0}}(x)$ *for all* $x \in \mathbb{R}_+$.

Proof. It is now clear that since the stopping time in (3.12) is arbitrary and the threshold $y^*_{x_0}$ exists and is unique under the conditions of our theorem, we have that $H(x) \geq F_{y^*_{x_0}}(x)$. To prove the opposite inequality, we first observe that the proposed value function $F_{y^*_{x_0}}(x)$ is nonnegative, that $F_{y^*_{x_0}} \in C^1(\mathbb{R}_+) \cap C^2(\mathbb{R}_+ \setminus \{y^*_{x_0}\})$, and that $\infty > \lim_{x \uparrow y^*_{x_0}} F''_{y^*_{x_0}}(x) = \frac{\psi''(y^*_{x_0})}{\psi'(y^*_{x_0})} \geq 0 = \lim_{x \downarrow y^*_{x_0}} F''_{y^*_{x_0}}(x)$ (i.e., $F_{y^*_{x_0}}(x)$ is stochastically of class $C^2(\mathbb{R}_+)$; cf. [12], [13], and [24, pp. 215–216]). Since $F_{y^*_{x_0}}(x)$ satisfies the ordinary second order linear differential equation $(\mathcal{A}F_{y^*_{x_0}})(x) = rF_{y^*_{x_0}}(x)$ on $(0, y^*_{x_0})$, it is sufficient to consider the sign of the mapping $D(x) = (\mathcal{A}F_{y^*_{x_0}})(x) - rF_{y^*_{x_0}}(x)$ on $(y^*_{x_0}, \infty)$. The convexity of the increasing fundamental solution implies that $r\psi(x) > \mu(x)\psi'(x)$ for all $x \in \mathbb{R}_+$. Thus we find that for all $x \in (y^*_{x_0}, \infty)$

$$\begin{aligned} D(x) &= \frac{1}{\psi'(x)} \left[\mu(x)\psi'(x) - r\psi'(x)(x-c) - r\frac{\psi(x_0)}{\psi'(y^*_{x_0})}\psi'(x) \right] \\ &\leq r \left[\frac{\psi(x)}{\psi'(x)} - (x-c) - \frac{\psi(x_0)}{\psi'(y^*_{x_0})} \right]. \end{aligned} \tag{3.13}$$

However, since $\psi(x) - \psi'(x)(x-c) < \psi(x_0)$ for all $x \in (y^*_{x_0}, \infty)$, we find that

$$D(x) \leq \frac{\psi(x_0)}{\psi'(x)} - \frac{\psi(x_0)}{\psi'(y^*_{x_0})} = \frac{\psi(x_0)(\psi'(y^*_{x_0}) - \psi'(x))}{\psi'(x)\psi'(y^*_{x_0})} \leq 0,$$

proving that $F_{y^*_{x_0}}(x)$ is r-superharmonic for the diffusion X (cf. [25, Chapter 10]). Since the class of nonnegative r-superharmonic mappings coincides with the class of r-excessive mappings, we find that $F_{y^*_{x_0}}(x)$ is r-excessive for the diffusion X. Consider now the difference

$$F_{y^*_{x_0}}(x) - (x - c + F_{y^*_{x_0}}(x_0)) = \begin{cases} 0, & x \geq y^*_{x_0}, \\ \frac{\psi(x) - \psi(x_0) - \psi'(y^*_{x_0})(x-c)}{\psi'(y^*_{x_0})}, & x < y^*_{x_0}, \end{cases}$$

and define the twice continuously differentiable mapping $\tilde{M} : \mathbb{R}_+ \mapsto \mathbb{R}$ as

$$\tilde{M}(x) = \frac{\psi(x) - \psi(x_0) - \psi'(y^*_{x_0})(x-c)}{\psi'(y^*_{x_0})}.$$

It is now clear that $\tilde{M}'(x) = (\psi'(x) - \psi'(y^*_{x_0}))/\psi'(y^*_{x_0})$ and that $\tilde{M}''(x) = \psi''(x)/\psi'(y^*_{x_0}) > 0$. Consequently, $y^*_{x_0}$ is a global minimum of the mapping $\tilde{M}(x)$. Since $\tilde{M}(y^*_{x_0}) = 0$, we find that $\tilde{M}(x) > 0$ for all $x \in \mathbb{R}_+$. Consequently, we find that $F_{y^*_{x_0}}(x)$ satisfies the conditions of Lemma 3.3 and, therefore, that

$$F_{y^*_{x_0}}(x) \geq \sup_{\tau < \tau(0)} \mathbf{E}_x \left[e^{-r\tau}(X_\tau - c + F_{y^*_{x_0}}(x_0)) \right].$$

However, since $F_{y^*_{x_0}}(x)$ satisfies the recursive relation

$$F_{y^*_{x_0}}(x) = \mathbf{E}_x \left[e^{-r\tau(0, y^*_{x_0})}(X_{\tau(0, y^*_{x_0})} - c + F_{y^*_{x_0}}(x_0)) \right],$$

we find that $\tau(0, y^*_{x_0})$ is optimal and that $F_{y^*_{x_0}}(x) \geq H(x)$, finally proving that $F_{y^*_{x_0}}(x) = H(x)$. It remains to prove that $F_{y^*_{x_0}}(x) = V(x)$. Since $X^\nu_t \in (0, y^*_{x_0})$ for all $t > 0$ under the proposed cutting policy, we find that

$$\lim_{k\to\infty} \mathbf{E}_x[e^{-r\tau_k} F_{y^*_{x_0}}(X^\nu_{\tau_k})] = 0.$$

Combining this finding with our previous results then proves that $F_{y^*_{x_0}}(x)$ satisfies the conditions of part (b) of Theorem 2.1 in [25] and, therefore, that $F_{y^*_{x_0}}(x) = V(x)$ as well. □

Theorem 3.4 states a set of conditions under which the stochastic impulse control problem (2.3) is solvable. In line with previous studies relying on explicit parametric models, we find that the optimal cutting strategy is such that there is a critical threshold $y^*_{x_0}$ at which the irreversible cutting decision should be exercised and at which the controlled diffusion is instantaneously driven to the lower state x_0. The results of Theorem 3.4 are general, since they show that an optimal cutting policy may exist even in the presence of extinction risk, i.e., in cases where the lower boundary may be attainable for the controlled diffusion X_t (this can be also interpreted as liquidation or default risk). This observation is of interest since *it demonstrates the surprising robustness of the modified Faustmann formula* (3.10). It is also worth pointing out that the proof of Theorem 3.4 also shows that $F_y(x)$ as defined in (3.5) is r-excessive for $\{X_t; t \in [0, \tau(0))\}$ and satisfies the inequality $F_y(x) \geq x - c + F_y(x_0)$ only when $y = y^*_{x_0}$. Consequently, we find that $F_{y^*_{x_0}}(x)$ is the unique r-excessive mapping for $\{X_t; t \in [0, \tau(0))\}$ satisfying the recursive condition (3.1) and the inequality $F_y(x) \geq x - c + F_y(x_0)$. An important consequence of the findings of Theorem 3.4 is now summarized in the following.

COROLLARY 3.5. *Assume that the conditions of Theorem* 3.4 *are met. Then the value of the ongoing rotation problem dominates the value of the single rotation problem; that is, then* $V(x) \geq \bar{V}(x)$.

Proof. As was demonstrated in the proof of Theorem 3.4, $V(x)$ is r-excessive for the diffusion $\{X_t; t \in [0, \tau(0))\}$ and satisfies the condition $V(x) \geq x - c + V(x_0)$ for all $x \in \mathbb{R}_+$. Consequently,

$$V(x) \geq \mathbf{E}_x\left[e^{-r\tau_n} V(X_{\tau_n})\right] \geq \mathbf{E}_x\left[e^{-r\tau_n}(X_{\tau_n} - c + V(x_0))\right] \geq \mathbf{E}_x\left[e^{-r\tau_n}(X_{\tau_n} - c)\right],$$

where $\tau_n = n \wedge \inf\{t \geq 0 : X_t \notin (n^{-1}, n)\} \wedge \tau$ is an almost surely finite $\mathcal{F}_t$-stopping time and τ is an arbitrary stopping time. Letting $n \to \infty$ and invoking Fatou's lemma then prove the alleged result. □

Corollary 3.5 shows the intuitively clear result that the value of the ongoing rotation (*Faustmann's tree cutting problem*) impulse control problem dominates the value of the associated single rotation (*Wicksell's tree cutting problem*) optimal stopping problem. The reason for this finding is clear in light of the recursive stopping problem (3.12). Interestingly, we find that the impulse control problem (2.3) is actually very closely related to the optimal stopping problem presented in Corollary 3.5. Our main result on this relationship is summarized in the following.

THEOREM 3.6. *Assume that the conditions of Theorem* 3.4 *are met and that the lower boundary* 0 *is exit, killing, or natural for* $\{X_t; t \in [0, \tau(0))\}$. *Then the following hold.*

(A) $\lim_{x_0 \downarrow 0} y^*_{x_0} = \bar{y} = \operatorname{argmax}\left\{\frac{x-c}{\psi(x)}\right\}$ *and* $\lim_{x_0\downarrow 0} V(x) = \bar{V}(x)$, *where* $\bar{V}(x)$ *is defined as in* (2.11).

(B) *The single rotation (Wicksellian) boundary* $\bar{y}$ *dominates the ongoing rotation (Faustmannian) boundary* $y^*_{x_0}$; *that is,* $\bar{y} > y^*_{x_0}$.

Proof. (A) If 0 is exit, killing, or natural for $\{X_t; t \in [0, \tau(0))\}$, then $\lim_{x_0 \downarrow 0} \psi(x_0) = 0$, implying the alleged results. The rest is a direct consequence of Proposition 1 in [8]. Part (B) then follows directly from Theorem 3.4. □

Theorem 3.6 shows that the associated optimal stopping problem (3.12) is solvable under the conditions of our Theorem 3.4 and under a set of suitable boundary conditions at 0. As one can also immediately observe from Theorem 3.6, the exercise threshold of the single rotation stopping problem dominates the threshold of the considered optimal rotation problem. Consequently, our results establish the intuitively clear result that the required exercise premium is higher in the single rotation problem than in the optimal rotation problem (2.3). *It is also worth observing that in line with the modified Faustmann rule* (3.10)*, the single rotation rule* (2.10) *can be rewritten as*

$$\frac{\mathbf{E}_x[e^{-r\tau(\bar{y})}B(X_{\tau(\bar{y})})]}{\mathbf{E}_x[e^{-r\tau(\bar{y})}(X_{\tau(\bar{y})} - c)]} = r, \tag{3.14}$$

implying that

$$\frac{\mathbf{E}_x[e^{-r\tau(\bar{y})}B(X_{\tau(\bar{y})})]}{\mathbf{E}_x[e^{-r\tau(\bar{y})}(X_{\tau(\bar{y})} - c)]} \leq \frac{r}{1 - \mathbf{E}_x[e^{-r\tau(\bar{y})}]}$$

and that given the conditions of Corollary 2.2 we have that

$$\frac{\mathbf{E}_x[e^{-r\tau(\bar{y})}\mu(X_{\tau(\bar{y})})]}{\mathbf{E}_x[e^{-r\tau(\bar{y})}(X_{\tau(\bar{y})} - c)]} \leq r \leq \frac{r}{1 - \mathbf{E}_x[e^{-r\tau(\bar{y})}]}. \tag{3.15}$$

Having considered the existence and uniqueness of a solution for the optimal rotation problem (2.3), it is our purpose to now consider the risk sensitivity of the optimal policy and its value. To accomplish this task, define the process

$$\tilde{X}_t^\nu = x + \int_0^t \mu(\tilde{X}_s^\nu)ds + \int_0^t \tilde{\sigma}(\tilde{X}_s^\nu)dW_s - \sum_{\tau_k \leq t} \zeta_k, \quad 0 \leq t \leq \tilde{\tau}^\nu(0), \tag{3.16}$$

where $\tilde{\tau}^\nu(0) = \inf\{t \geq 0 : \tilde{X}_t^\nu \leq 0\} \leq \infty$ and $\tilde{\sigma} : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is a continuous mapping satisfying the inequality $\tilde{\sigma}(x) \leq \sigma(x)$ for all $x \in \mathbb{R}_+$. In accordance with our previous notation, we denote as

$$\tilde{\mathcal{A}} = \frac{1}{2}\tilde{\sigma}^2(x)\frac{d^2}{dx^2} + \mu(x)\frac{d}{dx}$$

the differential operator representing the infinitesimal generator of $\{\tilde{X}_t; t \in [0, \tilde{\tau}(0))\}$ and as $\tilde{V}(x)$ the value of the impulse control problem (2.3) in the presence of the less volatile process $\tilde{X}_t$. Our main results on the risk sensitivity of the optimal policy and its value are now summarized in the following.

THEOREM 3.7. *Assume that the conditions of Theorem* 3.4 *are met. Then,* $\tilde{V}(x) \leq V(x)$ *for all* $x \in \mathbb{R}_+$*. That is, increased volatility increases the value of the optimal rotation problem* (2.3).

Proof. Theorem 3.4 shows that $V(x)$ is convex and r-excessive for $\{X_t; t \in [0, \tau(0))\}$. Moreover, for all $x \in \mathbb{R}_+ \backslash \{y_{x_0}^*\}$ we have that

$$(\tilde{\mathcal{A}}V)(x) - rV(x) = ((\tilde{\mathcal{A}} - \mathcal{A} + \mathcal{A})V)(x) - rV(x) \leq \frac{1}{2}(\tilde{\sigma}^2(x) - \sigma^2(x))V''(x) \leq 0.$$

Consequently, we find that $V(x)$ is r-excessive for $\{\tilde{X}_t; t \in [0, \tilde{\tau}(0))\}$ as well (cf. [3]). Since $V(x) \geq x - c + V(x_0)$ and $\lim_{k\to\infty} \mathbf{E}_x[e^{-r\tau_k} V(X^{\nu}_{\tau_k})] = 0$, we find that $V(x)$ satisfies the conditions of part (a) of Theorem 2.1 in [25] and, therefore, that $V(x) \geq \tilde{V}(x)$ for all $x \in \mathbb{R}_+$. □

Theorem 3.7 shows that given the conditions of Theorem 3.4, increased volatility increases the value of the optimal rotation problem (2.3). In other words, Theorem 3.7 demonstrates that increased volatility increases the expected NPV of the future harvests from the present up to an arbitrarily distant future. The impact of increased volatility on the optimal exercise threshold and rotation length is summarized in the following.

THEOREM 3.8. *Assume that the conditions of Theorem* 3.4 *are met. Then, increased volatility increases the optimal exercise threshold and, therefore, decelerates rational harvesting. More precisely,* $\tilde{y}_{x_0} \leq y^*_{x_0}$, *where* $\tilde{y}_{x_0}$ *denotes the optimal harvesting threshold in the presence of the less volatile value dynamics* $\tilde{X}^{\nu}_t$.

Proof. Define the continuous mappings $u : \mathbb{R}_+ \mapsto \mathbb{R}_+$ and $\tilde{u} : \mathbb{R}_+ \mapsto \mathbb{R}_+$ for all $y \in (x_0, x)$ as

$$u(y) = \frac{\psi(y)}{\psi(x)} - \frac{\psi(y) - \psi(x_0)}{\psi(x) - \psi(x_0)} \quad \text{and} \quad \tilde{u}(y) = \frac{\tilde{\psi}(y)}{\tilde{\psi}(x)} - \frac{\tilde{\psi}(y) - \tilde{\psi}(x_0)}{\tilde{\psi}(x) - \tilde{\psi}(x_0)},$$

where $\tilde{\psi}(x)$ denotes the increasing fundamental solution of the ordinary second order differential equation $(\tilde{\mathcal{A}}v)(x) = rv(x)$. Since $\lim_{y\uparrow x} u(y) = \lim_{y\uparrow x} \tilde{u}(y) = 0$, the mappings $\tilde{u}(y)$ and $u(y)$ are continuous, and $\psi(y)/\psi(x) \geq \tilde{\psi}(y)/\tilde{\psi}(x)$ for all $y \leq x$ and we find that for all $\varepsilon > 0$ there is an open neighborhood $(x - \delta, x)$ of x such that for all $y \in (x - \delta, x)$ we have that

$$\frac{\psi(y) - \psi(x_0)}{\psi(x) - \psi(x_0)} > \frac{\psi(y)}{\psi(x)} - \varepsilon > \frac{\tilde{\psi}(y)}{\tilde{\psi}(x)} - \varepsilon > \frac{\tilde{\psi}(y) - \tilde{\psi}(x_0)}{\tilde{\psi}(x) - \tilde{\psi}(x_0)} - \varepsilon.$$

However, since $\varepsilon > 0$ is arbitrary and $\lim_{y\uparrow x}(\psi(y) - \psi(x_0))/(\psi(x) - \psi(x_0)) = \lim_{y\uparrow x}(\tilde{\psi}(y) - \tilde{\psi}(x_0))/(\tilde{\psi}(x) - \tilde{\psi}(x_0)) = 1$, we find that

$$\frac{\psi'(x)}{\psi(x) - \psi(x_0)} \leq \frac{\tilde{\psi}'(x)}{\tilde{\psi}(x) - \tilde{\psi}(x_0)}$$

for all $x \in (x_0, \infty)$. Denote now as $\tilde{y}_{x_0}$ the optimal harvesting threshold in the presence of the less volatile value dynamics $\tilde{X}^{\nu}_t$. Then

$$\begin{aligned} f(\tilde{y}_{x_0}) &= (\psi(\tilde{y}_{x_0}) - \psi(x_0)) \left(1 - \frac{\psi'(\tilde{y}_{x_0})}{\psi(\tilde{y}_{x_0}) - \psi(x_0)}(\tilde{y}_{x_0} - c)\right) \\ &\geq (\psi(\tilde{y}_{x_0}) - \psi(x_0)) \left(1 - \frac{\tilde{\psi}'(\tilde{y}_{x_0})}{\tilde{\psi}(\tilde{y}_{x_0}) - \tilde{\psi}(x_0)}(\tilde{y}_{x_0} - c)\right) = 0, \end{aligned}$$

implying that $\tilde{y}_{x_0} \leq y^*_{x_0}$ and completing the proof of our theorem. □

Thus we find, in line with studies considering irreversible investment decisions, that increased volatility increases both the value and the optimal exercise threshold of the optimal rotation problem. Consequently, we observe that increased volatility increases the required exercise premium of a rational harvester and, therefore, postpones the irreversible harvesting decision.

4. Extensions. Although the results of our previous section are general, they rely heavily on the strict convexity of the increasing fundamental solution $\psi(x)$. Unfortunately, this condition is not always satisfied in models subject to *mean reversion or pure compensation* (for example, for the logistic model $\mu(x) = \mu x(1 - \gamma x)$ the auxiliary mapping $\theta(x)$ is decreasing only when $\mu \leq r$, that is, only when *the intrinsic growth rate at low densities is smaller than the discount rate* r). An important implication of Lemma 2.1 dealing with these cases is now summarized in the following (cf. [3] and [5]).

COROLLARY 4.1. *Assume that* $\lim_{x\downarrow 0}\theta(x) \geq 0 \geq \lim_{x\to\infty}\theta(x)$ *and that there is a threshold* $\bar{x} \in (0,\infty)$ *such that* $\theta(x)$ *is increasing on* $(0,\bar{x})$ *and decreasing on* $(\bar{x},\infty)$ *(i.e.,* $\bar{x} = \operatorname{argmax}\{\theta(x)\}$*). Then there is a unique threshold* $x^* \in (\bar{x}, \theta^{-1}(0))$ *for which* $\psi''(x^*) = 0$ *and* $\psi''(x) \lessgtr 0$, *when* $x \lessgtr x^*$. *Moreover,* $\lim_{x\downarrow 0}\psi'(x)x = 0$.

Proof. As was shown in the proof of Corollary 2.2, the mapping $I(x)$ is increasing on the set where $\theta(x)$ is decreasing. Proving then that $I(x)$ is decreasing on the set where $\theta(x)$ is increasing is completely analogous. Since $\lim_{x\downarrow 0}\psi'(x)/S'(x) \geq 0$, we find that

$$\lim_{x\downarrow 0} I(x) = -\lim_{x\downarrow 0}\theta(x)\frac{\psi'(x)}{S'(x)} \leq 0.$$

Consequently, $I(x)$ is negative on the set $(0,\bar{x})$. However, since $\theta(x)$ is decreasing on $(\bar{x},\infty)$ and satisfies the condition $\lim_{x\to\infty}\theta(x) \leq 0$, we find that $\theta(x)$ has a unique root $\theta^{-1}(0)$ on $(\bar{x},\infty)$. Since

$$I(\theta^{-1}(0)) = r\int_0^{\theta^{-1}(0)} \psi(y)\theta(y)m'(y)dy > 0$$

and $I(x)$ is monotonically increasing and continuous on $(\bar{x},\infty)$, we find that the equation $I(x) = 0$ has a unique root on $(\bar{x},\theta^{-1}(0))$, thus completing the proof of the first part of our corollary. To prove that $\lim_{x\downarrow 0}\psi'(x)x = 0$ we observe that the concavity of $\psi(x)$ on $(0,x^*)$ implies that $\psi(0) \leq \psi(x) - \psi'(x)x$ for all $x \in (0,x^*)$. Consequently, we find that $0 \leq \psi'(x)x \leq \psi(x) - \psi(0)$. Letting $x \downarrow 0$ then completes the proof of our corollary. □

Corollary 4.1 states a set of conditions under which the increasing fundamental solution $\psi(x)$ is strictly convex above a threshold x^* and concave below it (that is, $\psi'(x)$ has a unique global minimum on $\mathbb{R}_+$). An important consequence of this finding is now summarized in the following.

LEMMA 4.2. *Assume that* $x_0 < c$ *and that the conditions of Corollary* 4.1 *and Lemma* 2.3 *are met. Then there is a unique threshold* $y^*_{x_0} \in (\max(x^*,c),\bar{y})$ *satisfying* (3.7).

Proof. First, we observe that, given the conditions of Lemma 2.3, the mapping $(x-c)/\psi(x)$ attains its maximum on the set where $\psi(x)$ is convex, that is, on (x^*,∞). Assume first that $x^* < c$. Then we find that $f(y)$ is decreasing on $(0,x^*)\cup(c,\infty)$ and increasing on (x^*,c), thus implying that c constitutes an interior local maximum of $f(y)$. The concavity of $\psi(x)$ on $(0,x^*)$ and Taylor's theorem imply that

$$\psi(x_0) \leq \psi(x^*) + \psi'(x^*)(x_0 - x^*)$$

and, therefore, that

$$f(x^*) \geq \psi'(x^*)(c - x_0) > 0.$$

Since $f(c) = \psi(c) - \psi(x_0) > 0$ and $f'(y) = -\psi''(y)(y-c) < 0$ on (x^*, ∞) by the strict convexity of $\psi(y)$ on (x^*, ∞), the alleged result follows from the proof Lemma 3.2. The proof in the case $x^* > c$ is completely analogous (with the only exception being that in that case x^* constitutes an interior local maximum of $f(y)$). If $x^* = c$, then $f(y)$ is decreasing on $\mathbb{R}_+$. However, since $f(c) = \psi(c) - \psi(x_0) > 0$ and $f'(y) = -\psi''(y)(y-c) < 0$ on (x^*, ∞) by the strict convexity of $\psi(y)$ on (x^*, ∞), the alleged result follows again from the proof Lemma 3.2. □

Lemma 4.2 shows that given the conditions of Corollary 4.1 and Lemma 2.3 the necessary condition (3.7) also has a unique root on $(\max(x^*, c), \bar{y})$ at which the representation (3.4) is maximized on (x_0, ∞) since $g'(y) \gtreqless 0$ whenever $y \lesseqgtr y^*_{x_0}$. Consequently, we observe that the global convexity of the increased fundamental solution $\psi(x)$ is not necessarily required for the existence and uniqueness of an optimal cutting value $y^*_{x_0}$. Our main result is now summarized in the following.

THEOREM 4.3. *Assume that $x_0 < c$, that the conditions of Lemma* 4.2 *are met, and that $\psi'(y^*_{x_0})c > \psi(x_0)$ (that is, that $\psi'(y^*_{x_0})y^*_{x_0} > \psi(y^*_{x_0})$). Then, the optimal stopping time is $\tau(0, y^*_{x_0}) = \inf\{t \geq 0 : X_t \geq y^*_{x_0}\}$, and the value of the optimal stopping problem* (3.12) *reads as $H_c(x) = F_{y^*_{x_0}}(x)$, where $F_{y^*_{x_0}}(x)$ is defined as in* (3.8) *and $y^*_{x_0}$ is the optimal exercise threshold satisfying* (3.7) *and defined as in Lemma* 4.2. *Moreover, the optimal cutting value is $y^*_{x_0}$ and $V(x) = H(x) = F_{y^*_{x_0}}(x)$ for all $x \in \mathbb{R}_+$.*

Proof. Proving the r-excessivity of the mapping $F_{y^*_{x_0}}(x)$ for the diffusion $\{X_t; t \in [0, \tau(0)\}$ is analogous to the proof in Theorem 3.4. To prove that $F_{y^*_{x_0}}(x) \geq (x - c + F_{y^*_{x_0}}(x_0))$ for all $x \in \mathbb{R}_+$ we again observe that $\tilde{M}'(x) = (\psi'(x) - \psi'(y^*_{x_0}))/\psi'(y^*_{x_0})$ and that $\tilde{M}''(x) = \psi''(x)/\psi'(y^*_{x_0})$. Consequently, $y^*_{x_0}$ constitutes a local minimum of the mapping $\tilde{M}(x)$. Since $x^* = \operatorname{argmin}\{\psi'(x)\}$, we find that $\tilde{M}(x)$ is decreasing on $(x^*, y^*_{x_0})$. The concavity of $\psi(x)$ on $(0, x^*)$ then implies that $\tilde{M}(x) > 0$ on $(0, y^*_{x_0})$ provided that $\tilde{M}(0) > 0$, that is, provided that the condition $\psi'(y^*_{x_0})c > \psi(x_0)$ is met. Consequently, the proposed value function satisfies the conditions of the verification lemma and, therefore, $H(x) = F_{y^*_{x_0}}(x)$. The rest of the proof is analogous to the proof of Theorem 3.4. □

Theorem 4.3 extends the results of Theorem 3.4 to the case where the increasing fundamental solution $\psi(x)$ is not globally convex. As one can observe from Theorem 4.3, in order to guarantee the existence of an optimal cutting value, we have to make the extra assumption (in comparison with Theorem 3.4) $\psi'(y^*_{x_0})c > \psi(x_0)$ needed for the verification of the condition $V(x) > x - c + V(x_0)$. If this assumption is not satisfied, then the optimal rotation problem does not have a solution. It is also worth pointing out that the results of Corollary 3.5 and Theorem 3.6 are also satisfied in the case of this section, since the condition $\psi'(y^*_{x_0})c > \psi(x_0)$ is trivially satisfied when 0 is exit, killing, or natural for X_t and $x_0 \downarrow 0$. Unfortunately, the local concavity of the increasing fundamental solution makes it difficult to extend the results of Theorem 3.7 to the case of this section. In any case, it is a straightforward implication of the proof of Theorem 3.7 that increased volatility increases the value of the optimal policy at least locally on the set where the value is convex (cf. [9]). Moreover, since the increasing fundamental solution $\psi(x)$ is locally convex on a neighborhood of the optimal threshold $y^*_{x_0}$, we conjecture that the conclusion of Theorem 3.8 holds locally in the present case too. That is, we conjecture that increased volatility decelerates rational harvesting by increasing the optimal threshold $y^*_{x_0}$ in the case of this section as well.

5. Explicit example. In order to illustrate our results, we now plan to consider the optimal rotation problem when the underlying forest stand value process evolves according to the mean reverting diffusion (corresponding to logistic growth subject to a random intrinsic growth rate) considered in, among others, [4] and [21]. More precisely, we now assume that the stochastic forest stand value growth evolves according to the dynamics described by the stochastic differential equation

$$X_t^\nu = x + \int_0^t \mu X_s^\nu (1 - \gamma X_s^\nu) ds + \int_0^t \sigma X_s^\nu (1 - \gamma X_s^\nu) dW_s - \sum_{\tau_k \leq t} \zeta_k, \quad 0 \leq t \leq \tau^\nu(0),$$

where $\gamma, \mu, \sigma \in \mathbb{R}_+$ are known exogenously determined constants. It is now clear that in the present example $\theta(x) = \mu x(1 - \gamma x) - rx$ and consequently that the conditions of Lemma 3.2 or Corollary 4.1 are satisfied depending on whether $r \geq \mu$ or $r < \mu$, respectively. As was established in [4], the increasing fundamental solution of the ordinary second order differential equation

$$\frac{1}{2}\sigma^2 x^2 (1 - \gamma x)^2 u''(x) + \mu x(1 - \gamma x) u'(x) - ru(x) = 0$$

reads as

$$\psi(x) = \left(\frac{\gamma x}{1 - \gamma x}\right)^\alpha F\left(a, b, d; -\frac{\gamma x}{1 - \gamma x}\right),$$

where F is the standard hypergeometric function,

$$a = 1 + \sqrt{\left(\frac{1}{2} - \frac{\mu}{\sigma^2}\right)^2 + \frac{2r}{\sigma^2}} + \sqrt{\left(\frac{1}{2} + \frac{\mu}{\sigma^2}\right)^2 + \frac{2r}{\sigma^2}},$$

$$b = 1 + \sqrt{\left(\frac{1}{2} - \frac{\mu}{\sigma^2}\right)^2 + \frac{2r}{\sigma^2}} - \sqrt{\left(\frac{1}{2} + \frac{\mu}{\sigma^2}\right)^2 + \frac{2r}{\sigma^2}},$$

$$d = 1 + 2\sqrt{\left(\frac{1}{2} - \frac{\mu}{\sigma^2}\right)^2 + \frac{2r}{\sigma^2}},$$

and

$$\alpha = \frac{1}{2} - \frac{\mu}{\sigma^2} + \sqrt{\left(\frac{1}{2} - \frac{\mu}{\sigma^2}\right)^2 + \frac{2r}{\sigma^2}} > 0.$$

Although it is impossible to solve the threshold $y_{x_0}^*$ explicitly from the ordinary first order condition $\psi(y_{x_0}^*) - \psi(x_0) = (y_{x_0}^* - c)\psi'(y_{x_0}^*)$ it can usually be solved numerically when the values of the exogenous parameters are given. We illustrate graphically the optimal cutting values $y_{x_0}^*$ and $\bar{y}$ as functions of the underlying volatility coefficient σ under the assumption that $x_0 = 1, c = 2, \mu = 4.8\%, \gamma = 0.01$, and $r = 3\%$. In line with the findings of Lemma 3.2, Corollary 3.5, and Lemma 4.2, Figures 1 and 2 demonstrate that the value of the ongoing rotation problem dominates the value of the associated single rotation problem and that the exercise threshold is higher for the single rotation problem than for the ongoing rotation problem. Moreover, Figure 1 also shows that increased volatility increases the required exercise premium and prolongs the rotation period by increasing the optimal value at which the irreversible harvesting decision should be made.

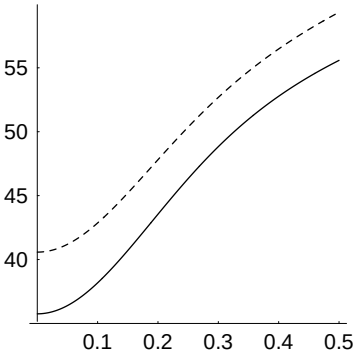


FIG. 1. *The optimal cutting values* $y^*_{x_0}(\sigma)$ *(uniform) and* $\bar{y}(\sigma)$ *(dashed).*

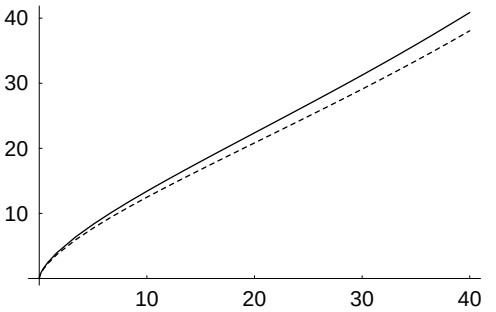


FIG. 2. *The values* $V(x)$ *(uniform) and* $\bar{V}(x)$ *(dashed) when* $\sigma = 0.1$.

6. Concluding comments. We considered the determination of the ongoing rotation policy maximizing the expected NPV of the future harvests from the present up to a potentially infinite future when the underlying stochastic forest stand value growth is modeled as a linear diffusion. By relying on a combination of stochastic calculus, the classical theory of diffusions, and ordinary nonlinear programming techniques, we presented a set of typically satisfied general sufficient conditions under which the optimal Faustmannian rotation problem is solvable and under which the optimal harvesting threshold can be determined from an algebraic equation. In accordance with previous studies considering single rotation problems and the optimal timing of irreversible investments, we found that at the optimum the project value has to be equal to its full costs and that the value of the optimal policy satisfies the familiar smooth-fit condition at the optimal harvesting threshold. We also considered the impact of increased volatility on the optimal rotation policy and stated a set of conditions under which increased volatility unambiguously increases the value of the optimal rotation policy and prolongs the rotation period by increasing the required exercise premium of the irreversible harvesting opportunity. Put somewhat differently, while increased volatility increases the (lost) option value of a single harvesting opportunity, it simultaneously increases the value of all future harvesting opportunities (i.e., the future harvesting potential). Since the latter effect dominates the former, we find that the impact of increased volatility on harvesting is unambiguously negative. This result is of interest since it illustrates the robustness of the qualitative findings obtained in studies relying on geometric Brownian motion.

Even though we considered a broad class of harvesting models, our analysis is subject to two significant constraints limiting at least to some extent the general applicability of our findings. First, by assuming that only timber revenues affect the optimal rotation policy, we overlooked the valuation of amenities which especially affect the rotation policies of preserved old forests. Second, while rotation periods are typically very long (especially in Nordic countries, where forestry plays an important role) we relied on a constant discount rate. Thus our analysis overlooked the potentially significant role of the intertemporal variability of the opportunity cost of investment. Unfortunately, an analysis of a model incorporating these important factors is out of the scope of this study and, therefore, is left for future research.

Acknowledgment. The author is grateful to an anonymous referee for constructive criticism and suggested improvements.

REFERENCES

[1] L. H. R. Alvarez, *A class of solvable singular stochastic control problems*, Stochastics Stochastics Rep., 67 (1999), pp. 83–122.
[2] L. H. R. Alvarez, *Singular stochastic control in the presence of a state-dependent yield structure*, Stochastic Process. Appl., 86 (2000), pp. 323–343.
[3] L. H. R. Alvarez, *On the properties of r-excessive mappings for a class of diffusions*, Ann. Appl. Probab., to appear.
[4] L. H. R. Alvarez, *On the option interpretation of rational harvesting planning*, J. Math. Biol., 40 (2000), pp. 383–405.
[5] L. H. R. Alvarez, *Singular stochastic control, linear diffusions, and optimal stopping: A class of solvable problems*, SIAM J. Control Optim., 39 (2001), pp. 1697–1710.
[6] L. H. R. Alvarez, *Does increased stochasticity speed up extinction?*, J. Math. Biol., 43 (2001), pp. 534–544.
[7] L. H. R. Alvarez, *Solving optimal stopping problems of linear diffusions by applying convolution approximations*, Math. Methods Oper. Res., 53 (2001), pp. 89–99.
[8] L. H. R. Alvarez, *Reward functionals, salvage values, and optimal stopping*, Math. Methods Oper. Res., 54 (2001), pp. 315–337.
[9] R. Bhattacharyya and D. L. Snyder, *Stumpage price uncertainty and the optimal rotation of a forest: An application of the Sandmo model*, J. Environmental Systems, 17 (1990), pp. 305–313.
[10] A. Borodin and P. Salminen, *Handbook on Brownian motion—facts and formulae*, Birkhäuser, Basel, 1996.
[11] K. A. Brekke and B. Øksendal, *The high contact principle as a sufficiency condition for optimal stopping*, in Stochastic Models and Option Values, D. Lund and B. Øksendal, eds., North–Holland, Amsterdam, 1991, pp. 187–208.
[12] K. A. Brekke and B. Øksendal, *Optimal switching in an economic activity under uncertainty*, SIAM J. Control Optim., 32 (1994), pp. 1021–1036.
[13] K. A. Brekke and B. Øksendal, *A verification theorem for combined stochastic control and impulse control*, in Stochastic Analysis and Related Topics Vol. 6, L. Decreusefond, J. Gjerde, B. Øksendal, and S. Ustunel, eds., Birkhäuser, Basel, 1997, pp. 211–220.
[14] C. W. Clark, *Mathematical Bioeconomics: The Optimal Management of Renewable Resources*, Wiley, New York, 1976.
[15] H. R. Clarke and W. J. Reed, *The tree-cutting problem in a stochastic environment*, J. Econom. Dynam. Control, 13 (1989), pp. 569–595.
[16] R. Gibson and E. S. Schwartz, *Stochastic convenience yield and the pricing of oil contingent claims*, J. Finance, 45 (1990), pp. 959–976.
[17] O. Gjolberg and A. G. Guttormsen, *Real options in the forest: What if prices are mean-reverting?*, Forest Policy and Economics, 4 (2002), pp. 13–20.
[18] K. Itô and H. P. McKean, Jr., *Diffusion Processes and Their Sample Paths*, Springer-Verlag, Berlin, 1965.
[19] P. O. Johansson and K. G. Löfgren, *The Economics of Forestry and Natural Resources*, Basil Blackwell, Oxford, UK, 1985.
[20] S. Karlin and H. Taylor, *A Second Course in Stochastic Processes*, Academic Press, Orlando, 1981.

[21] E. M. Lungu and B. Øksendal, *Optimal harvesting from a population in a stochastic crowded enviroment*, Math. Biosci., 145 (1997), pp. 47–75.
[22] R. McDonald and D. Siegel, *Investment and the valuation of firms when there is an option to shut down*, Internat. Econom. Rev., 26 (1985), pp. 331–349.
[23] R. A. Miller and K. Voltaire, *A stochastic analysis of the three paradigm*, J. Econom. Dynam. Control, 6 (1983), pp. 371–386.
[24] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, 5th Ed., Springer-Verlag, Berlin, 1998.
[25] B. Øksendal, *Stochastic control problems where small intervention costs have big effects*, Appl. Math. Optim., 40 (1999), pp. 355–375.
[26] A. J. Plantinga, *The optimal timber rotation: An option value approach*, Forest Sci., 44 (1998), pp. 192–202.
[27] W. J. Reed and H. R. Clarke, *Harvest decisions and asset valuation for biological resources exhibiting size-dependent stochastic growth*, Internat. Econom. Rev., 31 (1990), pp. 147–169.
[28] P. A. Samuelson, *Economics of forestry in an evolving society*, Economic Inquiry, 14 (1976), pp. 466–492.
[29] A. Sandmo, *The effect of uncertainty on saving decisions*, Rev. Econom. Stud., 37 (1970), pp. 353–360.
[30] E. S. Schwartz, *The stochastic behavior of commodity prices: Implications for valuation and hedging*, J. Finance, 52 (1997), pp. 923–973.
[31] D. L. Snyder and R. Bhattacharyya, *A more general dynamic economic model of the optimal rotation of multiple-use forests*, J. Environmental Economics and Management, 18 (1990), pp. 168–175.
[32] S. Sødal, *The stochastic rotation problem: A comment*, J. Econom. Dynam. Control, 26 (2002), pp. 509–515.
[33] T. A. Thomson, *Optimal forest rotation when stumpage prices follow a diffusion process*, Land Economics, 68 (1992), pp. 329–342.
[34] Y. Willassen, *The stochastic rotation problem: A generalization of Faustmann's formula to stochastic forest growth*, J. Econom. Dynam. Control, 22 (1998), pp. 573–596.
[35] R. Yin and D. Newman, *When to cut a stand of trees*, Natur. Resource Modeling, 10 (1997), pp. 251–261.
[36] A. Yoshimoto and I. Shoji, *Searching for an optimal rotation age for forest stand management under stochastic log prices*, European J. Oper. Res., 105 (1998), pp. 100–112.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 1994–2003

SURROGATE PROGRAMMING AND MULTIPLIERS IN QUASI-CONVEX PROGRAMMING*

JEAN-PAUL PENOT† AND MICHEL VOLLE‡

Abstract. A result due to Luenberger on the existence of multipliers in a quasi-convex programming problem is extended to the case of constraints given by an arbitrary convex cone under a constraint qualification condition more general than Slater's condition. The existence of solutions is not assumed. We point out links with even convexity in the sense of Fenchel and quasi subdifferentiability in the sense of Greenberg–Pierskalla, and we observe that the couples of primal-dual optimal solutions reduce to saddle-points of a suitable Lagrangian function.

Key words. even convexity, Lagrangian, multipliers, qualification conditions, quasi-convex functions, surrogate programming

AMS subject classifications. 49M45, 90C26, 90C31

DOI. 10.1137/S0363012902327819

1. Introduction. We consider the following mathematical programming problem:

$$(\mathcal{P}) \quad \text{minimize } f(x) \quad \text{subject to} \quad x \in B,\ g(x) \in C,$$

where B (resp., C) is a convex subset (resp., closed convex cone) of a Banach space X (resp., Z), $g : B \to Z$ is convex with respect to $-C$ (i.e., epi $g := \{(x, z) \in X \times Z : x \in B,\ z \in g(x) - C\}$ is a convex subset of $X \times Z$), and $f : B \to \mathbb{R}$ is quasi-convex on B (i.e., for each $r \in \mathbb{R}$, $\{x \in B : f(x) < r\}$ is convex). We denote by Y the topological dual of Z. Here the constraint $x \in B$ is considered as a basic constraint which is easy to deal with (for instance, $B = X$ or a box or an orthant).

Relaxing Slater's condition imposed in [20], we prove the existence of $\bar{y}$ in the negative polar cone $Y_+ := C^\circ = \{y \in Y : z \in C \Longrightarrow \langle y, z\rangle \le 0\}$ of C such that

$$(M) \qquad \inf\{f(x) : x \in B\ ,\ g(x) \in C\} = \inf\{f(x) : x \in B\ ,\ \langle \bar{y}, g(x)\rangle \le 0\}.$$

An element $\bar{y} \in Y_+$ such that (M) holds will be called a *surrogate multiplier* by analogy with the classical notion of Lagrange multiplier which satisfies the relation

$$(L) \qquad \inf\{f(x) : x \in B\ ,\ g(x) \in C\} = \inf\{f(x) + \langle \bar{y}, g(x)\rangle : x \in B\}.$$

The terminology made popular by [14] is justified as follows: when Lagrange multipliers do not exist, despite the fact that it is not possible to substitute $f + \langle \bar{y}, g(\cdot)\rangle$ for f in order to eliminate the constraint $g(x) \in C$ without changing the value of $(\mathcal{P})$, a surrogate multiplier enables one to replace problem $(\mathcal{P})$ by a simpler problem in which the constraint is a scalar one. In particular, when $B = X$ and g is linear, one is led to the minimization of f on a half-plane.

*Received by the editors July 22, 2002; accepted for publication (in revised form) June 27, 2003; published electronically January 22, 2004.
http://www.siam.org/journals/sicon/42-6/32781.html
†Department of Mathematics, University of Pau, Avenue de l'Université, 64000 Pau, France (jean-paul.penot@univ-pau.fr).
‡Department of Mathematics, University of Avignon, 33 rue Louis Pasteur, 84000 Avignon, France (michel.volle@univ-avignon.fr).

It is easy to see that relation (L) entails (M) while the converse is not true. Simple examples show that one cannot expect existence of usual Lagrange multipliers when f is just quasi-convex, even in the case when the Slater condition is satisfied.

Example 1.1. Let $X = B = \mathbb{R}$, $Z = \mathbb{R}$, $C = -\mathbb{R}_+$, $g(x) = -x$, $f(x) = x^3$. Then for any $y \in Y$ one has $\inf\{f(x) + y \cdot g(x) : x \in X\} = -\infty$, while for $\bar{y} = 1$ one has $\inf\{f(x) : x \in X, \bar{y} \cdot g(x) \leq 0\} = 0 = \inf\{f(x) : x \in g^{-1}(C)\}$.

Such a fact compels one to substitute the study of the set M_S of surrogate multipliers to the study of the set of Lagrange multipliers.

A different approach is adopted in [22], [28]. There one considers the set $M_<$ of $y \in Y_+$ such that

$$\inf\{f(x) : x \in B \cap g^{-1}(C)\} = \inf\{f(x) + \langle y, g(x)\rangle_+ : x \in B\},$$

where $\langle y, g(x)\rangle_+ = \max\{\langle y, g(x)\rangle, 0\}$. Since $f(x) + \langle y, g(x)\rangle_+ = f(x)$ for $x \in A_y := \{x \in B : \langle y, g(x)\rangle \leq 0\}$, we have

$$\inf\{f(x) + \langle y, g(x)\rangle_+ : x \in X\} \leq \inf\{f(x) : x \in A_y\};$$

hence any element of $M_<$ belongs to M_S. Again, the elements of M_S are surrogate to the elements of $M_<$ and lead to a problem with just one scalar convex inequality constraint. Such a problem is more tractable than the primal problem (see [34, Theorem 3.4], for example). The set M_S of surrogate multipliers can be interpreted as the set S_D of optimal solutions to the dual problem

$$(\mathcal{D}_S) \qquad \text{maximize} \quad d_S(y) := \inf\{f(x) : x \in B\ ,\ \langle y, g(x)\rangle \leq 0\} \quad \textit{over} \ \ Y_+ = C^\circ$$

when the values of $(\mathcal{P})$ and $(\mathcal{D}_S)$ are equal. We also provide a Lagrangian whose saddle-points give the optimal solution of $(\mathcal{P})$ and $(\mathcal{D}_S)$. Finally, we study some properties of the performance function p associated with $(\mathcal{P})$ in terms of even convexity and observe that the optimal solutions of $(\mathcal{D}_S)$ are linked with the quasi subdifferential of p at zero.

The constraint qualification condition we use can be considered as a classical condition ([1], [2], [3], [4], [13], [17], [26]); it is a general assumption as it does not suppose that the negative cone C of Z (ordered by $z_1 \leq z_2$ iff $z_1 - z_2 \in C$) has a nonempty interior, in contrast with Slater's condition, a concern common in [3], [4], [13], [17], [26]. It is well known that for many usual linear spaces of functional analysis the ordering cone has an empty interior. Moreover, in the existence result we present we do not assume that the primal problem $(\mathcal{P})$ has solutions. In the convex case it is known that this assumption is superfluous (see [8], [32], for instance). It appears that, in our quasi-convex case, we can face a similar situation: the dual problem has solutions whereas the primal has no solution. These duality questions are considered in Corollary 3.1 below; for a more systematic treatment we refer to [6], [14], [15], [23], [28], [35].

An abstract study of surrogate duality is conducted in [35], [36], [37, p. 28]. Surrogate duality methods are of particular importance for discrete optimization. Numerous studies and algorithms have been devised for such purposes; see [9], [11], [12], [14], [15], [18], [19] among many articles and monographs.

2. The main result. Our assumptions are as follows (we use the familiar convention $\sup \emptyset = -\infty$, and we write u.s.c. instead of upper semicontinuous).

(F) $\qquad f : B \to \mathbb{R}$ is quasi-convex and directionally u.s.c.:

$$\text{for any} \quad x \in B\ ,\ v \in X\backslash\{0\} : \quad \limsup_{\substack{(t,u) \to (0_+, v) \\ x+tu \in B}} f(x+tu) \le f(x).$$

This assumption is satisfied when f is quasi-convex and u.s.c. on B; on the other hand, it implies the upper semicontinuity of f along lines, which is the corresponding assertion of [20]. When X is finite dimensional (as in [20]), directional upper semicontinuity is equivalent to upper semicontinuity, as a compactness argument shows.

Our assumption on g is slightly reinforced (the closedness of the epigraph of g is not required in [20]):

$$(G) \qquad g \quad \text{has a closed convex epigraph} \quad \text{epi } g.$$

The qualification condition we choose among known ones has become classical (see [1], [2], [4], [13], [17]). It is as follows:

$$(H) \qquad W := \mathbb{R}_+(g(B) - C) \quad \text{is a closed vector subspace of} \quad Z.$$

This condition is obviously satisfied when the following assumption of [3], [26], [38], [40] holds:

$$(H_0) \qquad Z = \mathbb{R}_+(g(B) - C).$$

In turn, as $g(B)-C$ is convex (being the projection on Z of epi g), the latter condition is equivalent to

$$0 \in \text{ core } (g(B) - C),$$

i.e., $g(B) - C$ is absorbent. In fact, as the proof of the theorem below shows, under condition (H_0), the set $g(B)-C$ has a nonempty interior, so that (H_0) can be rewritten

$$(H_0') \qquad 0 \in \text{ int } (g(B) - C).$$

However, condition (H_0) is easier to check than (H_0') since it is just the algebraic condition: given $z \in Z$ there exist $r \in \mathbb{R}_+$, $b \in B$, $c \in C$ such that $z = r(g(b) - c)$. Both conditions are obviously satisfied when the following Slater condition holds:

$$(S) \qquad \text{there exists} \quad x_1 \in B \quad \text{such that} \quad g(x_1) \in \text{ int } C.$$

A discussion about the qualification conditions (H) and (S) is contained in [1], [3], [4], [13], [17], [26]. Let us make brief observations about these relationships. The fact that (H) is more general than (S) is clear.

Example 2.1. Taking for C a closed convex cone with empty interior in $Z = X = B$, g being the identity map I_X, we see that (H_0) is satisfied but (S) does not hold.

Condition (H) is called the generalized qualification condition in [13]; it amounts to $0 \in \text{sqri}(g(B) - C)$, where, for a convex subset D of Z, sqri (D) is the set of $z \in D$ such that $\mathbb{R}_+(D - z)$ is a closed vector subspace. Condition (H) is stronger than the purely algebraic condition

$$(H') \qquad W := \mathbb{R}_+(g(B) - C) \ \text{ is a vector subspace of } Z.$$

Observe that conditions $(F), (G), (H')$ do not entail relation (M), as shown by the following example.

Example 2.2. Let X be a normed vector space (n.v.s.) containing two closed vector subspaces B, C such that $B \cap C = \{0\}$ and $B + C$ is dense in X. As in [30, p. 77] and [13, Example 3.3], one can take $X = l_2$,

$$B = \{x \in l_2 : x_{2n-1} + x_{2n} = 0\ ,\ n = 1, 2, \dots\},$$
$$C = \{x \in l_2 : x_{2n-1} - x_{2n} = 0\ ,\ n = 1, 2, \dots\}.$$

Let $g = I_X$, and let $f(x) = x_1^3$ for $x = (x_1, x_2, \dots)$. Then, for any $y \in Y_+ = \{y \in Y : y|C = 0\}$ one has $\inf\,(f + y \circ g)(B) = -\infty$, but since $B \cap C = \{0\}$, we have $\inf\,\{f(x) : x \in B \cap g^{-1}(C)\} = 0$. Note that $(F), (G)$, and (H') are satisfied.

THEOREM 2.3. *Under assumptions* (F), (G), (H) *there exists a surrogate multiplier* $\bar{y} \in Y_+$*:*

$$\inf\{f(x) : x \in B\ ,\ g(x) \in C\} = \inf\{f(x) : x \in B,\ \langle \bar{y}, g(x)\rangle \le 0\}.$$

Moreover, if $\bar{x} \in B$ *is a solution to* $(\mathcal{P})$*, then it is a solution to*

$$(\mathcal{Q}_{\bar{y}}) \qquad \textit{minimize } f(x) \quad \textit{subject to} \quad x \in B, \langle \bar{y}, g(x)\rangle \le 0.$$

Conversely, if $\bar{x} \in B \cap g^{-1}(C)$ *is a solution to* $(\mathcal{Q}_{\bar{y}})$*, then it is a solution to* $(\mathcal{P})$*.*

Let us observe that for the proof we may assume (H_0) holds instead of (H): since $C \subset W$ and since g takes its values in W, we may replace Z by W and then extend $\bar{y} \in W_+^*$ to some element in $Z^* = Y$ which is still nonpositive on C.

The proof depends on the following result (which holds for spaces more general than Banach spaces, so that the preceding theorem is valid for such spaces; see [33], for instance).

LEMMA 2.4. (see [31], [38]). *Let* $G : X \to 2^Z$ *be a relation with closed convex graph such that* $Z = \mathbb{R}_+ G(X)$*. Then for any* $v \in G^{-1}(0)$*,* G *is open at* $(v, 0)$*; for each neighborhood* V *of* v*,* $G(V)$ *is a neighborhood of* 0 *in* Z*.*

Proof of Theorem 2.3. Let $G : X \to 2^Z$ be the relation given by $G(x) = g(x) - C$ if $x \in B$, and $G(x) = \emptyset$ if $x \in X \backslash B$. The graph of G is nothing but epi g and hence is closed and convex. Moreover, $Z = \mathbb{R}_+ G(X)$. Let $v_0 \in G^{-1}(0)$ i.e., $v_0 \in B$, and $g(v_0) \in C$ (such a v_0 exists by (H) and a well-known argument [1]) and let $x_0 \in S_m$, where

$$S_m = \{x \in B : f(x) < m\}\ ,$$

m being the value of $(\mathcal{P})$, $m = \inf\ f(B \cap g^{-1}(C))$. If such an x_0 does not exist then any $\bar{y} \in Y_+$ can be chosen since in that case $B \cap g^{-1}(C) \subset B \cap (\bar{y} \circ g)^{-1}(\mathbb{R}_-) \subset B$ and $f(x) \ge m$ for each $x \in B$. Now one has $f(x_0) < m \le f(v_0)$. Thus one has $v_0 - x_0 \ne 0$. As f is directionally u.s.c., there exist $t \in]0, 1[$ and a neighborhood V of v_0 in X such that for each $v \in V \cap B$, one has $f(x_0 + t(v - x_0)) < m$, i.e., $(1-t)x_0 + t(V \cap B) \subset S_m$.

Then, by the convexity of g,

$$(1-t)g(x_0) + tg(V \cap B) \subset g((1-t)x_0 + t(V \cap B)) - C \subset g(S_m) - C.$$

It follows that

$$(1-t)g(x_0) + t(g(V \cap B) - C) \subset g(S_m) - C,$$

i.e.,

$$(1-t)g(x_0) + tG(V) \subset g(S_m) - C,$$

so that by Lemma 2.4 $g(S_m) - C$ contains the neighborhood $(1-t)g(x_0) + tG(V)$ of $(1-t)g(x_0)$. By definition of m and S_m, the convex set $g(S_m) - C$ does not contain 0. As it has a nonempty interior, we can find $\bar{y} \in Y \backslash \{0\}$ such that

$$(*) \qquad \langle \bar{y}, z \rangle \geq 0 \quad \text{for each} \quad z \in g(S_m) - C.$$

It follows that $\langle \bar{y}, z \rangle \leq 0$ for each $z \in C : \bar{y} \in C^\circ$. Let us show that $\langle \bar{y}, z \rangle = 0$ for some $z \in g(S_m) - C$ is impossible. Otherwise, we can choose $x_0 \in S_m$, t, and V as above with, in addition, $g(x_0) - z \in C$. Then the linear functional $\bar{y}$ is nonnegative on the neighborhood $(1-t)g(x_0) + tG(V)$ of $(1-t)g(x_0)$ (and even on $g(S_m) - C$) and $\langle \bar{y}, (1-t)g(x_0)) = (1-t)\langle \bar{y}, g(x_0) - z \rangle \leq 0$. As a nonzero linear functional has no local minimum, we get a contradiction. Therefore, $\langle \bar{y}, z \rangle > 0$ for each $z \in g(S_m) - C$, and we have, for each $x \in B$,

$$\langle \bar{y}, g(x) \rangle \leq 0 \Longrightarrow g(x) \notin g(S_m) - C \Longrightarrow f(x) \geq m$$

so that

$$\inf\{f(x) : x \in B, \langle \bar{y}, g(x) \rangle \leq 0\} \geq m.$$

In fact, equality holds as $\{x \in B : \langle \bar{y}, g(x) \rangle \leq 0\}$ contains $B \cap g^{-1}(C)$. The two last assertions are obvious. □

The following result is a counterpart to [20, Theorem 2]. Here we do not assume that C has a nonempty interior, but f is supposed to be *semistrictly quasi-convex* in the following sense: for any $x_0, x_1 \in X$ with $f(x_0) < f(x_1)$ and for any $t \in [0,1[$ one has $f((1-t)x_0 + tx_1) < f(x_1)$. This assumption is satisfied when f is convex or strictly quasi-convex.

THEOREM 2.5. *Suppose with $(F), (G), (H)$ that f is semistrictly quasi-convex. Then, for any solution $\bar{x}$ to $(\mathcal{P})$ one can find $\bar{y} \in M_S$ such that $\langle \bar{y}, g(\bar{x}) \rangle = 0$.*

Proof. As observed in the proof of Theorem 2.3, when S_m is empty one has $M_S = Y_+$, so that we can take $\bar{y} = 0$. Let us suppose S_m is nonempty. For any $x_0 \in S_m$ and any $t \in [0,1[$ we have $x_t := (1-t)x_0 + t\bar{x} \in S_m$ since B is convex and f is semistrictly quasi-convex. It follows from relation $(*)$ that $\langle \bar{y}, g(x_t) \rangle \geq 0$, where $\bar{y}$ is as in the proof of Theorem 2.3. Since $t \longmapsto \langle \bar{y}, g(x_t) \rangle$ is convex, we have $\langle \bar{y}, g(\bar{x}) \rangle \geq 0$. Since $g(\bar{x}) \in C$ and $\bar{y} \in Y_+$, we get $\langle \bar{y}, g(\bar{x}) \rangle = 0$. □

3. Duality results. Theorem 2.3 can be interpreted in terms of dual problems. Here we set $r \vee s = \max(r, s)$.

COROLLARY 3.1. *Under assumptions $(F), (G), (H)$ the dual problems*

$$(\mathcal{D}_S) \qquad \textit{maximize } d_S(y) := \inf\{f(x) : x \in B \ , \ \langle y, g(x) \rangle \leq 0\} \textit{ over } Y_+ = C^\circ,$$

$$(\mathcal{D}'_S) \qquad \textit{maximize } d'_S(y) := \inf\{f(x) \vee (m + \langle y, g(x) \rangle) : x \in B\} \textit{ over } Y_+ = C^\circ$$

have solutions and their values are equal to the value of $(\mathcal{P})$.

Proof. This follows from the fact that $d_S(y) \leq m$, $d'_S(y) \leq m$ for each $y \in Y_+$ and $d_S(\bar{y}) = m$, and $d'_S(\bar{y}) = m$ for any surrogate multiplier $\bar{y}$. □

We now extend the Lagrangian introduced in [20, p. 1092] by setting

$$L_S(x,y) = \begin{cases} -\infty & \text{if } (x,y) \in X \times (Y \backslash Y_+), \\ f(x) & \text{if } (x,y) \in B \times Y_+ \text{ and } \langle y, g(x) \rangle \leq 0, \\ +\infty & \text{if } (x,y) \in B \times Y_+ \text{ and } \langle y, g(x) \rangle > 0 \text{ or } (x,y) \in (X \backslash B) \times Y_+. \end{cases}$$

We also extend the function d_S introduced above by setting

$$d_S(y) := \inf \{f(x) : x \in B,\ \langle y, g(x)\rangle \leq 0\} \ \text{ for } y \in Y_+ \ , \ d_S(y) = -\infty \text{ for } y \in Y\backslash Y_+.$$

It follows that

$$d_S(y) = \inf_{x\in X} L_S(x,y) \quad \text{for any } \ y \in Y,$$

and hence $\sup(\mathcal{D}_S) = \sup_{y\in Y} \ \inf_{x\in X} \ L_S(x,y)$.

LEMMA 3.2. *The function L_S is a Lagrangian for $(\mathcal{P})$: if i_A is the indicator function of the admissible set $A := B \cap g^{-1}(C)$ given by $i_A(x) = 0$ if $x \in A$, $i_A(x) = +\infty$ else, and if $f_A = f + i_A$, one has*

$$f_A(x) = \sup_{y\in Y} L_S(x,y) \quad \textit{for each } x \in X.$$

Proof. For $x \in A$, we have $\langle y, g(x)\rangle \leq 0$ for any $y \in Y_+$; hence $L_S(x,y) = f_A(x)$ and $\sup_{y\in Y} L_S(x,y) = f_A(x)$. For $x \in X\backslash A$, either $x \notin B$ or we have $g(x) \notin C$, and the bipolar theorem yields some $y \in Y_+$ such that $\langle y, g(x)\rangle > 0$. In both cases we get $L_S(x,y) = +\infty = f_A(x)$. □

It follows that

$$\inf(\mathcal{P}) = \inf_{x\in X} \sup_{y\in Y} L_S(x,\bar{y}).$$

Recall that $(\bar{x},\bar{y}) \in X \times Y$ is a saddle point of a function $L : X \times Y \to \mathbb{R}\cup\{-\infty,+\infty\}$ if $L(\bar{x},\bar{y})$ is finite and if for any $(x,y) \in X \times Y$ one has

$$L(\bar{x},y) \leq L(\bar{x},\bar{y}) \leq L(x,\bar{y}).$$

Since L_S is a Lagrangian for $(\mathcal{P})$, we can apply a general characterization of its saddle points; see, for instance, [28] in which L is the quasi-convex Lagrangian given by

$$L_<(x,y) = \begin{cases} f(x) + \langle y, g(x)\rangle_+ & \text{for} \quad (x,y) \in X \times Y_+, \\ -\infty & \text{for} \quad (x,y) \in X \times (Y\backslash Y_+). \end{cases}$$

Note that one has $L_< \leq L_S$ and $d_<(y) := \inf_{x\in X} L_<(x,y) \leq d_s(y)$ for each $y \in Y$. Moreover, L_S is quasi-convex in its first variable and quasi-concave (and even quasi-affine) in its second variable.

PROPOSITION 3.3. *Let $(\bar{x},\bar{y}) \in X \times Y$ be such that $L_S(\bar{x},\bar{y})$ is finite. Then the following assertions are equivalent:*

(a) *$(\bar{x},\bar{y})$ is a saddle point of L_S.*

(b) *$\bar{x}$ is a solution to $(\mathcal{P})$, $\bar{y}$ is a solution to $(\mathcal{D}_S)$, and $f(\bar{x}) = d_S(\bar{y})$ is finite.*

(c) *$\bar{x} \in A := B \cap g^{-1}(C)$ and $f(\bar{x}) = d_S(\bar{y})$ is finite.*

In view of this characterization and the analogous one for $L_<$, we get that any saddle point for $L_<$ is a saddle-point for L_S. Thus the search of saddle-points of L_S represents an easier task than the search of saddle-points of $L_<$.

Let us now consider the performance function p associated with the natural perturbation of $(\mathcal{P})$ given by

$$p(w) = \inf \{f(x) : x \in B, g(x) + w \in C\}$$

for $w \in Z$, with the usual convention $\inf \emptyset = +\infty$. As shown in [20, Lemma 3], the extended real-valued function p is quasi-convex and its domain is the convex set

$$\operatorname{dom} p := \{z \in Z : p(z) < +\infty\} = C - g(B).$$

Another important property of p is homotonicity (or isotonicity or monotonicity): for any $w, z \in Z$ with $w - z \in C$, one has $p(w) \leq p(z)$. While in [28] the set of multipliers $M_<$ associated with $L_<$ is related to the lower subdifferential of p at 0, here we relate the set M_S of surrogate multipliers of p to the quasi subdifferential $\partial^* p(0)$ of p at 0 in the sense of Greenberg and Pierskalla [15], where, given $\bar{z} \in p^{-1}(\mathbb{R})$,

$$\partial^* p(\bar{z}) = \{\bar{y} \in Y : \langle \bar{y}, z - \bar{z}\rangle < 0 \quad \forall z \in p^{-1}((-\infty, p(\bar{z})))\}\ .$$

COROLLARY 3.4. *Let $\hat{z} \in Z$. Under assumptions (F), (G), (H), for any $\hat{z} \in$ core $(C - g(B))$ there exists $\hat{y} \in Y_+$ such that*

$$p(\hat{z}) = \inf\{f(x) : x \in B\ ,\ \langle \hat{y}, g(x)\rangle \leq -\langle \hat{y}, \hat{z}\rangle\} \tag{1}$$

or, equivalently,

$$\text{for each} \quad z \in Z\ ,\ \langle \hat{y}, z\rangle \geq \langle \hat{y}, \hat{z}\rangle \Longrightarrow p(z) \geq p(\hat{z}). \tag{2}$$

When $p(\hat{z}) \in \mathbb{R}$, condition (2) amounts to saying that $\hat{y} \in \partial^* p(\hat{z})$.

Proof. Let us first show that the following condition is satisfied for any $\hat{z} \in$ core $(C - g(B))$:

$$\mathbb{R}_+(g(B) - C + \hat{z}) = W. \tag{$H_{\hat{z}}$}$$

Since $-\hat{z} \in$ core $(g(B) - C)$, we can find $s > 0$ such that $-\hat{z} = s(1+s)^{-1}\big(g(\hat{b}) - \hat{c}\big)$ for some $\hat{b} \in B$, $\hat{c} \in C$. Now, given $z \in W$ we can find $r \geq 0$, $b \in B$, $c \in C$ such that $z = r(g(b) - c)$; hence $z = r((1+s)\hat{z} + sg(\hat{b}) + g(b) - c - s\hat{c})$. Since g is convex, there is $c_1 \in C$ such that $sg(\hat{b}) + g(b) = (1+s)g((1+s)^{-1}(s\hat{b} + b)) - c_1$. Then

$$z = r(1+s)\left(\hat{z} + g\left(\frac{s\hat{b} + b}{1+s}\right) - \frac{c_1 + c + s\hat{c}}{1+s}\right) \in \mathbb{R}_+(\hat{z} + g(B) - C).$$

To prove (1) it suffices to replace the mapping $g : x \longmapsto g(x)$ by the mapping $\hat{g} : x \longmapsto g(x) + \hat{z}$ in the proof of the theorem. Then $(H_{\hat{z}})$ is the required assumption for the relation $\hat{G}$ given by $\hat{G}(x) = g(x) + \hat{z} - C$ if $x \in B$ and by $\hat{G}(x) = \emptyset$ if $x \in X \backslash B$ to be open at each point of $\hat{G}^{-1}(0)$. The existence of some multiplier $\hat{y}$ satisfying (1) follows. It remains to prove the equivalence between (1) and (2). Let us suppose that (1) holds. As $\hat{y} \in Y_+$, for each $z \in Z$ the definition of p ensures that

$$p(z) \geq \inf\{f(x) : x \in B\ ,\ \langle \hat{y}, g(x)\rangle \leq -\langle \hat{y}, z\rangle\}. \tag{3}$$

It follows that for each $z \in Z$ such that $\langle \hat{y}, z\rangle \geq \langle \hat{y}, \hat{z}\rangle$ we have

$$p(z) \geq \inf\{f(x) : x \in B\ ,\ \langle \hat{y}, g(x)\rangle \leq -\langle \hat{y}, \hat{z}\rangle\} = p(\hat{z}).$$

Conversely, suppose that (2) holds. Then for every $x \in B$ such that $\langle \hat{y}, g(x)\rangle \leq -\langle \hat{y}, \hat{z}\rangle$ we have $p(\hat{z}) \leq p(-g(x)) \leq f(x)$; this leads to the inequality

$$p(\hat{z}) \leq \inf\{f(x) : x \in B\ ,\ \langle \hat{y}, g(x)\rangle \leq -\langle \hat{y}, \hat{z}\rangle\}.$$

The opposite inequality is obtained by putting $z = \hat{z}$ in (3). □

COROLLARY 3.5. *Suppose $p(0)$ is finite and assumptions (F), (G), and (H) hold. Then the set M_S of surrogate multipliers is nonempty and coincides with the set of solutions to $(\mathcal{D}_S)$, and*

$$M_S = \partial^* p(0) \cap Y_+.$$

Moreover, if $\inf\ f(B) < p(0)$, *one has* $M_S = \partial^* p(0)$.

Proof. If $\inf f(A) = \inf\ f(B)$, then 0 is a global minimizer of p, so that $\partial^* p(0) = Y$ and then $M_S = Y_+$. If $\inf f(A) > \inf\ f(B)$, there exists some $x_0 \in B$ with $p(-g(x_0)) \leq f(x_0) < p(0)$. Let $z_0 = -g(x_0)$. For any $z \in C$ we have $p(z_0 + z) \leq p(z_0) < p(0)$; hence for any $\bar{y} \in \partial^* p(0)$ we get $\langle \bar{y}, z_0 + z\rangle \leq 0$, and since C is a cone, $\langle \bar{y}, z\rangle \leq 0$. Thus $\bar{y} \in Y_+$. Then we can apply Corollary 3.4. □

In the case when g is affine, one gets a simple multiplier rule.

COROLLARY 3.6. *Suppose B is closed and g is affine: $g(x) = \ell(x) - c$ with $c \in Z$ and ℓ linear and continuous. Under assumptions (F), (G), and (H), if $\bar{x}$ is a solution to $(\mathcal{P})$, then there exists $\bar{y} \in Y_+$ such that*

$$0 \in \partial^* f(\bar{x}) + \bar{y} \circ \ell(\bar{x}).$$

Proof. Let $\bar{y} \in Y_+$ be as in the statement of the theorem. For any $x \in X$ with $\bar{y} \circ \ell(x) \leq \bar{y} \circ \ell(\bar{x}) \leq \langle \bar{y}, c\rangle$ we have $\langle \bar{y}, g(x)\rangle \leq 0$; hence $f(x) \geq f(\bar{x})$. This shows that $-\bar{y} \circ \ell$ belongs to $\partial^* f(\bar{x})$. □

In the following result we show that the performance function p has some regularity property. This property is weaker than semicontinuity. Lower semicontinuity is a strong property which has been looked for by many researchers. Here we just get even quasi convexity; recall that p is *evenly quasi-convex* if for each $r \in \mathbb{R}$ the strict inferior level set $\{z \in Z : p(z) < r\}$ is an intersection of open half-spaces or the whole space. This property has been studied or used in various works [7], [10], [21], [25], [29].

COROLLARY 3.7. *Suppose that B is closed and conditions $(F), (G), (H_{\hat{z}})$ hold for each $\hat{z} \in C - g(B)$. Then the performance function is evenly quasi-convex.*

Proof. Given $r \in \mathbb{R}$, let $Z_r := \{z \in Z : p(z) < r\}$, and let $\hat{z} \in Z \backslash Z_r : p(\hat{z}) \geq r$. If $p(\hat{z}) = +\infty$, $\hat{z}$ does not belong to $D := C - g(B) = \operatorname{dom} p$. As D is the image by G (defined in the proof of Theorem 2.3) of the closed convex subset B, D is ideally convex [16]. Since (H_0) holds, $\operatorname{int} D = \operatorname{core} D$ is nonempty by [16] or [17, Proposition 3.2.] Since $\hat{z} \notin D$, there exists $y \in Y \backslash \{0\}$ (in fact $y \in Y_+ \backslash \{0\}$) such that

$$z \in Z_r \subset D = C - g(B) \Longrightarrow \langle y, z\rangle < \langle y, \hat{z}\rangle.$$

If $p(\hat{z}) < +\infty$, then $\hat{z} \in C - g(B)$. By hypothesis, $(H_{\hat{z}})$ is fulfilled; hence there exists $\hat{y} \in Y_+$ such that (2) holds. Then we have for each $z \in Z$

$$z \in Z_r \Longrightarrow p(z) < p(\hat{z}) \Longrightarrow \langle \hat{z}, z\rangle < \langle \hat{y}, \hat{z}\rangle.$$

In both cases there exists an open half space which includes Z_r and does not contain $\hat{z}$. □

Remark 3.8. If $C - g(B)$ is open, then $(H_{\hat{z}})$ holds for each $\hat{z} \in C - g(B)$.

REFERENCES

[1] H. Attouch and H. Brezis, *Duality for the sum of convex functions in general Banach spaces*, in Aspects of Mathematics and Its Applications, J. A. Barroso, ed., North–Holland, Amsterdam, 1986, pp. 125–133.

[2] D. Azé, *Duality for the sum of convex functions in general normed spaces*, Arch. Math. (Basel), 62 (1994), pp. 554–561.

[3] J. Borwein, *A Lagrange multiplier rule and a sandwich theorem for convex relations*, Math. Scand., 48 (1981), pp. 189–204.

[4] J. Borwein and A. S. Lewis, *Partially convex programming* I : *Quasi relative interiors and duality theory,* Math. Programming, 57 (1992), pp. 15–48.

[5] J. Borwein and A. S. Lewis, *Partially convex programming* II : *Explicit lattices models*, Math. Programming, 57 (1992), pp. 49–83.

[6] J.-P. Crouzeix, *Contributions à l'étude des fonctions quasiconvexes*, Thesis, University of Clermont-Ferrand, Clermont-Ferrand, France, 1977.

[7] A. Daniilidis and J.-E. Martinez-Legaz, *Characterizations of evenly convex sets and evenly quasiconvex functions*, J. Math. Anal. Appl., 273 (2002), pp. 58–66.

[8] I. Ekeland and Temam, *Analyse convexe et problèmes variationnels*, Dunod, Paris, 1974.

[9] M. L. Fisher, B. J. Lageweg, J. K. Lenstra, and A. H. G. Rinnooy Kan, *Surrogate duality relaxation for job shop scheduling*, Math. Cent. Amst. Afd. Math. Beslisk. BW 145/81, 17p; (1981).

[10] W. Fenchel, *A remark on convex sets and polarity*, Comm. Sém. Math. Univ. Lund [Medd. Lunds Univ. Mat. Sem.], 1952 (1952), pp. 82–89.

[11] A.M. Geoffrion, *Lagrangean relaxation for integer programming*, Math. Programming Stud., 2 (1974), pp. 82–114.

[12] F. Glover, *Surrogate constraint duality in mathematical programming*, Operations Res., 23 (1975), pp. 434–451.

[13] M. S. Gowda and M. Teboulle, *A comparison of constraint qualifications in infinite-dimensional convex programming*, SIAM J. Control Optim., 28 (1990), pp. 925–935.

[14] H. J. Greenberg and W. P. Pierskalla, *Surrogate mathematical programming*, Operations Res., 18 (1970), pp. 924–939.

[15] H. J. Greenberg and W. P. Pierskalla, *Quasiconjugate functions and surrogate duality*, Cahiers du Centre d'Etudes de Recherche Opérationnelle, 15 (1973), pp. 437–448.

[16] R. B. Holmes, *Geometric Functional Analysis and Its Applications*, Springer-Verlag, New York, 1975.

[17] V. Jeyakumar and H. Wolkowicz, *Generalization of Slater's constraint qualification for infinite convex programs*, Math. Programming, 57 (1992), pp. 85–101.

[18] H. M. Karvan and L. R. Rardin, *Some relationships between Lagrangian and surrogate duality in integer programming*, Math. Programming, 17 (1979), pp. 320–334.

[19] S. L. Kim and S. Kim, *Exact algorithm for the surrogate dual of an integer programming problem: Subgradient method approach*, J. Optim. Theory Appl., 96 (1998), pp. 363–375.

[20] D. G. Luenberger, *Quasiconvex programming*, SIAM J. Appl. Math., 16 (1968), pp. 1090–1095.

[21] J.-E. Martinez-Legaz, *A generalized concept of conjugation methods*, in Optimization, Theory and Algorithms, J.-B. Hiriart-Urruty, W. Oettli, and J. Stoer, eds., Marcel Dekker, New York, 1983, pp. 45–49.

[22] J.-E. Martinez-Legaz, *On lower subdifferentiable functions*, in Trends in Mathematical Optimization, K. H. Hoffmann et al., eds., Internat. Ser. Numer. Math. 84, Birkhäuser, Basel, 1988, pp. 197–232.

[23] J.-E. Martinez-Legaz, *quasiconvex duality theory by generalized conjugation methods*, Optimization, 19 (1988), pp. 603–652.

[24] J.-E. Martinez-Legaz, *Characterization of R-evenly quasiconvex function*, J. Optim. Theory Appl., 95 (1997), pp. 717–722.

[25] U. Passy and E. Z. Prisman, *Conjugacy in quasiconvex programming*, Math. Programming, 30 (1984), pp. 121–146.

[26] J.-P. Penot, *On regularity conditions in mathematical programming*, Math. Programming Stud., 19 (1982), pp. 167–199.

[27] J.-P. Penot, *Optimality conditions in mathematical programming and composite optimization*, Math. Programming, 67 (1994), pp. 225–245.

[28] J.-P. Penot, *A Lagrangian approach to quasiconvex programming*, J. Optim. Theory Appl., 117 (2003), pp. 637–647.

[29] J.-P. Penot and M. Volle, *On quasi-convex duality*, Math. Oper. Res., 15 (1990), pp. 597–625.
[30] J. Ponstein, *Approaches to the Theory of Optimization*, Cambridge University Press, Cambridge, UK, 1980.
[31] S. M. Robinson, *Regularity and stability for convex multivalued functions*, Math. Oper. Res., 1 (1976), pp. 130–143.
[32] R. T. Rockafellar, *Conjugate Duality and Optimization*, CRMS-NSF Regional Conf. Ser. in Appl. Math. 16, SIAM, Philadelphia, 1974.
[33] B. Rodrigues and S. Simons, *Conjugate functions and subdifferentials in nonnormed situations for operators with complete graphs*, Nonlinear Anal., 12 (1988), pp. 1069–1078.
[34] A. M. Rubinov, X. X. Huang, and X. Q. Yang, *The Zero Duality Gap Property, and Lower Semicontinuity of the Perturbation Function*, preprint, University of Ballarat, Victoria, Australia, 2001.
[35] I. Singer, *Surrogate dual problems and surrogate Lagrangians*, J. Math. Anal. Appl., 98 (1984), pp. 31–71.
[36] I. Singer, *A general theory of surrogate dual and perturbational extended surrogate dual optimization problems*, J. Math. Anal. Appl., 104 (1984), pp. 351–389.
[37] I. Singer, *Abstract Convex Analysis*, Wiley, New York, 1997.
[38] C. Ursescu, *Multifunctions with convex closed graph*, Czechoslovak Math. J., 25 (1975), pp. 438–441.
[39] M. Volle, *Some applications of the Attouch-Brézis condition to closedness criterious, optimization, and duality*, Sém. Anal. Convexe, 22 (1992).
[40] J. Zowe and S. Kurcyusz, *Regularity and stability for the mathematical programming problems in Banach spaces*, Appl. Math. Optim., 5 (1979), pp. 49–62.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 2004–2015

FIRST ORDER CONDITIONS FOR NONSMOOTH DISCRETIZED CONSTRAINED OPTIMAL CONTROL PROBLEMS*

XIAOJUN CHEN†

Abstract. This paper studies first order conditions (Karush–Kuhn–Tucker conditions) for discretized optimal control problems with nonsmooth constraints. We present a simple condition which can be used to verify that a local optimal point satisfies the first order conditions and that a point satisfying the first order conditions is a global or local optimal solution of the optimal control problem.

Key words. optimal control, first order condition, nonsmooth, discretization

AMS subject classifications. 49K20, 35J25

DOI. 10.1137/S0363012902414160

1. Introduction.

We consider the distributed optimal control problem

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\int_\Omega (y-y_d)^2 d\omega + \frac{\alpha}{2}\int_\Omega (u-u_d)^2 d\omega \\ &\text{subject to} \quad -\triangle y + \lambda \max(0,y) = u \quad \text{in} \quad \Omega, \quad y = g \quad \text{on}\,\Gamma, \\ &\qquad u \in U, \end{aligned} \tag{1.1}$$

where $y_d, u_d \in L^2(\Omega)$, $g \in C(\Gamma)$, $\alpha > 0$, and $\lambda > 0$ are constants, Ω is an open bounded convex subset of $R^N, N \le 3$, with smooth boundary Γ, and

$$U = \{u \in L^2(\Omega) \,|\, u(x) \le q(x) \text{ a.e in } \Omega\},$$

$q \in L^\infty(\Omega)$.

This problem is a special case of semilinear elliptic control problems whose constraints involve a semilinear elliptic equation [3]

$$\begin{aligned} \triangle y &= f(x,y,u) \quad \text{in} \quad \Omega, \\ y &= g \quad \text{on} \quad \Gamma, \end{aligned}$$

where $f : \Omega \times R^2 \to R$ is a continuous function. Optimality conditions for semilinear elliptic control problems have been studied extensively. However, most of papers assume that f is continuously differentiable with respect to the second and third variables [3]. These results are not applicable to (1.1) because the elliptic equation in (1.1) has a nonsmooth term $\lambda \max(0,y)$. Such nonsmooth equations can be found in equilibrium analysis of confined magnetohydrodynamics (MHD) plasmas [4, 5, 11], thin stretched membranes partially covered with water [9], or reaction-diffusion problems [1].

In this paper, we study first order conditions for the discretized nonsmooth constrained optimal control problems derived from a finite difference approximation or a finite element approximation of (1.1), which has the form

$$\text{minimize } \frac{1}{2}(y-y_d)^T H(y-y_d) + \frac{\alpha}{2}(u-u_d)^T M(u-u_d)$$

*Received by the editors September 9, 2002; accepted for publication (in revised form) June 26, 2003; published electronically January 22, 2004.
http://www.siam.org/journals/sicon/42-6/41416.html
†Department of Mathematical System Science, Hirosaki University, Hirosaki, Japan (chen@cc.hirosaki-u.ac.jp).

$$\begin{aligned} &\text{subject to } Ay + \lambda D \max(0, y) = Nu, \\ &\qquad\qquad u \le b. \end{aligned} \tag{1.2}$$

Here $y_d \in R^n$, $u_d, b \in R^m$, $H \in R^{n\times n}$, $M \in R^{m\times m}$, $A \in R^{n\times n}$,$D \in R^{n\times n}$, $N \in R^{n\times m}$, and $\max(\cdot)$ is understood coordinatewise. Moreover, H, M, A, D are symmetric positive definite matrices. We assume that D is a diagonal matrix. This assumption holds for finite difference discretization and finite element discretization with mass lumping.

For every $u \in R^m$, there is a unique vector y satisfying the equality constraints in (1.2), since $Ay + \lambda D \max(0, y)$ is strongly monotone. Therefore, (1.2) is equivalent to

$$\begin{aligned} &\text{minimize } \frac{1}{2}(y(u) - y_d)^T H(y(u) - y_d) + \frac{\alpha}{2}(u - u_d)^T M(u - u_d) \\ &\text{subject to} \qquad u \le b, \end{aligned} \tag{1.3}$$

where $y(u)$ is the solution function defined by the equations in the constraints of (1.2). We show that $y(\cdot)$ is a piecewise linear function in the next section.

Let $E(y)$ be an $n \times n$ diagonal matrix whose diagonal elements satisfy

$$E_{ii}(y) = \begin{cases} 1 & \text{if } y_i > 0, \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to verify that $E(y)$ is the Jacobian of the function $\max(0, \cdot)$ at y if y has no zero component.

Since A is a symmetric positive definite matrix and $\lambda DE(y)$ is a nonnegative diagonal matrix, the matrix $A + \lambda DE(y)$ is nonsingular and its inverse is symmetric positive definite.

We say (y, u) satisfies first order conditions for (1.2) or (y, u) is a Karush–Kuhn–Tucker (KKT) point of (1.2) if it together with some $(s, t) \in R^n \times R^m$ satisfies

$$\begin{pmatrix} H(y - y_d) + As + \lambda DE(y)s \\ \alpha M(u - u_d) - N^T s + t \\ Ay + \lambda D \max(0, y) - Nu \\ \min(t, b - u) \end{pmatrix} = 0. \tag{1.4}$$

The vectors $s \in R^n$ and $t \in R^m$ are referred to as Lagrange multipliers.

We say u satisfies first order conditions for (1.3) or u is a KKT point of (1.3) if it together with some $t \in R^m$ satisfies

$$\begin{pmatrix} ((A + \lambda DE(y(u)))^{-1}N)^T H(y(u) - y_d) + \alpha M(u - u_d) + t \\ \min(t, b - u) \end{pmatrix} = 0. \tag{1.5}$$

For $\lambda = 0$, the constraints in (1.1) involve only linear Dirichlet problems. In this case, problem (1.2) is a convex programming problem with linear constraints, and the function $y(\cdot)$ in problem (1.3) can be expressed explicitly as $y(u) = A^{-1}Nu$. Moreover, (1.4) and (1.5) are equivalent in the sense that if (y, u) is a KKT point of (1.2), then u is a KKT point of (1.3); conversely, if u is a KKT point of (1.3), then $(A^{-1}Nu, u)$ is a KKT point of (1.2). Furthermore, the convexity implies that (y, u) is a KKT point of (1.2) if and only if (y, u) is a global solution of (1.2). Therefore, problems (1.2), (1.3), (1.4), and (1.5) are equivalent in the case $\lambda = 0$. Many algorithms for solving

(1.1) based on the equivalent relation have been developed; we refer the reader to a comprehensive paper written by Bergounioux, Ito, and Kunisch [2].

For $\lambda > 0$, the constraints in (1.1) involve nonsmooth partial differential equations. In this case, problem (1.2) is a nonconvex programming problem with nondifferentiable constraints. It fails to satisfy the constraint qualification in mathematical programming [7] in the sense that the set of feasible directions and the set of feasible directions for the linearized constraint set are not same at points where the constraint is not differentiable. There are examples in [7] which show that a KKT point is not necessarily a minimizer for a nonconvex programming problem, and a minimizer is not necessarily a KKT point if the constraint qualification fails.

Recently many numerical methods for solving nonsmooth equations have been developed [4, 5, 8, 12]. We can find a solution of nonsmooth equation (1.4) or (1.5) by a fast (superlinearly convergent) algorithm. However, we do not know if the solution of (1.4) or (1.5) is a minimizer of (1.2) or (1.3). There are open questions in the relations between the four problems:

$$\begin{array}{ccc|ccc} (1.2) & \Leftrightarrow & (1.3) & & (1.2) & \Leftrightarrow & (1.3) \\ \Updownarrow & & \Updownarrow & | & ? & & ? \\ (1.4) & \Leftrightarrow & (1.5) & | & (1.4) & ? & (1.5) \\ & & & | & & & \\ \multicolumn{3}{c|}{\lambda = 0} & & \multicolumn{3}{c}{\lambda > 0} \end{array}$$

In this paper, we provide a necessary and sufficient condition for the solution function $y(\cdot)$ to be differentiable at a point u. By using the differentiability results, we show that problems (1.4) and (1.5) are equivalent. Moreover, we present a simple condition which can be used to verify that a local optimal solution of (1.3) is a solution of (1.5) and that a solution of (1.4) is a global or local optimal solution of (1.2).

We introduce our notation. For any matrix $B \in R^{m\times n}$, let $B_{\mathcal{K}\mathcal{J}}$ be the submatrix of B whose entries lie in the rows of B indexed by $\mathcal{K}$ and the columns indexed by $\mathcal{J}$. If $\mathcal{J} = \{1, 2, \ldots, n\}$, we simply denote $B_{\mathcal{K}\mathcal{J}}$ by $B_{\mathcal{K}}$. Let $e_i \in R^n$ be the ith column of the identity matrix $I \in R^{n\times n}$.

2. Differentibility. In this section, we study the function

$$F(y, u) = Ay + \lambda D \max(0, y) - Nu$$

and the solution function $y(u)$ defined by

$$F(y, u) = 0.$$

For a given $y \in R^n$, we define the index sets

$$\begin{aligned} \mathcal{J}(y) &:= \{\, i \,|\, y_i > 0\}, \\ \mathcal{K}(y) &:= \{\, i \,|\, y_i = 0\}, \\ \mathcal{L}(y) &:= \{\, i \,|\, y_i < 0\}. \end{aligned}$$

Note that $\mathcal{J}(y)$, $\mathcal{K}(y)$, and $\mathcal{L}(y)$ are mutually disjoint, and $\mathcal{J}(y) \cup \mathcal{K}(y) \cup \mathcal{L}(y) = \{1, 2, \ldots, n\}$. Using the function E, we can write the functions F and $y(\cdot)$ as follows:

$$F(y, u) = (A + \lambda DE(y))y - Nu$$

and

$$y(u) = (A + \lambda DE(y(u)))^{-1} Nu.$$

Moreover, for any $u, v \in R^m$, we have

$$\max(0, y(u)) - \max(0, y(v)) = V(y(u) - y(v)),$$

where V is an $n \times n$ diagonal matrix whose diagonal elements are defined by

$$V_{ii} = \begin{cases} 1, & y_i(u) > 0, \quad y_i(v) > 0, \\ \frac{y_i(u)}{y_i(u) - y_i(v)}, & y_i(u) > 0, \quad y_i(v) \le 0, \\ \frac{-y_i(v)}{y_i(u) - y_i(v)}, & y_i(u) \le 0, \quad y_i(v) > 0, \\ 0, & y_i(u) \le 0, \quad y_i(v) \le 0. \end{cases}$$

Since $V_{ii} \in [0, 1]$, $A + \lambda DV$ is symmetric positive definite and it holds that

$$\begin{aligned} \|y(u) - y(v)\|_2 &= \|(A + \lambda DV)^{-1} N(u - v)\|_2 \\ &\le \|A^{-1}\|_2 \|N\|_2 \|u - v\|_2. \end{aligned}$$

Hence y is a Lipschitz continuous function.

THEOREM 2.1. (i) *The function $F : R^n \times R^m \to R^n$ is differentiable at (y, u) if and only if $\mathcal{K}(y) = \emptyset$; in this case the derivative of F at (y, u) is given by*

$$F'(y, u) = (A + \lambda DE(y), -N).$$

(ii) *The function $y(\cdot) : R^m \to R^n$ is differentiable at u if and only if either $\mathcal{K}(y(u)) = \emptyset$ or*

$$((A + \lambda DE(y(u)))^{-1} N)_{\mathcal{K}(y(u))} = 0; \tag{2.1}$$

in this case the derivative of $y(\cdot)$ at u is given by

$$y'(u) = (A + \lambda DE(y(u)))^{-1} N.$$

Proof. (i) If $\mathcal{K}(y) = \emptyset$, then there is an open neighborhood $\mathcal{N}_y$ of y such that for all $z \in \mathcal{N}_y$, $\mathcal{J}(z) = \mathcal{J}(y), \mathcal{L}(z) = \mathcal{L}(y)$, and

$$F(z, u) \equiv (A + \lambda DE(y))z - Nu.$$

Hence F is differentiable at (y, u) and $F'(y, u) = (A + \lambda DE(y), -N)$.

Suppose that there is an $i \in \mathcal{K}(y)$. Then for any $\epsilon > 0$, we have

$$E(y + \epsilon e_i)(y + \epsilon e_i) = E(y)y + \epsilon e_i$$

and

$$E(y - \epsilon e_i)(y - \epsilon e_i) = E(y)y.$$

It follows that

$$\begin{aligned} &F(y + \epsilon e_i, u) - F(y, u) \\ &= \epsilon A e_i + \lambda D(E(y + \epsilon e_i)(y + \epsilon e_i) - E(y)y) \\ &= (A + \lambda D)(\epsilon e_i) \end{aligned}$$

and

$$\begin{aligned} &F(y - \epsilon e_i, u) - F(y, u) \\ &= -\epsilon A e_i + \lambda D(E(y - \epsilon e_i)(y - \epsilon e_i) - E(y)y) \\ &= -\epsilon A e_i. \end{aligned}$$

This shows that

$$\lim_{\epsilon\downarrow 0}\frac{F(y+\epsilon e_i,u)-F(y,u)}{\epsilon}\neq -\lim_{\epsilon\downarrow 0}\frac{F(y-\epsilon e_i,u)-F(y,u)}{\epsilon}.$$

Hence F is not differentiable with respect to y at (y,u). Consequently, if F is differentiable at (y,u), then $\mathcal{K}(y)=\emptyset$.

(ii) If $\mathcal{K}(y(u))=\emptyset$, the differentiability of $y(\cdot)$ follows directly from the first part of this theorem and the implicit function theorem, Theorem 5.2.4 in [10].

Now we consider the case that $\mathcal{K}(y(u))\neq\emptyset$ and (2.1) holds.

In order to simplify the notation, for a given $\bar{u}\in R^n$, we denote the unique vector $y(\bar{u})\in R^n$ by $\bar{y}$ and the associated index sets by

$$\mathcal{J}=\mathcal{J}(\bar{y}),\ \mathcal{K}=\mathcal{K}(\bar{y}),\ \mathcal{L}=\mathcal{L}(\bar{y}).$$

By the continuity of $y(\cdot)$, there is a neighborhood $\mathcal{N}$ of $\bar{u}$ such that for all $w\in\mathcal{N}$ we have

$$y_{\mathcal{J}}(w)>0\quad\text{and}\quad y_{\mathcal{L}}(w)<0,$$

which implies that $\mathcal{J}\subseteq\mathcal{J}(w)$ and $\mathcal{L}\subseteq\mathcal{L}(w)$. Assume that for a vector $w\in\mathcal{N}$ there is a nonempty subset $\mathcal{K}_1$ of $\mathcal{K}$ such that $\mathcal{J}\cup\mathcal{K}_1=\mathcal{J}(y(w))$. From (2.1) and the equality

$$N(w-\bar{u})=(A+\lambda DE(\bar{y}))(y(w)-\bar{y})+\lambda D(E(y(w))-E(\bar{y}))y(w),$$

we obtain

$$\begin{aligned}0&=((A+\lambda DE(\bar{y}))^{-1}N(w-\bar{u}))_{\mathcal{K}_1}\\&=(y(w)-\bar{y}+\lambda(A+\lambda DE(\bar{y}))^{-1}D(E(y(w))-E(\bar{y}))y(w))_{\mathcal{K}_1}\\&=(I_{\mathcal{K}_1\mathcal{K}_1}+\lambda(A+\lambda DE(\bar{y}))^{-1}_{\mathcal{K}_1\mathcal{K}_1}(DE(y(w)))_{\mathcal{K}_1\mathcal{K}_1})y(w)_{\mathcal{K}_1},\end{aligned}$$

where the last equality uses that $E_{ii}(\bar{y})=0$ for $i\in\mathcal{K}_1$ and

$$E_{ii}(y(w))-E_{ii}(\bar{y})=0\quad\text{for}\quad i\notin\mathcal{K}_1.$$

This implies that $y(w)_{\mathcal{K}_1}=0$, since $\lambda(A+\lambda DE(\bar{y}))^{-1}_{\mathcal{K}_1\mathcal{K}_1}$ and $(DE(y(w)))_{\mathcal{K}_1\mathcal{K}_1}$ are symmetric positive definite. This is a contradiction to $\mathcal{K}_1\subset\mathcal{J}(y(w))$. Hence we have $\mathcal{J}(y(w))=\mathcal{J}$, which gives $E(y(w))=E(\bar{y})$. The results ensure that the solution function $y(\cdot)$ in the neighborhood $\mathcal{N}$ can be expressed by

$$y(w)=(A+\lambda DE(\bar{y}))^{-1}Nw.$$

Hence $y(\cdot)$ is differentiable at $\bar{u}$ and

$$y'(\bar{u})=(A+\lambda DE(\bar{y}))^{-1}N.$$

Conversely, we assume that $y(\cdot)$ is differentiable at $\bar{u}$. According to the positive definite property of $A+\lambda DE(\bar{y})$, for any $h\in R^m$ the system

$$(A+\lambda DE(\bar{y}))z+\phi(z)=Nh \tag{2.2}$$

has a unique solution where

$$\phi_i(z)=\begin{cases}0, & i\in\mathcal{J}\cup\mathcal{L},\\ \lambda D_{ii}\max(0,z_i), & i\in\mathcal{K}.\end{cases}$$

Letting $t > 0$ be sufficiently small, we can therefore be assured that

$$(A + \lambda DE(\bar{y}))y(\bar{u} + th) + \phi(y(\bar{u} + th)) = N(\bar{u} + th).$$

Moreover, from $\bar{y}_{\mathcal{K}} = 0$, we have

$$\phi(y(\bar{u} + th)) - \phi(\bar{y}) = \phi(y(\bar{u} + th)) = \phi(y(\bar{u} + th) - \bar{y}).$$

Hence we find

$$(A + \lambda DE(\bar{y}))(y(\bar{u} + th) - \bar{y}) + \phi(y(\bar{u} + th) - \bar{y}) = N(\bar{u} + th) - N\bar{u} = tNh.$$

Note that $(A + \lambda DE(\bar{y}))(tz) + \phi(tz) = tNh$ for $t > 0$. This establishes that the unique solution of system (2.2) is the directional derivative $y'(\bar{y}; h)$ of $y(\cdot)$ along a direction $h \in R^m$ at $\bar{u}$, since

$$y(\bar{u} + th) - y(\bar{u}) = ty'(\bar{u}; h)$$

for sufficiently small $t > 0$, which gives

$$\lim_{t \downarrow 0} \frac{y(\bar{u} + th) - y(\bar{u})}{t} = y'(\bar{u}; h).$$

Moreover, the differentiability of $y(\cdot)$ at $\bar{u}$ implies

$$y'(\bar{u})h = y'(\bar{u}; h) = -y'(\bar{u}; -h),$$

from which we obtain that

$$(A + \lambda DE(\bar{y}))y'(\bar{u}; h) + \phi(y'(\bar{u}; h)) = Nh$$

and

$$-(A + \lambda DE(\bar{y}))y'(\bar{u}; h) + \phi(-y'(\bar{u}; h)) = -Nh.$$

It follows that

$$\phi(y'(\bar{u}; h)) + \phi(-y'(\bar{u}; h)) = 0.$$

Since ϕ is nonnegative, we have $\phi(y'(\bar{u}; h)) = 0$. Consequently, we obtain

$$y'(\bar{u}; h) = (A + \lambda DE(\bar{y}))^{-1}Nh = y'(\bar{u})h \quad \text{for all } h \in R^m,$$

which implies

$$y'(\bar{u}) = (A + \lambda DE(\bar{y}))^{-1}N.$$

Now we show $(y'(\bar{u}))_{\mathcal{K}} = 0$.

We have found that in a neighborhood $\mathcal{N}$ of $\bar{u}$, the function $y(\cdot)$ is linear and can be expressed by

$$y(w) = \bar{y} + (A + \lambda DE(\bar{y}))^{-1}N(w - \bar{u}) \quad \text{for all} \quad w \in \mathcal{N}. \tag{2.3}$$

From (2.3) and the equalities

$$N(w - \bar{u}) = (A + \lambda DE(\bar{y}))(y(w) - \bar{y}) + \lambda D(E(y(w)) - E(\bar{y}))y(w)$$

we obtain

$$\lambda(A+\lambda DE(\bar{y}))^{-1}D(E(y(w))-E(\bar{y}))y(w)=0.$$

Since $\lambda(A+\lambda DE(\bar{y}))^{-1}D$ is nonsingular, this implies that $(E(y(w))-E(\bar{y}))y(w)=0$, that is, $y_i(w)\le 0$ for all $i\in\mathcal{K}$.

If there is an $i\in\mathcal{K}$ such that $y_i(w)<0$, we can choose $t>0$ sufficiently small such that

$$\tilde{w}=\bar{u}-t(w-\bar{u})\in\mathcal{N}.$$

Then from the linearity of $y(\cdot)$ in the neighborhood $\mathcal{N}$, we get

$$y(\tilde{w})=\bar{y}-t(y(w)-\bar{y})$$

and

$$\lambda(A+\lambda DE(\bar{y}))^{-1}D(E(y(\tilde{w}))-E(\bar{y}))y(\tilde{w})=0.$$

However, $((E(y(\tilde{w}))-E(\bar{y}))y(\tilde{w}))_i=y_i(\tilde{w})>0$. This is a contradiction. Therefore, we have that $y_{\mathcal{K}}(\cdot)\equiv 0$ in the neighborhood $\mathcal{N}$, and thus $(y'(\bar{u}))_{\mathcal{K}}=0$.

Since $\bar{u}$ is arbitrarily chosen, we obtain the deserved result. □

A function $f:R^m\to R^n$ is called *piecewise linear* if there exists a finite number of linear functions $f^{(i)}:R^m\to R^n$, $i\in\{1,\dots,\ell\}$, such that the active index set $\{i|f(u)=f^{(i)}(u)\}$ is nonempty for every $u\in R^m$.

Theorem 2.1 states that there exists a finite number of linear functions

$$y^{(1)}u=A^{-1}Nu,\quad y^{(i)}u=\left(A+\lambda D\sum_{j\in\mathcal{I}_i}e_je_j^T\right)^{-1}Nu,$$

$\mathcal{I}_i\subseteq\{1,2,\dots,n\}$, such that the index set $\{i\mid y(u)=y^{(i)}(u)\}$ is nonempty for every $u\in R^m$. Hence $y(\cdot)$ is a piecewise linear function and the number of pieces is not more than 2^n.

The following example illustrates the differentiability of $y(\cdot)$ in the two cases of Theorem 2.1.

Example 2.1. Let $n=2,m=1,\lambda=1,b=1,D=I$,

$$A=\begin{pmatrix}2&-1\\-1&2\end{pmatrix},\qquad\text{and}\qquad N=\begin{pmatrix}3\\-1\end{pmatrix}.$$

The solution function $y(\cdot)$ can be given explicitly as

$$y(u)=\begin{cases}\begin{pmatrix}u\\0\end{pmatrix}, & u\ge 0,\\ \begin{pmatrix}5u/3\\u/3\end{pmatrix}, & u<0.\end{cases}$$

The solution function $y(\cdot)$ is differentiable at every point in R except $u=0$. Moreover, $2\in\mathcal{K}(y(u))$ for $u>0$. Let

$$y^{(1)}(u)=A^{-1}Nu=\begin{pmatrix}5/3\\1/3\end{pmatrix}u,\quad y^{(2)}(u)=(A+e_1e_1^T)^{-1}Nu=\begin{pmatrix}1\\0\end{pmatrix}u,$$

$$y^{(3)}(u) = (A + e_2 e_2^T)^{-1} N u = \begin{pmatrix} 8/5 \\ 1/5 \end{pmatrix} u, \quad y^{(4)}(u) = (A + I)^{-1} N u = \begin{pmatrix} 1 \\ 0 \end{pmatrix} u.$$

We have

$$\begin{array}{ll} y(u) = y^{(2)}(u) = y^{(4)}(u), & u > 0, \\ y(u) = y^{(1)}(u), & u < 0, \\ y(u) = y^{(1)}(u) = y^{(2)}(u) = y^{(3)}(u) = y^{(4)}(u), & u = 0. \end{array}$$

Hence y is a piecewise linear function and the number of pieces is less than 2^n. It is interesting to notice that $(A+I)^{-1}N = (A+E(y(u)))^{-1}N$ for $u > 0$. We can explain it theoretically.

The generalized Jacobian [6] of the function $\max(0, \cdot)$ at a point y is the convex hull of the set defined by a finite number of matrices:

$$\partial \max(0, y) = \operatorname{co} \left\{ \{E(y)\} \cup \left\{ E(y) + \sum_{i \in \mathcal{I}} e_i e_i^T \quad | \quad \mathcal{I} \subseteq \mathcal{K}(y) \right\} \right\}.$$

LEMMA 2.2. *If $\mathcal{K}(y) \neq \emptyset$ and $((A + \lambda DE(y))^{-1}N)_{\mathcal{K}(y)} = 0$, then for any $W \in \partial \max(0, y)$, we have*

$$(A + \lambda DW)^{-1} N = (A + \lambda DE(y))^{-1} N.$$

Proof. For a fixed point $y \in R^n$ which has r zero components, any element $W \in \partial \max(0, y)$ can be expressed by

$$W = E(y) + \sum_{j=1}^{r} \alpha_j \sum_{i \in \mathcal{I}_j} e_i e_i^T,$$

where $1 \geq \alpha_j \geq 0$ and $\mathcal{I}_j$ are subsets of $\mathcal{K}(y)$. Let

$$V = (A + \lambda DE(y))^{-1} N \quad \text{and} \quad U = (A + \lambda DW)^{-1} N. \tag{2.4}$$

We set $B = A + \lambda DE(y)$, $C = \lambda D(W - E(y))$, $\mathcal{K} = \mathcal{K}(y)$, $\mathcal{M} = \mathcal{J}(y) \cup \mathcal{L}(y)$. From (2.4), we get

$$\begin{pmatrix} B_{\mathcal{M}\mathcal{M}} & B_{\mathcal{M}\mathcal{K}} \\ B_{\mathcal{K}\mathcal{M}} & B_{\mathcal{K}\mathcal{K}} \end{pmatrix} \begin{pmatrix} V_{\mathcal{M}} \\ V_{\mathcal{K}} \end{pmatrix} = \begin{pmatrix} B_{\mathcal{M}\mathcal{M}} & B_{\mathcal{M}\mathcal{K}} \\ B_{\mathcal{K}\mathcal{M}} & (B + C)_{\mathcal{K}\mathcal{K}} \end{pmatrix} \begin{pmatrix} U_{\mathcal{M}} \\ U_{\mathcal{K}} \end{pmatrix}.$$

From $V_{\mathcal{K}} = 0$, we have

$$B_{\mathcal{M}\mathcal{M}}(V_{\mathcal{M}} - U_{\mathcal{M}}) - B_{\mathcal{M}\mathcal{K}} U_{\mathcal{K}} = 0 \tag{2.5}$$

and

$$B_{\mathcal{K}\mathcal{M}}(V_{\mathcal{M}} - U_{\mathcal{M}}) - (B + C)_{\mathcal{K}\mathcal{K}} U_{\mathcal{K}} = 0. \tag{2.6}$$

Since B is symmetric positive definite, $B_{\mathcal{M}\mathcal{M}}$ is nonsingular. Thus (2.5) and (2.6) yield

$$(B_{\mathcal{K}\mathcal{M}} B_{\mathcal{M}\mathcal{M}}^{-1} B_{\mathcal{M}\mathcal{K}} - (B + C)_{\mathcal{K}\mathcal{K}}) U_{\mathcal{K}} = 0. \tag{2.7}$$

The submatrix $B_{\mathcal{K}\mathcal{M}}B_{\mathcal{M}\mathcal{M}}^{-1}B_{\mathcal{M}\mathcal{K}} - (B+C)_{\mathcal{K}\mathcal{K}}$ is the Schur complement of the nonsingular matrix

$$\begin{pmatrix} B_{\mathcal{M}\mathcal{M}} & B_{\mathcal{M}\mathcal{K}} \\ B_{\mathcal{K}\mathcal{M}} & (B+C)_{\mathcal{K}\mathcal{K}} \end{pmatrix},$$

and thus it is nonsingular. It follows from (2.7) that $U_{\mathcal{K}} = 0$. Moreover, (2.5) and $U_{\mathcal{K}} = 0$ imply $V_{\mathcal{M}} = U_{\mathcal{M}}$. We complete the proof. □

LEMMA 2.3. *For any $W \in \partial \max(0, y)$, the following two systems are equivalent:*

$$\begin{pmatrix} H(y - y_d) + As + \lambda DWs \\ \alpha M(u - u_d) - N^T s + t \\ Ay + \lambda D \max(0, y) - Nu \\ \min(t, b - u) \end{pmatrix} = 0 \tag{2.8}$$

and

$$\begin{pmatrix} ((A + \lambda DW)^{-1}N)^T H(y(u) - y_d) + \alpha M(u - u_d) + t \\ \min(t, b - u) \end{pmatrix} = 0. \tag{2.9}$$

Proof. Note that $e_i^T \bar{y} = 0$ for $i \in \mathcal{K}(\bar{y})$, and any element $W \in \partial \max(0, \bar{y})$ can be expressed by

$$W = E(\bar{y}) + \sum_{j=1}^{r} \alpha_j \sum_{i \in \mathcal{I}_j} e_i e_i^T,$$

where $1 \le r \le n, 0 \le \alpha_j \le 1, \mathcal{I}_j \subseteq \mathcal{K}(\bar{y})$. Hence we have

$$\max(0, \bar{y}) = E(\bar{y})\bar{y} = W\bar{y} \quad \text{for any} W \in \partial \max(0, \bar{y}). \tag{2.10}$$

Suppose that $(\bar{y}, \bar{u}, s, t)$ satisfies (2.8). Then we have $y(\bar{u}) = \bar{y}$ and

$$\begin{aligned} &((A + \lambda DW)^{-1}N)^T H(y(\bar{u}) - y_d) + \alpha M(\bar{u} - u_d) + t \\ &= -((A + \lambda DW)^{-1}N)^T (A + \lambda DW)s + \alpha M(\bar{u} - u_d) + t \\ &= -N^T s + \alpha M(\bar{u} - u_d) + t \\ &= 0. \end{aligned}$$

Suppose that $(\bar{u}, t)$ satisfies (2.9). By the definition of $y(\cdot)$ and (2.10), we have $A\bar{y} + \lambda D \max(0, \bar{y}) - N\bar{u} = 0$ with $\bar{y} = y(\bar{u})$. Let $s = -(A + \lambda DW)^{-1}H(\bar{y} - y_d)$. We get

$$H(\bar{y} - y_d) + (A + \lambda DW)s = 0$$

and

$$\begin{aligned} \alpha M(\bar{u} - u_d) - N^T s + t = \alpha M(\bar{u} - u_d) + t \\ + N^T (A + \lambda DW)^{-1} H(\bar{y} - y_d) = 0. \end{aligned}$$

□

3. First order conditions. In this section we show the relations between (1.2), (1.3), (1.4), and (1.5). To simplify the notation, we denote, respectively, the objective functions of (1.2) and (1.3) by

$$f(y,u) = \frac{1}{2}(y-y_d)^T H(y-y_d) + \frac{\alpha}{2}(u-u_d)^T M(u-u_d)$$

and

$$\theta(u) = \frac{1}{2}(y(u)-y_d)^T H(y(u)-y_d) + \frac{\alpha}{2}(u-u_d)^T M(u-u_d).$$

From Lemma 2.3 and $E(y) \in \partial \max(0,y)$, we can see that (1.4) and (1.5) are equivalent.

THEOREM 3.1. *If $(\bar{y},\bar{u},\bar{s},\bar{t})$ is a solution of* (1.4)*, then $(\bar{u},\bar{t})$ is a solution of* (1.5)*. Conversely, if $(\bar{u},\bar{t})$ is a solution of* (1.5)*, then $(y(\bar{u}),\bar{u},-(A+\lambda DE(y(\bar{u})))^{-1}H(y(\bar{u})-y_d),\bar{t})$ is a solution of* (1.4)*.*

Now we show that (1.4) implies (1.2) and that (1.3) implies (1.5) under certain conditions.

THEOREM 3.2. *If (y^*,u^*,s^*,t^*) is a solution of* (1.4)*, then all feasible points (y,u) of* (1.2) *satisfy*

$$f(y,u) \geq f(y^*,u^*) + \lambda(D(E(y)-E(y^*))y)^T s^*.$$

Moreover, if either $\mathcal{K}(y^*) = \emptyset$ or $((A+\lambda DE(y^*))^{-1}N)_{\mathcal{K}(y^*)} = 0$, then (y^*,u^*) is a strict local optimal solution of (1.2)*.*

Proof. The objective function in (1.2) is quadratic. If (y^*,u^*,s^*,t^*) is a solution of (1.4), then we have

$$\begin{aligned}
&f(y,u) - f(y^*,u^*)\\
&= (y-y^*)^T H(y^*-y_d) + \alpha(u-u^*)^T M(u^*-u_d)\\
&\qquad + \frac{1}{2}(y-y^*)^T H(y-y^*) + \frac{\alpha}{2}(u-u^*)^T M(u-u^*)\\
&\geq (y-y^*)^T H(y^*-y_d) + \alpha(u-u^*)^T M(u^*-u_d)\\
&\geq (y-y^*)^T H(y^*-y_d) + \alpha(u-u^*)^T M(u^*-u_d) + (u-b)^T t^*\\
&= -(y-y^*)^T(A+\lambda DE(y^*))s^* + (u-u^*)^T(N^T s^* - t^*) + (u-u^*+u^*-b)^T t^*\\
&= -(y-y^*)^T(A+\lambda DE(y^*))s^* + (u-u^*)^T N^T s^*\\
&= -((A+\lambda DE(y^*))(y-y^*) - N(u-u^*))^T s^*\\
&= \lambda(D(E(y)-E(y^*))y)^T s^*,
\end{aligned}$$

where the second inequality uses $u \leq b$ and $t^* \geq 0$, the third equality uses $(u^*-b)^T t^* = 0$, and the fifth equality uses $(A+\lambda DE(y))y = Nu$.

If $\mathcal{K}(y^*) = \emptyset$ or $((A+\lambda DE(y^*))^{-1}N)_{\mathcal{K}(y^*)} = 0$, by Theorem 2.1, there is a neighborhood $\mathcal{N}_y$ of y^* such that for all *feasible* points $y \in \mathcal{N}_y$, we have $y = (A+\lambda DE(y^*))^{-1}Nu$ and

$$(E(y)-E(y^*))y = (E(y)-E(y^*))(A+\lambda DE(y^*))^{-1}Nu = 0.$$

Hence (y^*,u^*) is a local optimal solution of (1.2). Moreover, the first inequality above is strict if $y \neq y^*$, which implies that (y^*,u^*) is a strict local optimal solution. ∎

THEOREM 3.3. *Suppose that u^* is a local optimal solution of* (1.3). *If either $\mathcal{K}(y(u^*)) = \emptyset$ or $((A + \lambda DE(y(u^*)))^{-1}N)_{\mathcal{K}(y(u^*))} = 0$, then there is t^* such that (u^*, t^*) is a solution of* (1.5).

Proof. By Theorem 2.1, the function $y(\cdot)$ is differentiable at u^* and can be expressed by

$$y(u) = (A + \lambda DE(y^*))^{-1}Nu$$

in a neighborhood $\mathcal{N}_u$ of u^*. The gradient of θ at u^* is

$$\theta'(u^*) = ((A + \lambda DE(y^*))^{-1}N)^T H(y(u^*) - y_d) + \alpha M(u^* - u_d).$$

Let $W = (A + \lambda DE(y^*))^{-1}N$. For all $u \in \mathcal{N}_u$ we have

$$\theta(u) = \theta(u^*) + \theta'(u^*)^T(u - u^*) + (u - u^*)(W^T HW + \alpha M)(u - u^*) \geq \theta(u^*),$$

which implies that there is a sufficiently small $\epsilon > 0$ such that if $\|u - u^*\| \leq \epsilon$, then

$$\theta'(u^*)^T(u - u^*) \geq 0. \tag{3.1}$$

By choosing m feasible points

$$u_j^{(i)} = \begin{cases} u_i^* - \epsilon, & j = i, \\ u_j^*, & j \neq i, \end{cases}$$

for $i = 1, 2, \ldots m$, we find $\theta'(u^*) \leq 0$. From $u^* \leq b$, this gives

$$\theta'(u^*)^T(b - u^*) \leq 0.$$

If $\theta'(u^*)^T(b - u^*) < 0$, then there is an $i (1 \leq i \leq m)$ such that $\theta_i'(u^*) < 0$ and $b_i - u_i^* > 0$. We define a feasible point $\tilde{u}$ by

$$\tilde{u}_i = u_i^* + \min(\epsilon, b_i - u_i^*), \quad \tilde{u}_j = u_j^* \ (j \neq i).$$

Then we have

$$\theta'(u^*)^T(\tilde{u} - u^*) < 0.$$

This contradicts (3.1). Hence we obtain $\theta'(u^*)^T(b - u^*) = 0$. Let $t^* = -\theta'(u^*)$. Then (u^*, t^*) is a solution of (1.5). We complete the proof. ∎

COROLLARY 3.4.

1. *Let (y^*, u^*, s^*, t^*) be a solution of* (1.4). *If $y^* = y_d$, then (y^*, u^*) is a global optimal solution of* (1.2).
2. *Let u^* be a local optimal solution of* (1.3). *If $y(u^*) = y_d$, then $(u^*, -\alpha M(u^* - u_d))$ is a solution of* (1.5). *Moreover, u^* is a global optimal solution of* (1.3).

Proof. 1. From the first equality of (1.4), we find that $s^* = 0$. Then the result is derived from Theorem 3.2.

2. By the relation between (1.2) and (1.3), (y_d, u^*) is a local optimal solution of (1.2) and

$$f'(y_d, u^*)^T(u - u^*) = \alpha(u - u^*)^T M(u^* - u_d) \geq 0$$

for u being sufficiently close to u^*. Let $t^* = \alpha M(u^* - u_d)$. We can show that $\max(t^*, b - u^*) = 0$ by using the same technique in the proof for Theorem 3.3. Moreover, by Theorem 3.1 and part 1 of this corollary, u^* is a global optimal solution of (1.3). ∎

4. Final remarks. This paper presents new theoretical results on the first order conditions for the discretization problem arising from nonsmooth constrained optimal control problems. The results generalize existing results for smooth discretized constrained optimal control problems and answer some open questions on first order conditions for two kinds of discretization problems. Theorem 3.1 states that the first order conditions for problem (1.2) are equivalent to the first order conditions for problem (1.3). Theorems 3.2 and 3.3 show that the condition $\mathcal{K}(y) = \emptyset$ and $((A + \lambda DE(y))^{-1}N)_{\mathcal{K}(y)} = 0$ can be used to verify a local optimal point satisfies the first order conditions and that a point satisfying the first order conditions is a local optimal solution. Notice that $((A + \lambda DE(y))^{-1}N)_{\mathcal{K}(y)} = 0$ does not imply that $F(y, u) = Ay + \lambda D \max(0, y) - Nu$ is differentiable at (y, u).

Acknowledgments. The author is very grateful to the anonymous associate editor and referees for their helpful comments and suggestions. The support of the Japan Society of Promotion and Science is also acknowledged.

REFERENCES

[1] A. K. Aziz, A. B. Stephens, and M. Suri, *Numerical methods for reaction-diffusion problems with non-differentiable kinetics*, Numer. Math., 51 (1988), pp. 1–11.
[2] M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.
[3] E. Casas and M. Mateos, *Second order optimality conditions for semilinear elliptic control problems with finitely many state constraints*, SIAM J. Control Optim., 40 (2002), pp. 1431–1454.
[4] X. Chen, N. Matsunaga, and T. Yamamoto, *Smoothing Newton methods for nonsmooth Dirichlet problems,* in Reformulation—Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 65–79.
[5] X. Chen, Z. Nashed, and L. Qi, *Smoothing methods and semismooth methods for nondifferentiable operator equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1200–1216.
[6] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.
[7] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., John Wiley & Sons, New York, 1987.
[8] J. B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.
[9] F. Kikuchi, K. Nakazato, and T. Ushijima, *Finite element approximation of a nonlinear eigenvalue problem related to MHD equilibria*, Japan J. Appl. Math., 1 (1984), pp. 369–403.
[10] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[11] J. Rappaz, *Approximation of a nondifferentiable nonlinear problem related to MHD equilibria*, Numer. Math., 45 (1984), pp. 117–133.
[12] M. Ulbrich, *Nonmonotone trust-region methods for bound-constrained semismooth equations with applications to nonlinear mixed complementarity problems*, SIAM J. Optim., 11 (2001), pp. 889–917.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 2016–2042

A CONVEX APPROACH TO ROBUST STABILITY FOR LINEAR SYSTEMS WITH UNCERTAIN SCALAR PARAMETERS*

PIERRE-ALEXANDRE BLIMAN†

Abstract. In this paper, robust stability for linear systems with several uncertain (complex and/or real) scalar parameters is studied. A countable family of conditions sufficient for robust stability is given, in terms of solvability of some simple linear matrix inequalities (LMIs). These conditions are of increasing precision, and it is shown conversely that robust stability implies solvability of these LMIs from a certain rank and beyond. This result constitutes an extension of the characterization by solvability of Lyapunov inequality of the asymptotic stability for usual linear systems. It is based on the search of parameter-dependent quadratic Lyapunov functions, polynomial of increasing degree in the parameters.

Key words. robust stability, real and complex parametric uncertainty, polytopic uncertainty, parameter-dependent Lyapunov functions, linear matrix inequalities, μ-analysis, structured singular values, Kalman–Yakubovich–Popov lemma

AMS subject classifications. 93D05, 93D09, 90C22, 15A39

DOI. 10.1137/S0363012901398691

1. Introduction.

In this paper we study the robust asymptotic stability of finite-dimensional linear systems subject to several scalar parametric uncertainties, namely

$$\dot{x} = A(z)x, \quad z \stackrel{\text{def}}{=} (z_1, \ldots, z_m), \quad A(z) \stackrel{\text{def}}{=} A_0 + z_1 A_1 + \cdots + z_m A_m, \tag{1}$$

where the fixed matrices $A_0, A_1, \ldots, A_m$ are elements of $\mathbb{C}^{n\times n}$. Here, the uncertain scalar parameters z_i may be complex or real numbers. In the latter case, for the sake of clarity, we shall rather write r_i.

It is a well-known fact that asymptotic stability of system (1) without uncertainty ($z_1 = \cdots = z_m = 0$) is equivalent to the existence of a hermitian matrix $P \in \mathbb{C}^{n\times n}$ such that

$$P > 0_n, \ A_0^H P + P A_0 < 0_n.$$

This is the well-known Lyapunov inequality. This approach is related to the search for a Lyapunov function of the form $x(t)^H P x(t)$, positive definite and decreasing along the trajectories of $\dot{x} = A_0 x$.

This approach has been extended in different ways in order to consider uncertain systems (1). In the various existing variants, one usually considers a set of constant systems, typically compact and convex: the task is to establish whether all the systems in this set are asymptotically stable or not. Various types of parameter sets are in consequence associated to (1), usually elliptic or polytopic. In the present paper, we mainly face the case of constant noncorrelated parameters, with values in closed unit balls of $\mathbb{R}$ or $\mathbb{C}$. In other words, we wish to test the existence of a hermitian matrix $P(z)$ such that

$$P(z) > 0_n, \ A(z)^H P(z) + P(z) A(z) < 0_n \tag{2}$$

*Received by the editors November 23, 2001; accepted for publication (in revised form) May 6, 2003; published electronically January 28, 2004.
http://www.siam.org/journals/sicon/42-6/39869.html
†INRIA, Domaine de Voluceau, Rocquencourt BP 105, 78153 Le Chesnay cedex, France (pierre-alexandre.bliman@inria.fr).

for any $z \in \mathbb{C}^m$ with $|z_i| \leq 1$, $z_i \in \mathbb{R}$, or $\mathbb{C}$, $i = 1, \ldots, m$. This problem appears as a *parameter-dependent linear matrix inequality* (LMI).

This problem is decidable but NP-hard. Indeed, it amounts to evaluating some particular structured singular values [15, 46]. Generally speaking, computing and approximating μ is a hard task [9, 39, 18], and the gap with its usual upper bound is infinite [40, 37]. The more specific problem studied here may be seen equivalently [10] as checking delay-independent stability [27, 28, 23] of a delay system, which has been proved to be NP-hard too [38].

A first method to cope with uncertainty consists of looking for a simultaneous Lyapunov function, i.e., for a constant hermitian positive definite P such that $x(t)^H P x(t)$ decreases along the trajectories of (1), for any value of z in the convenient set; see the bibliography on quadratic stability in [8, pp. 72–73]. Subsequent developments have led to consider *parameter-dependent Lyapunov functions*: sufficient conditions for existence of affine parameter-dependent functions $P(z)$ in (2) are provided in [21, 17, 12, 33] and in [41, 42] for functions quadratic in the parameters. Methods involving piecewise quadratic Lyapunov functions [43, 35] and LMIs with augmented numbers of variables [22, 32] may also be found.

Another approach is based on the use of *scaling* or *multiplier* in an input/output stability framework. The use of diagonal scaling (D-scaling) [15] permits us to obtain upper bound for μ, whereas DG-scaling [16] plays an analog role for real parametric uncertainty. Contributions based on the larger class of LFT-scaling [1] and on multiplier technique [19] have provided less conservative results. Some results are based on mixed methods [13, 20].

The contributions presented previously provide sufficient conditions for *robust stability* of (1), that is, for asymptotic stability for any value of the parameters in the adequate set. However, they are far from being necessary and, due to their conservatism, may fail to detect robust stability. On the other hand, they may be checked easily. Indeed, most of them reduce to testing the solvability of LMI problems, a standard convex optimization problem [8], achievable in polynomial-time. Efficient interior-point methods have been developed and are available as toolboxes in widely spread control-oriented scientific softwares, such as MATLAB or SCILAB.

From a theoretical point of view, the connection between the two methods has been enlightened by Iwasaki [24] and Iwasaki and Hara [26]. Both may be interpreted as special cases of the *quadratic separator*, separating in an appropriate space a graph associated to the "system" from a graph associated to the "perturbation," here the parameters. Roughly speaking, the previous results are obtained when looking for such a separator with prespecified "simple" dependency, with respect to either the frequency (frequency-dependent scaling matrix in μ-analysis) or to parameters (parameter-dependent Lyapunov functions). Clearly, taking small separator classes yields gain in computational simplicity. On the other hand, increasing the separator class size reduces the conservatism of the obtained criterion.

The existing exact methods of resolution of the problem are based on the use of upper and lower bounds on (smaller and smaller) subdomains of the parameter space; see [11, 3, 45]. Due to the computational complexity of the task, they lead to prohibitive growth in computation cost with the problem size, at least in the worst case. The main problem is to find an acceptable trade-off between precision and computational burden.

The results in the present paper provide a systematic way for the use of parameter-dependent Lyapunov functions and their related LMI criteria. The general principle

for their derivation may be explained as follows.

First, the solution $P(z)$ of the Lyapunov equation $A(z)^H P(z) + P(z)A(z) = -I_n$ is *analytic* with respect to the vector of parameters z and its conjugate $\bar{z}$ (this fact may be checked from the explicit form $P(z) = \int_0^{+\infty} e^{A(z)^H t} e^{A(z)t}\, dt$). This suggests that for systems which are robustly stable, there always exists a parameter-dependent Lyapunov function $x(t)^H P(z)x(t)$ with P fulfilling (2) which is *polynomial with respect to* $z, \bar{z}$. Basically, this comes from the fact that one may truncate the latter expansion, due to convergence uniform in z. One hence takes as new unknowns of the problem a positive integer k such that $k-1$ represents the maximal power in the variables $z, \bar{z}$ of the polynomial $P(z)$, plus the k^m coefficients themselves, which are hermitian matrices of size $n \times n$. Second, it turns out that the conditions that must be verified by the previous coefficients (including the global condition of positivity of $P(z)$ for all z) may be transformed into a set of LMIs in a total of $m+1$ unknown hermitian matrices. The main tool for this operation is the application, repeated m times, of the discrete-time Kalman–Yakubovich–Popov lemma.

This two-step procedure motivates the form of the results presented in the core of the paper, which we now summarize. A *family of LMIs* is exhibited, indexed by the positive integer k (roughly speaking, the degree in the $z, \bar{z}$ of a solution of (2)), and whose solvability implies robust stability of system (2). Also, it is shown that solvability for rank k implies solvability for $k' \geq k$, so these sufficient conditions are more and more precise (less and less conservative) as the degree of the polynomial solution increases. A key issue is that a *necessity* property also holds in the precise sense that if robust stability holds, then the corresponding LMIs are fulfilled *from a certain rank k and beyond*. Thus, the conservatism vanishes asymptotically. Robust stability of system (1) is hence *characterized* by solvability of LMI problems. The originality of the proposed method is to associate to a sequence of increasing classes of candidate parameter-dependent Lyapunov functions, whose existence for a precise problem may be checked by solving a LMI, a *completeness* result, ensuring that robust stability implies existence of a Lyapunov function in at least one of the classes. A related idea for generation of parameter-dependent Lyapunov functions based on nonminimal state is used in [25] without, however, insight into the necessity part.

The paper is organized as follows. In section 2 some notation is given necessary for the statement of the results. In section 3 a class of parameter-dependent Lyapunov functions is presented which plays a central role in the paper. In section 4 the two results corresponding to m complex parameters (Theorem 4.1) and m real parameters (Theorem 4.3) are stated. The mixed case may be written down easily and is not extensively developed here. In what follows, we provide as a straightforward consequence a result on robust stability of systems with polytopic uncertainties (Corollary 4.4). A numerical example is presented further on in section 5. Comments on the status of the results are given in section 6. Complete proofs of Theorems 4.1 and 4.3 are given in section 7. Last, concluding remarks are made in section 8.

2. Notation. The matrices I_n, 0_n, $0_{n\times p}$ are the $n \times n$ identity matrix and the $n \times n$ and $n \times p$ zero matrices, respectively. The symbol $\otimes$ denotes Kronecker product, the power of Kronecker products being used with the natural meaning $M^{0\otimes} = 1$, $M^{p\otimes} \stackrel{\text{def}}{=} M^{(p-1)\otimes} \otimes M$. Recall the important property that $(A \otimes B)(C \otimes D) = (AC \otimes BD)$ for any matrices with compatible size. The spectrum of a square matrix M is written $\sigma(M)$, and, applying the operation Re to this set, one denotes by $\mathrm{Re}\,\sigma(M)$ the set $\{\mathrm{Re}\, s \ : \ s \in \sigma(M)\}$. For example, $\mathrm{Re}\,\sigma(M) < 0$ means that M is Hurwitz. The spectral radius of a square matrix M is written $\rho(M)$. The conjugate

and transconjugate of M are denoted M^T and M^H. $\mathbb{N}$ is the set of positive integers. $\overline{\mathbb{D}}$ is denoted the closed unit ball in $\mathbb{C}$. The unit circle is denoted as the boundary $\partial\mathbb{D}$. $\overline{\mathbb{C}^+}$ is the closed set of complex numbers with nonnegative real part. Last, the set of complex hermitian matrices of size $n \times n$ is denoted by $\mathcal{H}^n$.

Let $\hat{J}_k, \check{J}_k \in \mathbb{R}^{k\times(k+1)}$ be defined by

$$\hat{J}_k \stackrel{\text{def}}{=} \begin{pmatrix} I_k & 0_{k\times 1} \end{pmatrix}, \ \check{J}_k \stackrel{\text{def}}{=} \begin{pmatrix} 0_{k\times 1} & I_k \end{pmatrix}.$$

These matrices will prove essential for polynomial manipulation. In particular, a key property is that, for $u^{[k]} \in \mathbb{C}^k$ defined, for $u \in \mathbb{C}$, by

$$u^{[k]} \stackrel{\text{def}}{=} \begin{pmatrix} 1 \\ u \\ \vdots \\ u^{k-1} \end{pmatrix}, \tag{3}$$

one has

$$\hat{J}_k u^{[k+1]} = u^{[k]}, \ \check{J}_k u^{[k+1]} = u u^{[k]}. \tag{4}$$

Also, we will use the fact that, for any $k \in \mathbb{N}$,

$$\hat{J}_k \check{J}_{k+1} = \check{J}_k \hat{J}_{k+1} = \begin{pmatrix} 0_{k\times 1} & I_k & 0_{k\times 1} \end{pmatrix}. \tag{5}$$

Finally, one shows directly that, for any matrix $M \in \mathbb{C}^{p\times q}$, for any $u \in \mathbb{C}$,

$$(u^{[k]} \otimes I_p)M = (I_k \otimes M)(u^{[k]} \otimes I_q). \tag{6}$$

3. Polynomially parameter-dependent quadratic functions and their evolution. In the study of system (1), a crucial role will be played here by the search for parameter-dependent Lyapunov functions chosen within the following class.

DEFINITION 3.1. *We call a* polynomially parameter-dependent quadratic function *(PPDQ function for short) any quadratic function $x^H P(z)x$ on $\mathbb{C}^n$ such that*

$$P(z) \stackrel{\text{def}}{=} (z_m^{[k]} \otimes \cdots \otimes z_1^{[k]} \otimes I_n)^H P_k (z_m^{[k]} \otimes \cdots \otimes z_1^{[k]} \otimes I_n) \tag{7}$$

for a certain $P_k \in \mathcal{H}^{k^m n}$. The integer $k-1$ is called the degree of the PPDQ function P.

Notice that the expression $(z_m^{[k]} \otimes \cdots \otimes z_1^{[k]})$ gathers in a column all the monomials with degree at most $k-1$ in each of the components of z.

The following auxiliary result provides the derivative of a PPDQ function along the trajectories of (1).

PROPOSITION 3.2. *The derivative of the PPDQ function* (7) *of degree $k-1$ along the trajectories of the system $\dot{x} = A(z)x$ is a PPDQ function $R(z)$ of degree k given by*

$$R(z) \stackrel{\text{def}}{=} (z_m^{[k+1]} \otimes \cdots \otimes z_1^{[k+1]} \otimes I_n)^H R_k (z_m^{[k+1]} \otimes \cdots \otimes z_1^{[k+1]} \otimes I_n), \tag{8}$$

where $R_k \in \mathcal{H}^{(k+1)^m n}$ *is defined as*

(9)

$$R_k \stackrel{\text{def}}{=} \left(\left(\hat{J}_k^{m\otimes}\otimes A_0\right)+\sum_{i=1}^{m}\left(\hat{J}_k^{(m-i)\otimes}\otimes \check{J}_k\otimes \hat{J}_k^{(i-1)\otimes}\otimes A_i\right)\right)^H P_k\left(\hat{J}_k^{m\otimes}\otimes I_n\right)$$
$$+\left(\hat{J}_k^{m\otimes}\otimes I_n\right)^T P_k\left(\left(\hat{J}_k^{m\otimes}\otimes A_0\right)+\sum_{i=1}^{m}\left(\hat{J}_k^{(m-i)\otimes}\otimes \check{J}_k\otimes \hat{J}_k^{(i-1)\otimes}\otimes A_i\right)\right)$$

and depends linearly upon $P_k \in \mathcal{H}^{k^m n}$.

Proof of Proposition 3.2. Clearly, $R(z) = A(z)^H P(z) + P(z)A(z)$. As an example, let us evaluate $P(z)A(z)$. One has

$$\begin{aligned}
P(z)A(z) &= (z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n)^H P_k (z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n)A(z)\\
&= (z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n)^H P_k (I_{k^m}\otimes A(z))(z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n)\\
&\qquad \text{(due to (6))}\\
&= (z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n)^H P_k \Big[(I_{k^m}\otimes A_0)(z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n)\\
&\quad +(I_{k^m}\otimes A_1)(z_m^{[k]}\otimes\cdots\otimes z_1 z_1^{[k]}\otimes I_n)\\
&\quad +\cdots+(I_{k^m}\otimes A_m)(z_m z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n)\Big],
\end{aligned}$$

and the second term in (9) is obtained by repeated use of the two formulas in (4). □

To study systems with real parameters $\dot{x} = A(r)x$, we use the change of variables $r = \frac{z+\bar{z}}{2}$, which maps $\overline{\mathbb{D}}^m$ onto $[-1;+1]^m$. It turns out that the formulas are of smaller size when one is directly looking for a Lyapunov function parametrized by z and not by r. The analogue of Proposition 3.2 for this case is given below, and its proof, using the same techniques, is left to the reader.

Proposition 3.3. *The derivative of the PPDQ function* (7) *of degree* $k-1$ *along the trajectories of the system* $\dot{x} = A(\frac{z+\bar{z}}{2})x$ *is a PPDQ function* $R(z)$ *of degree* k *given as* (8)*, where* $R_k \in \mathcal{H}^{(k+1)^m n}$ *is now defined by*

(10)

$$\begin{aligned}
R_k \stackrel{\text{def}}{=}\ & \frac{1}{2}\left(\left(\hat{J}_k^{m\otimes}\otimes A_0\right)+\sum_{i=1}^{m}\left(\hat{J}_k^{(m-i)\otimes}\otimes \check{J}_k\otimes \hat{J}_k^{(i-1)\otimes}\otimes A_i\right)\right)^H P_k\left(\hat{J}_k^{m\otimes}\otimes I_n\right)\\
+\ & \frac{1}{2}\left(\left(\hat{J}_k^{m\otimes}\otimes A_0\right)^H P_k\left(\hat{J}_k^{m\otimes}\otimes I_n\right)+\sum_{i=1}^{m}\left(\hat{J}_k^{m\otimes}\otimes A_i\right)^H P_k\left(\hat{J}_k^{(m-i)\otimes}\otimes \check{J}_k\otimes \hat{J}_k^{(i-1)\otimes}\otimes I_n\right)\right)\\
+\ & \frac{1}{2}\left(\hat{J}_k^{m\otimes}\otimes I_n\right)^T P_k\left(\left(\hat{J}_k^{m\otimes}\otimes A_0\right)+\sum_{i=1}^{m}\left(\hat{J}_k^{(m-i)\otimes}\otimes \check{J}_k\otimes \hat{J}_k^{(i-1)\otimes}\otimes A_i\right)\right)\\
+\ & \frac{1}{2}\left(\left(\hat{J}_k^{m\otimes}\otimes I_n\right)^T P_k\left(\hat{J}_k^{m\otimes}\otimes A_0\right)+\sum_{i=1}^{m}\left(\hat{J}_k^{(m-i)\otimes}\otimes \check{J}_k\otimes \hat{J}_k^{(i-1)\otimes}\otimes I_n\right)^T P_k\left(\hat{J}_k^{m\otimes}\otimes A_i\right)\right)
\end{aligned}$$

and depends linearly upon $P_k \in \mathcal{H}^{k^m n}$.

4. Main results. We are now in a position to state the main results of the paper.

Theorem 4.1 (robust stability of systems with complex parameters). *The following three properties are* equivalent.

(i) *The matrix* $A(z)$ *in* (1) *is Hurwitz for any* $z \in \overline{\mathbb{D}}^m$.

(ii) *There exists a PPDQ Lyapunov function* $x^H P(z)x$ *for the class of systems* $\dot{x} = A(z)x$ *with* $A(z)$ *defined by* (1), *i.e., such that*

$$\forall z \in \overline{\mathbb{D}}^m,\ P(z) > 0,\ R(z) < 0,$$

where $R(z)$ *is defined by* (8), (9).

(iii) *There exist a positive integer* k *and* $(m+1)$ *matrices*

$$P_k \in \mathcal{H}^{k^m n} \text{ and } Q_{k,i} \in \mathcal{H}^{k^{m-i+1}(k+1)^{i-1}n},\ i = 1, \dots, m,$$

which solve the following LMI:

(LMI_k)

$$\begin{cases} P_k > 0_{k^m n}, \\ R_k + \displaystyle\sum_{i=1}^{m} \left(\hat{J}_k^{(m-i+1)\otimes} \otimes I_{(k+1)^{i-1}n}\right)^T Q_{k,i} \left(\hat{J}_k^{(m-i+1)\otimes} \otimes I_{(k+1)^{i-1}n}\right) \\ - \displaystyle\sum_{i=1}^{m} \left(\hat{J}_k^{(m-i)\otimes} \otimes \check{J}_k \otimes I_{(k+1)^{i-1}n}\right)^T Q_{k,i} \left(\hat{J}_k^{(m-i)\otimes} \otimes \check{J}_k \otimes I_{(k+1)^{i-1}n}\right) < 0_{(k+1)^m n}, \end{cases}$$

with $R_k = R_k(P_k)$ *defined in* (9).

Moreover, if (LMI_k) *with* (9) *is solvable for the index* k, *then it is also solvable for all indices* $k' \geq k$. *Finally, if the matrices* A_i, $0 \leq i \leq m$, *are real, then the statement holds with real, symmetric, matrices* $P_k, Q_{k,i}$, $1 \leq i \leq m$.

The proof of Theorem 4.1 is given in section 7.1.

From Theorem 4.1, one deduces in particular that, for any positive integer k,

$$\begin{aligned} &(\mathrm{LMI}_k) \text{ is solvable} \Rightarrow (\mathrm{LMI}_{k'}) \text{ is solvable for } k' \geq k \\ &\qquad\qquad \Rightarrow \text{system (1) is robustly stable against any } z \in \overline{\mathbb{D}}^m. \end{aligned} \tag{11}$$

In other words, any of the conditions (LMI_k) is sufficient for robust stability, and they are more and more precise. Necessity of the condition is obtained asymptotically for large enough k.

The sufficiency result (11) is central and turns out to be the "easy" part of the proof. Before commenting further on Theorem 4.1, we provide indications on its demonstration, leaving the details for the complete proof in section 7.1.

Sketch of proof for (11). Left- and right-multiplication of the second inequality in (LMI_k) by $(z_m^{[k+1]} \otimes \cdots \otimes z_1^{[k+1]} \otimes I_n)$ and its transconjugate yields $R(z) + \sum_{i=1}^m (1 - |z_i|^2)(z_m^{[k]} \otimes \cdots \otimes z_i^{[k]} \otimes z_{i-1}^{[k+1]} \otimes \cdots \otimes z_1^{[k+1]} \otimes I_n)^H Q_{k,i} (z_m^{[k]} \otimes \cdots \otimes z_i^{[k]} \otimes z_{i-1}^{[k+1]} \otimes \cdots \otimes z_1^{[k+1]} \otimes I_n) < 0_n$. Indeed, this is a direct consequence of (4). Thus, $R(z) < 0_n$ if $|z_1| = \cdots = |z_m| = 1$, so the matrix $A(z)$ is Hurwitz for all $z \in (\partial\mathbb{D})^m$, and this may be extended to the whole $\overline{\mathbb{D}}^m$; see the details in section 7.1.1. This proves that solvability of (LMI_k) is sufficient for robust stability.

To prove that solvability of (LMI_k) implies solvability of (LMI_{k+1}), one constructs

directly a new solution $P_{k+1}, Q_{k+1,1}, \dots, Q_{k+1,m}$, by taking

$$P_{k+1} \stackrel{\text{def}}{=} \sum_{M_i \in \{\hat{J}_k, \check{J}_k\},\ i=1,\dots,m} (M_m \otimes \cdots \otimes M_1 \otimes I_n)^T P_k (M_m \otimes \cdots \otimes M_1 \otimes I_n),$$

$$Q_{k+1,i} \stackrel{\text{def}}{=} \sum_{\substack{M_l \in \{\hat{J}_{k+1}, \check{J}_{k+1}\},\ l=1,\dots,i-1,\\ M_l \in \{\hat{J}_k, \check{J}_k\},\ l=i,\dots,m}} (M_m \otimes \cdots \otimes M_1 \otimes I_n)^T Q_{k,i} (M_m \otimes \cdots \otimes M_1 \otimes I_n),$$

for $i = 1, \dots, m$. One then shows that the matrix R_{k+1} obtained from P_{k+1} by formula (9) verifies

$$R_{k+1} \stackrel{\text{def}}{=} \sum_{M_i \in \{\hat{J}_{k+1}, \check{J}_{k+1}\},\ i=1,\dots,m} (M_m \otimes \cdots \otimes M_1 \otimes I_n)^T R_k (M_m \otimes \cdots \otimes M_1 \otimes I_n).$$

As a matter of fact, one may show that this amounts to multiplying $P(z)$ and $R(z)$ by $(1+|z_1|^2)\dots(1+|z_m|^2)$ in (ii). Based on properties (5), (6), the two inequalities of (LMI_{k+1}) are then deduced from the inequalities of (LMI_k); see details in section 7.1.3 below. □

Remark 4.2. Paradoxically, the positive definite PPDQ function $x^H P(z) x$ of degree $k-1$ formed from a solution of (LMI_k) is not ensured to decrease along the trajectories of the system. The argument developed above consisted of showing that $R(z) < 0_n$ for $z \in (\partial\mathbb{D})^m$. As said before, this yields Hurwitzness of $A(z)$ for $z \in (\partial\mathbb{D})^m$, which implies the same property in $\overline{\mathbb{D}}^m$, basically by an analyticity result; see section 7.1.1. However, in general $R(z) \not< 0_n$ for $z \in \overline{\mathbb{D}}^m$, unless $Q_{k,i} > 0_{k^{m-i+1}(k+1)^{i-1}n}$ for all $i = 1, \dots, m$. In the case of a *unique* scalar uncertainty ($m = 1$), there is no loss of generality to add this positivity condition on $Q_{k,1}$ in the LMI; see [5]. We conjecture that the same remains true for $m > 1$.

In the case $m = 0$, the problem (LMI_k) simply states that $\exists P \in \mathcal{H}^n$, $P > 0_n$, $A_0^H P + P A_0 < 0_n$. For $m = 1$, one gets the following family of LMIs indexed by $k \in \mathbb{N}$: $\exists P_k \in \mathcal{H}^{kn}$, $P_k > 0_{kn}$, $\exists Q_k \in \mathcal{H}^{kn}$,

$$\begin{aligned}(\hat{J}_k \otimes A_0)^H P_k (\hat{J}_k \otimes I_n) + (\check{J}_k \otimes A_1)^H P_k (\hat{J}_k \otimes I_n)\\ + (\hat{J}_k \otimes I_n)^T P_k (\hat{J}_k \otimes A_0) + (\hat{J}_k \otimes I_n)^T P_k (\check{J}_k \otimes A_1)\\ + (\hat{J}_k \otimes I_n)^T Q_k (\hat{J}_k \otimes I_n) - (\check{J}_k \otimes I_n)^T Q_k (\check{J}_k \otimes I_n) < 0_{(k+1)n};\end{aligned}$$

that is,

(12)
$$\begin{pmatrix} \hat{J}_k \otimes I_n \\ \check{J}_k \otimes I_n \end{pmatrix}^T \begin{pmatrix} (I_k \otimes A_0)^H P_k + P_k (I_k \otimes A_0) + Q_k & P_k (I_k \otimes A_1) \\ (I_k \otimes A_1)^H P_k & -Q_k \end{pmatrix} \begin{pmatrix} \hat{J}_k \otimes I_n \\ \check{J}_k \otimes I_n \end{pmatrix} < 0_{(k+1)n}.$$

For two parameters ($m = 2$), one obtains $\exists P_k \in \mathcal{H}^{k^2 n}, P_k > 0_{k^2 n}$, $\exists Q_{k,1} \in \mathcal{H}^{k^2 n}$, $\exists Q_{k,2} \in \mathcal{H}^{k(k+1)n}$,

$$\begin{aligned}(\hat{J}_k^{2\otimes} \otimes A_0)^H P_k (\hat{J}_k^{2\otimes} \otimes I_n) + (\hat{J}_k^{2\otimes} \otimes I_n)^T P_k (\hat{J}_k^{2\otimes} \otimes A_0)\\ + (\hat{J}_k \otimes \check{J}_k \otimes A_1)^H P_k (\hat{J}_k^{2\otimes} \otimes I_n) + (\hat{J}_k^{2\otimes} \otimes I_n)^T P_k (\hat{J}_k \otimes \check{J}_k \otimes A_1)\\ + (\check{J}_k \otimes \hat{J}_k \otimes A_2)^H P_k (\hat{J}_k^{2\otimes} \otimes I_n) + (\hat{J}_k^{2\otimes} \otimes I_n)^T P_k (\check{J}_k \otimes \hat{J}_k \otimes A_2)\\ + (\hat{J}_k^{2\otimes} \otimes I_n)^T Q_{k,1} (\hat{J}_k^{2\otimes} \otimes I_n) - (\hat{J}_k \otimes \check{J}_k \otimes I_n)^T Q_{k,1} (\hat{J}_k \otimes \check{J}_k \otimes I_n)\\ + (\hat{J}_k \otimes I_{(k+1)n})^T Q_{k,2} (\hat{J}_k \otimes I_{(k+1)n}) - (\check{J}_k \otimes I_{(k+1)n})^T Q_{k,2} (\check{J}_k \otimes I_{(k+1)n})\\ < 0_{(k+1)^2 n},\end{aligned}$$

or again

(13)
$$\begin{pmatrix} \hat{J}_k \otimes \hat{J}_k \otimes I_n \\ \hat{J}_k \otimes \check{J}_k \otimes I_n \\ \check{J}_k \otimes \hat{J}_k \otimes I_n \end{pmatrix}^T \begin{pmatrix} (I_{k^2} \otimes A_0)^H P_k + P_k (I_{k^2} \otimes A_0) + Q_{k,1} & P_k (I_{k^2} \otimes A_1) & P_k (I_{k^2} \otimes A_2) \\ (I_{k^2} \otimes A_1)^H P_k & -Q_{k,1} & 0_{k^2 n} \\ (I_{k^2} \otimes A_2)^H P_k & 0_{k^2 n} & 0_{k^2 n} \end{pmatrix}$$
$$\times \begin{pmatrix} \hat{J}_k \otimes \hat{J}_k \otimes I_n \\ \hat{J}_k \otimes \check{J}_k \otimes I_n \\ \check{J}_k \otimes \hat{J}_k \otimes I_n \end{pmatrix} + \begin{pmatrix} \hat{J}_k \otimes I_{(k+1)n} \\ \check{J}_k \otimes I_{(k+1)n} \end{pmatrix}^T \begin{pmatrix} Q_{k,2} & 0_{k(k+1)n} \\ 0_{k(k+1)n} & -Q_{k,2} \end{pmatrix} \begin{pmatrix} \hat{J}_k \otimes I_{(k+1)n} \\ \check{J}_k \otimes I_{(k+1)n} \end{pmatrix} < 0_{(k+1)^2 n}.$$

An interesting comparison may be made, concerning the simplest criterion, obtained for $k = 1$. In the case $m = 1$ (see (12)), (LMI_1) writes

$$P_1 = P_1^H > 0,\ Q_{1,1} = Q_{1,1}^H,\ \begin{pmatrix} A_0^H P_1 + P_1 A_0 + Q_{1,1} & P_1 A_1 \\ A_1^H P_1 & -Q_{1,1} \end{pmatrix} < 0,$$

which matches the conditions for quadratic stability with D-scalings. For the case $m = 2$ of two parameters (see (13)), the inequalities are

$$P_1 = P_1^H > 0,\ Q_{1,1} = Q_{1,1}^H,\ Q_{1,2} = Q_{1,2}^H,$$
$$\begin{pmatrix} A_0^H P_1 + P_1 A_0 + Q_{1,1} & P_1 A_1 & P_1 A_2 & 0_n \\ A_1^H P_1 & -Q_{1,1} & 0_n & 0_n \\ A_2^H P_1 & 0_n & 0_n & 0_n \\ 0_n & 0_n & 0_n & 0_n \end{pmatrix} + \begin{pmatrix} Q_{1,2} & 0_{2n} \\ 0_{2n} & -Q_{1,2} \end{pmatrix} < 0,$$

where here the size of $Q_{1,2}$ is twice that of $Q_{1,1}$. This is clearly less restrictive than the conditions obtained with D-scalings, namely,

$$P_1 = P_1^H > 0,\ Q_{1,1} = Q_{1,1}^H,\ Q_{1,2} = Q_{1,2}^H,$$
$$\begin{pmatrix} A_0^H P_1 + P_1 A_0 + Q_{1,1} + Q_{1,2} & P_1 A_1 & P_1 A_2 \\ A_1^H P_1 & -Q_{1,1} & 0_n \\ A_2^H P_1 & 0_n & -Q_{1,2} \end{pmatrix} < 0.$$

For larger values of m, (LMI_1) is obtained by introduction of the remaining multipliers $Q_{1,i}$, along the same principles. The obtained conditions are related to, but less conservative than, the ones obtained with D-scalings.

The result for systems with real parameters is analogous to Theorem 4.1.

THEOREM 4.3 (robust stability of systems with real parameters). *The following three properties are* equivalent.

(i) *The matrix $A(r)$ in* (1) *is Hurwitz for any $r \in [-1;+1]^m$.*

(ii) *There exists a PPDQ Lyapunov function $x^H P(r) x$ for the class of systems $\dot{x} = A(r)x$ with $A(r)$ defined by* (1)*, i.e., such that*

$$\forall r \in [-1;+1]^m,\ P(r) > 0,\ R(r) < 0,$$

where $R(r)$ is defined as in (8)*, with R_k given by* (10).

(iii) *There exist a positive integer k and $(m+1)$ matrices*

$$P_k \in \mathcal{H}^{k^m n} \text{ and } Q_{k,i} \in \mathcal{H}^{k^{m-i+1}(k+1)^{i-1} n}, \quad i = 1, \ldots, m,$$

which solve the (LMI_k) *with $R_k = R_k(P_k)$ defined in* (10).

Moreover, if (LMI$_k$) *with* (10) *is solvable for the index* k*, then it is also solvable for all indices* $k' \geq k$*. Finally, if the matrices* A_i*,* $0 \leq i \leq m$*, are real, then the statement holds with real, symmetric, matrices* $P_k, Q_{k,i}$*,* $1 \leq i \leq m$*.*

The proof of Theorem 4.3 is given in section 7.2.

For $m = 1$ and $m = 2$, respectively, the two following families of LMIs are obtained: $\exists P_k \in \mathcal{H}^{kn}$, $P_k > 0_{kn}$, $\exists Q_k \in \mathcal{H}^{kn}$,

$$(\hat{J}_k \otimes A_0)^H P_k (\hat{J}_k \otimes I_n) + \tfrac{1}{2}\left((\hat{J}_k \otimes A_1)^H P_k (\check{J}_k \otimes I_n) + (\check{J}_k \otimes A_1)^H P_k (\hat{J}_k \otimes I_n)\right)$$
$$+ (\hat{J}_k \otimes I_n)^T P_k (\hat{J}_k \otimes A_0) + \tfrac{1}{2}\left((\check{J}_k \otimes I_n)^T P_k (\hat{J}_k \otimes A_1) + (\hat{J}_k \otimes I_n)^T P_k (\check{J}_k \otimes A_1)\right)$$
$$+ (\hat{J}_k \otimes I_n)^T Q_k (\hat{J}_k \otimes I_n) - (\check{J}_k \otimes I_n)^T Q_k (\check{J}_k \otimes I_n) < 0_{(k+1)^n}$$

and $\exists P_k \in \mathcal{H}^{k^2 n}, P_k > 0_{k^2 n}$, $\exists Q_{k,1} \in \mathcal{H}^{k^2 n}$, $\exists Q_{k,2} \in \mathcal{H}^{k(k+1)n}$,

$$(\hat{J}_k^{2\otimes} \otimes A_0)^H P_k (\hat{J}_k^{2\otimes} \otimes I_n) + (\hat{J}_k^{2\otimes} \otimes I_n)^T P_k (\hat{J}_k^{2\otimes} \otimes A_0)$$
$$+ \tfrac{1}{2}\left((\hat{J}_k^{2\otimes} \otimes A_1)^H P_k (\hat{J}_k \otimes \check{J}_k \otimes I_n) + (\hat{J}_k \otimes \check{J}_k \otimes A_1)^H P_k (\hat{J}_k^{2\otimes} \otimes I_n)\right)$$
$$+ \tfrac{1}{2}\left((\hat{J}_k^{2\otimes} \otimes A_2)^H P_k (\check{J}_k \otimes \hat{J}_k \otimes I_n) + (\check{J}_k \otimes \hat{J}_k \otimes A_2)^H P_k (\hat{J}_k^{2\otimes} \otimes I_n)\right)$$
$$+ \tfrac{1}{2}\left((\hat{J}_k \otimes \check{J}_k \otimes I_n)^T P_k (\hat{J}_k^{2\otimes} \otimes A_1) + (\hat{J}_k^{2\otimes} \otimes I_n)^T P_k (\hat{J}_k \otimes \check{J}_k \otimes A_1)\right)$$
$$+ \tfrac{1}{2}\left((\check{J}_k \otimes \hat{J}_k \otimes I_n)^T P_k (\hat{J}_k^{2\otimes} \otimes A_2) + (\hat{J}_k^{2\otimes} \otimes I_n)^T P_k (\check{J}_k \otimes \hat{J}_k \otimes A_2)\right)$$
$$+ (\hat{J}_k^{2\otimes} \otimes I_n)^T Q_{k,1} (\hat{J}_k^{2\otimes} \otimes I_n) - (\hat{J}_k \otimes \check{J}_k \otimes I_n)^T Q_{k,1} (\hat{J}_k \otimes \check{J}_k \otimes I_n)$$
$$+ (\hat{J}_k \otimes I_{(k+1)n})^T Q_{k,2} (\hat{J}_k \otimes I_{(k+1)n}) - (\check{J}_k \otimes I_{(k+1)n})^T Q_{k,2} (\check{J}_k \otimes I_{(k+1)n}) < 0_{(k+1)^2 n}.$$

Forms similar to (12) and (13) may be obtained. For $k = 1$, $m = 1$, one gets for the LMI defined in Theorem 4.3

$$P_1 = P_1^H > 0,\ Q_{1,1} = Q_{1,1}^H,\ \begin{pmatrix} A_0^H P_1 + P_1 A_0 + Q_{1,1} & \frac{1}{2}(A_1^H P_1 + P_1 A_1) \\ \frac{1}{2}(A_1^H P_1 + P_1 A_1) & -Q_{1,1} \end{pmatrix} < 0,$$

to be compared to the condition obtained by *DG*-scaling:

$$P_1 = P_1^H > 0,\ D = D^H,\ G + G^H = 0,\ \begin{pmatrix} A_0^H P_1 + P_1 A_0 + D & P_1 A_1 + G \\ A_1^H P_1 + G^H & -D \end{pmatrix} < 0.$$

One may verify that there is no loss of generality to take $G = \frac{1}{2}(A_1^H P_1 - P_1 A_1)$ in the latter inequality, so the two criteria are equivalent. The formulas for larger m are obtained similarly to the complex case; they provide also tighter sufficient conditions than the ones based on *DG*-scaling.

Theorems 4.1 and 4.3 are easily adapted to treat the mixed complex/real case. The result is not stated completely here for the sake of space. As an example, for stability analysis of $A_0 + zA_1 + rA_2$, for a complex parameter z and a real parameter

r, both of norm less than or equal to 1, the criterion is based on the following family of LMIs: $\exists P_k \in \mathcal{H}^{k^2 n}$, $P_k > 0_{k^2 n}$, $\exists Q_{k,1} \in \mathcal{H}^{k^2 n}$, $\exists Q_{k,2} \in \mathcal{H}^{k(k+1)n}$,

$$
\begin{aligned}
&(\hat{J}_k^{2\otimes} \otimes A_0)^H P_k (\hat{J}_k^{2\otimes} \otimes I_n) + (\hat{J}_k \otimes \check{J}_k \otimes A_1)^H P_k (\hat{J}_k^{2\otimes} \otimes I_n) \\
&\quad + \tfrac{1}{2}\left((\hat{J}_k^{2\otimes} \otimes A_2)^H P_k (\check{J}_k \otimes \hat{J}_k \otimes I_n) + (\check{J}_k \otimes \hat{J}_k \otimes A_2)^H P_k (\hat{J}_k^{2\otimes} \otimes I_n) \right) \\
&\quad + (\hat{J}_k^{2\otimes} \otimes I_n)^T P_k (\hat{J}_k^{2\otimes} \otimes A_0) + (\hat{J}_k^{2\otimes} \otimes I_n)^T P_k (\hat{J}_k \otimes \check{J}_k \otimes A_1) \\
&\quad + \tfrac{1}{2}\left((\check{J}_k \otimes \hat{J}_k \otimes I_n)^T P_k (\hat{J}_k^{2\otimes} \otimes A_2) + (\hat{J}_k^{2\otimes} \otimes I_n)^T P_k (\check{J}_k \otimes \hat{J}_k \otimes A_2) \right) \\
&\quad + (\hat{J}_k^{2\otimes} \otimes I_n)^T Q_{k,1} (\hat{J}_k^{2\otimes} \otimes I_n) - (\hat{J}_k \otimes \check{J}_k \otimes I_n)^T Q_{k,1} (\hat{J}_k \otimes \check{J}_k \otimes I_n) \\
&+ (\hat{J}_k \otimes I_{(k+1)n})^T Q_{k,2} (\hat{J}_k \otimes I_{(k+1)n}) - (\check{J}_k \otimes I_{(k+1)n})^T Q_{k,2} (\check{J}_k \otimes I_{(k+1)n}) < 0_{(k+1)^2 n}.
\end{aligned}
$$

The robust stability of a polytope of matrices is now addressed by reduction to the problems solved in Theorem 4.3. Consider, for $m+1$ fixed matrices $B_1, \ldots, B_{m+1} \in \mathbb{C}^{n\times n}$, the class of systems

$$
\dot{x} = B(\beta)x, \quad \beta \stackrel{\text{def}}{=} (\beta_0, \beta_1, \ldots, \beta_m), \quad B(\beta) \stackrel{\text{def}}{=} \beta_0 B_0 + \beta_1 B_1 + \cdots + \beta_m B_m. \tag{14}
$$

Define the polytope $\mathbb{P}^{m+1} \stackrel{\text{def}}{=} \{\beta \in \mathbb{R}^{m+1} \ : \ \beta_i \geq 0, \ \beta_0 + \cdots + \beta_m = 1\}$.

COROLLARY 4.4 (robust stability of real convex polytopic systems). *The following three properties are* equivalent.

(i) *The matrix $B(\beta)$ in* (14) *is Hurwitz for any $\beta \in \mathbb{P}^{m+1}$.*

(ii) *There exists $m+1$ PPDQ functions $x^H P_i(\beta) x$, $i = 0, \ldots, m$ such that*

$$
\forall \beta \in \mathbb{P}^{m+1}, \ \begin{array}{l} P_i(\beta) > 0, i = 0, \ldots, m, \\ B(\beta)^H P_{\arg\max \beta_i}(\beta) + P_{\arg\max \beta_i}(\beta) B(\beta) < 0. \end{array}
$$

(iii) *For each value of $i = 0, \ldots, m$, there exists a positive integer k for which* (LMI$_k$) *with R_k defined in* (10) *is solvable, with*

$$
A_0 \stackrel{\text{def}}{=} B_i + \frac{1}{2} \sum_{j=0, j\neq i}^{m} B_j, \quad \{A_1, \ldots, A_m\} = \left\{ \frac{1}{2} B_j \ : \ j \neq i \right\}. \tag{15}
$$

Proof of Corollary 4.4. Let, for example, $\max_{0\leq i\leq m} \beta_i = \beta_0 > 0$, and write

$$
B(\beta) = \beta_0 \left(B_0 + \frac{1}{2} \sum_{i=1}^{m} B_i + \frac{1}{2} \sum_{i=1}^{m} \left(2\frac{\beta_i}{\beta_0} - 1 \right) B_i \right).
$$

Remark that the map $[0; +1] \to [-1; +1]$, $u \mapsto 2u - 1$ is one-to-one. For any fixed value of $i = \arg\max \beta_j$ (take any value if the maximum is attained for more than one index), property (i) is thus equivalent to robust stability of $\dot{x} = A(r)x$ with the definition of $A_0, A_1, \ldots, A_m$ given in (15), and it is possible to apply Theorem 4.3. For any fixed $i = \arg\max \beta_j$, a PPDQ Lyapunov function is found as a function of $r_j \stackrel{\text{def}}{=} -1 + 2\beta_j / \max \beta_i$, $j \neq i$, and may be expressed with respect to β_j after adequate change of the coefficients. ∎

Application of Theorem 4.3 thus provides for this problem too a family of sufficient conditions for robust stability, whose conservatism vanishes asymptotically: $m+1$ families of LMIs are found such that robust stability of (14) is equivalent to the solvability of at least one LMI in each family. Clearly, this approach amounts to the search for a piecewise PPDQ Lyapunov function, chosen, in $m+1$ quadrants of the parameter space, according to the value of $\max \beta_i$.

5. Numerical example. Consider the following example. Let $n = 3$,

$$A_0 = \begin{pmatrix} -12 & -7 & 7 \\ -11 & -13 & -5 \\ -2 & 9 & -8 \end{pmatrix},\ A_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 2 \\ 0 & 3 & 0 \end{pmatrix},\ A_2 = \begin{pmatrix} 1 & 2 & 0 \\ -3 & -1 & 0 \\ -1 & 0 & 0 \end{pmatrix}.$$

We evaluate the following robustness margins:

$$\begin{aligned}
\alpha_{zz} &\stackrel{\text{def}}{=} \sup_{(z_1,z_2)\in\overline{\mathbb{D}}^2} \operatorname{Re}\sigma(A_0 + z_1A_1 + z_2A_2),\\
\alpha_{zr} &\stackrel{\text{def}}{=} \sup_{(z,r)\in\overline{\mathbb{D}}\times[-1;+1]} \operatorname{Re}\sigma(A_0 + zA_1 + rA_2),\\
\alpha_{rz} &\stackrel{\text{def}}{=} \sup_{(r,z)\in[-1;+1]\times\overline{\mathbb{D}}} \operatorname{Re}\sigma(A_0 + rA_1 + zA_2),\\
\alpha_{z} &\stackrel{\text{def}}{=} \sup_{z\in\overline{\mathbb{D}}} \operatorname{Re}\sigma(A_0 + z(A_1 + A_2)),\\
\alpha_{rr} &\stackrel{\text{def}}{=} \sup_{(r_1,r_2)\in[-1;+1]^2} \operatorname{Re}\sigma(A_0 + r_1A_1 + r_2A_2),\\
\alpha_{r} &\stackrel{\text{def}}{=} \sup_{r\in[-1;+1]} \operatorname{Re}\sigma(A_0 + r(A_1 + A_2)).
\end{aligned}$$

Clearly, the latter quantities are linked by the inequalities:

$$\alpha_{zz} \geq \alpha_{zr}, \quad \alpha_{rz} \geq \alpha_{rr} \geq \alpha_r, \quad \text{and } \alpha_{zz} \geq \alpha_z \geq \alpha_r. \tag{16}$$

We use the previously presented LMIs to find, for each uncertainty structure, the least real number α such that $A(z) - \alpha I_n$ is robustly stable. For each integer k and for each value of α, a convex problem is solved, but the problem is not jointly convex in the four unknowns $P_k, Q_{k,1}, Q_{k,2}$, and α, so a bisection process is achieved.

The computations presented here have been performed using the package `lmitool` of the free software Scilab. The successive (upper) estimates of the robustness margins, according to the value of k, are given in Table 1. Between parentheses is given the CPU time necessary for the solution of the LMIs (for the corresponding values of α and k), measured on a computer equipped with a Pentium III 800MHz.

The values are compared to those obtained by checking directly the robust stability by means of gridding of the parameter space, which are presented in Table 2. Due to the small size of the problem, small computation times are required.

One verifies that, for each margin, the successive estimates are nonincreasing functions of k and that the inequalities corresponding to (16) are fulfilled for any value of k. In the present case, the tests achieved for $k = 2$ provide these true values up to three digits, except for α_r ($k = 3$).

In principle, the previous numbers may also be determined using the fact that

$$\alpha_{zz} = \inf\left\{\alpha \in \mathbb{R} \ :\ \forall \omega \in \mathbb{R}, \forall \alpha' \in (\alpha; +\infty),\ \mu_\Delta(G(j\omega + \alpha')) < 1\right\},$$

TABLE 1
Successive estimates of the margins and corresponding CPU times.

	α_{zz}	α_{zr}	α_{rz}
$k=1$	-2.42 (0.2s)	-3.17 (0.2s)	-2.42 (0.19s)
$k=2$	-3.87 (7.63s)	-4.57 (9.73s)	-4.46 (7.84s)
	α_z	α_{rr}	α_r
$k=1$	-3.24 (0.06s)	-3.17 (0.2s)	-3.24 (0.07s)
$k=2$	-4.14 (0.23s)	-5.24 (10.1s)	-5.39 (0.26s)
$k=3$			-5.41 (0.65s)

TABLE 2
Successive estimates of the margins by gridding and corresponding CPU times.

Number of nodes in parameter space	α_{zz}	α_{zr}	α_{rz}
10×10	-3.93 (0.02s)	-4.63 (0.02s)	-4.48 (0.02s)
100×100	-3.88 (1.67s)	-4.57 (1.68s)	-4.47 (1.66s)
	α_z	α_{rr}	α_r
10×10	-4.15 (0.01s)	-5.24 (0.01s)	-5.42 (0.01s)
100×100	-4.15 (0.02s)	-5.24 (0.84s)	-5.42 (0.01s)

where $G(s) \stackrel{\text{def}}{=} \binom{I_n}{I_n} \left(sI_n - A_0\right)^{-1} \begin{pmatrix} A_1 & A_2 \end{pmatrix}$, and for the uncertainty structure $\Delta = \{\text{diag}\{z_1 I_n; z_2 I_n\} : z_i \in \mathbb{C}\}$. Similar formulas hold for the other margins. Define the constants

$$\overline{\alpha} \stackrel{\text{def}}{=} \inf\left\{\alpha \in \mathbb{R} \ : \ \forall \omega \in \mathbb{R}, \forall \alpha' \in (\alpha; +\infty), \ \overline{\sigma}(G(j\omega + \alpha')) < 1\right\},$$

$$\overline{\alpha}_{zz} \stackrel{\text{def}}{=} \inf\left\{\alpha \in \mathbb{R} \ : \ \forall \omega \in \mathbb{R}, \forall \alpha' \in (\alpha; +\infty), \ \nu_\Delta(G(j\omega + \alpha')) < 1\right\},$$

where Δ is the same set as above, and $\overline{\sigma}$ and ν_Δ denote, respectively, the largest singular value and the usual upper bound of μ_Δ [16]. Based on the properties of ν_Δ [16], one has $\alpha_{zz} \leq \overline{\alpha}_{zz} \leq \overline{\alpha}$. The results of the estimation of the previous constants, based on the underlying LMI problems, are summarized in Table 3. As before, the CPU time necessary to check that, for a fixed value of α, $\overline{\sigma}(G(j\omega+\alpha)) < 1$ (respectively, $\nu_\Delta(G(j\omega+\alpha)) < 1$) for a discretized sample of frequencies ω on the real axis is indicated between parentheses. Tighter discretization, not reproduced here, shows that the values obtained for 1000 gridding points are the true values of the extrema $\overline{\alpha}$ and $\overline{\alpha}_{zz}$.

Recall that the estimate α_{zz} obtained by use of Theorem 4.1 for $k = 2$ is exact (up to the precision considered), while $\overline{\alpha}$ and $\overline{\alpha}_{zz}$ provide conservative robust stability margins. For this simple example, the gain in precision is clear for comparable computation time.

6. Comments on the results. The results stated in section 4 permit a systematic approach to the study of parameter-dependent quadratic Lyapunov functions for robust stability: a class of candidate Lyapunov functions is exhibited (given in Definition 3.1), rich enough to *characterize* robust stability but structured enough to permit the use of LMI tests. In our opinion, this offers a useful insight into the powerfulness of quadratic Lyapunov functions for stability analysis. Similarly, it provides information on the kind of problems solvable by LMIs: the issue of robust stability analysis is located "on the boundary" of these problems, as it may be relaxed with

TABLE 3
Successive estimates of the margins by upper bounds and corresponding CPU times.

Number of nodes in frequency domain	$\overline{\alpha}$	$\overline{\alpha}_{zz}$
100	-0.241 (0.43s)	-1.30 (0.84s)
1000	-0.151 (5.22s)	-1.21 (7.96s)

arbitrary precision into a standard LMI, obtained explicitly. In this sense, the results given here constitute an attempt to investigate in more detail the abilities of the LMIs, which have become, in the last decade, a unifying framework for expressing and solving many problems in control theory.

We believe that, beyond their theoretical interest, the proposed results may offer attractive numerical alternatives for robust stability analysis, at least for problems of low order. The construction of the LMIs involved is reasonably simple, using only elementary algebraic operations. For a given value of k, their complexity is polynomial with respect to the dimension n of the matrices and exponential with respect to the number m of scalar parameters. More precisely, the total number of scalar elements of the unknowns $P_k, Q_{k,1}, \ldots, Q_{k,m}$ in (LMI_k) is $\frac{1}{2}[k^m n(k^m+1) + \sum_{1\leq i\leq m} k^{m-i+1}(k+1)^{i-1}(k^{m-i+1}(k+1)^{i-1}+1)]$, which is equivalent to $\frac{m+1}{2}k^{2m}n^2$ when $k \to +\infty$, while the number of rows of the inequalities involved is $[k^m + (k+1)^m]n$, which is of the order of $2k^m n$ when $k \to +\infty$.

A quantitative evaluation of the relationship between the size of k and the precision of the criteria should be considered as a natural next step in forthcoming research. In the general case, however, when no special matrix structure exists for the system under study, the growth of the value of k needed to check robust stability of a system cannot be polynomial in the worst case. The effective use of large values of k hinges upon the possibility of intensive computation and use of large memory.

The method for robust stability analysis proposed here may be compared to the one consisting of checking stability in every node of a grid of the parameter space. Both methods are able to provide less and less conservative criteria—when the discretization step goes to zero or when the degree, with respect to the parameters, of the underlying parameter-dependent Lyapunov function goes to infinity. Both methods are exact, in the sense that they provide asymptotically arbitrarily precise estimates of the true stability margins. The gridding method, however, offers successive (less and less) optimistic estimates, whereas the other one provides (less and less) pessimistic indications, usually more useful in practice.

Also, both methods are, in the present state of knowledge, computationally *undecidable*: no information is known on the size of the least k, if any, for which the LMIs are solvable (in other words, of the largest k which is necessary to test numerically to decide whether the system is robustly stable or not). This is an important question, both from theoretical and practical points of view. Some numerical experiments (see section 5) indicate that small values of k often yield correct answers.

Finally, recall that ensuring robust stability analysis is equivalent to checking that a certain structured singular value is less than 1. In consequence, the results presented above may also be seen as providing a family of more and more precise *upper bounds* for these special-structured singular values. In a future work, the extension of this to the general case of structured singular values with repeated scalar blocks will be investigated. Indeed, it may be reasonable to employ these new upper bounds in a

branch and bound algorithm [3] in place of the usual ones.

Alternatively, in the case of real parameters, a possible way to consider problems of larger size rests in the combination of decomposition of the parameter domain and resolution on each subdomain by use of a low-order test. This should permit us to find a compromise between the number of independent LMIs to be solved (equal to the number of subdivisions) and the computational complexity (due to the high degree of the underlying parameter-dependent Lyapunov functions). When coupled with decomposition in the parameter space, one may expect a better fit of the proposed method than the direct use of Theorems 4.1 and 4.3, at least for problems of medium size.

7. Proofs. The following decomposition of the matrix $\hat{J}_k$ defined in section 2 will be used:

$$\hat{J}_k = \begin{pmatrix} I_k & 0_{k\times 1} \end{pmatrix} = \begin{pmatrix} f_k & F_k \end{pmatrix},$$

where

$$f_k \stackrel{\text{def}}{=} \begin{pmatrix} 1 \\ 0_{(k-1)\times 1} \end{pmatrix}, \quad F_k \stackrel{\text{def}}{=} \begin{pmatrix} 0_{1\times(k-1)} & 0 \\ I_{k-1} & 0_{(k-1)\times 1} \end{pmatrix}. \tag{17}$$

The size of the previous matrices is $f_k : k \times 1$, $F_k : k \times k$, and the spectrum of F_k is $\{0\}$. Simple computation shows that, for any $z \in \mathbb{C}$, $(I_k - zF_k)z^{[k]} = f_k$, that is,

$$(I_k - zF_k)^{-1} f_k = z^{[k]}. \tag{18}$$

Another useful property is the fact that, for any $i \in \mathbb{N}$, $0 \le i \le k-1$,

$$F_k^{iT} z^{[k]} = \begin{pmatrix} 0_{(k-i)\times i} & I_{k-i} \\ 0_i & 0_{i\times(k-i)} \end{pmatrix} \begin{pmatrix} z^{[i]} \\ z^i z^{[k-i]} \end{pmatrix} = z^i \begin{pmatrix} z^{[k-i]} \\ 0_{i\times 1} \end{pmatrix}. \tag{19}$$

7.1. Proof of Theorem 4.1. We now prove Theorem 4.1. The proof of the equivalence between the properties (i), (ii), and (iii) consists of three main stages that we now present and which are detailed below in sections 7.1.1–7.1.3.

1*st stage.* We detail here the ideas given in the sketch of the proof of formula (11). We show that the computations proposed there permit us to establish that solvability of (LMI_k) implies solvability of (2) for all $z \in (\partial\mathbb{D})^m$. It remains to show that this implies, however, Hurwitzness of $A(z)$ for any z *in the whole set* $\overline{\mathbb{D}}^m$. This gives the implication (iii) $\Rightarrow$ (i).

2*nd stage.* The second step establishes that the robust stability property (i) implies that the parameter-dependent Lyapunov inequality (2) admits a solution $P(z)$ of the form (7), for a certain $k \in \mathbb{N}$, with P_k *positive definite*. For this, one shows essentially that the associated Lyapunov *equation* $A(z)^H P(z) + P(z)A(z) = -I_n$ admits as a solution an infinite sum of powers of $z, \bar{z}$, converging uniformly in $\overline{\mathbb{D}}^m$. It then suffices to truncate this expansion to obtain a polynomial solution (of unknown degree) to *inequality* (2). The corresponding coefficient matrix P_k is positive *semidefinite* by construction, and some more work is necessary to obtain an expression with a *positive* definite matrix. This gap is filled in Lemma 7.1. As a by-product, the implication (i) $\Rightarrow$ (ii) is obtained here.

3*rd stage.* At this point, (i) has been shown to imply existence, for large enough k, of a certain $P_k > 0$ such that $R(z)$ given by (8), (9) is negative definite for any $z \in \overline{\mathbb{D}}^m$. The next step (Lemma 7.2) is the key part of the necessity proof. It consists

of showing that $R(z) < 0$ for all $z \in (\partial\mathbb{D})^m$ *if and only if* the second inequality in (LMI_k) holds. This is done by applying recursively D-scaling with respect to each of the parameters z_i. At each step, a new matrix, depending upon the remaining parameters $z_{i+1}, \ldots, z_m$, is introduced. The latter may be assumed polynomial in the previous parameters and their conjugates (this is deduced from a general result on existence of polynomial solutions to parameter-dependent LMIs, Theorem 7.3), with coefficients defined by a constant matrix, which is precisely the variable $Q_{k,i}$ of (LMI_k). The transformation carried out by this procedure is not restrictive, as the scaling technique (the Kalman–Yakubovich–Popov lemma, recalled in Appendix A) is lossless for one complex parameter. This yields the implication (ii) $\Rightarrow$ (iii). Incidentally, we prove at this stage that the solvability of (LMI_k) implies the same property for largest indices.

When the coefficients are real, then the polynomial solutions exhibited above are easily proved to be real too, basically due to the remark on realness given after the version of the Kalman–Yakubovich–Popov lemma recalled in Appendix A.

7.1.1. First stage. Suppose (LMI_k) holds. As suggested in the sketch of the proof of formula (11), left- and right-multiplication of its second inequality by $(z_m^{[k+1]} \otimes \cdots \otimes z_1^{[k+1]} \otimes I_n)$ and its transconjugate yields $R(z) + \sum_{i=1}^m (1-|z_i|^2)(z_m^{[k]} \otimes \cdots \otimes z_i^{[k]} \otimes z_{i-1}^{[k+1]} \otimes \cdots \otimes z_1^{[k+1]} \otimes I_n)^H Q_{k,i} (z_m^{[k]} \otimes \cdots \otimes z_i^{[k]} \otimes z_{i-1}^{[k+1]} \otimes \cdots \otimes z_1^{[k+1]} \otimes I_n) < 0_n$. Indeed, this comes directly from the fact that, due to (4), for any $i = 1, \ldots, m$,

$$\left(\hat{J}_k^{(m-i+1)\otimes} \otimes I_{(k+1)^{i-1}n}\right)(z_m^{[k+1]} \otimes \cdots \otimes z_1^{[k+1]} \otimes I_n) = (z_m^{[k]} \otimes \cdots \otimes z_i^{[k]} \otimes z_{i-1}^{[k+1]} \otimes \cdots \otimes z_1^{[k+1]} \otimes I_n),$$

$$\left(\hat{J}_k^{(m-i)\otimes} \otimes \check{J}_k \otimes I_{(k+1)^{i-1}n}\right)(z_m^{[k+1]} \otimes \cdots \otimes z_1^{[k+1]} \otimes I_n) = z_i (z_m^{[k]} \otimes \cdots \otimes z_i^{[k]} \otimes z_{i-1}^{[k+1]} \otimes \cdots \otimes z_1^{[k+1]} \otimes I_n).$$

Thus, $R(z) < 0_n$ if $|z_1| = \cdots = |z_m| = 1$, so the matrix $A(z)$ is Hurwitz for all $z \in (\partial\mathbb{D})^m$.

The remaining argument is based on a subharmonicity and continuity argument. Using the fact that the map $\overline{\mathbb{C}^+} \cup \{\infty\} \to \overline{\mathbb{D}}, s_z \mapsto (1-s_z)/(1+s_z)$ is one-to-one, one proves [7] that

$$\begin{aligned}\max_{z\in\overline{\mathbb{D}}^m} \rho(e^{A(z)}) &= \sup_{s\in\overline{\mathbb{C}^+}^m} \rho(e^{A_0+(1-s_1)/(1+s_1)A_1+\cdots+(1-s_m)/(1+s_m)A_m}) \\ &= \sup_{s\in(j\mathbb{R})^m} \rho(e^{A_0+(1-s_1)/(1+s_1)A_1+\cdots+(1-s_m)/(1+s_m)A_m}) \\ &= \max_{z\in(\partial\mathbb{D})^m} \rho(e^{A(z)}).\end{aligned}$$

As a consequence, if all the matrices $A(z)$ are Hurwitz for $z \in (\partial\mathbb{D})^m$, then the previous expression is less than 1, and the same property holds on the whole $\overline{\mathbb{D}}^m$. This shows that (iii) implies (i).

7.1.2. Second stage. Property (i) implies solvability of (2) for each $z \in \overline{\mathbb{D}}^m$, which, as is well known, is equivalent [29] to the solvability of the (Lyapunov) *equation*

$$P(z) > 0_n, \ A(z)^H P(z) + P(z)A(z) = -I_n. \tag{20}$$

Now, when (i) holds, the latter has a solution *analytic in* z, $\bar{z}$ *in* $\overline{\mathbb{D}}^m$. Indeed, when $A(z)$ is Hurwitz, the explicit form of the solution of (20) is given by

$$P(z) = \int_0^{+\infty} e^{A(z)^H t} e^{A(z)t}\, dt.$$

When $A(z)$ is Hurwitz for any z in the compact set $\overline{\mathbb{D}}^m$, the convergence of this integral in $t = +\infty$ is uniform with respect to z, so there exists $T > 0$ independent of z such that $P(z)$ defined now by

$$P(z) = \int_0^{T} e^{A(z)^H t} e^{A(z)t}\, dt \tag{21}$$

is positive definite and solves inequality (2) in $\overline{\mathbb{D}}^m$.

Expanding the integrand in powers of the $z_i, \bar{z}_i$, $1 \le i \le m$, and interverting the sum and the integral, one exhibits an expansion of $P(z)$ in powers of $z, \bar{z}$, converging uniformly for $(z,t) \in \overline{\mathbb{D}}^m \times [0;T]$. More precisely, let $M_k : [0;T] \to \mathbb{C}^{n \times k^m n}$ be such that $M_k(t)(z_m^{[k]} \otimes \cdots \otimes z_1^{[k]} \otimes I_n)$ represents the terms of degree less than k in each of the z_i in the expansion of $e^{A(z)t}$. Then

$$\int_0^{T} e^{A(z)^H t} e^{A(z)t}\, dt = \lim_{k \to +\infty} (z_m^{[k]} \otimes \cdots \otimes z_1^{[k]} \otimes I_n)^H \tilde{P}_k (z_m^{[k]} \otimes \cdots \otimes z_1^{[k]} \otimes I_n) \tag{22}$$

with uniform convergence in $\overline{\mathbb{D}}^m$, where $\tilde{P}_k \in \mathcal{H}^{k^m n}$ is defined by

$$\tilde{P}_k \stackrel{\text{def}}{=} \int_0^{T} M_k(t)^H M_k(t)\, dt \ge 0. \tag{23}$$

Now this implies that, for large enough k, $(z_m^{[k]} \otimes \cdots \otimes z_1^{[k]} \otimes I_n)^H \tilde{P}_k (z_m^{[k]} \otimes \cdots \otimes z_1^{[k]} \otimes I_n)$ solves inequality (2) in $\overline{\mathbb{D}}^m$. This provides a PPDQ Lyapunov function for (1), but not the desired one, as the matrices $\tilde{P}_k$ are only positive *semidefinite* (except $\tilde{P}_1 = P(0)$, which is positive definite).

Let instead, for the matrix F_k defined in (17),

$$P_k \stackrel{\text{def}}{=} \sum_{i_1,\ldots,i_m=0}^{k-1} (F_k^{i_m} \otimes \cdots \otimes F_k^{i_1} \otimes I_n) \tilde{P}_k (F_k^{i_m} \otimes \cdots \otimes F_k^{i_1} \otimes I_n)^T. \tag{24}$$

LEMMA 7.1. *The matrix* $P_k \in \mathcal{H}^{k^m n}$ *defined in* (24) *is positive definite and, for large enough* $k \in \mathbb{N}$, $(z_m^{[k]} \otimes \cdots \otimes z_1^{[k]} \otimes I_n)^H P_k (z_m^{[k]} \otimes \cdots \otimes z_1^{[k]} \otimes I_n)$ *solves* (2) *in* $\overline{\mathbb{D}}^m$.

Proof. We begin with the positivity property. Note first that $P_k \ge 0$, because $\tilde{P}_k \ge 0$. Let $u \in \mathbb{C}^{k^m n}$ such that $u^H P_k u = 0$, and let us establish that this implies $u = 0$. In view of (24), and thanks to the fact that $\tilde{P}_k \ge 0$, this implies that for all $0 \le i_1, \ldots, i_m \le k-1$,

$$u^H (F_k^{i_m} \otimes \cdots \otimes F_k^{i_1} \otimes I_n) \tilde{P}_k (F_k^{i_m} \otimes \cdots \otimes F_k^{i_1} \otimes I_n)^T u = 0. \tag{25}$$

First, notice that for any integer i, k, $i \le k$, all the terms of degree less than $k-i$ in $e^{A(z)t}$, whose total sum is $M_{k-i}(t)(z_m^{[k-i]} \otimes \cdots \otimes z_1^{[k-i]} \otimes I_n)$ by definition, are also

present in $M_k(t)(z_m^{[k]} \otimes \cdots \otimes z_1^{[k]} \otimes I_n)$. At the level of the matrices $\tilde{P}_k$, this property reads as

(26)
$$\tilde{P}_{k-i} = \left(\begin{pmatrix} I_{k-i} \\ 0_{i\times(k-i)} \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} I_{k-i} \\ 0_{i\times(k-i)} \end{pmatrix} \otimes I_n\right)^T \tilde{P}_k \left(\begin{pmatrix} I_{k-i} \\ 0_{i\times(k-i)} \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} I_{k-i} \\ 0_{i\times(k-i)} \end{pmatrix} \otimes I_n\right).$$

Indeed, for $z \in \mathbb{C}$, $\begin{pmatrix} I_{k-i} \\ 0_{i\times(k-i)} \end{pmatrix} z^{[k-i]}$ is equal to $z^{[k]}$, except for the terms of degree larger than $k-i-1$, which are replaced by zero. Remark also that, for $i' \leq i \leq k$,

$$F_k^{iT} = \begin{pmatrix} 0_{(k-i)\times i} & I_{k-i} \\ 0_i & 0_{i\times(k-i)} \end{pmatrix} = \begin{pmatrix} I_{k-i'} \\ 0_{i'\times(k-i')} \end{pmatrix} \begin{pmatrix} 0_{(k-i)\times i} & I_{k-i} \\ 0_{(i-i')\times i} & 0_{(i-i')\times(k-i)} \end{pmatrix}. \tag{27}$$

Putting now $i_1 = \cdots = i_m = k-1$ in (25) and using identity (27) with $i = i' = k-1$ and (26), one deduces first that

$$\left\| \tilde{P}_1^{1/2} \left(\begin{pmatrix} 0_{1\times(k-1)} & 1 \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 0_{1\times(k-1)} & 1 \end{pmatrix} \otimes I_n \right) u \right\|^2 = 0;$$

i.e., as $\tilde{P}_1 > 0$,

$$\left(\begin{pmatrix} 0_{1\times(k-1)} & 1 \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 0_{1\times(k-1)} & 1 \end{pmatrix} \otimes I_n \right) u = 0. \tag{28}$$

Taking then $i_1 = k-2, i_2 = \cdots = i_m = k-1$ in (25), using (27) with $i = i' = k-2$ and $i' = k-2$, $i = k-1$, and then using (26), yields

$$\left\| \tilde{P}_2^{1/2} \left(\begin{pmatrix} 0_{2\times(k-2)} & I_2 \end{pmatrix} \otimes \begin{pmatrix} 0_{1\times(k-1)} & 1 \\ 0_{1\times(k-1)} & 0 \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 0_{1\times(k-1)} & 1 \\ 0_{1\times(k-1)} & 0 \end{pmatrix} \otimes I_n \right) u \right\|^2 = 0.$$

Now

$$\begin{aligned}
&\tilde{P}_2^{1/2} \left(\begin{pmatrix} 0_{2\times(k-2)} & I_2 \end{pmatrix} \otimes \begin{pmatrix} 0_{1\times(k-1)} & 1 \\ 0_{1\times(k-1)} & 0 \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 0_{1\times(k-1)} & 1 \\ 0_{1\times(k-1)} & 0 \end{pmatrix} \otimes I_n \right) u \\
&= \tilde{P}_2^{1/2} \left(\begin{pmatrix} 0_{1\times(k-2)} & 1 & 0 \\ 0_{1\times(k-2)} & 0 & 0 \end{pmatrix} \otimes \begin{pmatrix} 0_{1\times(k-1)} & 1 \\ 0_{1\times(k-1)} & 0 \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 0_{1\times(k-1)} & 1 \\ 0_{1\times(k-1)} & 0 \end{pmatrix} \otimes I_n \right) u \\
&\quad + \tilde{P}_2^{1/2} \left(\begin{pmatrix} 0_{1\times(k-1)} & 0 \\ 0_{1\times(k-1)} & 1 \end{pmatrix} \otimes \begin{pmatrix} 0_{1\times(k-1)} & 1 \\ 0_{1\times(k-1)} & 0 \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 0_{1\times(k-1)} & 1 \\ 0_{1\times(k-1)} & 0 \end{pmatrix} \otimes I_n \right) u \\
&\qquad \text{(by linearity)} \\
&= \tilde{P}_2^{1/2} \left(\begin{pmatrix} 0_{1\times(k-2)} & 1 & 0 \\ 0_{1\times(k-2)} & 0 & 0 \end{pmatrix} \otimes \begin{pmatrix} 0_{1\times(k-1)} & 1 \\ 0_{1\times(k-1)} & 0 \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 0_{1\times(k-1)} & 1 \\ 0_{1\times(k-1)} & 0 \end{pmatrix} \otimes I_n \right) u \\
&\qquad \text{(due to (28))},
\end{aligned}$$

and, thanks to (26),

$$\begin{aligned}
&\left\| \tilde{P}_2^{1/2} \left(\begin{pmatrix} 0_{1\times(k-2)} & 1 & 0 \\ 0_{1\times(k-2)} & 0 & 0 \end{pmatrix} \otimes \begin{pmatrix} 0_{1\times(k-1)} & 1 \\ 0_{1\times(k-1)} & 0 \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 0_{1\times(k-1)} & 1 \\ 0_{1\times(k-1)} & 0 \end{pmatrix} \otimes I_n \right) u \right\| \\
&\quad = \left\| \tilde{P}_1^{1/2} \left(\begin{pmatrix} 0_{1\times(k-2)} & 1 & 0 \end{pmatrix} \otimes \begin{pmatrix} 0_{1\times(k-1)} & 1 \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 0_{1\times(k-1)} & 1 \end{pmatrix} \otimes I_n \right) u \right\|.
\end{aligned}$$

The last term is thus null, due as before to the definiteness of $\tilde{P}_1$, so one concludes that

$$\left(\begin{pmatrix}0_{1\times(k-2)} & 1 & 0\end{pmatrix}\otimes\begin{pmatrix}0_{1\times(k-1)} & 1\end{pmatrix}\otimes\cdots\otimes\begin{pmatrix}0_{1\times(k-1)} & 1\end{pmatrix}\otimes I_n\right)u=0.$$

Carrying on in this way, one shows that all the components of u, taken n by n, are null. Thus, $u^H P_k u = 0$ implies $u = 0$, so $P_k > 0$ for any $k \in \mathbb{N}$.

We now show that, for large enough values of k, P_k defined in (24) generates a PPDQ function fulfilling the requirement of (ii). For this, let us first establish that

$$\lim_{k\to+\infty}\frac{1}{\|z_1^{[k]}\|^2\dots\|z_m^{[k]}\|^2}(z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n)^H P_k(z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n) = \lim_{k\to+\infty}(z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n)^H\tilde{P}_k(z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n), \tag{29}$$

where both limits are uniform in $\overline{\mathbb{D}}^m$. The second limit is already known to exist and to be equal to $P(z)$ in (21).

From identity (24) one deduces, thanks to (19), that

$$\begin{aligned}&(z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n)^H P_k(z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n)\\ &\quad=\sum_{i_1,\dots,i_m=0}^{k-1}|z_1|^{2i_1}\dots|z_m|^{2i_m}\left(\begin{pmatrix}z_m^{[k-i_m]}\\0_{i_1\times1}\end{pmatrix}\otimes\cdots\otimes\begin{pmatrix}z_1^{[k-i_1]}\\0_{i_m\times1}\end{pmatrix}\otimes I_n\right)^H\\ &\qquad\times\tilde{P}_k\left(\begin{pmatrix}z_m^{[k-i_m]}\\0_{i_m\times1}\end{pmatrix}\otimes\cdots\otimes\begin{pmatrix}z_1^{[k-i_1]}\\0_{i_1\times1}\end{pmatrix}\otimes I_n\right).\end{aligned}$$

As

$$\sum_{i_1,\dots,i_m=0}^{k-1}|z_1|^{2i_1}\dots|z_m|^{2i_m}=\|z_1^{[k]}\|^2\dots\|z_m^{[k]}\|^2,$$

we get

$$\begin{aligned}&\|z_1^{[k]}\|^2\dots\|z_m^{[k]}\|^2P(z)-(z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n)^H P_k(z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n)\\ &=\sum_{i_1,\dots,i_m=0}^{k-1}|z_1|^{2i_1}\dots|z_m|^{2i_m}\left[P(z)\right.\\ &\quad\left.-\left(\begin{pmatrix}z_m^{[k-i_m]}\\0_{i_m\times1}\end{pmatrix}\otimes\cdots\otimes\begin{pmatrix}z_1^{[k-i_1]}\\0_{i_1\times1}\end{pmatrix}\otimes I_n\right)^H\tilde{P}_k\left(\begin{pmatrix}z_m^{[k-i_m]}\\0_{i_m\times1}\end{pmatrix}\otimes\cdots\otimes\begin{pmatrix}z_1^{[k-i_1]}\\0_{i_1\times1}\end{pmatrix}\otimes I_n\right)\right].\end{aligned} \tag{30}$$

Now, uniform convergence of the right-hand side of (29) yields the following: for any $\varepsilon > 0$, there exists k_ε such that, for any $k > k_\varepsilon$, for any $z \in \overline{\mathbb{D}}^m$,

$$\|P(z)-(z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n)^H\tilde{P}_k(z_m^{[k]}\otimes\cdots\otimes z_1^{[k]}\otimes I_n)\|<\varepsilon.$$

Distinguishing between the terms for which $\max\{i_1,\dots,i_m\} < k-k_\varepsilon$ and $\max\{i_1,\dots,i_m\} \geq k-k_\varepsilon$ allows us to show that the norm of the left-hand side of (30) is bounded

from above by

$$\varepsilon\ \|z_1^{[k]}\|^2 \dots \|z_m^{[k]}\|^2 + 2c \sum_{\substack{i_1,\dots,i_m=0 \\ \max\{i_1,\dots,i_m\} \geq k-k_\varepsilon}}^{k-1} |z_1|^{2i_1} \dots |z_m|^{2i_m}.$$

In the previous expression, c is defined as

$$c \stackrel{\text{def}}{=} \max \left\{ \left\| \left(\begin{pmatrix} z_m^{[k-i_m]} \\ 0_{i_m \times 1} \end{pmatrix} \otimes \dots \otimes \begin{pmatrix} z_1^{[k-i_1]} \\ 0_{i_1 \times 1} \end{pmatrix} \otimes I_n \right)^H \right.$$
$$\left. \times \tilde{P}_k \left(\begin{pmatrix} z_m^{[k-i_m]} \\ 0_{i_m \times 1} \end{pmatrix} \otimes \dots \otimes \begin{pmatrix} z_1^{[k-i_1]} \\ 0_{i_1 \times 1} \end{pmatrix} \otimes I_n \right) \right\| :$$
$$\left. z \in \overline{\mathbb{D}}^m, i_1, \dots, i_m \leq k, k \in \mathbb{N} \right\}.$$

The constant c is finite, because, when $i_j \to +\infty$, $1 \leq j \leq m$, the expression inside the norm converges uniformly in $\overline{\mathbb{D}}^m$ towards $P(z)$, which, being continuous, is itself bounded. On the other hand, for any $z \in \mathbb{C}^m$,

$$\sum_{\substack{i_1,\dots,i_m=0 \\ \max\{i_1,\dots,i_m\} \geq k-k_\varepsilon}}^{k-1} |z_1|^{2i_1} \dots |z_m|^{2i_m}$$
$$= \sum_{i_1,\dots,i_m=0}^{k-1} |z_1|^{2i_1} \dots |z_m|^{2i_m} - \sum_{i_1,\dots,i_m=0}^{k-k_\varepsilon-1} |z_1|^{2i_1} \dots |z_m|^{2i_m}$$
$$= \|z_1^{[k]}\|^2 \dots \|z_m^{[k]}\|^2 - \|z_1^{[k-k_\varepsilon]}\|^2 \dots \|z_m^{[k-k_\varepsilon]}\|^2$$
$$= \left(\|z_1^{[k]}\|^2 - \|z_1^{[k-k_\varepsilon]}\|^2 \right) \|z_2^{[k]}\|^2 \dots \|z_m^{[k]}\|^2$$
$$+ \dots + \|z_1^{[k-k_\varepsilon]}\|^2 \dots \|z_{m-1}^{[k-k_\varepsilon]}\|^2 \left(\|z_m^{[k]}\|^2 - \|z_m^{[k-k_\varepsilon]}\|^2 \right)$$
$$\leq m \|z_1^{[k]}\|^2 \dots \|z_m^{[k]}\|^2 \max_{i=1,\dots,m} \frac{\|z_i^{[k]}\|^2 - \|z_i^{[k-k_\varepsilon]}\|^2}{\|z_i^{[k]}\|^2}$$
$$= m \|z_1^{[k]}\|^2 \dots \|z_m^{[k]}\|^2 \max_{i=1,\dots,m} |z_i|^{2(k-k_\varepsilon)} \frac{1 + |z_i|^2 + \dots + |z_i|^{2(k_\varepsilon-1)}}{1 + |z_i|^2 + \dots + |z_i|^{2(k-1)}}.$$

It turns out that, uniformly in $\overline{\mathbb{D}}^m$, the following estimate holds:

$$\left\| P(z) - \frac{1}{\|z_1^{[k]}\|^2 \dots \|z_m^{[k]}\|^2} (z_m^{[k]} \otimes \dots \otimes z_1^{[k]} \otimes I_n)^H P_k (z_m^{[k]} \otimes \dots \otimes z_1^{[k]} \otimes I_n) \right\|$$
$$\leq \varepsilon + 2mc \sup_{r \in [0;1)} r^{k-k_\varepsilon} \frac{1 - r^{k_\varepsilon}}{1 - r^k},$$

provided that $k > k_\varepsilon$. Notice that, for any fixed k_ε, the quantity

$$\sup_{r \in [0;1)} r^{k-k_\varepsilon} \frac{1 - r^{k_\varepsilon}}{1 - r^k}$$

vanishes when k goes to infinity. Thus, for large enough k, it is smaller than $\varepsilon/2mc$ and then, for any $z \in \overline{\mathbb{D}}^m$,

$$\left\| P(z) - \frac{1}{\|z_1^{[k]}\|^2 \dots \|z_m^{[k]}\|^2} (z_m^{[k]} \otimes \cdots \otimes z_1^{[k]} \otimes I_n)^H P_k (z_m^{[k]} \otimes \cdots \otimes z_1^{[k]} \otimes I_n) \right\| < 2\varepsilon.$$

This achieves the proof of the announced convergence property (29).

As a consequence of (29), the truncated expression $\frac{1}{\|z_1^{[k]}\|^2 \dots \|z_m^{[k]}\|^2} (z_m^{[k]} \otimes \cdots \otimes z_1^{[k]} \otimes I_n)^H P_k (z_m^{[k]} \otimes \cdots \otimes z_1^{[k]} \otimes I_n)$ solves (2) for large enough k, and $x^H (z_m^{[k]} \otimes \cdots \otimes z_1^{[k]} \otimes I_n)^H P_k (z_m^{[k]} \otimes \cdots \otimes z_1^{[k]} \otimes I_n) x$ with $P_k > 0$ is also a PPDQ Lyapunov function for (1). This achieves the proof of Lemma 7.1. ∎

As a conclusion of this second stage of the proof of Theorem 4.1, we have shown until now that property (i) is equivalent to (ii), in which, moreover, the hermitian P_k defining $P(z)$ may be supposed *positive definite* without loss of generality.

7.1.3. Third stage. This part is achieved by induction. Consider for any $i = 0, \dots, m$ the following property.

Property $(\mathcal{P}_i)$: $\exists k \in \mathbb{N}$, $\exists P_k \in \mathcal{H}^{k^m n}$, $P_k > 0$, $\exists Q_{k,j} \in \mathcal{H}^{k^{m-j+1}(k+1)^{j-1} n}$, $j = 1, \dots, i$, for all $(z_{i+1}, \dots, z_m) \in (\partial\mathbb{D})^{m-i}$,

$$\begin{aligned} &\left(z_m^{[k+1]} \otimes \cdots \otimes z_{i+1}^{[k+1]} \otimes I_{(k+1)^i n}\right)^H \\ &\quad \left[R_k + \sum_{j=1}^{i} \left(\hat{J}_k^{(m-j+1)\otimes} \otimes I_{(k+1)^{j-1} n}\right)^T Q_{k,j} \left(\hat{J}_k^{(m-j+1)\otimes} \otimes I_{(k+1)^{j-1} n}\right) \right. \\ &\quad \left. - \sum_{j=1}^{i} \left(\hat{J}_k^{(m-j)\otimes} \otimes \check{J}_k \otimes I_{(k+1)^{j-1} n}\right)^T Q_{k,j} \left(\hat{J}_k^{(m-j)\otimes} \otimes \check{J}_k \otimes I_{(k+1)^{j-1} n}\right) \right] \\ &\qquad \left(z_m^{[k+1]} \otimes \cdots \otimes z_{i+1}^{[k+1]} \otimes I_{(k+1)^i n}\right) < 0_{(k+1)^i n}. \end{aligned}$$

In the previous expression, the matrix $R_k = R_k(P_k)$ is defined in (9). One verifies easily that $(\mathcal{P}_0)$ may also be expressed as follows: there exists $P(z)$ as in (7) such that $P_k > 0$ and $R(z)$ defined in (8), (9) is negative definite for all $z \in (\partial\mathbb{D})^m$. Property $(\mathcal{P}_0)$ is thus a consequence of (ii) (see section 7.1.2), while in parallel $(\mathcal{P}_m)$ writes simply the following: there exists $k \in \mathbb{N}$ such that (LMI_k) holds, that is, (iii).

In order to prove that $(\mathcal{P}_0)$ implies $(\mathcal{P}_m)$, we establish the slightly stronger following result.

LEMMA 7.2. *For all $i = 0, \dots, m-1$, $(\mathcal{P}_i) \Leftrightarrow (\mathcal{P}_{i+1})$.*

Proof of Lemma 7.2. First, remark that

$$\begin{aligned} &\left(z_m^{[k+1]} \otimes \cdots \otimes z_{i+1}^{[k+1]} \otimes I_{(k+1)^i n}\right) \\ &= \left(z_m^{[k+1]} \otimes \cdots \otimes z_{i+2}^{[k+1]} \otimes I_{(k+1)^{i+1} n}\right) \left(z_{i+1}^{[k+1]} \otimes I_{(k+1)^i n}\right) \\ &= \left(z_m^{[k+1]} \otimes \cdots \otimes z_{i+2}^{[k+1]} \otimes I_{(k+1)^{i+1} n}\right) \\ &\quad \times \begin{pmatrix} I_{(k+1)^i n} \\ z_{i+1} \left(I_{k(k+1)^i n} - z_{i+1}(F_k \otimes I_{(k+1)^i n})\right)^{-1} (f_k \otimes I_{(k+1)^i n}) \end{pmatrix}, \end{aligned}$$

the last identity being obtained after writing $z_{i+1}^{[k+1]} = \begin{pmatrix} 1 \\ z_{i+1} z_{i+1}^{[k]} \end{pmatrix}$ and using (18).

Applying the Kalman–Yakubovich–Popov lemma as recalled in Appendix A, with $p = k(k+1)^i n$, $q = (k+1)^i n$, $A = F_k \otimes I_{(k+1)^i n}$, $B = f_k \otimes I_{(k+1)^i n}$, and remarking that the following identities hold:

$$\begin{pmatrix} B & A \end{pmatrix} = \hat{J}_k \otimes I_{(k+1)^i n}, \ \begin{pmatrix} 0_{p \times q} & I_p \end{pmatrix} = \check{J}_k \otimes I_{(k+1)^i n},$$

property $(\mathcal{P}_i)$ is proved to be equivalent to $\exists k \in \mathbb{N}$, $\exists P_k \in \mathcal{H}^{k^m n}$, $P_k > 0$, $\exists Q_{k,j} \in \mathcal{H}^{k^{m-j+1}(k+1)^{j-1} n}$, $j = 1, \ldots, i$, for all $(z_{i+2}, \ldots, z_m) \in (\partial \mathbb{D})^{m-i-1}$, $\exists \tilde{Q}_{k,i+1}(z_{i+2}, \ldots, z_m) \in \mathcal{H}^{k(k+1)^i n}$,

$$\begin{aligned}
&\left(z_m^{[k+1]} \otimes \cdots \otimes z_{i+2}^{[k+1]} \otimes I_{(k+1)^i n} \right)^H \\
&\quad \left[R_k + \sum_{j=1}^{i} \left(\hat{J}_k^{(m-j+1)\otimes} \otimes I_{(k+1)^{j-1} n} \right)^T Q_{k,j} \left(\hat{J}_k^{(m-j+1)\otimes} \otimes I_{(k+1)^{j-1} n} \right) \right. \\
&\quad \left. - \sum_{j=1}^{i} \left(\hat{J}_k^{(m-j)\otimes} \otimes \check{J}_k \otimes I_{(k+1)^{j-1} n} \right)^T Q_{k,j} \left(\hat{J}_k^{(m-j)\otimes} \otimes \check{J}_k \otimes I_{(k+1)^{j-1} n} \right) \right] \\
&\qquad \left(z_m^{[k+1]} \otimes \cdots \otimes z_{i+2}^{[k+1]} \otimes I_{(k+1)^i n} \right) \\
&+ \left(\hat{J}_k \otimes I_{(k+1)^i n} \right)^T \tilde{Q}_{k,i+1} \left(\hat{J}_k \otimes I_{(k+1)^i n} \right) - \left(\check{J}_k \otimes I_{(k+1)^i n} \right)^T \tilde{Q}_{k,i+1} \left(\check{J}_k \otimes I_{(k+1)^i n} \right) \\
&< 0_{(k+1)^{i+1} n}.
\end{aligned}$$

The next step consists of assigning polynomial form to $\tilde{Q}_{k,i+1}$. This is done with the help of the following general result, proved in Appendix B, and which, up to our knowledge, is original.

THEOREM 7.3. *Suppose $G_0, G_1, \ldots, G_p$ are continuous mappings defined in a* compact *subset K of $\mathbb{R}^m$ and taking values in the set of symmetric matrices of $\mathbb{R}^{n \times n}$. If, for any $\delta \in K$, there exists a solution $x(\delta) \in \mathbb{R}^p$ to the parameter-dependent LMI*

$$\exists x \in \mathbb{R}^p, \ G(x, \delta) \stackrel{\text{def}}{=} G_0(\delta) + x_1 G_1(\delta) + \cdots + x_p G_p(\delta) > 0_n, \tag{31}$$

then there exists a polynomial *function $x^* : K \to \mathbb{R}^p$ such that, for any $\delta \in K$, $G(x^*(\delta), \delta) > 0_n$.*

Remark 7.4. Incidentally, one may wonder why Theorem 7.3 was not used in section 7.1.2 in order to get a polynomial expansion of $P(z)$; see formula (22) above. The reason is that semidefiniteness of the matrices $\tilde{P}_k$ as given by (23), which cannot be obtained by Theorem 7.3, was a crucial point to carry on the second stage.

Notice that any LMI depending upon a finite number of scalar parameters may be put under the form (31).

By use of the previous result, $\tilde{Q}_{k,i+1}(z_{i+2}, \ldots, z_m)$, being the solution of a LMI continuous with respect to the parameters $(z_{i+2}, \ldots, z_m)$ in $(\partial \mathbb{D})^{m-i-1}$ (seen as a compact set in $\mathbb{R}^{2(m-i-1)}$), may be chosen polynomial in the real and imaginary parts of the z_i, or also in the $z_i, \bar{z}_i$; that is,

$$\tilde{Q}_{k,i+1} = \left(z_m^{[\tilde{k}]} \otimes \cdots \otimes z_{i+2}^{[\tilde{k}]} \otimes I_{k(k+1)^i n} \right)^H Q_{\tilde{k},i+1} \left(z_m^{[\tilde{k}]} \otimes \cdots \otimes z_{i+2}^{[\tilde{k}]} \otimes I_{k(k+1)^i n} \right) \tag{32}$$

for certain degree $\tilde{k}-1$ and coefficient matrix $Q_{\tilde{k},i+1} \in \mathcal{H}^{k\tilde{k}^{m-i-1}(k+1)^i n}$.

A priori, the integers k and $\tilde{k}$ are different. If $\tilde{k} < k$, one may also suppose that $\tilde{k} = k$, enlarging the coefficient matrix $Q_{\tilde{k},i+1}$ by addition of zeros. If $\tilde{k} > k$, one shows now that k may also be replaced by $k+1$. For this, define

$$P_{k+1} \stackrel{\text{def}}{=} \sum_{M_i \in \{\hat{J}_k, \check{J}_k\},\ i=1,\dots,m} (M_m \otimes \cdots \otimes M_1 \otimes I_n)^T P_k (M_m \otimes \cdots \otimes M_1 \otimes I_n),$$

and, for $j = 1, \dots, i$,

$$Q_{k+1,j} \stackrel{\text{def}}{=} \sum_{\substack{M_l \in \{\hat{J}_{k+1}, \check{J}_{k+1}\},\ l=1,\dots,j-1, \\ M_l \in \{\hat{J}_k, \check{J}_k\},\ l=j,\dots,m}} (M_m \otimes \cdots \otimes M_1 \otimes I_n)^T Q_{k,j} (M_m \otimes \cdots \otimes M_1 \otimes I_n),$$

$$\tilde{Q}_{k+1,i+1} \stackrel{\text{def}}{=} \sum_{\substack{M_j \in \{\hat{J}_{k+1}, \check{J}_{k+1}\},\ j=1,\dots,i, \\ M_j \in \{\hat{J}_k, \check{J}_k\},\ j=i+1,\dots,m}} (M_m \otimes \cdots \otimes M_1 \otimes I_n)^T \tilde{Q}_{k,i+1} (M_m \otimes \cdots \otimes M_1 \otimes I_n).$$

One first shows that the positivity of P_k implies positivity of P_{k+1}: for any $u \in \mathbb{C}^{(k+1)^m n}$ such that $u^H P_{k+1} u = 0$, one has $P_k^{1/2}(M_m \otimes \cdots \otimes M_1 \otimes I_n)u = 0$ for any $M_i \in \{\hat{J}_k, \check{J}_k\}, i = 1, \dots, m$, and this implies that $u = 0$, whence the positivity of P_{k+1}. One then shows that the matrix R_{k+1} obtained from P_{k+1} by formula (9) verifies

$$R_{k+1} \stackrel{\text{def}}{=} \sum_{M_i \in \{\hat{J}_{k+1}, \check{J}_{k+1}\},\ i=1,\dots,m} (M_m \otimes \cdots \otimes M_1 \otimes I_n)^T R_k (M_m \otimes \cdots \otimes M_1 \otimes I_n).$$

This requires cumbersome but straightforward calculations, using property (5). A new set of matrices verifying property $(\mathcal{P}_i)$ has thus been generated, with index $k+1$ instead of k. Remark that otherwise the degree $\tilde{k}-1$ in the unknowns $z_{i+2}, \dots, z_m$ of the new matrix $\tilde{Q}_{k+1,i+1}$ is the same as for $\tilde{Q}_{k,i+1}$. It thus suffices to repeat this operation to obtain a solution with $k = \tilde{k}$. Finally, *up to a possible increase of* k, one may always suppose that $k = \tilde{k}$ in the decomposition (32) of $\tilde{Q}_{k,i+1}$.

Remark 7.5. Applying the previous argument to $(\mathcal{P}_m)$ proves that solvability of (LMI_k) implies the same property for the larger values of the index, as announced in the sketch of the proof of formula (11).

It now remains to achieve some matrix manipulations. Using the following formula, obtained by use of (6),

$$\begin{aligned}
&\left(z_m^{[k]} \otimes \cdots \otimes z_{i+2}^{[k]} \otimes I_{k(k+1)^i n}\right)\left(\hat{J}_k \otimes I_{(k+1)^i n}\right) \\
&= \left(I_{k^{m-i-1}} \otimes \hat{J}_k \otimes I_{(k+1)^i n}\right)\left(z_m^{[k]} \otimes \cdots \otimes z_{i+2}^{[k]} \otimes I_{(k+1)^{i+1} n}\right) \\
&= \left(\hat{J}_k^{(m-i)\otimes} \otimes I_{(k+1)^i n}\right)\left(z_m^{[k+1]} \otimes \cdots \otimes z_{i+2}^{[k+1]} \otimes I_{(k+1)^{i+1} n}\right),
\end{aligned}$$

and similarly

$$\begin{aligned}
&\left(z_m^{[k]} \otimes \cdots \otimes z_{i+2}^{[k]} \otimes I_{k(k+1)^i n}\right)\left(\check{J}_k \otimes I_{(k+1)^i n}\right) \\
&\qquad = \left(\hat{J}_k^{(m-i-1)\otimes} \otimes \check{J}_k \otimes I_{(k+1)^i n}\right)\left(z_m^{[k+1]} \otimes \cdots \otimes z_{i+2}^{[k+1]} \otimes I_{(k+1)^{i+1} n}\right),
\end{aligned}$$

one finally proves that $(\mathcal{P}_i)$ is equivalent to

$\exists k \in \mathbb{N}, \exists P_k \in \mathcal{H}^{k^m n}, P_k > 0, \exists Q_{k,j} \in \mathcal{H}^{k^{m-j+1}(k+1)^{j-1}n}$, $j = 1, \ldots, i+1$, $\forall (z_{i+2}, \ldots, z_m) \in (\partial\mathbb{D})^{m-i-1}$,

$$\left(z_m^{[k+1]} \otimes \cdots \otimes z_{i+2}^{[k+1]} \otimes I_{(k+1)^{i+1}n}\right)^H \left[R_k + \sum_{j=1}^{i+1} \left(\hat{J}_k^{(m-j+1)\otimes} \otimes I_{(k+1)^{j-1}}\right)^T Q_{k,j} \left(\hat{J}_k^{(m-j+1)\otimes} \otimes I_{(k+1)^{j-1}}\right) - \sum_{j=1}^{i+1} \left(\hat{J}_k^{(m-j)\otimes} \otimes \check{J}_k \otimes I_{(k+1)^{j-1}}\right)^T Q_{k,j} \left(\hat{J}_k^{(m-j)\otimes} \otimes \check{J}_k \otimes I_{(k+1)^{j-1}}\right)\right] \left(z_m^{[k+1]} \otimes \cdots \otimes z_{i+2}^{[k+1]} \otimes I_{(k+1)^{i+1}n}\right) < 0_{(k+1)^{i+1}n}.$$

One recognizes property $(\mathcal{P}_{i+1})$. Hence, $(\mathcal{P}_i) \Leftrightarrow (\mathcal{P}_{i+1})$, and Lemma 7.2 is proved. □

The equivalence between $(\mathcal{P}_0)$ and $(\mathcal{P}_m)$ shows in particular that (ii) implies (iii). This achieves the proof of Theorem 4.1.

7.2. Proof of Theorem 4.3. The proof proceeds by using the change of variables $r = (z + \bar{z})/2$, $z \in \overline{\mathbb{D}}^m$, already introduced to get Proposition 3.3, and by achieving the slight necessary adaptations of the proof of Theorem 4.1, using R_k defined in (10) and not in (9). The argument used in the first stage is here trivial, as the sets $\{A((z+\bar{z})/2) : z \in \overline{\mathbb{D}}^m\}$ and $\{A((z+\bar{z})/2) : z \in (\partial\mathbb{D})^m\}$ are identical.

Remark 7.6. Notice that the change of variables which is used leads to D-scaling as in the complex parameter case, and not DG-scaling, although the parameters involved here are real.

8. Conclusion. Robust stability of linear systems with several scalar (complex or real) parameters has been studied. For each problem, a family of LMIs, indexed by a positive integer k, is provided. Their solvability is sufficient for robust stability, and the corresponding conditions are becoming less conservative with increasing k. Conversely, if robust stability holds, then the corresponding LMI problems are solvable from a certain k and beyond. The method involves the search for a quadratic Lyapunov function depending polynomially on the parameters and their conjugates.

The LMIs are obtained in a constructive and systematic way, resulting from a limited set of elementary algebraic matrix operations. In consequence, the derived algorithms are immediately implementable in a MATLAB/SCILAB-like environment. In practice, the accuracy of the approximation is only limited by computation time and available memory size.

Further research includes the following aspects.

1. Determination of the degree of accuracy needed to test the robust stability of any specific system, that is, of an a priori (upper) estimate on the least k, if any, for which the LMIs are solvable. More generally, the complexity and numerical aspects have to be analyzed.

2. Extension of the results to robust input/output performance evaluation for systems with scalar parameters and to systems with polynomial and LFT dependency (see the first results in [6] and [5, 4], respectively). Application to μ-analysis.

Appendix A. Discrete-time version of the Kalman–Yakubovich–Popov lemma.

Initially appearing in [44], the result has been first published under its discrete-time form by Szegö and Kalman [36]. We use the statement as expressed, e.g., in [34]. A proof of the result in the complex case (and for the continuous-time case) may be found in [30, Theorem 1.11.1 and Remark 1.11.1].

Let $A \in \mathbb{C}^{p\times p}, B \in \mathbb{C}^{p\times q}, M \in \mathcal{H}^{p+q}$.

LEMMA A.1. *If* $\det(I_n - zA) \neq 0$ *for any* $z \in \partial\mathbb{D}$, *then the following two statements are equivalent.*

(i) *There exists* $Q \in \mathcal{H}^p$ *such that*

$$0_{p+q} > \begin{pmatrix} B & A \end{pmatrix}^H Q \begin{pmatrix} B & A \end{pmatrix} - \begin{pmatrix} 0_{p\times q} & I_p \end{pmatrix}^H Q \begin{pmatrix} 0_{p\times q} & I_p \end{pmatrix} + M.$$

(ii) *For any* $z \in \partial\mathbb{D}$,

$$\begin{pmatrix} I_p \\ z(I_p - zA)^{-1}B \end{pmatrix}^H M \begin{pmatrix} I_p \\ z(I_p - zA)^{-1}B \end{pmatrix} < 0_p.$$

When in the statements the matrices A, B, M are real, then Q is real symmetric.

Appendix B. Proof of Theorem 7.3.

Under the hypothesis of solvability of (31) for any $\delta \in K$, there exists, by continuity and compactness, a real number $\alpha > 0$ such that

$$\forall \delta \in K,\ \{x \in \mathbb{R}^p\ :\ G_0(\delta) + x_1 G_1(\delta) + \cdots + x_p G_p(\delta) \geq 2\alpha I_n\} \neq \emptyset.$$

Define

$$\begin{aligned} F\ :\ & K \to 2^{\mathbb{R}^p}, \\ & \delta \mapsto F(\delta) = \{x \in \mathbb{R}^p\ :\ G_0(\delta) + x_1 G_1(\delta) + \cdots + x_p G_p(\delta) \geq \alpha I_n\}. \end{aligned} \tag{33}$$

The set-valued map F maps K into the nonvoid closed convex subsets of $\mathbb{R}^p$.

Let us first establish that F fulfils the following property of *lower semicontinuity*; see, e.g., [2].

DEFINITION B.1. *Let* X *be a topological space and* Y *a metric space. A set-valued map* F *from* X *to* Y *is said to be* lower semicontinuous *at* $x^0 \in X$ *if for any* $y^0 \in F(x^0)$ *and any neighborhood* $N(y^0)$ *of* y^0 *there exists a neighborhood* $N(x^0)$ *such that*

$$\forall x \in N(x^0),\ F(x) \cap N(y^0) \neq \emptyset.$$

F *is said to be* lower semicontinuous *if it is lower semicontinuous at every point* $x^0 \in X$.

Let $\delta^0 \in K$, $x^0 \in F(\delta^0)$, $\varepsilon > 0$. To prove lower semicontinuity of F at δ^0, we exhibit $\eta > 0$ such that for any $\delta \in K$ with $\|\delta - \delta^0\|_m < \eta$, there exists $x \in F(\delta)$, $\|x - x^0\|_p < \varepsilon$.

Indeed, by assumption, there exists $x^{\delta^0} \in \mathbb{R}^p$ such that $G(x^{\delta^0}, \delta^0) \geq 2\alpha I_n$. For $\lambda \in (0, 1]$ to be defined afterwards, let $x \stackrel{\text{def}}{=} (1-\lambda)x^0 + \lambda x^{\delta^0}$. Then the fact that G is

affine with respect to x implies for any $\eta > 0$, any $\delta \in K$ such that $\|\delta - \delta^0\|_m < \eta$:

$$\begin{aligned} G(x,\delta) &= (1-\lambda)G(x^0,\delta) + \lambda G(x^{\delta^0},\delta) \\ &= (1-\lambda)G(x^0,\delta^0) + \lambda G(x^{\delta^0},\delta^0) \\ &\quad +(1-\lambda)\left(G(x^0,\delta) - G(x^0,\delta^0)\right) + \lambda\left(G(x^{\delta^0},\delta) - G(x^{\delta^0},\delta^0)\right) \\ &\geq \alpha(1+\lambda)I_n \\ &\quad -\left(\sup_{\|\delta-\delta^0\|_m<\eta} \|G(x^0,\delta) - G(x^0,\delta^0)\|_n + \sup_{\|\delta-\delta^0\|_m<\eta} \|G(x^{\delta^0},\delta) - G(x^{\delta^0},\delta^0)\|_n\right) I_n. \end{aligned}$$

On the other hand,

$$\|x - x^0\|_p = \lambda \|x^{\delta^0} - x^0\|_p.$$

So take $\lambda \in (0,1]$ such that

$$\lambda \leq \frac{\varepsilon}{2\|x^{\delta^0} - x^0\|_p},$$

and choose $\eta > 0$ such that

$$\sup_{\|\delta-\delta^0\|_m<\eta} \|G(x^0,\delta) - G(x^0,\delta^0)\|_n + \sup_{\|\delta-\delta^0\|_m<\eta} \|G(x^{\delta^0},\delta) - G(x^{\delta^0},\delta^0)\|_n \leq \alpha\lambda.$$

With these choices, one has $\|x - x^0\|_p \leq \varepsilon/2 < \varepsilon$, and $G(x,\delta) \geq \alpha(1+\lambda)I_n - \alpha\lambda I_n = \alpha I_n$, so $x \in F(\delta)$, provided that $\delta \in K$ and $\|\delta - \delta^0\|_m < \eta$. One concludes that F is lower continuous at δ^0. This achieves the proof of lower semicontinuity of F.

We now apply to F defined in (33) Michael's selection theorem [31]; see also [2].

THEOREM B.2 (Michael's selection theorem). *Let X be a metric space and Y a Banach space. Let F from X into the closed convex subsets of Y be lower semicontinuous. Then there exists $f : X \to Y$, a continuous selection from F.*

This yields existence of a continuous selection $f : K \to \mathbb{R}^p$ from F defined in (33). This function is such that

$$\forall \delta \in K,\ G(f(\delta),\delta) \geq \alpha I_n.$$

It remains to apply to each of the p coefficients of f the following result; see, e.g., [14].

THEOREM B.3 (Weierstrass approximation theorem). *Every continuous real-valued function defined on a compact subset K of $\mathbb{R}^m$ is the limit of a sequence of polynomials, which converges uniformly in K.*

Thus, the selection f previously exhibited is the uniform limit in K of a sequence of (matrix-valued) polynomials in x. In particular, there exists a polynomial function $x^* : K \to \mathbb{R}^p$ such that

$$\forall \delta \in K,\ G(x^*(\delta),\delta) \geq \frac{\alpha}{2} I_n > 0_n.$$

One concludes that there exists a polynomial solution to the parameter-dependent LMI (31), and this achieves the proof of Theorem 7.3.

Acknowledgments. The revised form of this paper has benefited from many comments and encouragements during seminaries given in 2002 in St. Petersburg, Russia. The author wishes to express his indebtedness to his colleagues of this community, especially to Prof. V. A. Yakubovich and to Prof. A. L. Fradkov. He also wishes to thank Prof. P. Tsiotras for his attentive reading and anonymous reviewers whose constructive criticism allowed precious improvements of the manuscript.

REFERENCES

[1] T. Asai, S. Hara, and T. Iwasaki, *Simultaneous modeling and synthesis for robust control by LFT scaling*, in Proceedings of the IFAC World Congress, part G, San Francisco, 1996, pp. 309–314.

[2] J.-P. Aubin and A. Cellina, *Differential Inclusions. Set-Valued Maps and Viability Theory*, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1984.

[3] V. Balakrishnan, S. Boyd, and S. Balemi, *Branch and bound algorithm for computing the minimum stability degree of parameter-dependent linear systems*, Internat. J. Robust Nonlinear Control, 1 (1992), pp. 295–317.

[4] P.-A. Bliman, *LMI approach to spectral stabilizability of linear delay systems and stabilizability of linear systems with complex parameter*, in Proceedings of the 40th IEEE CDC, Orlando, FL, 2001.

[5] P.-A. Bliman, *Lyapunov equation for the stability of linear delay systems of retarded and neutral type*, IEEE Trans. Automat. Control, 47 (2002), pp. 327–335.

[6] P.-A. Bliman, *LMIs for delay-independent properties of delay systems and stability/input-output analysis of systems with complex parameter*, in Proceedings of the 41st IEEE CDC, Las Vegas, NV, 2002.

[7] S. Boyd and C. A. Desoer, *Subharmonic functions and performance bounds on linear time-invariant feedback systems*, IMA J. Math. Control Inform., 2 (1985), pp. 153–170.

[8] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.

[9] R. P. Braatz, P. M. Young, J. C. Doyle, and M. Morari, *Computational complexity of μ calculation*, IEEE Trans. Automat. Control, 39 (1994), pp. 1000–1002.

[10] J. Chen and H. A. Latchman, *Frequency sweeping tests for stability independent of delay*, IEEE Trans. Automat. Control, 40 (1995), pp. 1640–1645.

[11] R. R. De Gaston and M. G. Safonov, *Exact calculation of the multiloop stability margin*, IEEE Trans. Automat. Control, 33 (1988), pp. 156–171.

[12] M. Dettori and C. W. Scherer, *Robust stability analysis for parameter dependent systems using full block S-procedure*, in Proceedings of the 37th IEEE CDC, Tampa, FL, 1998, pp. 2798–2799.

[13] M. Dettori and C. W. Scherer, *New robust stability and performance conditions based on parameter dependent multipliers*, in Proceedings of the 39th IEEE CDC, Sydney, Australia, 2000.

[14] J. Dieudonné, *Foundations of modern analysis* (enlarged and corrected printing), Pure Appl. Math. 10-I, Academic Press, New York, London, 1969.

[15] J. C. Doyle, *Analysis of feedback systems with structured uncertainties*, IEE Proc. Part D, 129 (1982), pp. 242–250.

[16] M. K. H. Fan, A. L. Tits, J. C. Doyle, *Robustness in the presence of mixed parametric uncertainty and unmodeled dynamics*, IEEE Trans. Automat. Control, 36 (1991), pp. 25–38.

[17] E. Feron, P. Apkarian, and P. Gahinet, *Analysis and synthesis of robust control systems via parameter-dependent Lyapunov functions*, IEEE Trans. Automat. Control, 41 (1996), pp. 1041–1046.

[18] M. Fu, *The real structured singular value is hardly approximable*, IEEE Trans. Automat. Control, 42 (1997), pp. 1286–1288.

[19] M. Fu and N. E. Barabanov, *Improved upper bounds for the mixed structured singular value*, IEEE Trans. Automat. Control, 42 (1997), pp. 1447–1452.

[20] M. Fu and S. Dasgupta, *Parametric Lyapunov functions for uncertain systems: The multiplier approach*, in Advances in Linear Matrix Inequality Methods in Control, L. El Ghaoui and S.-I. Niculescu, eds., SIAM, Philadelphia, 2000, pp. 95–108.

[21] P. Gahinet, P. Apkarian, and M. Chilali, *Affine parameter-dependent Lyapunov functions and real parametric uncertainty*, IEEE Trans. Automat. Control, 41 (1996), pp. 436–442.

[22] J. C. Geromel, M. C. de Oliveira, and L. Hsu, *LMI characterization of structural and robust stability*, Linear Algebra Appl., 285 (1998), pp. 69–80.
[23] J. K. Hale, E. F. Infante, and F. S. P. Tsen, *Stability in linear delay equations*, J. Math. Anal. Appl., 115 (1985), pp. 533–555.
[24] T. Iwasaki, *LPV system analysis with quadratic separator*, in Proceedings of the 37th IEEE CDC, Tampa, FL, 1998.
[25] T. Iwasaki, *Generalized quadratic Lyapunov functions for nonlinear/uncertain systems analysis*, in Perspectives in Robust Control, S. O. Reza Moheimani, ed., Lecture Notes in Control and Inform. Sci. 268, Springer-Verlag, London, 2001, pp. 149–174.
[26] T. Iwasaki and S. Hara, *Well-posedness of feedback systems: Insights into exact robustness analysis and approximate computations*, IEEE Trans. Automat. Control, 43 (1998), pp. 619–630.
[27] E. W. Kamen, *Linear systems with commensurate time delays: Stability and stabilization independent of delay*, IEEE Trans. Automat. Control, 27 (1982), pp. 367–375.
[28] E. W. Kamen, *Correction to "Linear systems with commensurate time delays: stability and stabilization independent of delay,"* IEEE Trans. Automat. Control, 28 (1983), pp. 248–249.
[29] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Clarendon Press, Oxford University Press, New York, 1995.
[30] G. A. Leonov, D. V. Ponomarenko, and V. B. Smirnova, *Frequency-Domain Methods for Nonlinear Analysis. Theory and Applications*, World Sci. Ser. Nonlinear Sci. Ser. A Monogr. Treatises, 9, World Scientific, River Edge, NJ, 1996.
[31] E. Michael, *Continuous selections* I, Annals of Math., 63 (1956), pp. 361–381.
[32] D. Peaucelle, D. Arzelier, O. Bachelier, and J. Bernussou, *A new robust D-stability condition for real convex polytopic uncertainty*, Systems Control Lett., 40 (2000), pp. 21–30.
[33] D. C. W. Ramos and P. L. D. Peres, *An LMI approach to compute robust stability domains for uncertain linear systems*, in Proceedings of the American Control Conference, Arlington, VA, 2001, pp. 4073–4078.
[34] A. Rantzer, *On the Kalman-Yakubovich-Popov lemma*, Systems Control Lett., 28 (1996), pp. 7–10.
[35] A. Rantzer and M. Johansson, *Piecewise linear quadratic optimal control*, IEEE Trans. Automat. Control, 45 (2000), pp. 629–637.
[36] G. Szegö and R. E. Kalman, *Sur la stabilité absolue d'un système d'équations aux différences finies*, Comp. Rend. Acad. Sci., 257 (1963), pp. 338–390.
[37] M. Sznaier and P. A. Parrilo, *On the gap between μ and its upper bound for systems with repeated uncertainty blocks*, in Proceedings of the 38th IEEE CDC, Phoenix, AZ, 1999, pp. 4511–4516.
[38] O. Toker and H. Özbay, *Complexity issues in robust stability of linear delay-differential systems*, Math. Control Signals Systems, 9 (1996), pp. 386–400.
[39] O. Toker and H. Özbay, *On the complexity of purely complex μ computation and related problems in multidimensional systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 409–414.
[40] S. Treil, *The Gap between Complex Structured Singular Value μ and Its Upper Bound Is Infinite*, http://www.math.msu.edu/treil/papers/mu/mu-abs.html (1999).
[41] A. Trofino, *Parameter dependent Lyapunov functions for a class of uncertain linear systems: A LMI approach*, in Proceedings of the 38th IEEE CDC, Phoenix, AZ, 1999, pp. 2341–2346.
[42] A. Trofino and C. E. de Souza, *Bi-quadratic stability of uncertain linear systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 1303–1307.
[43] L. Xie, S. Shishkin, and M. Fu, *Piecewise Lyapunov functions for robust stability of linear time-varying systems*, Systems Control Lett., 31 (1997), pp. 165–171.
[44] V. A. Yakubovich, *Solution of certain matrix inequalities in the stability theory of nonlinear control systems*, Dokl. Akad. Nauk. SSSR, 143 (1962), pp. 1304–1307 (in Russian); Soviet Math. Dokl., 3 (1962), pp. 620–623 (in English).
[45] P. M. Young, M. P. Newlin, and J. C. Doyle, *Let's get real*, in Robust Control Theory, IMA Vol. Math. Appl. 66, Springer, New York, 1995, pp. 143–173.
[46] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice–Hall, Upper Saddle River, NJ, 1996.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 2043–2061

NECESSARY OPTIMALITY CONDITIONS IN MULTIOBJECTIVE DYNAMIC OPTIMIZATION*

SAÏD BELLAASSALI† AND ABDERRAHIM JOURANI†

Abstract. We consider a nonsmooth multiobjective optimal control problem related to a general preference. Both differential inclusion and endpoint constraints are involved. Necessary conditions and Hamiltonian necessary conditions expressed in terms of the limiting Fréchet subdifferential are developed. Examples of useful preferences are given.

Key words. multiobjective optimal control, necessary conditions, Hamiltonian necessary conditions, preference, utility function, differential inclusions

AMS subject classifications. 49K24, 90C29

DOI. 10.1137/S0363012902406576

1. Introduction. This paper is mainly concerned with the following multiobjective dynamic optimization problem with the dynamic governed by a differential inclusion:

$$\text{(P)} \qquad \begin{aligned} &\min f(x(a), x(b)), \\ &(x(a), x(b)) \in S, \\ &\dot{x}(t) \in F(t, x(t)) \quad \text{a.e.} \quad t \in [a, b], \end{aligned}$$

where $f\colon \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}^m$ is a mapping, $S \subset \mathbb{R}^n \times \mathbb{R}^n$ is a closed nonempty set, and $F\colon [a, b] \times \mathbb{R}^n \mapsto \mathbb{R}^n$ is a closed-valued multivalued mapping which is measurable in $t \in [a, b]$.

These problems naturally arise, for example, in economics (economic growth models) (see [16] and references therein), in chemical engineering (polymerization processes) (see [3], [4], and references therein), and in multiobjective control design (see [45], [9], and references therein). Problems considered in this paper use preferences determined by cones (Pareto and weak Pareto optimum), use preferences determined by utility function, or use the concept of Nash equilibrium.

Our aim in this paper is to use a general preference including the previous ones in order to state necessary and Hamiltonian necessary conditions for multiobjective optimal control problems (P).

The concept of preference appeared in the value theory in economics. Many authors in the early studies often defined the preference by a utility function, i.e., given a preference whether it is always possible to find a utility function that can determine the preference.

In [17] the author proved that a preference $\prec$ can be determined by a continuous utility function if and only if for any x the sets

$$\text{(1)} \qquad \{y : x \prec y\} \quad \text{and} \quad \{y : y \prec x\} \quad \text{are closed.}$$

This theorem is not general and besides this it is an existence theorem (i.e., does provide methods for determing a utility function), and there are some useful preferences

*Received by the editors April 30, 2002; accepted for publication (in revised form) May 7, 2003; published electronically January 28, 2004.
http://www.siam.org/journals/sicon/42-6/40657.html
†Université de Bourgogne, Analyse Appliquée et Optimisation, B.P. 47870, 21078 Dijon Cedex, France (bellaass@u-bourgogne.fr, jourani@u-bourgogne.fr).

that do not satisfy (1) (like the preference determined by lexicographical order). There are different approaches and various results on necessary conditions for (P). Several researches have been devoted to the weak Pareto solution and its generalization (see [5], [11], [15], [32], [42], [46], [47], and references therein). Other research gets refinements of necessary optimality conditions for real-valued objective optimal control problems (see [23], [29], [30], [31], [44], [43], and [26]) or Hamiltonian necessary conditions (see [37], [19], [20], [13], [35], [36], [39], [48], and [49]).

These results are expressed in terms of various generalized derivatives including Clarke's generalized subgradient [11], a limiting subgradient which is also known under other names: limiting subgradient set in [12], approximate subdifferential in Ioffe [22], subdifferential in Mordukhovich [36], and subgradient set in the general sense in Rockafellar [40]. Most of these results are obtained for Lipschitz, integrably sub-Lipschitz, bounded, or unbounded differential inclusions.

In [23], Ioffe used results of [40] and [24] to obtain general necessary optimality conditions and Hamiltonian optimality conditions for single-objective optimal control problems.

In [49], Zhu used recent progress in nonsmooth analysis, in particular calculus for smooth subdifferentials of lower semicontinuous (l.s.c.) functions (see [6], [7], [14], [24]), the methods for proving the extremal principle (see [27], [28], [33], [38]), and techniques in handling the Hamiltonian for a differential inclusion, to prove Hamiltonian necessary conditions that extend the classical Hamiltonian necessary conditions for optimal control problems that had previously been derived for uniformly Lipschitz, bounded, and convex-valued differential inclusions related to a general preference. The obtained conditions are expressed in terms of Clarke's generalized gradient which is larger than the limiting Fréchet subdifferential. The regularity conditions $(A3)$ imposed in [49], which use the usual limiting normal cone, are too strong to include the preference defined by a utility function (see Example 3).

In this paper we propose a different approach. We introduce a definition of regularity modified from that introduced in [49]. To solve the problem of regularity of preference determined by a utility function, we define a larger limiting normal cone to replace the usual one in [49]. Under our regularity condition of the general preference and a sub-Lipschitz property of multivalued mappings, introduced by Loewen and Rockafellar in [29], we obtain Euler–Lagrange necessary optimality conditions for multiobjective optimal control problems with nonconvex differential inclusion constraints in terms of the limiting Fréchet subdifferential. Necessary optimality conditions for the weak Pareto solution and its generalization can be derived and refined by using our necessary conditions.

Our main result extends the necessary optimality condition of Ioffe (see Theorem 1 in [23]) from a single objective optimal control of differential inclusion problem to a multiobjective one. This is also an extension of the Hamiltonian necessary optimality conditions for convex differential inclusions obtained in [49].

The paper is organized as follows. Section 2 contains the key definitions, normals, subgradients, and coderivatives used in what follows. In section 3 we state our main result and establish necessary optimality conditions for multiobjective control problems with some examples and discussions. Then we derive necessary conditions for these examples of preferences. In section 4 we give a technical proof of the main result.

2. Background. Now we state basic tools of generalized differentiation that are more appropriate for our main purpose. Details may be found in [33].

Let C be a closed subset of $\mathbb{R}^n$ containing some point c. The ε-normal cone to C at c is the set

$$\hat{N}_\varepsilon(C,c) := \left\{ \zeta \in \mathbb{R}^n \ : \ \liminf_{x \in C \to c} \frac{\langle -\zeta, x-c\rangle}{\|x-c\|} \geq -\varepsilon \right\}.$$

The normal cone to C at c is the set

$$N(C;c) := \limsup_{\substack{x \in C \to c \\ \varepsilon \to 0^+}} \hat{N}_\varepsilon(C,c).$$

Now let $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{\infty\}$ be an l.s.c. function, and let $c \in \mathbb{R}^n$ such that $f(c) < \infty$. The limiting Fréchet subdifferential of f at c is the set

$$\partial f(c) = \{\zeta \in \mathbb{R}^n \ : \ (\zeta, -1) \in N(epi\, f; (c, f(c)))\},$$

where *epi* f denotes the epigraph of f. We have the following analytic characterization of $\partial f(c)$:

$$\partial f(c) = \limsup_{\substack{x \to c \\ f(x) \to f(c) \\ \varepsilon \to 0^+}} \partial_\varepsilon f(x),$$

where

$$\partial_\varepsilon f(x) = \left\{ x^* \in X^* \ : \ \liminf_{h \to 0} \frac{f(x+h) - f(x) - \langle x^*, h\rangle}{\|h\|} \geq -\varepsilon \right\}.$$

The singular subdifferential of f at c is the set

$$\partial^\infty f(c) = \{\zeta \in \mathbb{R}^n \ : \ (\zeta, 0) \in N(epi\, f; (c, f(c)))\}.$$

Next we consider a multivalued mapping F from $\mathbb{R}^n$ to $\mathbb{R}^m$ of the closed graph

$$\mathrm{Gr}F := \{(x,y) : \ y \in F(x)\}.$$

The multivalued mapping $D^*F(x,y) : \mathbb{R}^m \mapsto \mathbb{R}^n$ defined by

$$D^*F(x,y)(y^*) := \{x^* \in \mathbb{R}^n : \ (x^*, -y^*) \in N(\mathrm{Gr}F; (x,y))\}$$

is called the coderivative of F at the point $(x,y) \in \mathrm{Gr}F$.

The domain over which our study occurs is typically one of the functions $W^{1,1}([a,b], \mathbb{R}^n)$ (abbreviated $W^{1,1}$) consisting of all absolutely continuous functions $x\colon [a,b] \mapsto \mathbb{R}^n$ for which $|\dot{x}|$ is integrable on $[a,b]$ ($\dot{x}$ denotes the derivative (a.e.) of x). An *arc* is a function in $W^{1,1}$. The space $W^{1,1}$ is endowed with the norm

$$\|x\| = |x(a)| + \int_a^b |\dot{x}(t)| dt,$$

where $|\cdot|$ denotes the Euclidean norm of $\mathbb{R}^n$. Here $\mathbb{B}$ stands for the closed unit ball in $\mathbb{R}^n$ and

$$B(z,r) = \{x \in W^{1,1} \colon \ \|x - z\| \leq r\}.$$

The distance function on $W^{1,1}$, $\mathbb{R}^n$ or $\mathbb{R}^n \times \mathbb{R}^n$ will be denoted by $d(\cdot\,,\cdot)$. The convex hull and the closed convex hull are denoted by co and $\bar{\mathrm{co}}$, respectively.

The following lemma is needed.

LEMMA 2.1. *Let G be pseudo-Lipschitzian* [1], [41] *around $(x_0, y_0) \in \mathrm{Gr}G$ with modulus K; i.e., there exists $r > 0$ such that for all $x, u \in x_0 + r\mathbb{B}$*

$$G(x) \cap (y_0 + r\mathbb{B}) \subset G(u) + K|x - u|\mathbb{B}.$$

Then for all $y^* \in \mathbb{R}^n$, with $D^*G(x_0, y_0)(y^*) \neq \emptyset$, one has

$$\sup\{|x^*| : \quad x^* \in D^*G(x_0, y_0)(y^*)\} \leq K|y^*|.$$

If in addition G is closed-valued, then for all $(x, y) \in (x_0 + \frac{r}{12}\mathbb{B}) \times (y_0 + \frac{r}{12}\mathbb{B})$, with $(x, y) \notin \mathrm{Gr}G$, and all $(x^*, y^*) \in \partial d(\cdot\,; G(\cdot))(x, y)$ we have

$$|y^*| = 1 \text{ and } |x^*| \leq K|y^*|.$$

Proof. It suffices to establish the second part; the first one follows from the definition of limiting Fréchet subdifferential. Let $(x, y) \in (x_0 + \frac{r}{12}\mathbb{B}) \times (y_0 + \frac{r}{12}\mathbb{B})$, with $(x, y) \notin \mathrm{Gr}G$, and let $(x^*, y^*) \in \partial d(\cdot; G(\cdot))(x, y)$. Then there are sequences $x_k \to x$, $y_k \to y$, $x_k^* \to x^*$, $y_k^* \to y^*$, $\varepsilon_k \to 0^+$, and $r_k \to 0^+$ such that

$$d(v; G(u)) - d(y_k; G(x_k)) - \langle x_k^*, u - x_k\rangle - \langle y_k^*, v - y_k\rangle + \varepsilon_k[|u - x_k| + |v - y_k|] \geq 0$$

for all $u \in x_k + r_k\mathbb{B}$ and $v \in y_k + r_k\mathbb{B}$. For each integer k, there exists $v_k \in G(x_k)$ such that

$$d(y_k; G(x_k)) = |y_k - v_k|.$$

So

$$|y' - v| - |y_k - v_k| - \langle x_k^*, u - x_k\rangle - \langle y_k^*, v - y_k\rangle + \varepsilon_k[|u - x_k| + |v - y_k|] \geq 0$$

for all $u \in x_k + r_k\mathbb{B}$, $v \in y_k + r_k\mathbb{B}$, and $y' \in G(u)$.

Consider the function g defined by

$$g(u, y', v) = |y' - v| - \langle x_k^*, u - x_k\rangle - \langle y_k^*, v - y_k\rangle + \varepsilon_k[|u - x_k| + |v - y_k|].$$

Then [34]

$$(0, 0, 0) \in \partial g(x_k, v_k, y_k) + N(\mathrm{Gr}G; (x_k, v_k)) \times \{0\}.$$

As for k large enough $y_k \neq v_k$, then

$$\partial g(x_k, v_k, y_k) \subset \{(0, v^*, -v^*) \colon |v^*| = 1\} + (-x_k^*, 0, -y_k^*) + \varepsilon_k\mathbb{B} \times \{0\} \times \varepsilon_k\mathbb{B},$$

and hence we obtain $(u_k^*, v_k^*) \in N(\mathrm{Gr}G; (x_k, v_k))$, with $|v_k^*| = 1$, such that

$$|x_k^* - u_k^*| \leq \varepsilon_k \text{ and } |y_k^* - v_k^*| \leq \varepsilon_k.$$

Now since $d(y_k; G(x_k))) = |y_k - v_k|$, we get for k sufficiently large

$$|y_k - v_k| \leq \frac{r}{2},$$

and hence

$$|x_0 - x_k| + |y_0 - v_k| \leq \frac{5r}{6}.$$

Thus for all $u, u' \in x_k + \frac{r}{6}\mathbb{B}$

$$G(u) \cap \left(v_k + \frac{r}{6}\mathbb{B}\right) \subset G(u') + K|u - u'|\mathbb{B}.$$

So the first part of the lemma ensures that

$$|u_k^*| \leq K|v_k^*|,$$

and since $u_k^* \to x^*$ and $v_k^* \to y^*$ we get $|x^*| \leq K|y^*|$, and the proof is complete.

LEMMA 2.2. *Let $G\colon V \mapsto \mathbb{R}^m$ be a multivalued mapping, where V is a nonempty set in $\mathbb{R}^n$. Suppose that*

(i) GrG *is closed and*

(ii) *there exists a compact set K in $\mathbb{R}^m$ such that*

$$G(x) \subset K \quad \forall x \in V.$$

Then G is upper semicontinuous (u.s.c.) on V; that is, for all $u \in V$ and all $\varepsilon > 0$ there exists a neighborhood U of u in V such that

$$G(x) \subset G(u) + \varepsilon\mathbb{B} \quad \forall x \in U.$$

With the help of the last lemma, we can prove the following one.

LEMMA 2.3. *Suppose that the mapping $f\colon (x_0, y_0) + r\mathbb{B} \mapsto \mathbb{R}$ is Lipschitzian with constant K. Define the multivalued mapping $\Gamma\colon (x_0, y_0) + r\mathbb{B} \times \mathbb{R}^n \times \mathbb{R} \mapsto \mathbb{R}^n$ by*

$$\Gamma(x, y, p, s) = co\{q\colon (q, p) \in \partial f(x, y) + s\mathbb{B}\}.$$

Then for all $\lambda \in]0, 1[$, all $(x, y, s) \in (x_0, y_0, 0) + \lambda r\mathbb{B}$, and all $p \in \mathbb{R}^n$, with $\Gamma(x, y, p, s) \neq \emptyset$, Γ is u.s.c. at (x, y, p, s) in the sense of Lemma 2.2.

Proof. Note that (ii) of Lemma 2.2 is satisfied. It is not difficult to show that Γ is of closed graph and to apply Lemma 2.2.

LEMMA 2.4 (see [11]). *Let $\varepsilon > 0$ and $\Gamma : [a, b] \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \mapsto \mathbb{R}^n$ be a multivalued mapping such that for almost all $t \in [a, b]$, $\Gamma(t, \cdot)$ has nonempty, compact, and convex values around $(z(t), \dot{z}(t), p, s)$, with $s \in [0, \varepsilon]$ and $\Gamma(t, z(t), \dot{z}(t), p, s) \neq \emptyset$. For sequences (z_k) and (p_k) in $W^{1,1}$, (ϕ_k) in $L^1([a, b],]0, +\infty[)$, (α_k) and (s_k) in $\mathbb{R}_+$ with $z_k \to z$ in $W^{1,1}$, $\phi_k \to \phi$ in $L^1([a, b],]0, +\infty[)$ for some integrable function ϕ, $\alpha_k \to 0$, and $s_k \to 0$ we suppose the following:*

(i) *For every (x, y, p, s) in the interior of the set*

$$\{(x', y', p', s') : t \in [a, b],\ x' \in z(t) + \varepsilon\mathbb{B},\ y' \in \dot{z}(t) + \varepsilon\mathbb{B},\ s' \in [0, \varepsilon],\ \Gamma(t, x', y', p', s') \neq \emptyset\}$$

the multivalued mapping $t' \mapsto \Gamma(t', x, y, p, s)$ is measurable.

(ii) *For all k, $|\dot{p}_k(t)| \leq \phi_k(t)$ for almost all $t \in [a, b]$.*

(iii) *For all k, $\dot{p}_k(t) \in \Gamma(t, z_k(t), \dot{z}_k(t), p_k(t), s_k) + \alpha_k\mathbb{B}$ a.e. $t \in [a, b]$.*

(iv) *For almost all $t \in [a, b]$ for every $p \in \mathbb{R}^n$ with $\Gamma(t, z(t), \dot{z}(t), p, 0) \neq \emptyset$, the multivalued mapping $(x', y', p', s') \mapsto \Gamma(t, x', y', p', s')$ is u.s.c. at $(z(t), \dot{z}(t), p, 0)$.*

(v) *The sequence $(p_k(a))$ is bounded.*

(vi) *There exists an integrable function ψ such that*

$$\sup_{\{(p', s') : s' \in [0, \varepsilon],\ \Gamma(t, z(t), \dot{z}(t), p', s') \neq \emptyset\}} \max_{y \in \Gamma(t, z(t), \dot{z}(t), p', s')} |y| \leq \psi(t) \text{ a.e.}$$

Then there is a subsequence of (p_k) *which converges uniformly to an arc* p *satisfying*

$$\dot{p}(t) \in \Gamma(t, z(t), \dot{z}(t), p(t), 0) \quad \text{a.e. } t \in [a, b].$$

We conclude this section by recalling necessary optimality conditions for the following generalized problem of Bolza:

$$(\mathrm{P_B}) \qquad \min \left\{ \ell(x(a), x(b)) + \int_a^b L(t, x(t), \dot{x}(t))\, dt \right\},$$

where the functions $L : [a, b] \times \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ and $\ell : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ are such that for each $t \in [a, b]$, the functions $L(t, \cdot, \cdot)$ and ℓ are l.s.c. on $\mathbb{R}^n \times \mathbb{R}^n$.

The function L is *epi-Lipschitz* [10] at an arc z if there exist an integrable function $k : [a, b] \mapsto \mathbb{R}$ and a positive ε satisfying the following conditions: for almost all $t \in [a, b]$, given two points z_1 and z_2 within ε of $z(t)$ and $u_1 \in \mathbb{R}^n$ such that $L(t, z_1, u_1)$ is finite, there exist a point $u_2 \in \mathbb{R}^n$ and $\delta \geq 0$ such that $L(t, z_2, u_2)$ is finite and

$$|u_1 - u_2| + |L(t, z_1, u_1) - L(t, z_2, u_2) - \delta| \leq k(t)|z_1 - z_2|.$$

This is equivalent to saying that the multivalued mapping

$$E(t, s) = \{(u, r) \in \mathbb{R}^n \times \mathbb{R} : L(t, s, u) \leq r\}$$

is Lipschitzian in s on $z(t) + \varepsilon\mathbb{B}$ with constant $k(t)$ (i.e., for all $s, s' \in z(t) + \varepsilon\mathbb{B}$ we have $E(t, s') \subset E(t, s) + k(t) \mid s' - s \mid \mathbb{B}$).

L is said to be epimeasurable (in t) [10] if for each $s \in \mathbb{R}^n$ the multivalued mapping $E(t, s)$ is Lebesgue measurable in t.

The notation ∂L will denote the limiting Fréchet subdifferential of the function $L(t, \cdot, \cdot)$.

Now we may state a variant of the necessary conditions for the generalized Bolza problem established in Jourani [26].

THEOREM 2.1. *Let* z *solve locally the generalized problem of Bolza* $(\mathrm{P_B})$ *(in* $W^{1,1}$*). Suppose that* $L(t, z, u)$ *is epimeasurable in* t*, and* $L(t, \cdot, \cdot)$ *is epi-Lipschitzian at* z*, and* ℓ *is locally Lipschitzian around* $(z(a), z(b))$*. Then there exists an arc* p *such that one has*

$$\dot{p}(t) \in co\{q : \quad (q, p(t)) \in \partial L(t, z(t), \dot{z}(t))\} \quad \text{a.e. } t \in [a, b],$$

$$(p(a), -p(b)) \in \partial \ell(z(a), z(b)),$$

$$\langle p(t), \dot{z}(t) \rangle - L(t, z(t), \dot{z}(t)) = \max\{\langle p(t), v \rangle - L(t, z(t), v) : v \in \mathbb{R}^n\}.$$

3. The main result.

DEFINITION 3.1. *F is said to be sub-Lipschitzian in the sense of Loewen and Rockafellar* [29] *at z if there exist $\beta > 0$, $\varepsilon > 0$, and a summable function $k : [a, b] \mapsto \mathbb{R}$ such that for almost all $t \in [a, b]$, for all $N > 0$, for all $x, x' \in z(t) + \varepsilon\mathbb{B}$, and $y \in \dot{z}(t) + N\mathbb{B}$ one has*

$$d(y, F(t, x)) - d(y, F(t, x')) \leq (k(t) + \beta N)|x - x'|.$$

Let $\prec$ be a (nonreflexive) preference for vectors in $\mathbb{R}^m$. We consider the following multiobjective optimization problem:

$$\text{(P)} \qquad \begin{gathered} \min f(x(a), x(b)), \\ (x(a), x(b)) \in S, \\ \dot{x}(t) \in F(t, x(t)) \quad \text{a.e.} \quad t \in [a, b], \end{gathered}$$

where $f\colon \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}^m$ is a mapping, $S \subset \mathbb{R}^n \times \mathbb{R}^n$ is a closed nonempty set, and $F\colon [a, b] \times \mathbb{R}^n \mapsto \mathbb{R}^n$ is a closed-valued multivalued mapping which is measurable in $t \in [a, b]$.

We say that an arc $x \in W^{1,1}$ is a feasible trajectory for problem (P) if x satisfies $(x(a), x(b)) \in S$ and $\dot{x}(t) \in F(t, x(t))$ a.e. $t \in [a, b]$.

z is a solution to (P), provided that it is feasible and there does not exist any feasible trajectory x of (P) such that $f(x(a), x(b)) \prec f(z(a), z(b))$. For all $r \in \mathbb{R}^m$, we denote

$$\mathcal{L}(r) := \{s \in \mathbb{R}^m \,:\, s \prec r\}.$$

We will need the following regularity assumptions on the preference modified from [49].

Definition 3.2. *We say that a preference $\prec$ is regular at $r \in \mathbb{R}^m$, provided that*
(A_1) for any $s \in \mathbb{R}^m$, $s \in \mathrm{cl}\mathcal{L}(s)$;
(A_2) for any $r \prec s$, $t \in \mathrm{cl}\mathcal{L}(r)$ implies that $t \prec s$.

Remark 3.1. The preference determined by the lexicographical order $\prec$ is defined by $r \prec s$ if there exists an integer $q \in \{0, 1, \dots, m-1\}$ such that $r_i = s_i,\ i = 1, \dots, q$, and $r_{q+1} < s_{q+1}$. This preference is not regular. Indeed we consider in $\mathbb{R}^3$ the vectors $r = (1, 1, 3)$, $s = (1, 1, 5)$, and $t = (1, 1, 6)$. We have $r \prec s$ and $t \in \mathrm{cl}\mathcal{L}(r)$, but $s \prec t$; then (A_2) does not hold so that $\prec$ is not regular at r.

Note that a preference determined by the lexicographical order does not correspond to any real utility function [16].

Remark 3.2. Our definition of regularity is different from that given by Zhu in [49], where the following third condition is in force: for any sequences $r_k, \theta_k \mapsto r$ in $\mathbb{R}^m$

$$\limsup_{k \to +\infty} N(\mathrm{cl}\mathcal{L}(r_k); \theta_k) \subset N(\mathrm{cl}\mathcal{L}(r); r).$$

But with this condition, preferences defined by a utility function (e.g., u) are not regular at any $r \in \mathbb{R}^m$ even if

$$\lim_{s \to r} d(0, \partial u(s)) > 0.$$

For more details, see Example 3.

We consider the following enlargement cone of the limiting Fréchet normal cone:

$$\tilde{N}(\mathrm{cl}\mathcal{L}(x), x) = \limsup_{y, x' \to x} N(\mathrm{cl}\mathcal{L}(y); x').$$

Before stating our main result we recall that the Hamiltonian associated with F is defined by

$$H(t, x, y) = \sup_{v \in F(t,x)} \langle y, v \rangle.$$

THEOREM 3.1. *Let z be a local solution to the multiobjective optimal control problem* (P). *Suppose that F is sub-Lipschitzian at z and that the preference $\prec$ is regular at $f(z(a), z(b))$. Then there exist $p \in W^{1,1}$, $\lambda \geq 0$, and $w \in \tilde{N}(\mathrm{cl}\mathcal{L}(f(z(a), z(b))), f(z(a), z(b)))$, with $|\omega| = 1$ such that $(\lambda, p) \neq 0$ and*

$$\dot{p}(t) \in coD^*F(t, z(t), \dot{z}(t))(-p(t)) \ \text{a.e. } t \in [a, b]; \tag{2}$$

$$(p(a), -p(b)) \in \lambda\partial(\langle\omega, f(\cdot,\cdot)\rangle)(z(a), z(b)) + N(S; (z(a), z(b))); \tag{3}$$

$$\langle p(t), \dot{z}(t)\rangle = H(t, z(t), p(t)) \ \text{a.e. } t \in [a, b]. \tag{4}$$

If in addition F is convex-valued, then (2) *may be replaced by the following one:*

$$\dot{p}(t) \in co\ \{q :\ (-q, \dot{z}(t)) \in \partial H(t, (z(t), p(t)))\} \quad \text{a.e. } t \in [a, b]. \tag{5}$$

The aim of Theorem 3.1 is to extend the necessary optimality conditions of Ioffe (Theorem 1 in [23]) from a single objective optimal control of differential inclusion problem to a multiobjective one. By using the large class of sub-Lipschitz differential inclusion, Theorem 3.1 also extends the Hamiltonian necessary optimality conditions for convex-valued differential inclusions obtained in [49].

In the remainder of this section we now examine a few examples. The proof of Theorem 3.1 is postponed to the next section.

Example 1 (a generalized Pareto optimal). Let K be a pointed convex cone ($K \cap (-K) = \{0\}$). We define the preference $\prec$ by $r \prec s$ if and only if $r - s \in K$ and $r \neq s$. A multiobjective optimal control problem with this preference is called a generalized Pareto optimal control problem. Notice that if $K = \mathbb{R}^m_-$ (resp., $K = \operatorname{int}\mathbb{R}^m_-$, where $\mathbb{R}^m_- = \{(x_1, x_2, \dots, x_m) \in \mathbb{R}^m : x_i \leq 0 \text{ for all } i = 1, \dots, m\}$) we get Pareto (resp., weak Pareto) optimal control problems. This preference is regular at any $r \in \mathbb{R}^m$. Moreover, for any $r \in \mathbb{R}^m$ we have $\tilde{N}(\mathrm{cl}\mathcal{L}(r), r) = K^0$ with $K^0 = \{s \in \mathbb{R}^m : \langle s, q\rangle \leq 0 \ \text{ for all } \ q \in K\}$.

COROLLARY 3.1. *Let z be a local solution to the generalized Pareto multiobjective optimal control problem* (P). *Then there exist $p \in W^{1,1}$, $\lambda \geq 0$, and $\omega \in K^0$ with $|\ \omega\ | = 1$ such that $(\lambda, p) \neq 0$ and*

$$\dot{p}(t) \in coD^*F(t, z(t), \dot{z}(t))(-p(t)) \ \text{a.e. } t \in [a, b]; \tag{6}$$

$$(p(a), -p(b)) \in \lambda\partial(\langle\omega, f(\cdot,\cdot)\rangle)(z(a), z(b)) + N(S; (z(a), z(b))); \tag{7}$$

$$\langle p(t), \dot{z}(t)\rangle = H(t, z(t), p(t)) \ \text{a.e. } t \in [a, b]. \tag{8}$$

Example 2 (a preference determined by a utility function). Let u be a continuous function; we define the preference $\prec$ determined by utility function u by $r \prec s$ if and only if $u(r) < u(s)$.

LEMMA 3.1. *Let u be a continuous utility function determining the preference $\prec$. Suppose that*

$$\liminf_{s \to r} d(0, \partial u(s)) > 0. \tag{9}$$

Then the preference $\prec$ is regular at r and

$$\tilde{N}(\mathrm{cl}\mathcal{L}(r), r) = \limsup_{r' \to r} N(\mathrm{cl}\mathcal{L}(r'); r') = \partial^{\infty} u(r) \bigcup \Big(\bigcup_{a>0} a \partial u(r) \Big).$$

Proof. The proof of Lemma 3.1 is similar to that given in [49]. From (9), $\mathcal{L}(r)$ is nonempty, and from the continuity of u it follows that $\prec$ satisfies (A_1) and (A_2) in Definition 3.2, and thus $\prec$ is regular. Now for $r^{'}$ sufficiently close to r, $\mathrm{cl}\mathcal{L}(r^{'}) = \{s \in \mathbb{R}^m : u(s) - u(r^{'}) \leq 0\}$. Then

$$\partial_{\varepsilon} u(r^{'}) \subset \hat{N}_{\varepsilon}(\mathrm{cl}\mathcal{L}(r^{'}), r^{'}).$$

By passing to the limits we have

$$\partial^{\infty} u(r) \bigcup \Big(\bigcup_{a>0} a \partial u(r) \Big) \subset \limsup_{r' \to r} N(\mathrm{cl}\mathcal{L}(r'); r') \subset \tilde{N}(\mathrm{cl}\mathcal{L}(r), r).$$

Conversely, let $\zeta \in \tilde{N}(\mathrm{cl}\mathcal{L}(r), r)$ such that $\zeta \neq 0$. Then there are sequences $\zeta_k \to \zeta$, $r_k, r_k^{'} \to r$ such that $\zeta_k \in N(\mathrm{cl}\mathcal{L}(r_k); r_k^{'})$. By the definition of limiting Fréchet normal cone, we may assume that $\zeta_k \in \hat{N}_{\varepsilon_k}(\mathrm{cl}\mathcal{L}(r_k), r_k^{'})$. We must have $u(r_k) = u(r_k^{'})$. Indeed, $\hat{N}_{\varepsilon_k}(\mathrm{cl}\mathcal{L}(r_k), r_k^{'}) = \{0\}$ when $u(r_k^{'}) < u(r_k)$ and is empty when $u(r_k^{'}) > u(r_k)$. Then $\hat{N}_{\varepsilon_k}(\mathrm{cl}\mathcal{L}(r_k), r_k^{'}) = \hat{N}_{\varepsilon_k}(\mathrm{cl}\mathcal{L}(r_k), r_k)$. From $\hat{N}_{\varepsilon_k}(\mathrm{cl}\mathcal{L}(r_k), r_k) = \hat{N}_{\varepsilon_k}(\{s : u(s) - u(r_k) \leq 0\}, r_k)$ and [8], there exist $a_k > 0$ and $\theta_k \in \partial_{\varepsilon_k} u(r)$ such that $\mid a_k \theta_k - \zeta_k \mid < \frac{1}{k}$ so that

$$\lim_{k \to \infty} a_k \theta_k = \zeta.$$

We claim that (a_k) is bounded. Indeed, suppose the contrary. Then (a_k) has a subsequence going to infinity. But in this case (θ_k) must have a subsequence converging to zero, and this contradicts (9). So (a_k) is bounded, and we can assume that $a_k \to a$. If $a \neq 0$, then $\zeta \in a \partial u(r)$. If $a = 0$, then $\zeta \in \partial^{\infty} u(r)$, and the proof is complete.

From Lemma 3.1 and Theorem 3.1 we have the following corollary.

COROLLARY 3.2. *Let $\prec$ be a preference determined by a utility function u and z be a local solution to the multiobjective optimal control problem* (P). *Suppose that*

$$\liminf_{s \to f(z(a), z(b))} d(0, \partial u(s)) > 0.$$

Then there exist $p \in W^{1,1}$, $\lambda \geq 0$, and

$$\omega \in \partial^{\infty} u(f(z(a), z(b))) \bigcup \Big(\bigcup_{a>0} a \partial u(f(z(a), z(b))) \Big)$$

with $\mid \omega \mid = 1$ such that $(\lambda, p) \neq 0$, and

$$\dot{p}(t) \in co D^* F(t, z(t), \dot{z}(t))(-p(t)) \ \ a.e.\ t \in [a, b]; \tag{10}$$

$$(p(a), -p(b)) \in \lambda \partial(\langle \omega, f(\cdot, \cdot) \rangle)(z(a), z(b)) + N(S; (z(a), z(b))); \tag{11}$$

$$\langle p(t), \dot{z}(t) \rangle = H(t, z(t), p(t)) \ \ a.e.\ t \in [a, b]. \tag{12}$$

In [49], the author showed that, for a preference $\prec$ defined by a continuous utility function u, $N(\mathrm{cl}\mathcal{L}(r);r) = \partial^\infty u(r) \bigcup \big(\bigcup_{a>0} a\partial u(r)\big)$, provided that $\lim_{s\to r} d(0,\partial u(s)) > 0$. This could give him the regularity and the explicit shape of $N(\mathrm{cl}\mathcal{L}(r);r)$. But there is a gap in the proof. The following example shows that Zhu's regularity does not hold.

Example 3. Consider the function $u : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$u(x,y) = \mid x \mid - \mid y \mid .$$

Then u is Lipschitz continuous and satisfies $\partial u(0,0) = [-1,1] \times \{-1,1\}$, so that $(0,0) \notin \partial u(0,0)$, $\partial^\infty u(0,0) = \{(0,0)\}$, and

$$N(\mathrm{cl}\mathcal{L}(0,0);(0,0)) = \{(x,y) \in \mathbb{R}^2 : |y| = |x|\}.$$

Then it is clear that

$$N(\mathrm{cl}\mathcal{L}(0,0);(0,0)) \neq \partial^\infty u(0,0) \bigcup \Big(\bigcup_{a>0} a\partial u(0,0) \Big).$$

4. Proof of Theorem 3.1. Since F is sub-Lipschitzian at z there exist $\beta > 0$, $\varepsilon > 0$, and a summable function $k : [a,b] \mapsto \mathbb{R}$ such that for almost all $t \in [a,b]$, for all $N > 0$, for all $x, x' \in z(t) + \varepsilon\mathbb{B}$, and $y \in \dot{z}(t) + N\mathbb{B}$ one has

$$d(y, F(t,x)) - d(y, F(t,x')) \leq (k(t) + \beta N)|x - x'|.$$

Let G be the solution set of the system

$$\dot{x}(t) \in F(t,x(t)) \, a.e., \ (x(a),x(b)) \in S. \tag{13}$$

Let ε be as above. We say that the system (13) is seminormal [25] at z if there exist $\alpha > 0$ and $r > 0$ such that for all $x \in B(z,r)$

$$d(x, G \cap B(z,\varepsilon)) \leq \alpha \left\{ d((x(a),x(b));S) + \int_a^b d(\dot{x}(t); F(t,x(t)))dt \right\}. \tag{14}$$

Set $G_\varepsilon = G \cap B(z,\varepsilon)$.

We divide the proof into two parts and each part is divided into two steps.

Part 1 (when system (13) is not seminormal at z). The proof of this part is similar to that given in [23].

Step 1 (*application of Ekeland's variational principle* [18] *and Theorem* 2.1). Consider the function h defined by

$$h(x) = d((x(a),x(b));S) + \int_a^b d(\dot{x}(t); F(t,x(t)))dt.$$

Since F is sub-Lipschitzian at z, then h is l.s.c. on the set $B(z,\varepsilon)$ and G_ε is closed (see the appendix). If system (13) is not seminormal at z, then there is a sequence $x_k \to z$ in $W^{1,1}$ such that for k large enough

$$d(x_k, G_\varepsilon) > kh(x_k).$$

Set $\varepsilon_k = \sqrt{h(x_k)} > 0$, $\lambda_k = \min(\varepsilon_k, k\varepsilon_k^2)$, and $s_k = \frac{\varepsilon_k^2}{\lambda_k}$. Then $\varepsilon_k \to 0^+$ and $s_k \to 0^+$. Therefore one has

$$h(x_k) \leq \inf_{x \in B(z,\varepsilon)} h(x) + \varepsilon_k^2.$$

By Ekeland variational principle we get $z_k \in B(z,\varepsilon)$ satisfying

$$\|z_k - x_k\| < \lambda_k, \tag{15}$$

$$h(z_k) \le h(x) + s_k\|x - z_k\| \quad \forall x \in B(z,\varepsilon). \tag{16}$$

Observe that for k sufficiently large $\| z_k - z \| \le \frac{\varepsilon}{2}$. By the closedness of G_ε and relation (15) $z_k \notin G$, and by (16) z_k is a local solution to the following Bolza problem:

$$\min\left\{\ell_k(x(a),x(b)) + \int_a^b L_k(t,x(t),\dot{x}(t))dt\right\},$$

where

$$\ell_k(u,v) = d((u,v);S) + s_k|u - z_k(a)|$$

and

$$L_k(t,x,y) = \begin{cases} d(y;F(t,x)) + s_k|y - \dot{z}_k(t)| & \text{if } (x,y) \in A(t), \\ +\infty & \text{otherwise,} \end{cases}$$

where $A(t) = (z(t) + \varepsilon\mathbb{B}) \times (\dot{z}(t) + (N + |\dot{z}(t) - \dot{z}_k(t)|)\mathbb{B})$ and $N > 0$ is an arbitrary integer.

Since $L_k(t,\cdot,\cdot)$ is l.s.c, epi-Lipschitzian at z_k (see the appendix) and epimeasurable in t and since ℓ_k is locally Lipschitzian around $(z_k(a), z_k(b))$, then Theorem 2.1 yields the existence of an arc p_k in $W^{1,1}$ satisfying

$$\dot{p}_k(t) \in \operatorname{co}\{q : \quad (q, p_k(t)) \in \partial L_k(t, z_k(t), \dot{z}_k(t))\} \quad a.e. \quad t \in [a,b] \tag{17}$$

$$(p_k(a), -p_k(b)) \in \partial\ell_k(z_k(a), z_k(b)), \tag{18}$$

$$\langle p_k(t), \dot{z}_k(t)\rangle - L_k(t, z_k(t), \dot{z}_k(t)) = \max_{v\in\mathbb{R}^n}\{\langle p_k(t), v\rangle - L_k(t, z_k(t), v)\}. \tag{19}$$

From (17), (18), and (19) we have

$$(p_k(a), -p_k(b)) \in \partial d((z_k(a), z_k(b)); S) + s_k\mathbb{B} \times \{0\}, \tag{20}$$

$$\dot{p}_k(t) \in \text{ co } \{q : (q, p_k(t)) \in \partial d(\cdot; F(t,\cdot))(z_k(t), \dot{z}_k(t)) + \{0\} \times s_k\mathbb{B}\} \text{ a.e.}, \tag{21}$$

$$\begin{aligned} &\langle p_k(t), \dot{z}_k(t)\rangle - d(\dot{z}_k(t); F(t, z_k(t))) \\ &\quad = \max_{v\in\dot{z}(t)+(N+|\dot{z}(t)-\dot{z}_k(t)|)\mathbb{B}} \{\langle p_k(t), v\rangle - d(v; F(t, z_k(t))) - s_k|v - \dot{z}_k(t)|\} \text{ a.e.} \end{aligned}$$

Step 2 (*application of Lemmas* 2.1–2.4). By (20) there exists $\zeta_k \in \partial d((z_k(a), z_k(b)); S)$ such that

$$(p_k(a), -p_k(b)) - \zeta_k \in s_k\mathbb{B} \times \{0\}. \tag{22}$$

Since $z_k \notin G$, we have either

$$|\zeta_k| = 1 \text{ if } (z_k(a), z_k(b)) \notin S \tag{23}$$

or (because of Lemma 2.1 and (21)) on a set of positive measure on which $\dot{z}_k(t) \notin F(t, z_k(t))$ we have

$$1 - s_k \leq |p_k(t)| \leq 1 + s_k. \tag{24}$$

It follows from (22)–(24) that

$$\frac{1}{\sqrt{2}} - s_k \leq \max_{t\in[a,b]} |p_k(t)| \leq 1 + s_k. \tag{25}$$

Now let $\Gamma : [a, b] \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}_+ \mapsto \mathbb{R}^n$ be the multivalued mapping defined by

$$\Gamma(t, x, y, w, s) = \text{ co } \{q : (q, w) \in \partial d(\cdot; F(t, \cdot))(x, y) + \{0\} \times s\mathbb{B}\}.$$

Then

$$(p_k(a), -p_k(b)) - \zeta_k \in s_k\mathbb{B} \times \{0\}, \tag{26}$$

$$\dot{p}_k(t) \in \Gamma(t, z_k(t), \dot{z}_k(t), p_k(t), s_k) \quad \text{a.e.}, \tag{27}$$

$$\begin{aligned} &\langle p_k(t), \dot{z}_k(t)\rangle - d(\dot{z}_k(t); F(t, z_k(t))) \\ &= \max_{v\in\dot{z}(t)+(N+|\dot{z}(t)-\dot{z}_k(t)|)\mathbb{B}} \{\langle p_k(t), v\rangle - d(v; F(t, z_k(t))) - s_k|v - \dot{z}_k(t)|\} \quad \text{a.e.} \end{aligned} \tag{28}$$

Extracting a subsequence if necessary we may suppose that $\zeta_k \to \zeta$ for some ζ in $\partial d((z(a), z(b)); S)$ with

$$|\zeta| = 1 \text{ if } (z_k(a), z_k(b)) \notin S \text{ for infinite number of } k.$$

On the other hand, by Lemma 2.2, the multivalued mapping $\Gamma(t, \cdot)$ is u.s.c. with compact convex values and by the definition of the limiting Fréchet subdifferential and the sub-Lipschitz condition we have (via Lemma 2.1 and (21)) for all k

$$|\dot{p}_k(t)| \leq 1 + k(t) + \beta(1 + |\dot{z}(t) - \dot{z}_k(t)|) \quad \text{a.e.}$$

Note that $\Gamma(t, x, y, w, s)$ is measurable in t (see the appendix). By Lemma 2.4 there exists a subsequence of (p_k) converging uniformly to an arc p satisfying

$$\dot{p}(t) \in \Gamma(t, z(t), \dot{z}(t), p(t), 0) \quad \text{a.e.}, \tag{29}$$

and hence we obtain, by passing to the limit in (26) and (28),

$$(p(a), -p(b)) \in \partial d((z(a), z(b)); S), \tag{30}$$

$$\langle p(t), \dot{z}(t)\rangle = \max_{v\in F(t,z(t))\cap(\dot{z}(t)+N\mathbb{B})} \langle p(t), v\rangle \text{ a.e.} \tag{31}$$

Now because of (25) the pair (ζ, p) must be nonzero. In fact we have

$$\frac{1}{\sqrt{2}} \leq \max_{t\in[a,b]} |p(t)| \leq 1. \tag{32}$$

As p depends on N, we obtain a sequence (p_N) satisfying (29)–(32) and

$$|\dot{p}_N(t)| \leq 1 + k(t) + \beta \quad \text{a.e.}$$

Again Lemma 2.4 produces a subsequence of (p_N) converging uniformly to some p which satisfies the following:

$$\dot{p}(t) \in \Gamma(t, z(t), \dot{z}(t), p(t), 0) \quad \text{a.e.},$$

$$(p(a), -p(b)) \in \partial d((z(a), z(b)); S),$$

$$\langle p(t), \dot{z}(t)\rangle = \max_{v \in F(t,z(t))} \langle p(t), v\rangle,$$

$$\frac{1}{\sqrt{2}} \leq \max_{t\in[a,b]} |p(t)| \leq 1.$$

Finally we have

$$\dot{p}(t) \in \mathrm{co}D^*F(t, z(t), \dot{z}(t))(-p(t)) \ \ a.e.\ \ t \in [a, b],$$

$$(p(a), -p(b)) \in N(S; (z(a), z(b))),$$

$$\langle p(t), \dot{z}(t)\rangle = H(t, z(t), p(t)) \ \ a.e.\ \ t \in [a, b].$$

Part 2 (when system (13) is seminormal).

Step 1 (*application of Ekeland's variational principle*). Let k be a positive integer, choose $\theta_k \prec f(z(a), z(b))$ such that $|\theta_k - f(z(a), z(b))| < \frac{1}{k^2}$, and define $\Theta := \mathrm{cl}\mathcal{L}(\theta_k)$. Define the function

$$h(x, \theta) = \begin{cases} |f(x(a), x(b)) - \theta| & \text{if } x \in B(z, s_1), \\ +\infty & \text{otherwise,} \end{cases}$$

where s_1 is such that f is Lipschitzian on $(z(a), z(b)) + s_1\mathbb{B}$ with constant k_f. From (A_1) we have $(z, \theta_k) \in G_\varepsilon \times \Theta$, and hence

$$h(z, \theta_k) \leq \inf_{(x,\theta)\in G_\varepsilon\times\Theta} h(x, \theta) + \frac{1}{k^2}.$$

Note that G_ε and Θ are closed in $W^{1,1}$ and $\mathbb{R}^m$, respectively, and that h is l.s.c. on $G_\varepsilon \times \Theta$. Then by Ekeland variational principle there exists $(z_k, \gamma_k) \in G_\varepsilon \times \Theta$ such that

$$\|z_k - z\| + |\gamma_k - \theta_k| \leq \frac{1}{k} \tag{33}$$

and

$$h(z_k, \gamma_k) \leq h(x, \theta) + \frac{1}{k}\left[\|z_k - x\| + |\gamma_k - \theta|\right] \forall (x, \theta) \in G_\varepsilon \times \Theta. \tag{34}$$

From (34) one gets

$$h(z_k, \gamma_k) \leq h(x, \gamma_k) + \frac{1}{k}\|z_k - x\| \forall x \in G_\varepsilon \tag{35}$$

and

$$h(z_k,\gamma_k) \le h(z_k,\theta) + \frac{1}{k}|\gamma_k - \theta| \,\forall \theta \in \Theta. \tag{36}$$

Since z is an optimal local solution to problem (P), then, by (A_2) and the choice of θ_k, one has $\gamma_k \neq f(z_k(a), z_k(b))$. Set $w_k = \frac{f(z_k(a),z_k(b))-\gamma_k}{|f(z_k(a),z_k(b))-\gamma_k|}$. Extracting a subsequence we may assume that (w_k) converges to some w, with $|w| = 1$ so that by (36) one has

$$w \in \limsup_{k\to+\infty} N(\mathrm{cl}\mathcal{L}(\theta_k);\gamma_k)$$

and then

$$\omega \in \tilde{N}(\mathrm{cl}\mathcal{L}(f(z(a),z(b))), f(z(a),z(b))).$$

Now from (35) and the seminormality of (13) there exist $\alpha > 0$ and $\min(s_1, r, \varepsilon) > s > 0$ (both not depending on k) such that

$$h(z_k,\gamma_k) \le h(x,\gamma_k) + \frac{1}{k}\|z_k - x\| + \alpha(k_f+1)\left[d((x(a),x(b));S) + \int_a^b d(\dot{x}(t), F(t,x(t))dt\right]$$

for all $x \in B(z,s)$, where r and α are as in (14).

Define the functions

$$\ell_k(u,v) = |f(u,v) - \gamma_k| + \frac{1}{k}|u - z_k(a)| + \alpha(k_f+1)d((u,v);S)$$

and

$$L_k(t,x,y) = \begin{cases} \alpha(k_f+1)d(y;F(t,x)) + \frac{1}{k}|y - \dot{z}_k(t)| & \text{if } (x,y) \in A(t), \\ +\infty & \text{otherwise,} \end{cases}$$

where $A(t) = (z(t) + s\mathbb{B}) \times (\dot{z}(t) + (N + |\dot{z}(t) - \dot{z}_k(t)|)\mathbb{B})$ so that z_k is a local solution to the Bolza problem

$$\min\left\{\ell_k(x(a),x(b)) + \int_a^b L_k(t,x(t),\dot{x}(t))dt\right\}.$$

Step 2 (*application of Theorem* 2.1 *and Lemmas* 2.1–2.4). It is easy to check that ℓ_k is l.s.c and locally Lipschitzian around $(z_k(a), z_k(b))$, $L_k(t,\cdot,\cdot)$ is l.s.c, and L_k is epimeasurable in t and epi-Lipschitzian at z_k (see the appendix). Then by Theorem 2.1 there exists an arc p_k in $W^{1,1}$ satisfying

$$\dot{p}_k(t) \in \mathrm{co}\{q : \quad (q, p_k(t)) \in \partial L_k(t, z_k(t), \dot{z}_k(t))\} \quad a.e.\ t \in [a,b], \tag{37}$$

$$(p_k(a), -p_k(b)) \in \partial \ell_k(z_k(a), z_k(b)), \tag{38}$$

$$\langle p_k(t), \dot{z}_k(t)\rangle - L_k(t, z_k(t), \dot{z}_k(t)) = \max_{v\in\mathbb{R}^n}\{\langle p_k(t), v\rangle - L_k(t, z_k(t), v)\}. \tag{39}$$

Consider the multivalued mapping defined by

$$\Gamma(t,x,y,w,s) = \text{ co } \{q : (q,w) \in \alpha(k_f+1)\partial d(\cdot; F(t,\cdot))(x,y) + \{0\} \times s\mathbb{B}\}.$$

From (37)–(39) we have

$$(40)\quad (p_k(a), -p_k(b)) \in \partial(|f(\cdot) - \gamma_k|)(z_k(a), z_k(b)) + N(S; (z_k(a), z_k(b))) + \frac{1}{k}\mathbb{B} \times \{0\},$$

$$(41)\qquad \dot{p}_k(t) \in \Gamma\left(t, z_k(t), \dot{z}_k(t), p_k(t), \frac{1}{k}\right) \quad \text{a.e.},$$

$$(42)\quad \langle p_k(t), \dot{z}_k(t)\rangle - \alpha(k_f+1)d(\dot{z}_k(t); F(t, z_k(t))) = \max_{v \in z_k(t) + (N + |\dot{z}(t) - z_k(t)|)\mathbb{B}} \{\langle p_k(t), v\rangle - \alpha(k_f+1)d(v; F(t, z_k(t))) - s_k|v - \dot{z}_k(t)|\} \text{ a.e.}$$

By Lemma 2.2, the multivalued mapping $\Gamma(t,\cdot)$ is u.s.c. with compact convex values, and by the definition of the limiting Fréchet subdifferential and the sub-Lipschitz condition we have (via Lemma 2.1 and (41)) for all k

$$|\dot{p}_k(t)| \le \alpha(k_f+1)(1 + k(t) + \beta(1 + |\dot{z}(t) - \dot{z}_k(t)|)) \quad \text{a.e.}$$

By Lemma 2.4 there exists a subsequence of (p_k) converging uniformly to an arc p satisfying

$$(43)\qquad \dot{p}(t) \in \Gamma(t, z(t), \dot{z}(t), p(t), 0) \quad \text{a.e.}$$

Note that

$$\partial(|f(\cdot,\cdot) - \gamma_k|)(z_k(a), z_k(b)) \subset \partial(\langle w_k, f(\cdot,\cdot)\rangle)(z_k(a), z_k(b)),$$

and hence, by passing to the limit in (40) and (42) and using the same argument as in Part 1, Step 2, we have

$$(p(a), -p(b)) \in \partial(\langle \omega, f(\cdot,\cdot)\rangle)(z(a), z(b)) + N(S; (z(a), z(b))),$$

$$\langle p(t), \dot{z}(t)\rangle = H(t, z(t), p(t)) \text{ a.e.}$$

Now if we assume that F is convex-valued, then, by (29) and/or (43) and Rockafeller result [40], we obtain

$$\dot{p}(t) \in \text{co}\,\{q : (-q, \dot{z}(t)) \in \partial H(t, z(t), p(t))\} \quad \text{a.e.} \quad t \in [a,b],$$

which completes the proof.

5. Appendix.

- $h(x) = d((x(a), x(b)); S) + \int_a^b d(\dot{x}(t); F(t, x(t)))dt$ is l.s.c. on $B(z, \varepsilon)$.
Since F is sub-Lipschitzian at z, then there exist $\beta > 0$, $\varepsilon > 0$, and a summable function $k : [a, b] \mapsto \mathbb{R}$ such that for almost all $t \in [a, b]$, for all $N > 0$, for all $x, x' \in z(t) + \varepsilon\mathbb{B}$, and $y \in \dot{z}(t) + N\mathbb{B}$ one has

$$d(y, F(t, x)) - d(y, F(t, x')) \leq (k(t) + \beta N)|x - x'|.$$

Let $x \in B(z, \varepsilon)$ and $\varepsilon' > 0$, and set $\delta < \frac{\varepsilon'}{1+\int_a^b k(t)\,dt+\beta(\varepsilon+b-a)}$.
Let $x' \in B(z, \varepsilon)$ such that $\| x - x' \| < \delta$, and set $N = | \dot{x'}(t) - \dot{z}(t) | +1$. We have

$$\begin{aligned}\int_a^b d(\dot{x}(t), F(t, x(t)))\,dt - \int_a^b d(\dot{x'}(t), F(t, x'(t)))\,dt \,| &\leq | \dot{x}(t) - \dot{x'}(t) \,| \\ &+ \int_a^b d(\dot{x'}(t), F(t, x(t)))\,dt - \int_a^b d(\dot{x'}(t), F(t, x'(t)))\,dt \\ &\leq \delta + \int_a^b (k(t) + \beta N) \mid x(t) - x'(t) \mid dt \\ &\leq \delta + \delta \left(\int_a^b k(t)\,dt + \beta(\varepsilon + b - a) \right) \leq \varepsilon'.\end{aligned}$$

Thus h is l.s.c on $B(z, \varepsilon)$.

- G_ε is closed.
Let (x_n) be a subsequence in G_ε such that $x_n \longrightarrow x$ in $W^{1,1}$. Since S is closed $(x(a), x(b)) \in S$. Set $N' = | \dot{x}(t) - \dot{z}(t) | +1$; since F is sub-Lipschitzian at z we have

$$d(\dot{x}(t), F(t, x(t))) \leq (k(t) + \beta N) \mid x(t) - x_n(t) \mid + d(\dot{x}(t), F(t, x_n(t)))$$

so that

$$\begin{aligned}\int_a^b d(\dot{x}(t), F(t, x(t)))\,dt \leq &\| x - x_n \| \int_a^b (k(t) + \beta N)\,dt \\ &+ \int_a^b d(\dot{x}(t), F(t, x_n(t)))\,dt \\ \leq &\| x - x_n \| \left(\beta \| x - z \| + \beta(b - a) + \int_a^b k(t)\,dt \right) \\ &+ \| x - x_n \| .\end{aligned}$$

Then $d(\dot{x}(t), F(t, x(t))) = 0$ a.e., and since F is closed-valued $x \in G_\varepsilon$.

- $L_k(t, \cdot, \cdot)$ is epi-Lipschitzian at z_k.
We have

$$L_k(t, x, y) = \begin{cases} \alpha(k_f + 1)d(y; F(t, x)) + \frac{1}{k}|y - \dot{z}_k(t)| & \text{if } (x, y) \in A(t), \\ +\infty & \text{otherwise,} \end{cases}$$

where $A(t) = (z(t) + s\mathbb{B}) \times (\dot{z}(t) + (N + |\dot{z}(t) - \dot{z}_k(t)|)\mathbb{B})$.
For k large enough we can suppose that $| z_k(t) - z(t) | < \frac{s}{2}$. Let $x_1, x_2 \in B(z_k(t), \frac{s}{2})$ and $y \in \mathbb{R}^n$ such that $L_k(t, x_1, y)$ is finite. Then

$$| x_1 - z(t) | \leq s \text{ and } | y - \dot{z}(t) | \leq N + |\dot{z}(t) - \dot{z}_k(t)|.$$

Since $\mid x_2 - z(t) \mid < s$, $L_k(t, x_2, y)$ is finite, and using the fact that F is sub-Lipschitzian at z we get

$$L_k(t, x_2, y) - L_k(t, x_1, y) = \alpha(k_f + 1)[d(y; F(t, x_2)) - d(y; F(t, x_1))]$$

$$\leq \alpha(k_f + 1)(k(t) + \beta(N + |\dot{z}(t) - \dot{z}_k(t)|)) \mid x_1 - x_2 \mid .$$

Then $L_k(t, \cdot, \cdot)$ is epi-Lipschitzian at z_k.

- $\Gamma(t, x, y, w, s)$ is measurable in t.
  The measurability of the multivalued mapping $\Gamma(t, x, y, w, s)$ in t follows from the two following lemmas.
  LEMMA 5.1. *Let $G : [a, b] \to \mathbb{R}^n$ be a measurable multivalued mapping, and let K be a compact set in $\mathbb{R}^n$. Then the multivalued mapping $G(\cdot) + K$ is also measurable.*
  Proof. It suffices to see that for any set A in $\mathbb{R}^n$ we have

$$(G(\cdot) + K)^{-1}(A) = G^{-1}(A - K),$$

  where $G^{-1}(A) = \{t : G(t) \cap A \neq \emptyset\}$.
  LEMMA 5.2. *Let $f : [a, b] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be a l.s.c. function in (x, y) and measurable in (t, x, y). Consider the multivalued mapping*

$$R(t, x, y, p) = \{q : (q, -p) \in \partial f(t, x, y) + \{0\} \times s\mathbb{B}\} .$$

  Then R and $\bar{co}\, R$ are measurable in t.
  Proof. It follows from Lemma 2 in [21] that the graph of the multivalued mapping $t \to \partial f(t, x, y)$ is measurable. As this multivalued mapping is closed-valued, Theorem 8.1.4 in [2] implies that it is measurable in t. Now Lemma 5.1 asserts that the multivalued mapping

$$t \longrightarrow \partial f(t, x, y) + \{0\} \times s\mathbb{B}$$

  is measurable in t. The measurability of $t \to R(t, x, y, p)$ follows from the formula

$$(\partial f(\cdot, x, y) + \{0\} \times s\mathbb{B})^{-1}(A \times \{-p\}) = R^{-1}(\cdot, x, y, p)(A).$$

  The measurability of $\bar{co}R$ follows from Theorem 8.2.2 in [2].

Acknowledgment. We wish to thank the anonymous referees for their careful reading of the manuscript. Their helpful comments are greatly appreciated.

REFERENCES

[1] J. P. AUBIN, *Lipschitz behaviour of solutions to convex minimization problems*, Math. Oper. Res., 8 (1984), pp. 87–111.
[2] J. P. AUBIN AND H. FRANKOWSKA, *Set-valued Analysis, Systems and Control: Foundations and Applications*, 2, Birkhäuser Boston, Cambridge, MA, 1990.
[3] V. BHASKAR, S. K. GUPTA, AND A. K. RAY, *Multiobjective optimization of an industrial wiped film PET reactor*, American Institute Chem. Eng. J., 46 (2000), pp. 1046–1058.
[4] V. BHASKAR, S. K. GUPTA, AND A. K. RAY, *Applications of multiobjective optimization in chemical engineering*, Reviews Chem. Eng., 16 (2000), pp. 1–54.
[5] J. M. BORWEIN, *Proper efficient points for maximizations with respect to cones*, SIAM J. Control. Optim., 15 (1977), pp. 57–63.

[6] J. Borwein and A. Ioffe, *Proximal analysis in smooth spaces*, Set-Valued Anal., 4 (1996), pp. 1–24.
[7] J. M. Borwein and Q. J. Zhu, *Viscosity solutions and viscosity subderivatives in smooth Banach spaces with applications to metric regularity*, SIAM J. Control. Optim., 34 (1996), pp. 1568–1591.
[8] J. M. Borwein and Q. J. Zhu, *A survey of subdifferential calculus with applications*, Nonlinear Anal., 49 (2002), pp. 295–296.
[9] M. Chilali, P. Gahinet, and C. Scherer, *Multiobjective output-feedback control via LMI optimization*, in Proceedings of the IFAC World Congress, San Francisco, CA, 1996, pp. 249–254.
[10] F. H. Clarke, *The generalized problem of Bolza*, SIAM J. Control Optim., 14 (1976), pp. 682–699.
[11] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
[12] F. H. Clarke, *Methods of Dynamic and Nonsmooth Optimization*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 57, SIAM, Philadelphia, 1989.
[13] F. H. Clarke and P. R. Wolenski, *Necessary conditions for functional differential inclusions*, Appl. Math. Optim., 34 (1996), pp. 51–78.
[14] F. H. Clarke, Yu. S. Ledyayev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Grad. Texts in Math. 178, Springer-Verlag, New York, 1998.
[15] B. D. Craven, *Nonsmooth multiobjective programming*, Numer. Funct. Anal. Optim., 10 (1989), pp. 49–64.
[16] G. Debreu, *Theory of Value*, John Wiley and Sons, New York, 1959.
[17] G. Debreu, *Mathematical Economics: Twenty Papers of Gerard Debreu*, Cambrige University Press, UK, 1983, pp. 163–172.
[18] I. Ekeland, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.
[19] H. Frankowska, *The maximum principle for an optimal solution to a differential inclusion with end points constraints*, SIAM J. Control Optim., 25 (1987), pp. 145–157.
[20] H. Frankowska, *Local controllability and infinitesimal generators of semigroups of set-valued maps*, SIAM J. Control. Optim., 25 (1987), pp. 412–432.
[21] A. D. Ioffe, *Absolutely continuous subgradients of nonconvex integral functions*, Nonlinear Anal., 11 (1987), pp. 245–257.
[22] A. D. Ioffe, *Approximate subdifferentials and applications* I: *The finite dimensional theory*, Trans. Amer. Math. Soc., 281 (1984), pp. 389–416.
[23] A. D. Ioffe, *Euler-Lagrange and Hamiltonian formalisms in dynamic optimization*, Trans. Amer. Math. Soc., 349 (1997), pp. 2871–2900.
[24] A. D. Ioffe and R. T. Rockafellar, *The Euler and Weierstrass conditions for nonsmooth variational problems*, Calc. Var. Partial Differential Equations, 4 (1996), pp. 59–87.
[25] A. Jourani, *Normality, local controllability and NOC for multiobjective optimal control problems*, in Contemporary Trends in Nonlinear Geometric Control Theory and Its Applications, A. Anzaldo-Meneses, B. Bonnard, J.-P. Gauthier, and F. Monroy-Pérez, eds., World Scientific, River Edge, NJ, 2002.
[26] A. Jourani, *Lagrangian and Hamiltonian Necessary Conditions for the Generalized Bolza Problem and Applications*, Préprint, Université de Bourgogne, Dijon, France, 2000.
[27] B. Kaskosz and S. Lojasiewicz, Jr., *Lagrange-type extremal trajectories in differential inclusions*, Systems Control Lett., 19 (1992), pp. 241–247.
[28] A. Y. Kruger and B. S. Mordukhovich, *Extremal points and Euler equations in nonsmooth optimization*, Dokl. Akad. Nauk. BSSR, 24 (1980), pp. 684–687 (in Russian).
[29] P. D. Loewen and R. T. Rockafellar, *Optimal control of unbounded differential inclusions*, SIAM J. Control Optim., 32 (1994), pp. 442–470.
[30] P. D. Loewen and R. T. Rockafellar, *New necessary conditions for the generalized problem of Bolza*, SIAM J. Control Optim., 34 (1996), pp. 1496–1511.
[31] P. D. Loewen and R. T. Rockafellar, *Bolza problems with general time constraints*, SIAM J. Control Optim., 35 (1997), pp. 2050–2069.
[32] M. Minami, *Weak Pareto optimal necessary conditions in a nondifferential multiobjective program on a Banach space*, J. Optim. Theory Appl., 41 (1983), pp. 451–461.
[33] B. S. Mordukhovich, *Maximum principle in problems of time optimal control with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.
[34] B. S. Mordukhovich, *Metric approximations and necessary optimality conditions for general classes of nonsmooth extremal problems*, Soviet Math. Dokl., 22 (1980), pp. 526–530.
[35] B. S. Mordukhovich, *Optimal control of difference, differential and differential-difference inclusions*, J. Math. Sci. (New York), 100 (2000), pp. 2613–2632.
[36] B. S. Mordukhovich, *Discrete approximations and refined Euler–Lagrange conditions for*

nonconvex differential inclusions, SIAM J. Control Optim., 33 (1995), pp. 882–915.
[37] B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988 (in Russian); English translation to appear in Wiley-Interscience.
[38] B. S. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.
[39] R. T. Rockafellar, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Anal. Appl., 32 (1970), pp. 174–222.
[40] R. T. Rockafellar, *Equivalent subgradient versions of Hamiltonian and Euler–Lagrange equations in variational analysis*, SIAM J. Control Optim., 34 (1996), pp. 1300–1314.
[41] R. T. Rockafellar, *Lipschitzian properties of multifunctions*, Nonlinear Anal., 9 (1985), pp. 867–885.
[42] C. Singh, *Optimality conditions in multiobjective differentiable programming*, J. Optim. Theory Appl., 53 (1987), pp. 115–123.
[43] H. Sussmann, *A strong version of the Lojasiewicz maximum principle*, in Optimal Control of Differential Equations, N. H. Pavel, ed., Marcel Dekker, New York, 1994, pp. 293–309.
[44] R. Vinter and H. Zheng, *The extended Euler–Lagrange condition for nonconvex variational problems*, SIAM J. Control Optim., 35 (1997), pp. 56–77.
[45] B. Vroemen and B. De Jager, *Multiobjective control: An overview*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 440–445.
[46] L. Wang, J. Dong, and Q. Liu, *Optimality conditions in nonsmooth multiobjective programming*, System Sci. Math. Sci., 7 (1994), pp. 250–255.
[47] L. Wang and Q. Lui, *Nonsmooth multiobjective programming*, System Sci. Math. Sci., 7 (1994), pp. 362–366.
[48] Q. Zhu, *Necessary optimality conditions for nonconvex differential inclusions with endpoint constraints*, J. Differential Equations, 124 (1996), pp. 186–204.
[49] Q. J. Zhu, *Hamiltonian necessary conditions for a multiobjective optimal control problem with endpoint constraints*, SIAM J. Control Optim., 39 (2000), pp. 97–112.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 2062–2077

ON LINEAR-QUADRATIC OPTIMAL CONTROL PROBLEMS FOR TIME-VARYING DESCRIPTOR SYSTEMS*

GALINA A. KURINA† AND ROSWITHA MÄRZ‡

Abstract. We deal with linear-quadratic optimal control problems for time-varying descriptor systems in a Hilbert space setting. A sufficient solvability condition is given by means of an appropriately stated linear boundary value problem (BVP) and by discussing the special structure of the regular differential system inherent in this BVP.

Key words. linear-quadratic control problem, abstract implicit differential equation, variable coefficient descriptor system

AMS subject classifications. 49J27, 49J15, 49N10

DOI. 10.1137/S0363012900380991

1. Introduction.

This paper aims at minimizing quadratic cost functionals over solutions of singular linear differential equations or differential-algebraic equations (DAEs) of the form

$$(A(t)x(t))' = C(t)x(t) + B(t)u(t), \quad t \in [0,T],$$

with fixed $T > 0$ and continuous coefficients A, C, and B. Especially in the case of constant coefficients, with singular A, the equation

$$Ax'(t) = Cx(t) + Bu(t), \quad t \in [0,T],$$

is often called a descriptor system. We keep this name for more general cases, too. There is an extensive literature (see [4], and [1], [2], [10] for an overview) on optimal control problems for time-invariant descriptor systems, and there are some papers investigating the case of smooth time-varying coefficients [9] and that of C and B being variable while A remains constant, respectively, [5], [6].

In [5], [6] the linear quadratic optimization problem—given in a Hilbert space setting—is traced back to the solution of a linear boundary value problem (BVP), where the appropriate formulation of the adjoint system as well as the invertibility properties of a specially structured linear operator play an essential role. Let us remark that this approach is fundamentally different from that in [1], [4], [10], for instance, where the matrix pencil $\lambda A - C$ is assumed to be regular, the descriptor system is subject to a reduction, and, finally, an approach based on solving a Riccati equation, which, however, may not be solvable in [1], [10] (cf. [12]), is studied.

The new insights concerning the structure of linear DAEs and their adjoint systems obtained in [7] allow for further generalizations of the results from [5], [6], in particular in the case of time-varying coefficients A. The main results of the present paper are the given sufficient solvability condition by means of an appropriate linear

*Received by the editors November 10, 2000; accepted for publication (in revised form) May 27, 2003; published electronically January 28, 2004.

http://www.siam.org/journals/sicon/42-6/38099.html

†Voronezh State Forestry Academy, ul. Timirjazeva, 8, Voronezh, 394613, Russia (kurina@kma.vsu.ru). The work of this author was supported by Russian Fundamental Research Foundation grant 99-01-00968.

‡Humboldt University, Institute of Mathematics, site: Rudower Chaussee 25, D-10099 Berlin, Germany (maerz@mathematik.hu-berlin.de).

BVP in section 2 as well as solvability statements for the BVP and, hence, for the optimal control problem in section 5. We do not need any assumptions on the regularity and index of the descriptor system itself. Properties of a special operator

$$F = \begin{pmatrix} F_1 & 0 & F_2 \\ F_3 & -F_1^* & F_4 \\ -F_4^* & F_2^* & -F_5 \end{pmatrix}, \tag{1.1}$$

whose invertibility turns out to be equivalent to the index-1 property of the DAE contained in the BVP, are essential here. The necessary characteristics of the operator F itself are provided in section 4.

A special aspect of the linear BVP under consideration is the question regarding the structure of the inherent regular differential equation. This question is answered in Theorem 5.1. It turns out that this structure is not nonnegative Hamiltonian in each case but only if the image of $A(t)$ does not vary with t.

In section 6 we discuss two transparent simple examples. The first one is characterized by a time-varying image of $A(t)$; in the second example the descriptor system is time-invariant but infinite-dimensional. In both examples, the optimization problems and the BVPs, respectively, are uniquely solvable, whereas, for a given control, the system state is not uniquely determined by the respective descriptor system to be controlled.

2. Sufficient conditions of control optimality. First of all, let us collect some notation and well-known facts on linear operators in real Hilbert spaces, which we want to use in the following.

For given Hilbert spaces X, Y we denote by $L(X,Y)$ the Banach space of linear bounded mappings of X with values in Y. The kernel $\mathrm{Ker}A$ of an operator $A \in L(X,Y)$ is a closed linear manifold in X, i.e., a subspace. We speak of a nullspace then. A subspace of a Hilbert space equipped with the same scalar product is a Hilbert space again. Note that $A \in L(X,Y)$ leads to $A^* \in L(Y,X)$ for the adjoint operator A^* to A. Recall that a mapping $A \in L(X,Y)$ with closed image $\mathrm{Im}A \subset Y$ is said to be normally solvable [8].

For a normally solvable operator $A \in L(X,Y)$ we can make use of orthogonal decompositions

$$X = \mathrm{Ker}A \oplus \mathrm{Im}A^*, \quad Y = \mathrm{Ker}A^* \oplus \mathrm{Im}A.$$

In particular, there are uniquely determined orthogonal projectors $P \in L(X,X) =: L(X), Q \in L(Y)$, such that $\mathrm{Im}P = \mathrm{Ker}A, \mathrm{Im}Q = \mathrm{Ker}A^* = (\mathrm{Im}A)^\perp$. The Moore inverse A^+ of a normally solvable operator $A \in L(X,Y)$ is again a bounded map, i.e., $A^+ \in L(Y,X), AA^+A = A, A^+AA^+ = A^+, AA^+ = I - Q, A^+A = I - P$ hold true (cf. [8]).

By I we mean the identity operator in the given spaces. For scalar products we shall use the uniform notation $\langle .,. \rangle$, even if we have different spaces.

We say that an operator $A \in L(X)$ is positive definite (semidefinite) if $\langle Ax, x\rangle > 0$ (≥ 0) for all nonzero $x \in X$.

Now we turn to our main topic, the problem of minimizing the quadratic cost functional

$$\begin{aligned} J(u,x) = \frac{1}{2}\langle x(T), Vx(T)\rangle + \frac{1}{2}\int_0^T (\langle x(t), W(t)x(t)\rangle \\ + 2\langle x(t), S(t)u(t)\rangle + \langle u(t), R(t)u(t)\rangle)dt \end{aligned} \tag{2.1}$$

on trajectories of the linear system

$$(A(t)x(t))' = C(t)x(t) + B(t)u(t), \quad t \in [0,T], \tag{2.2}$$

$$A(0)x(0) = y_0. \tag{2.3}$$

Here $T > 0$ is fixed, $x(t) \in X, u(t) \in U, y_0 \in Y$. X, Y, U are real Hilbert spaces, $V, W(t) \in L(X), S(t) \in L(U,X)$, $R(t) \in L(U), A(t), C(t) \in L(X,Y)$, $B(t) \in L(U,Y)$, $W(t) = W^*(t)$, $R(t) = R^*(t)$ for all $t \in [0,T], V = V^*$, the operators $W(t), S(t), R(t)$, $A(t), C(t), B(t)$ are continuous with respect to t. $A(t)$ is assumed to be normally solvable for all $t \in [0,T]$.

Admissible controls are continuous functions with values in U for which there is a solution of the problem (2.2), (2.3). A solution of (2.2) is a continuous function $x : [0,T] \to X$ that has a continuously differentiable product $A(t)x(t)$ and satisfies (2.2) pointwise.

Since the operator $A(t)$ is normally solvable for all $t \in [0,T]$, the spaces X and Y are decomposed into the orthogonal sums $X = \mathrm{Ker}A(t) \oplus \mathrm{Im}A^*(t), Y = \mathrm{Ker}A^*(t) \oplus \mathrm{Im}A(t)$. Denote by $P(t)$ the orthogonal projector of the space X onto $\mathrm{Ker}A(t)$ and by $Q(t)$ the orthogonal projector of the space Y onto $\mathrm{Ker}A^*(t)$.

Remark 2.1. From the relation (2.3) it follows that $y_0 \in \mathrm{Im}A(0)$; that is, for some $\tilde{x}_0 \in X$ the equality $y_0 = A(0)\tilde{x}_0$ should hold. Later on we will assume that $y_0 \in \mathrm{Im}A(0)$. Note that differential-algebraic systems (2.2) and initial value problems (2.2), (2.3) are considered, e.g., in [7].

The following conditions shall be used as basic assumptions throughout this paper:

I. The operator V is positive semidefinite.

II. For all $t \in [0,T]$ the operator $\begin{pmatrix} W(t) & S(t) \\ S^*(t) & R(t) \end{pmatrix}$ is positive semidefinite.

III. The projector $P(t)$ is continuous in t, and the projector $Q(t)$ depends continuously differentiably on t.

LEMMA 2.2. *Let conditions* I, II, *and* III *be given. If the triple of continuous functions* $(x_*, \psi_*, u_*) : [0,T] \to X \times Y \times U$ *has continuously differentiable parts* $Ax_* : [0,T] \to Y$, $(I-Q)\psi_* : [0,T] \to Y$ *and satisfies the system*

$$(Ax)'(t) = C(t)x(t) + B(t)u(t), \quad A(0)x(0) = y_0, \tag{2.4}$$

$$\begin{gathered} A^*(t)((I-Q)\psi)'(t) = W(t)x(t) - (C^*(t) + A^*(t)Q'(t))\psi(t) + S(t)u(t), \\ A^*(T)\psi(T) = -Vx(T), \end{gathered} \tag{2.5}$$

$$0 = -S^*(t)x(t) + B^*(t)\psi(t) - R(t)u(t), \tag{2.6}$$

then $u_*(t)$ *is an optimal control for the problem* (2.1)–(2.3).

Proof. Let $x_*(t), u_*(t), \psi_*(t)$ be a solution of the system (2.4)–(2.6), $u(t)$ be an arbitrary admissible control, and $x(t)$ be a corresponding solution of the problem (2.2), (2.3).

Taking into account the relations (2.1)–(2.6), straightforward computations yield the equality

$$\begin{aligned} J(u,x) - J(u_*,x_*) = \mathfrak{A} &+ \frac{1}{2}\langle x(T) - x_*(T), V(x(T) - x_*(T))\rangle \\ &+ \frac{1}{2}\int_0^T \left\langle \begin{pmatrix} x(t) - x_*(t) \\ u(t) - u_*(t) \end{pmatrix}, \begin{pmatrix} W(t) & S(t) \\ S^*(t) & R(t) \end{pmatrix} \begin{pmatrix} x(t) - x_*(t) \\ u(t) - u_*(t) \end{pmatrix} \right\rangle dt \end{aligned}$$

with $\mathfrak{A} = \langle x(T) - x_*(T), Vx_*(T)\rangle + \int_0^T \{\langle x(t) - x_*(t), W(t)x_*(t) + S(t)u_*(t)\rangle + \langle u(t) - u_*(t), S^*(t)x_*(t) + R(t)u_*(t)\rangle\}dt$. The term $\mathfrak{A}$ will be shown to vanish. Then conditions I and II yield the inequality

$$J(u, x) - J(u_* x_*) \geq 0;$$

hence, u_* is an optimal control. Now we show $\mathfrak{A} = 0$ to be true in fact. We have

$$\begin{aligned}
\mathfrak{A} &= \langle x(T) - x_*(T), Vx_*(T)\rangle \\
&\quad + \int_0^T \{\langle x(t) - x_*(t), A^*(t)((I-Q)\psi_*)'(t) + C^*(t)\psi_*(t) + A^*(t)Q'(t)\psi_*(t)\rangle \\
&\quad + \langle u(t) - u_*(t), B^*(t)\psi_*(t)\rangle\}dt \\
&= \langle x(T) - x_*(T), Vx_*(T)\rangle + \int_0^T \{\langle A(t)(x(t) - x_*(t)), ((I-Q)\psi_*)'(t)\rangle \\
&\quad + \langle C(t)(x(t) - x_*(t)) + B(t)(u(t) - u_*(t)) + Q'(t)A(t)(x(t) - x_*(t)), \psi_*(t)\rangle\}dt \\
&= \langle x(T) - x_*(T), Vx_*(T)\rangle + \int_0^T \{\langle A(t)(x(t) - x_*(t)), ((I-Q)\psi_*)'(t)\rangle \\
&\quad + \langle (A(x - x_*))'(t) + Q'(t)A(t)(x(t) - x_*(t)), \psi_*(t)\rangle\}dt.
\end{aligned}$$

Taking into account that $A = (I-Q)A$, $Q' = -(I-Q)'$, and therefore $(A(x-x_*))' + Q'A(x-x_*) = ((I-Q)A(x-x_*))' - (I-Q)'A(x-x_*) = (I-Q)(A(x-x_*))'$, we derive

$$\begin{aligned}
\mathfrak{A} &= \langle x(T) - x_*(T), Vx_*(T)\rangle + \int_0^T \{\langle A(t)(x(t) - x_*(t)), ((I-Q)\psi_*)'(t)\rangle \\
&\quad + \langle (I - Q(t))(A(x - x_*))'(t), \psi_*(t)\rangle\}dt \\
&= \langle x(T) - x_*(T), Vx_*(T)\rangle + \int_0^T \{\langle A(t)(x(t) - x_*(t)), ((I-Q)\psi_*)'(t)\rangle \\
&\quad + \langle (A(x - x_*))'(t), (I - Q(t))\psi_*(t)\rangle\}dt \\
&= \langle x(T) - x_*(T), Vx_*(T)\rangle + \int_0^T \frac{d}{dt}\langle A(t)(x(t) - x_*(t)), (I - Q(t))\psi_*(t)\rangle dt \\
&= \langle x(T) - x_*(T), Vx_*(T)\rangle + \langle x(T) - x_*(T), A^*(T)\psi(T)\rangle = 0. \qquad \square
\end{aligned}$$

Remark 2.3. The boundary condition $A^*(T)\psi(T) = -Vx(T)$ can be rewritten in two parts as

$$(I - Q(T))\psi(T) = -A^{*+}(T)Vx(T),\ P(T)Vx(T) = 0. \tag{2.7}$$

If the inclusion

$$\operatorname{Im} V \subseteq \operatorname{Im} A^*(T) \tag{2.8}$$

is valid, then it holds that $P(T)V = 0$ and $V = V(I - P(T)) = VA^+(T)A(T)$. In the consequence, the boundary condition (2.7) simplifies to

$$(I - Q(T))\psi(T) = -A^{+*}(T)VA^+(T)A(T)x(T) \tag{2.9}$$

such that the boundary conditions in the BVP (2.4)–(2.6) are exclusively directed to the smooth components Ax and $(I-Q)\psi$. We will exploit this fact in section 5 in order to derive the solvability of (2.4)–(2.6) from the solvability of BVPs in inherent Hamiltonian systems. Condition (2.8) is a helpful means for formulating sufficient solvability conditions of control problems. If it is not satisfied, the BVP (2.4)–(2.6) may be solvable nevertheless, or it is unsolvable (cf. Remark 6.1 below). Let us stress here once more that we are concerned with sufficient solvability conditions but not with necessary conditions.

Let us mention that in the standard reference [4], assumption (2.8) is built in the problem at the very beginning by using a cost functional with $A^*\tilde{V}A$ instead of V.

3. Equations in subspaces. For each $t \in [0,T]$ the operators

$$\begin{aligned} A(t) &= (I-Q(t))A(t)(I-P(t)) &&: \text{Im } A^*(t) \to \text{Im } A(t), \\ A^*(t) &= (I-P(t))A^*(t)(I-Q(t)) &&: \text{Im } A(t) \to \text{Im } A^*(t) \end{aligned}$$

have the Moore inverses $A^+(t) = (I-P(t))A^+(t)(I-Q(t))$ and $A^{*+}(t) = (I-Q(t))A^{+*}(t)(I-P(t)) = A^{+*}(t)$.

Using the projectors P, Q, and the identities $Q(t)(Ax)'(t) = -Q(t)Q'(t)(Ax)(t)$, $Q(t)((I-Q)\psi)'(t) = -Q(t)Q'(t)((I-Q)\psi)(t)$, $((I-P)x)(t) = (A^+Ax)(t)$, $Q'(t)(I-Q(t)) = Q(t)Q'(t)$ we obtain the following two relations from the system (2.4)–(2.6):

$$0 = L \begin{pmatrix} Ax \\ (I-Q)\psi \end{pmatrix} + G \begin{pmatrix} Px \\ Q\psi \\ u \end{pmatrix}, \tag{3.1}$$

$$\begin{pmatrix} Ax \\ (I-Q)\psi \end{pmatrix}' = M \begin{pmatrix} Ax \\ (I-Q)\psi \end{pmatrix} + K \begin{pmatrix} Px \\ Q\psi \\ u \end{pmatrix}, \tag{3.2}$$

where

$$L = \begin{pmatrix} QQ'(I-Q) + QCA^+ & 0 \\ PWA^+ & -PC^*(I-Q) \\ -S^*A^+ & B^*(I-Q) \end{pmatrix},$$

$$G = \begin{pmatrix} QCP & 0 & QB \\ PWP & -PC^*Q & PS \\ -S^*P & B^*Q & -R \end{pmatrix}, \tag{3.3}$$

$$\hat{M} = \begin{pmatrix} (I-Q)CA^+ & 0 \\ A^{*+}WA^+ & -A^{*+}C^*(I-Q) \end{pmatrix},$$

$$M = \hat{M} - \begin{pmatrix} QQ'(I-Q) & 0 \\ 0 & QQ'(I-Q) \end{pmatrix},$$

$$K = \begin{pmatrix} (I-Q)CP & 0 & (I-Q)B \\ A^{*+}WP & -A^{*+}C^*Q - (I-Q)Q'Q & A^{*+}S \end{pmatrix}.$$

For brevity we often omit the argument t. With

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \tag{3.4}$$

and

$$\hat{J} = \mathrm{diag}(J, I)$$

it holds that

$$K = JL^*\hat{J}.$$

By construction it holds that

$$\mathrm{Ker}G(t) \supseteq \mathrm{Ker}P(t) \times \mathrm{Ker}Q(t) \times 0 = \mathrm{Im}A^*(t) \times \mathrm{Im}A(t) \times 0,$$
$$\mathrm{Im}G(t) \subseteq \mathrm{Im}Q(t) \times \mathrm{Im}P(t) \times U = \mathrm{Ker}A^*(t) \times \mathrm{Ker}A(t) \times U.$$

Following the lines of [7], [11] one may easily check that (2.4)–(2.6) represents an index-1-tractable DAE if and only if the mapping $G(t)$ acts bijectively from $\mathrm{Ker}A(t) \times \mathrm{Ker}A^*(t) \times U$ onto $\mathrm{Ker}A^*(t) \times \mathrm{Ker}A(t) \times U$. Consequently, the respective properties of the operator $G(t)$ are of great interest. It is evident that the operator G has a matrix representation of the form (1.1), which shall be studied in the next section.

If $G(t) : \mathrm{Ker}A(t) \times \mathrm{Ker}A^*(t) \times U \to \mathrm{Ker}A^*(t) \times \mathrm{Ker}A(t) \times U$ is actually a bijection and $G(t)^{-1}$ denotes its inverse, then (3.1) yields

$$\begin{pmatrix} Px \\ Q\psi \\ u \end{pmatrix} = -G^{-1}L \begin{pmatrix} Ax \\ (I-Q)\psi) \end{pmatrix}, \tag{3.5}$$

and (3.2) leads to a regular differential equation concerning the components Ax, $(I-Q)\psi$. This implies the solutions x, ψ, u of the optimality BVP (2.4)–(2.6) a priori to be continuous functions having continuously differentiable components Ax, $(I-Q)\psi$. More regularity, e.g., a fully continuously differentiable ψ, can be obtained via (3.5) by assuming the corresponding entries of the operator $G^{-1}L$ to be continuously differentiable.

4. Properties of the operator (1). We begin with the statement of sufficient conditions for the invertibility of operators $F : X_1 \times X_2 \times X_3 \to X_2 \times X_1 \times X_3$ having a matrix representation of the form (1.1), namely,

$$F = \begin{pmatrix} F_1 & 0 & F_2 \\ F_3 & -F_1^* & F_4 \\ -F_4^* & F_2^* & -F_5 \end{pmatrix},$$

where $X_i,\ i = 1, 2, 3$, are real Hilbert spaces, $F_i,\ i = 1, \ldots, 5$, denote bounded linear operators acting in the corresponding spaces and, additionally, F_3, F_5 are symmetric. Thereby, our special interest is directed to the particular case where, for fixed $t \in [0, T]$, $X_1 = \mathrm{Ker}A(t) = \mathrm{Im}P(t)$, $X_2 = \mathrm{Ker}A^*(t) = \mathrm{Im}Q(t)$, $X_3 = U$, $F_1 = (Q(t)C(t)P(t))^r$, $F_2 = (Q(t)B(t))^r$, $F_3 = (P(t)W(t)P(t))^r$, $F_4 = (P(t)S(t))^r$, $F_5 = R(t)$, where $(.)^r$ indicates the restriction of the operators inside the brackets to the

corresponding spaces, i.e., $F_1 = (Q(t)C(t)P(t))^r \in L(X_1, X_2)$ and so on. Clearly, if the operator $G(t)$ given by (3.3) acts bijectively on $\mathrm{Ker}A(t) \times \mathrm{Ker}A^*(t) \times U$ onto $\mathrm{Ker}A^*(t) \times \mathrm{Ker}A(t) \times U$, then the corresponding F is invertible and vice versa.

LEMMA 4.1. *If the operator*

$$\begin{pmatrix} F_3 & F_4 \\ F_4^* & F_5 \end{pmatrix} \tag{4.1}$$

is positive definite and has a bounded inverse, and if the operator $(F_1F_2)^*$ is injective and normally solvable, then F has a bounded inverse.

Proof. First we show that there is no sequence $\{x_n\}$ such that $x_n \in X_1 \times X_2 \times X_3$, $\|x_n\| = 1$ for all $n \in N$ and $Fx_n \to 0$. We assume the contrary, namely,

$$\begin{aligned} F_1x_n^1 \quad &+ \quad F_2x_n^3 \quad &&\to \quad 0, \\ F_3x_n^1 \quad &- \quad F_1^*x_n^2 \quad + \quad F_4x_n^3 \quad &&\to \quad 0, \\ -F_4^*x_n^1 \quad &+ \quad F_2^*x_n^2 \quad - \quad F_5x_n^3 \quad &&\to \quad 0, \end{aligned}$$

where

$$x_n = \begin{pmatrix} x_n^1 \\ x_n^2 \\ x_n^3 \end{pmatrix}, \quad x_n^i \in X_i, i = \overline{1,3}, \ \|x_n\| = 1 \text{ for all } n \in N.$$

We scalarly multiply the left-hand side of the first relation of this system by x_n^2, the second relation by x_n^1, and the third one by $(-x_n^3)$. Adding the results we obtain the relation

$$\left\langle \begin{pmatrix} x_n^1 \\ x_n^3 \end{pmatrix}, \begin{pmatrix} F_3 & F_4 \\ F_4^* & F_5 \end{pmatrix} \begin{pmatrix} x_n^1 \\ x_n^3 \end{pmatrix} \right\rangle \to 0.$$

By virtue of the assumptions we have $x_n^1 \to 0, x_n^3 \to 0$. Then the relation $(F_1\ F_2)^* x_n^2 \to 0$ follows from the last system. Hence, we obtain $x_n^2 \to 0$. This contradicts the equality $\|x_n\| = 1$. Therefore, there is a constant $k > 0$ such that the inequality $\|Fx\| \geq k$ is valid for all $x \in X_1 \times X_2 \times X_3$ with $\|x\| = 1$. Consequently, the inequality

$$\|Fx\| \geq k\|x\| \tag{4.2}$$

is valid for all $x \in X_1 \times X_2 \times X_3$.

Next we prove that $\mathrm{Ker}F^* = 0$ is valid. The operator F^* has the matrix representation

$$F^* = \begin{pmatrix} F_1^* & F_3 & -F_4 \\ 0 & -F_1 & F_2 \\ F_2^* & F_4^* & -F_5 \end{pmatrix}.$$

Assume that there is an $x \in X_2 \times X_1 \times X_3$ such that $F^*x = 0$. This means in detail

$$\begin{aligned} F_1^*x_2 \quad &+ \quad F_3x_1 \quad - \quad F_4x_3 \quad = \quad 0, \\ -F_1x_1 \quad &+ \quad F_2x_3 \quad = \quad 0, \\ F_2^*x_2 \quad &+ \quad F_4^*x_1 \quad - \quad F_5x_3 \quad = \quad 0, \end{aligned}$$

where

$$x = \begin{pmatrix} x_2 \\ x_1 \\ x_3 \end{pmatrix}, \quad x_i \in X_i,\ i = \overline{1,3}.$$

We scalarly multiply the first equation of this system by x_1, the second equation by x_2, and the third equation by $(-x_3)$. Adding the results yields the equality

$$\left\langle \begin{pmatrix} -x_1 \\ x_3 \end{pmatrix}, \begin{pmatrix} F_3 & F_4 \\ F_4^* & F_5 \end{pmatrix} \begin{pmatrix} -x_1 \\ x_3 \end{pmatrix} \right\rangle = 0.$$

By virtue of the assumption on (4.1) we have $x_1 = 0, x_3 = 0$. Hence, the equality $(F_1\ F_2)^* x_2 = 0$ follows. We obtain $x_2 = 0$; therefore $\mathrm{Ker} F^* = 0$.

It remains to prove that the range of the operator F is a closed set. We take a sequence $\{y_n\}$, where $y_n = Fx_n, x_n \in X_1 \times X_2 \times X_3, y_n \to y_0$. The sequence $\{Fx_n\}$ is fundamental. From (4.2) it follows that the sequence $\{x_n\}$ is also fundamental. It tends to an element $x_0 \in X_1 \times X_2 \times X_3$ due to the completeness of X_i. As F is a bounded operator, $Fx_n \to Fx_0$. Hence, $y_0 \in \mathrm{Im}\ F$, i.e., $\mathrm{Im}\ F$ is a closed set.

Therefore, the following decomposition is realized: $X_2 \times X_1 \times X_3 = \mathrm{Im} F \oplus \mathrm{Ker} F^* = \mathrm{Im} F$, since $\mathrm{Ker} F^* = 0$. Using the inequality (4.2) we are done (see, e.g., Theorem (2) in [3, p. 204]). □

Remark 4.2. In the finite-dimensional case, if $X = Y$ is valid and A, C, B, W, S, R are constant matrices, the conditions of Lemma 4.1 for the operator G coincide with Assumption 2 in [4].

LEMMA 4.3. *If the operator* (4.1) *is positive semidefinite and the operators*

$$(F_1\ F_2)^* \quad \textit{and} \quad \begin{pmatrix} F_1 & F_2 \\ F_3 & F_4 \\ F_4^* & F_5 \end{pmatrix}$$

are injective and normally solvable, then F has a bounded inverse operator.

The proof of Lemma 4.3 is similar to the proof of Lemma 4.1.

Remark 4.4. The conditions of Lemma 4.1 and Lemma 4.3 are not necessary for the invertibility of an operator F of the form (1.1). For example, the specific operator

$$F = \begin{pmatrix} 1 & 0 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

is invertible, but neither the conditions of Lemma 4.1 nor those of Lemma 4.3 are satisfied. For the operator

$$F = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

(corresponding to the example in [4, p. 677]) the conditions of Lemma 4.1 (Assumption 2 from [4]) fail to be valid, but Lemma 4.3 applies.

From section 3 recall the operator

$$\hat{J} = \mathrm{diag}(J, I) = \begin{pmatrix} 0 & I & 0 \\ -I & 0 & 0 \\ 0 & 0 & I \end{pmatrix}.$$

LEMMA 4.5. *If the operator F is invertible, then the operator $\hat{J}F^{-1}$ is symmetric.*

Proof. The identity $F = -\hat{J}\tilde{F}$, where

$$\tilde{F} = \begin{pmatrix} F_3 & -F_1^* & F_4 \\ -F_1 & 0 & -F_2 \\ F_4^* & -F_2^* & F_5 \end{pmatrix},$$

is easily checked. $\tilde{F}$ is obviously symmetric and invertible. Thus, $\hat{J}F^{-1} = -\hat{J}\tilde{F}^{-1}\hat{J}^{-1}$ is a symmetric operator, as $\tilde{F}^{-1}$ is symmetric and $\hat{J}$ is unitary. ∎

COROLLARY 4.6. *If the operator F is invertible, then the inverse operator F^{-1} has a matrix representation of the form*

$$F^{-1} = \begin{pmatrix} D_1 & D_2 & D_3 \\ D_4 & -D_1^* & D_5 \\ D_5^* & -D_3^* & D_6 \end{pmatrix}, \tag{4.3}$$

where D_2, D_4, D_6 are symmetric operators.

The proof of this corollary follows from Lemma 4.5 *and the matrix representation of the operator $\hat{J}^{-1}$.*

LEMMA 4.7. *If the operator F is invertible and the operator* (4.1) *is positive semidefinite, then the operators $\begin{pmatrix} D_2 & D_3 \\ D_3^* & -D_6 \end{pmatrix}$ and D_4 in the representation* (4.3) *are positive semidefinite.*

Proof. The proof follows the lines of [5], where the special case of $F_4 = 0$ is considered. Introduce the operator $\tilde{J}$ having the matrix representation

$$\begin{pmatrix} 0 & I & 0 \\ I & 0 & 0 \\ 0 & 0 & -I \end{pmatrix}.$$

For every

$$z = \begin{pmatrix} z_2 \\ z_1 \\ z_3 \end{pmatrix} \in X_2 \times X_1 \times X_3$$

the identity

$$\langle \tilde{J}F^{-1}z, z\rangle = \langle D_4 z_2, z_2\rangle + \left\langle \begin{pmatrix} D_2 & D_3 \\ D_3^* & -D_6 \end{pmatrix} \begin{pmatrix} z_1 \\ z_3 \end{pmatrix}, \begin{pmatrix} z_1 \\ z_3 \end{pmatrix} \right\rangle \tag{4.4}$$

is valid. With

$$v = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = F^{-1}z$$

we also have

$$\langle \tilde{J}F^{-1}z, z\rangle = \langle \tilde{J}v, Fv\rangle = \left\langle \begin{pmatrix} F_3 & F_4 \\ F_4^* & F_5 \end{pmatrix} \begin{pmatrix} v_1 \\ v_3 \end{pmatrix}, \begin{pmatrix} v_1 \\ v_3 \end{pmatrix} \right\rangle. \tag{4.5}$$

Here z, v are represented in a form that corresponds to the matrix representation of the operator F. Since the operator (4.1) is positive semidefinite, (4.5) yields $\langle \tilde{J}F^{-1}z, z\rangle \geq 0$ for all z. Hence, the expression on the right-hand side of formula (4.4) is nonnegative. Setting in (4.4) at first $z_2 = 0$ and second $z_1 = 0$, $z_3 = 0$ we obtain the statement. ∎

5. Solvability of the problem. Now we turn back to the system (3.1), (3.2) and to the BVP (2.4)–(2.6), respectively.

The operator $G(t)$ defined by (3.3) maps $X \times Y \times U$ into $Y \times X \times U$. By construction, it holds that $\mathrm{Im}G(t) \subseteq \mathrm{Ker}A^*(t) \times \mathrm{Ker}A(t) \times U = \mathrm{Im}Q(t) \times \mathrm{Im}P(t) \times U$, $\mathrm{Ker}G(t) \supseteq \mathrm{Ker}P(t) \times \mathrm{Ker}Q(t) \times 0$. It is natural to consider the restriction $(G(t))^r : \mathrm{Im}P(t) \times \mathrm{Im}Q(t) \times U \to \mathrm{Im}Q(t) \times \mathrm{Im}P(t) \times U$. Below we also use the shorter denotation $G(t)$ for $(G(t))^r$.

THEOREM 5.1. *Let conditions* I, II, *and* III *be satisfied. If, for every* $t \in [0, T]$, *the operator* $G(t) : \mathrm{Ker}\ A(t) \times \mathrm{Ker}\ A^*(t) \times U \to \mathrm{Ker}\ A^*(t) \times \mathrm{Ker}\ A(t) \times U$ *is bijective, then the system* (3.1), (3.2) *provides an explicit differential equation for the pair* $(Ax, (I-Q)\psi)$, *which is of the form*

$$\begin{pmatrix} Ax \\ (I-Q)\psi \end{pmatrix}' = E \begin{pmatrix} Ax \\ (I-Q)\psi \end{pmatrix} - \begin{pmatrix} QQ'(I-Q) & 0 \\ 0 & QQ'(I-Q) \end{pmatrix} \begin{pmatrix} Ax \\ (I-Q)\psi \end{pmatrix}, \tag{5.1}$$

where

$$E = \begin{pmatrix} E_1 & E_2 \\ E_3 & -E_1^* \end{pmatrix}, \quad E_2 = E_2^*,\ E_3 = E_3^*, \tag{5.2}$$

and the operators E_2 *and* E_3 *are positive semidefinite.*

Proof. Since $G(t)$ is invertible, (3.1), (3.2) immediately imply, in view of the relation for K, that (5.1) holds true with $E = \hat{M} - JL^*\hat{J}G^{-1}L$. The term $\hat{J}G^{-1}$ is symmetric (cf. Lemma 4.5); thus $L^*\hat{J}G^{-1}L$ is also symmetric. Due to the structure of $\hat{M}$, we arrive at (5.2).

It remains to show that the operators E_2 and E_3 in (5.2) are positive semidefinite. Expressions for the operators $E_2,\ E_3$ are

$$\begin{aligned} E_2 &= ((I-Q)CPD_2 - (I-Q)BD_3^*)PC^*(I-Q) - ((I-Q)CPD_3 \\ &\quad + (I-Q)BD_6)B^*(I-Q), \\ E_3 &= A^{*+}WA^+ + ((I-Q)Q'Q + A^{*+}C^*Q)(D_4(QQ'(I-Q) + QCA^+) \\ &\quad - D_1^*PWA^+ - D_5S^*A^+) + A^{*+}WP(-D_1(QQ'(I-Q) + QCA^+) \\ &\quad - D_2PWA^+ + D_3S^*A^+) - A^{*+}S(D_5^*(QQ'(I-Q) + QCA^+) \\ &\quad - D_3^*PWA^+ - D_6S^*A^+). \end{aligned}$$

Here the operators $D_i,\ i = \overline{1,6}$, belong to the matrix representation (4.3) of G^{-1} by Corollary 4.6.

For any $y \in Y$ we compute

$$\langle E_2(I-Q)y, (I-Q)y\rangle = \left\langle \begin{pmatrix} D_2 & D_3 \\ D_3^* & -D_6 \end{pmatrix} \begin{pmatrix} PC^*(I-Q)y \\ -B^*(I-Q)y \end{pmatrix}, \begin{pmatrix} PC^*(I-Q)y \\ -B^*(I-Q)y \end{pmatrix} \right\rangle.$$

Due to Lemma 4.7, E_2 is positive semidefinite.

Next we turn to the operator E_3.

Taking into account the symmetry of the operators $D_2,\ D_4,\ D_6$ (see Corollary

4.6) it is not difficult to verify the identities

$$\begin{aligned}
&\langle E_3(I-Q)y,(I-Q)y\rangle = \langle WA^+(I-Q)y, A^+(I-Q)y\rangle - \langle D_2z_1,z_1\rangle\\
&+\langle D_4z_2,z_2\rangle + \langle D_6z_3,z_3\rangle - 2\langle D_1^*z_1 + D_5z_3,z_2\rangle + 2\langle D_3z_3,z_1\rangle\\
&= \left\langle \begin{pmatrix} W & S\\ S^* & R\end{pmatrix}\begin{pmatrix} z_4\\ z_5\end{pmatrix}, \begin{pmatrix} z_4\\ z_5\end{pmatrix}\right\rangle\\
&+\langle (D_2 - D_2PWPD_2 + D_2PSD_3^* + D_3S^*PD_2 - D_3RD_3^*)z_1,z_1\rangle\\
&+\langle (D_4 - D_1^*PWPD_1 - D_1^*PSD_5^* - D_5S^*PD_1 - D_5RD_5^*)z_2,z_2\rangle\\
&+\langle (-D_6 - D_3^*PWPD_3 - D_3^*PSD_6 - D_6S^*PD_3 - D_6RD_6)z_3,z_3\rangle\\
&+2\langle (-D_1^*PWPD_2 + D_1^*PSD_3^* - D_5S^*PD_2 + D_5RD_3^*)z_1,z_2\rangle\\
&+2\langle (-D_3^* + D_3^*PWPD_2 - D_3^*PSD_3^* + D_6S^*PD_2 - D_6RD_3^*)z_1,z_3\rangle\\
&+2\langle (D_3^*PWPD_1 + D_3^*PSD_5^* + D_6S^*PD_1 + D_6RD_5^*)z_2,z_3\rangle,
\end{aligned}$$

where $z_1 = PWA^+(I-Q)y,\ z_3 = (QQ'(I-Q)+QCA^+)(I-Q)y,\ z_3 = S^*A^+(I-Q)y,$ $z_4 = A^+(I-Q)y + P(-D_2z_1 - D_1z_2 + D_3z_3),\ z_5 = D_3^*z_1 - D_5^*z_2 + D_6z_3.$

Due to Corollary 4.6 and the identities

$$\begin{aligned}
&QCPD_1 + QBD_5^* = I,\ QCPD_2 - QBD_3^* = 0,\ QCPD_3 + QBD_5 = 0,\\
&PWPD_1 - (QCP)^*D_4 + PSD_5^* = 0,\ PWPD_2 + (QCP)^*D_1^* - PSD_3^* = I,\\
&PWPD_3 - (QCP)^*D_5 + PSD_6 = 0,\\
&-S^*PD_1 + B^*QD_4 - RD_5^* = 0,\ -S^*PD_2 - B^*QD_1^* + RD_3^* = 0,\\
&-S^*PD_3 + B^*QD_5 - RD_6 = I,
\end{aligned}$$

all of which follow from the representation (4.3) of the operator G^{-1}, we obtain that all scalar products in the expression for $\langle E_3(I-Q)y,(I-Q)y\rangle$, except for the first one, vanish identically. For example, we have $D_2 - D_2PWPD_2 + D_2PSD_3^* + D_3S^*PD_2 - D_3RD_3^* = D_2 - D_2(I-(QCP)^*D_1^*) + D_3(-B^*QD_1^*) = (D_2(QCP)^* - D_3B^*Q)D_1^* = 0.$ □

Remark 5.2. If the continuously differentiable functions $y:[0,T]\to Y, z:[0,T]\to Y$ satisfy the explicit differential equation (cf. (5.1))

$$\begin{pmatrix} y\\ z\end{pmatrix}' = E\begin{pmatrix} y\\ z\end{pmatrix} - \begin{pmatrix} QQ'(I-Q) & 0\\ 0 & QQ'(I-Q)\end{pmatrix}\begin{pmatrix} y\\ z\end{pmatrix}, \tag{5.3}$$

then, multiplying (5.3) by diag (Q,Q) and taking into account that diag $(Q,Q)E = 0,\ Q'(I-Q) = QQ'$, we find that

$$\begin{pmatrix} Qy\\ Qz\end{pmatrix}' = \begin{pmatrix} (I-Q)Q'Q & 0\\ 0 & (I-Q)Q'Q\end{pmatrix}\begin{pmatrix} Qy\\ Qz\end{pmatrix} \tag{5.4}$$

holds true. Consequently, if $Q(t_0)y(t_0) = 0,\ Q(t_0)z(t_0) = 0$ for a certain $t_0\in[0,T]$, then $Q(t)y(t), Q(t)z(t)$ vanish identically, i.e., Im $A(t)\times$ Im $A(t)\subseteq Y\times Y$ becomes a time-varying invariant subspace of the differential equation (5.3). Since (5.4) decouples into $(Qy)' = (I-Q)Q'Q(Qy), (Qz)' = (I-Q)Q'Q(Qz),\ y(0)\in \mathrm{Im}A(0), z(T)\in \mathrm{Im}A(T)$ also imply $y(t)\in \mathrm{Im}A(t),\ z(t)\in \mathrm{Im}A(t),\ t\in[0,T]$, and vice versa.

If this invariant subspace does not vary with t, i.e., if $\mathrm{Im}A(t) = \mathrm{Im}A(0)$ for all $t\in[0,T]$ and, hence, if the projector $Q(t)$ is constant, things become easier. Then

(5.3) simplifies to the nonnegative Hamiltonian system

$$\begin{pmatrix} y \\ z \end{pmatrix}' = E \begin{pmatrix} y \\ z \end{pmatrix}$$

with the constant invariant subspace $\mathrm{Im}A(0) \times \mathrm{Im}A(0)$.

THEOREM 5.3. *Under the conditions of Theorem* 5.1*, if* Im $A(t)$ *is time-invariant, the system* (3.1), (3.2) *provides an explicit nonnegative Hamiltonian system.*

Remark 5.4. For operators $A(t),\ R(t)$ that have bounded inverses $A^{-1}(t),\ R^{-1}(t)$ on $Y,\ U$, respectively, and if $S(t) = 0,\ t \in [0,T]$, then (5.1) is of the special form

$$\begin{pmatrix} Ax \\ \psi \end{pmatrix}' = \begin{pmatrix} CA^{-1} & BR^{-1}B^* \\ A^{*-1}WA^{-1} & -A^{*-1}C^* \end{pmatrix} \begin{pmatrix} Ax \\ \psi \end{pmatrix}. \tag{5.5}$$

Obviously, this is a nonnegative Hamiltonian system. If, additionally, $A(t)$ depends continuously differentiably on t, we can turn to an explicit differential equation with respect to the unknowns x, ψ, but the resulting system

$$\begin{pmatrix} x \\ \psi \end{pmatrix}' = \begin{pmatrix} A^{-1}C - A^{-1}A' & A^{-1}BR^{-1}B^* \\ A^{*-1}W & -A^{*-1}C^* \end{pmatrix} \begin{pmatrix} x \\ \psi \end{pmatrix}$$

is no longer Hamiltonian in general. This is why we should prefer the form (5.5) also in this case. It should be mentioned that if we consider the new variable $y = A^*\psi$, then the latter system provides again a nonnegative Hamiltonian system with respect to the unknowns x, y.

If V satisfies the inclusion (2.8), the boundary conditions in the BVP (2.4)–(2.6) apply to the components $Ax = y$ and $(I - Q)\psi = z$ only (cf. (2.9)). Then Theorem 5.3 allows us to apply results concerning the unique solvability of two-point boundary values problems, which are known for nonnegative Hamiltonian systems, to systems of the form (2.4)–(2.6). In particular, we have the theorem stated below.

THEOREM 5.5. *Let conditions* I, II, *and* III *be satisfied. Let the inclusion* (2.8) *be valid. If, for every* $t \in [0,T]$*, the operator* $G(t)$ *is invertible on the space* $\mathrm{Ker}A^*(t) \times \mathrm{Ker}A(t) \times U$*, and* $\mathrm{Im}A(t)$ *does not depend on* t*, then the problem* (2.1)–(2.3) *is solvable.*

Now we take a closer look at the BVP for (5.3) with the boundary conditions

$$y(0) = y_0,\ z(T) = -A^{+*}(T)VA^+(T)y(T). \tag{5.6}$$

THEOREM 5.6. *If the space* Y *is finite-dimensional, then the BVP* (5.3), (5.6) *is uniquely solvable.*

Proof. It is enough to prove that the problem

$$\begin{aligned} x_1' &= E_1x_1 + E_2x_2 - QQ'(I-Q)x_1, \\ x_2' &= E_3x_1 - E_1^*x_2 - QQ'(I-Q)x_2, \\ x_1(0) &= 0, \\ x_2(T) &= -\tilde{V}x_1(T), \end{aligned} \tag{5.7}$$

where $\tilde{V} = A^{*+}(T)VA^+(T)$, has only the zero solution.

We scalarly multiply the first equation of (5.7) by x_2 and the second equation by x_1. Adding the results and taking into account Remark 5.2 we obtain the relation $\langle x_1, x_2\rangle' = \langle E_2x_2, x_2\rangle + \langle E_3x_1, x_1\rangle$.

Integrating the last equation on the interval $[0,T]$ due to the boundary conditions we have $\langle x_1(T), \tilde{V}x_1(T)\rangle + \int_0^T (\langle E_2x_2, x_2\rangle + \langle E_3x_1, x_1\rangle)dt = 0$.

Since the self-adjoint operators $\tilde{V}, E_2,\ E_3$ are positive semidefinite (see Theorem 5.1) and the operators $E_2,\ E_3$ are continuous with respect to t, it follows from the previous relation that $E_2x_2 \equiv 0,\ E_3x_1 \equiv 0$. By this, system (5.7) implies $x_1 \equiv 0$, $x_2 \equiv 0$. □

Lemma 2.2 and Theorems 5.1 and 5.6 provide the next statement, which applies to the case of time-varying Im $A(t)$.

THEOREM 5.7. *Let conditions* I, II, *and* III *be satisfied. Let the inclusion* (2.8) *be valid. If the space Y is finite-dimensional and if, for every $t \in [0,T]$, the operator $G(t)$ is invertible on the space* $\mathrm{Ker}A^*(t) \times \mathrm{Ker}A(t) \times U$, *then the problem* (2.1)–(2.3) *is solvable.*

Remark 5.8. For $A(t)$ being time-invariant, results of the present paper are given in [5] ($S = 0$) and in [6] ($S \neq 0$).

Comparing with [4], [10], where $X = Y$ are finite-dimensional spaces and the time-invariant problem is considered, we stress that we do not use conditions on the regularity of the pencil $\lambda A - C$. Further, we do not diagonalize or transform A and factorize the coefficients in the cost functional. In particular, Assumption 2 in [4] turns out to be no longer mandatory.

6. Small but illustrative examples.

Example 1. Let us consider the problem of minimizing the functional (2.1) on trajectories of the system (2.2), (2.3) with

$$V = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},\ W(t) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},\ S(t) = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ R(t) = 1,\ A(t) = \begin{pmatrix} 0 & t \\ 0 & 1 \end{pmatrix},$$

$$C(t) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},\ B(t) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Here $Q(t) = \frac{1}{1+t^2}\begin{pmatrix} 1 & -t \\ -t & t^2 \end{pmatrix}, P = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. The system (2.4)–(2.6) is specified to

$$\left(\begin{pmatrix} 0 & t \\ 0 & 1 \end{pmatrix} x(t)\right)' = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x(t) + \begin{pmatrix} 1 \\ 0 \end{pmatrix} u(t),$$

$$\begin{pmatrix} 0 & 0 \\ t & 1 \end{pmatrix} ((I - Q(t))\psi(t))' = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} x(t) - \left(\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ t & 1 \end{pmatrix} Q'(t)\right)\psi(t),$$

$$0 = (1\ \ 0)\psi(t) - u(t),$$

with the boundary conditions

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} x(0) = y_0 := \begin{pmatrix} 0 \\ y_{02} \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ T & 1 \end{pmatrix} \psi(T) = 0,\ y_{02} \neq 0.$$

Theorem 5.1 provides the special explicit differential system (5.1)

$$(Ax)' = \frac{1}{1+t^2} \begin{pmatrix} t & 1 \\ 0 & 0 \end{pmatrix} Ax, \tag{6.1}$$

$$((I - Q)\psi)' = -\frac{1}{(1+t^2)^2} \begin{pmatrix} t^3 - t & t^2 - 1 \\ 2t^2 & 2t \end{pmatrix} (I - Q)\psi. \tag{6.2}$$

In particular, (6.1) immediately leads to $x_2'(t) = 0$. Taking into account the initial condition, we obtain $x_2(t) \equiv y_{0_2}$. Due to the boundary condition $(I - Q(T))\psi(T) = 0$, the function $(I - Q(t))\psi(t)$ vanishes identically, i.e., $\psi(t) = Q(t)\psi(t)$, $A^*(t)\psi(t) = 0$; hence, $t\psi_1(t) + \psi_2(t) = 0$.

Next we derive from (2.5) that

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} x(t) = 0,$$

i.e., $x_1(t) \equiv 0$, and from (2.4), (2.6) that $u(t) \equiv 0$, $\psi_1(t) \equiv 0$; hence, $\psi(t) \equiv 0$.

Consequently, the BVP (2.4)–(2.6) has the unique solution

$$x_*(t) \equiv y_0, \quad \psi_*(t) \equiv 0, \quad u_*(t) \equiv 0,$$

and, by Lemma 2.2, $u_*(\cdot)$ is the optimal control for (2.1)–(2.3).

This confirms the result that we could have in this special case by much simpler considerations. Namely, by looking at the problem in detail,

$$J(u, x) = \frac{1}{2} \int_0^T \{x_1(t)^2 + u(t)^2\} dt,$$

$(tx_2(t))' = x_2(t) + u(t)$, $(x_2(t))' = 0, t \in [0, T], x_2(0) \neq 0$ given, we find immediately that

$$x_{*2}(t) \equiv x_2(0), \quad u_*(t) \equiv 0.$$

For obtaining a minimum we obviously have to put $x_{*1}(t) \equiv 0$. Let us stress once more that this special DAE (2.2) has even a singular pencil $\lambda A(t) - C(t)$, but the resulting DAE in (2.4)–(2.6) is index-1 tractable. In this case, Lemmas 4.1 and 4.7, and further Theorems 5.1 and 5.6 apply. However, since $\mathrm{Im} A(t) = \mathrm{Im}(I - Q(t))$ varies with time, the resulting system of explicit differential equations (6.1), (6.2) is obviously no longer a Hamiltonian one.

Remark 6.1. If the matrix $V = 0$ in Example 1 is replaced by

$$V = \alpha \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \ \alpha > 0,$$

then the functional

$$J(u, x) = \frac{\alpha}{2}(x_1(T) + x_2(T))^2 + \frac{1}{2} \int_0^T \{x_1(t)^2 + u(t)^2\} dt$$

is to be minimized on the trajectories as above. The resulting BVP (2.4)–(2.6) differs from that in Example 1 only by the condition

$$\begin{pmatrix} 0 & 0 \\ T & 1 \end{pmatrix} \psi(T) = -\alpha \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} x(T)$$

(which replaces the condition $\left(\begin{smallmatrix} 0 & 0 \\ T & 1 \end{smallmatrix}\right)\psi(T) = 0$). Now, because of $\alpha > 0$, the inclusion (2.8) fails to be true, whereas it is valid in Example 1. In particular, the relation

$$x_1(T) + x_2(T) = 0$$

has to be taken into account. On the other hand, as in Example 1, we have $x_2(t) \equiv x_2(0),\ x_1(t) \equiv 0$; thus $x_1(T) + x_2(T) = x_2(0) \neq 0$. Because of the resulting contradiction this BVP (2.4)–(2.6) is no longer solvable.

Considering the control problem itself, say for $x_2(0) = 1,\ T = 1$, we realize immediately that we have to put $x_2(t) \equiv 1, u(t) \equiv 0$ and then minimize $\tilde{J}(x_1) := \frac{\alpha}{2}(x_1(1)+1)^2 + \frac{1}{2}\int_0^1 x_1(t)^2 dt$ on the continuous functions $x_1(.)$. However, $\tilde{J}$ has the zero infimum, but there is no minimal element in this setting.

Another situation arises if we choose $V = \alpha I, \alpha > 0$, instead. Condition (2.8) fails again, but the boundary condition

$$\begin{pmatrix} 0 & 0 \\ T & 1 \end{pmatrix} \psi(T) = -\alpha x(T),$$

and, in particular, condition $x_1(T) = 0$ do not contradict (2.5). The resulting BVP (2.4)–(2.6) is solvable. The corresponding control problem with

$$J(u,x) = \frac{\alpha}{2}\{(x_1(T)^2 + x_2(T)^2\} + \frac{1}{2}\int_0^T \{x_1(t)^2 + u(t)^2\} dt$$

obviously has the solution $u_*(t) \equiv 0,\ x_{*1}(t) \equiv 0,\ x_{*2}(t) \equiv x_2(0)$.

Example 2. The functional $J(u,x) = \frac{1}{2}\int_0^T (x_1(t)^2 + u(t)^2) dt$ is to be minimized subject to the infinite-dimensional system

$$\begin{aligned} x_2'(t) &= x_1(t) + u(t), \\ x_i'(t) &= 0 \quad \text{for } i \geq 3, \end{aligned}$$

and the initial condition $x_i(0) = \frac{1}{i}$ for $i \geq 2$. Obviously, the minimum is obtained by

$$u_*(t) \equiv 0, \quad x_{*1}(t) \equiv 0, \quad x_{*i}(t) \equiv \frac{1}{i} \quad \text{for } i \geq 2.$$

Choosing $X = Y = l^2, U = \mathbb{R}$, as well as

$$A = \begin{pmatrix} 0\,1\,0\,\cdots \\ 0\,0\,1\,\cdots \\ \cdots \end{pmatrix},\ C = \begin{pmatrix} 1\,0\,\cdots \\ 0\,0\,\cdots \\ \cdots \end{pmatrix},\ B = \begin{pmatrix} 1 \\ 0 \\ \cdots \end{pmatrix},\ W = \begin{pmatrix} 1\,0\cdots \\ 0\,0\cdots \\ \cdots \end{pmatrix},\ y_0 = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{3} \\ \cdots \end{pmatrix},$$

$R = 1,\ S = 0,\ V = 0$, we put this problem into the form (2.1)–(2.3) with infinite-dimensional Hilbert spaces. Next we compute

$$P = \begin{pmatrix} 1\,0\,\cdots \\ 0\,0\,\cdots \\ \cdots \end{pmatrix} \quad \text{and} \quad Q = 0.$$

The resulting BVP (2.4), (2.5), (2.6) is

$$\begin{aligned} x_2'(t) &= x_1(t) + u(t), \\ x_i'(t) &= 0 \quad \text{for } i \geq 3,\ x_i(0) = \tfrac{1}{i} \ \text{ for } i \geq 2, \\ 0 &= x_1(t) - \psi_1(t), \\ \psi_i'(t) &= 0 \quad \text{for } i \geq 1,\ \psi_i(T) = 0 \ \text{ for } i \geq 1, \\ 0 &= \psi_1(t) - u(t). \end{aligned} \tag{6.3}$$

The explicit differential system (5.1) is also an infinite-dimensional one, namely,

$$(Ax(t))' = \begin{pmatrix} 2 & 0 & \dots \\ 0 & 0 & \dots \\ & \dots & \end{pmatrix} \psi(t), \quad \psi'(t) = 0, \tag{6.4}$$

or, equivalently, $x_2'(t) = 2\psi_1(t)$, $x_i'(t) = 0$ for $i \geq 3$, $\psi_i'(t) = 0$ for $i \geq 1$. This is a nonnegative Hamiltonian system. Due to the boundary conditions we obtain

$$x_{*i}(t) \equiv \frac{1}{i} \text{ for } i \geq 2, \quad \psi_{*i}(t) \equiv 0 \text{ for } i \geq 1,$$

as solutions of (6.4). Finally, (6.3) gives $u_*(t) \equiv \psi_{*1}(t) \equiv 0$ and $x_{*1}(t) \equiv \psi_{*1}(t) \equiv 0$.

7. Concluding remarks. Our paper refers to descriptor systems with coefficients $A(t), C(t)$, and $B(t)$ that are bounded linear operators for all t and depend continuously on t in the norm sense. In the meantime, certain ideas about abstract DAEs with unbounded coefficients $C(t)$ have evolved; they take into consideration, among other things, the case of so-called partial differential-algebraic systems (cf. [11]). A corresponding generalization of this case does not seem to be impossible; however, it will require further elaborate investigations.

Acknowledgments. The idea to write this paper arose during the visit of the first author to Humboldt University; therefore she is grateful to all persons that supported this visit. Our thanks are due to the unknown referees for their valuable comments which helped to improve our paper.

REFERENCES

[1] F. L. Lewis, *A survey of linear singular systems*, Circuits Systems Signal Process., 5 (1986), pp. 3–36.
[2] G. A. Kurina, *Singular perturbations of control problems with state equation unresolved with respect to derivative. A survey*, Izv. Ross. Akad. Nauk Tekhn. Kibernet., 4 (1992), pp. 20–48 (in Russian).
[3] L. V. Kantorovich and G. P. Akilov, *Funkcional'nyj analiz*, Nauka, Moscow, 1984 (in Russian).
[4] D. J. Bender and A. J. Laub, *The linear-quadratic optimal regulator for descriptor systems*, IEEE Trans. Automat. Control, 32 (1987), pp. 672–688.
[5] G. A. Kurina, *To theory of linear-quadratic control problems unresolved with respect to derivative*, in Funkcional'nye prostranstva i uravnenija matematicheskoi fiziki. Sbornik statei. Voronezh, izdatel'stvo VGU, 1988, pp. 25–32. See also the unpublished paper VINITI 755-B86 (1986) (in Russian).
[6] G. A. Kurina, *On linear-quadratic control problems for descriptor systems*, in Proceedings of the International Conference "Dynamical Systems: Stability, Control, Optimization" (DSSCO'98), Minsk, Belarus, 1998, pp. 170–172 (in Russian).
[7] K. Balla and R. März, *Linear differential algebraic equations of index* 1 *and their adjoint equations*, Results Math., 37 (2000), pp. 13–35.
[8] W. V. Petryshin, *On generalized inverses and on the uniform convergence of* $(I - \beta K)^n$ *with application to iterative methods*, J. Math. Anal. Appl., 18 (1967), pp. 417–439.
[9] P. Kunkel and V. Mehrmann, *The linear quadratic optimal control problem for linear descriptor systems with variable coefficients*, Math. Control Signals Systems, 10 (1997), pp. 247–264.
[10] V. Mehrmann, *The autonomous linear quadratic control problem*, Lecture Notes in Control and Inform. Sci. 163, Springer-Verlag, Berlin, 1991.
[11] R. März, *Differential algebraic systems anew*, Appl. Numer. Math., 42 (2002), pp. 315–335.
[12] G. A. Kurina, *Feed-back control for time-varying descriptor systems*, Systems Sci., 26 (2000), pp. 47–59.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 2078–2093

POLE PLACEMENT AND MATRIX EXTENSION PROBLEMS: A COMMON POINT OF VIEW*

MEEYOUNG KIM†, JOACHIM ROSENTHAL‡, AND XIAOCHANG ALEX WANG§

Dedicated to the memory of Meeyoung Kim, 1966–2001

Abstract. This paper studies a general inverse eigenvalue problem which generalizes many well-studied pole placement and matrix extension problems. It is shown that the problem corresponds geometrically to a so-called central projection from some projective variety. The degree of this variety represents the number of solutions the inverse problem has in the critical dimension. We are able to compute this degree in many instances, and we provide upper bounds in the general situation.

Key words. pole placement and inverse eigenvalue problems, matrix completion problems, Grassmann varieties, degree of a projective variety

AMS subject classifications. 14M15, 15A18, 93B55, 93B60

DOI. 10.1137/S0363012999354429

1. Introduction and motivational examples. Let $\mathbb{K}$ be an arbitrary field and consider matrices A of size $n \times n$ and matrices G, H of size $n \times m$. Let $Mat_{m\times n}$ be the vector space of all $m \times n$ matrices over $\mathbb{K}$, and let $\mathcal{L} \subset Mat_{m\times n}$ be a linear subspace of dimension d. This paper will be devoted to the following question.

Problem 1.1. Given an arbitrary monic polynomial $\varphi(s) \in \mathbb{K}[s]$ of degree n, is there a $F \in \mathcal{L}$ such that

$$\det[s(I + HF) - (A + GF)] = \varphi(s)? \tag{1.1}$$

We can think of the triple (A, G, H) representing a dynamical system in various ways, and we will say more about it in a moment. The set of monic polynomials of degree n can be identified with the vector space $\mathbb{K}^n$. If Problem 1.1 has a positive answer for all monic polynomials $\varphi(s) \in \mathbb{K}^n$ of degree n, then we will say that the system (A, G, H) is arbitrarily pole assignable in the class of feedback compensators $\mathcal{L}$. If for a generic set of monic polynomials $\varphi(s) \in \mathbb{K}^n$ of degree n Problem 1.1 has a positive answer, then we will say that system (A, G, H) is generically pole assignable in the class of feedback compensators $\mathcal{L}$.

Problem 1.1 covers a large set of "constrained" state and output pole placement problems. It also covers some important matrix extension problems. The following three examples will illustrate this.

Example 1.2 (constrained state feedback pole placement problem). Consider a linear system having the following form:

$$\dot{x} = Ax + Bu + H\dot{u}, \quad x \in \mathbb{C}^n,\ u \in \mathbb{C}^m. \tag{1.2}$$

*Received by the editors April 1, 1999; accepted for publication (in revised form) May 23, 2003; published electronically January 28, 2004.

http://www.siam.org/journals/sicon/42-6/35442.html

†The author is deceased. Former address: Department of Mathematics, Michigan State University, East Lansing, MI 48824.

‡Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556 (Rosenthal.1@nd.edu). The work of this author was supported in part by NSF grants DMS-96-10389 and DMS-00-72383.

§Department of Mathematics, Texas Tech University, Lubbock, TX 79409-1024 (alex.wang@ttu.edu).

If a state feedback law of the form $u = -Fx$ is applied, then the closed loop characteristic polynomial has the form $\det[s(I + HF) - (A - BF)]$ which is exactly the form of (1.1). Assume that not every feedback law of the form $u = -Fx$ can be applied, and that a valid feedback matrix must satisfy some linear constraints of its parameters, i.e., $F \in \mathcal{L}$ for some linear subspace $\mathcal{L} \subset Mat_{m\times n}$. Such constraints occur, e.g., if the ith input channel has to be kept zero all the time. Problem 1.1 covers such situations.

Of course if $H = 0$ and the feedback laws are unconstrained, then we simply deal with the classical state feedback problem, and it is well known that this problem always has a positive solution as soon as rank $\left[B, AB, \ \ldots \ A^{n-1}B\right] = n$.

Example 1.3 (constrained output feedback pole placement problem). Consider the linear system

$$\dot{x} = Ax + Bu, \ y = Cx, \quad x \in \mathbb{C}^n, \ u \in \mathbb{C}^m, \text{ and } y \in \mathbb{C}^p. \tag{1.3}$$

The problem asks for a static feedback law $u = Ky$ such that the closed loop system has some desired closed loop characteristic polynomial. Once more assume that the feedback laws have to satisfy some linear constraints. For this we assume that $\mathcal{U} \subset Mat_{m\times p}$ is some linear subspace, and we do require that $K \in \mathcal{U}$. One immediately verifies that Problem 1.1 covers this situation if one chooses $H := 0$, $G := B$, and $\mathcal{L} := \{KC \mid K \in \mathcal{U}\}$.

For the unconstrained problem ($\mathcal{U} = Mat_{m\times p}$) the main result in this area of research was given by Brockett and Byrnes [2]. It states the following.

THEOREM 1.4. *If $\mathcal{U} = Mat_{m\times p}$ and if $n \leq mp = \dim\mathcal{U}$, then for a generic set of matrices A, B, C, the system* (1.3) *is arbitrarily pole assignable. Moreover, if $n = mp$, then when counted with multiplicities there are exactly as many solutions as the degree of the complex Grassmann variety* $\mathrm{Grass}(m, \mathbb{C}^{m+p})$ *once embedded via the Plücker embedding.*

The importance of the matrix H becomes apparent if we deal with general proper systems of the form:

$$\dot{x} = Ax + Bu, \ y = Cx + Du, \quad x \in \mathbb{C}^n, \ u \in \mathbb{C}^m, \text{ and } y \in \mathbb{C}^p.$$

Assume C has rank p, and let C^+ be a right inverse of C, i.e., an $n \times p$ matrix such that $CC^+ = I_p$. Assume again that the feedback laws are constrained to some subspace $\mathcal{U} \subset Mat_{m\times p}$. If we apply the feedback law $u = Ky$, $K \in \mathcal{U}$, then we obtain the closed loop system:

$$\dot{x} = Ax + BKy, \ y = Cx + DKy, \quad x \in \mathbb{C}^n, \text{ and } y \in \mathbb{C}^p.$$

The closed loop characteristic polynomial is therefore computed as

$$\begin{aligned}
&\det \begin{bmatrix} sI_n - A & -BK \\ -C & I_p - DK \end{bmatrix} \\
&= \det(sI_n - A) \det\left(I_p - DK - C(sI_n - A)^{-1}BK\right) \\
&= \det(sI_n - A) \det\left(I_n - C^+DKC - (sI_n - A)^{-1}BKC\right) \\
&= \det\left((sI_n - A)(I_n - C^+DKC) - BKC\right) \\
&= \det\left(s(I_n - C^+DKC) - (A + (B - AC^+D)KC\right).
\end{aligned}$$

If one defines $H := -C^+D$, $G := (B - AC^+D)$, and $\mathcal{L} := \{KC \mid K \in \mathcal{U}\}$ one immediately verifies that the output feedback pole placement problem involving proper transfer functions is also covered by Problem 1.1 and that in this case it is of importance to have the matrix H in the formulation of Problem 1.1.

Example 1.5 (matrix extension problems). Let $m = n$, $H = 0$, and $G = I_n$. In this case Problem 1.1 asks for conditions which guarantee that the characteristic map

$$\chi_A : \mathcal{L} \longrightarrow \mathbb{K}^n, \quad F \longmapsto \det(sI + A + F), \tag{1.4}$$

is surjective or at least "generically" surjective. This general matrix extension problem itself contains many of the matrix completion problems as they were studied in [1, 4, 5, 7].

The main result in the situation of Example 1.5 has been derived in [11]. It states the following.

THEOREM 1.6. *If the base field $\mathbb{K}$ is algebraically closed, then for a generic set of matrices $A \in Mat_{n\times n}$ the characteristic map* (1.4) *is dominant (generically surjective) if and only if*

1. $\dim \mathcal{L} \geq n$;
2. *there must be at least one element $L \in \mathcal{L}$ whose trace $tr(L) \neq 0$, i.e., $\mathcal{L} \not\subset sl_n$.*

The main results of this paper will show that if $\mathbb{K}$ is algebraically closed, then for a generic set of matrices (A, G, H) Problem 1.1 is solvable in a projective closure of $\mathcal{L}$ for every $\varphi(s)$ if and only if $\dim \mathcal{L} \geq n$ (Theorem 2.7), and if $\dim \mathcal{L} = n$, then there are at most $\min(m, n)^n$ solutions for each $\varphi(s)$ for each subspace $\mathcal{L}$ (Theorems 2.8, 4.3, and 4.8).

The paper is structured as follows: In section 2 we will introduce a natural compactification of the linear space $\mathcal{L}$ which we will denote by $\bar{\mathcal{L}}$. In order to prove the main theorems we will show that one has a characteristic map χ defined on a Zariski open set of the variety $\bar{\mathcal{L}}$. Geometrically χ describes a central projection from the variety $\bar{\mathcal{L}}$ to the projective space $\mathbb{P}^n$. As a consequence the number of solutions in the critical dimension, i.e., in the situation where $\dim \mathcal{L} = n$, is equal to $\deg \bar{\mathcal{L}}$ when counted with multiplicities and when some possible "infinite solutions" are taken into account. The results in section 2 generalize mathematical ideas which have been developed for the static pole placement problem by Brockett and Byrnes [2] and for the dynamic pole placement problem by Ravi, Rosenthal, and Wang [14], Rosenthal [15], and Rosenthal and Wang [16].

The degree of the variety $\bar{\mathcal{L}}$ is of crucial importance for the understanding of the characteristic map χ. In section 3 we compute the degree of $\bar{\mathcal{L}}$ in many special cases. As a corollary we will rediscover several matrix completion results as they were derived earlier in [3, 4, 5, 7].

In section 4 we will be concerned with the value of the "generic degree"; this is the largest possible degree a variety $\bar{\mathcal{L}}$ of a fixed dimension can have. We determine an upper bound for the generic degree in the case when $d = n$ and prove that this bound is reached when $m = n < 5$.

2. Compactification of the problem. The inverse eigenvalue problem formulated in Problem 1.1 describes an intersection problem in the linear variety $\mathcal{L}$. In order to invoke results from intersection theory [6] it is important to understand the intersection at the "boundary" of $\mathcal{L}$. What is needed is a good compactification of $\mathcal{L}$. It turns out that Problem 1.1 induces in a natural way a compactification, and we will explain this in what follows.

Recall [8, 9] that the Grassmannian $\mathrm{Grass}(m, \mathbb{K}^q)$, $m \leq q$, is the set of all m-dimensional subspaces in $\mathbb{K}^q$. For each $X \in \mathrm{Grass}(m, \mathbb{K}^q)$, let $X = \mathrm{span}\,\{\alpha_1, \ldots, \alpha_m\}$; then

$$\alpha_1 \wedge \alpha_2 \wedge \cdots \wedge \alpha_m$$

is a vector in the exterior product $\wedge^m \mathbb{K}^q \cong \mathbb{K}^{\binom{q}{m}}$, and

$$\mathrm{span}\,\{\alpha_1, \ldots, \alpha_m\} = \mathrm{span}\,\{\beta_1, \ldots, \beta_m\}$$

if and only if

$$\alpha_1 \wedge \cdots \wedge \alpha_m = k(\beta_1 \wedge \cdots \wedge \beta_m)$$

for some $k \neq 0$ in $\mathbb{K}$. Therefore, by considering the components of $\alpha_1 \wedge \cdots \wedge \alpha_m$ as homogeneous coordinates of a point in $\mathbb{P}^{\binom{q}{m}-1}$, we have an embedding

$$\mathrm{Grass}(m, \mathbb{K}^q) \subset \mathbb{P}\left(\wedge^m \mathbb{K}^q\right) = \mathbb{P}^{\binom{q}{m}-1},$$

which is called Plücker embedding, and $\alpha_1 \wedge \cdots \wedge \alpha_m$ are called Plücker coordinates of $X \in \mathrm{Grass}(m, \mathbb{K}^q)$. If $\{\alpha_i\}$ are row vectors, then the Plücker coordinates of $\mathrm{span}\,\{\alpha_1, \ldots, \alpha_m\}$ are given by all the full size minors of the $m \times q$ matrix

$$\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix}.$$

The closed loop characteristic polynomial can be written as

$$\det\left[s(I + HF) - (A + GF)\right] = \det \begin{bmatrix} I_m & F \\ -sH + G & sI - A \end{bmatrix}. \tag{2.1}$$

Following an idea introduced by Brockett and Byrnes [2] for the static output pole placement problem we will identify $\mathrm{rowsp}\,[I_m\ F]$ with an element of $\mathrm{Grass}(m, \mathbb{K}^{m+n})$. In this way we have natural embeddings

$$\mathcal{L} \subset \mathrm{Grass}(m, \mathbb{K}^{m+n}) \subset \mathbb{P}\left(\wedge^m \mathbb{K}^{m+n}\right) = \mathbb{P}^N, \quad N = \binom{m+n}{m} - 1.$$

Definition 2.1. *Let $\bar{\mathcal{L}}$ be the projective closure of $\mathcal{L}$.*

By definition $\bar{\mathcal{L}}$ is a projective variety of dimension $\dim \bar{\mathcal{L}} = \dim \mathcal{L} = d$. The remainder of the paper will be devoted to a large extent to the study of this variety. In order to have a general idea of how the projective closure of $\mathcal{L}$ is defined, we start with an illustrative example.

Example 2.2. Let $m = n = d = 3$, and let $\mathcal{L} \subset Mat_{3\times 3}$ be defined by

$$\mathcal{L} = \left\{ \begin{bmatrix} a & b & 0 \\ c & a & b \\ 0 & c & a \end{bmatrix} \right\}, \tag{2.2}$$

where $a,\ b,\ c \in \mathbb{K}$ are arbitrary elements. Then for fixed a, b, c

$$\mathrm{rowsp} \begin{bmatrix} 1 & 0 & 0 & a & b & 0 \\ 0 & 1 & 0 & c & a & b \\ 0 & 0 & 1 & 0 & c & a \end{bmatrix} \tag{2.3}$$

is a point in $\mathrm{Grass}(3,\mathbb{K}^6)$. Let z_{ijk} be the full size minor of (2.3) consisting of the ith, jth, kth columns. Then $\{z_{ijk}\}$ are the Plücker coordinates of $\mathrm{Grass}(3,\mathbb{K}^6)$ in $\mathbb{P}^{\binom{6}{3}-1}=\mathbb{P}^{19}$. $\mathcal{L}$ is defined by 6 linear equations of its entries. In terms of the Plücker coordinates, they become

$$\begin{aligned} &z_{234}=-z_{135}, \quad z_{234}=z_{126}, \quad z_{235}=-z_{136},\\ &z_{125}=-z_{134}, \quad z_{124}=0, \qquad z_{236}=0. \end{aligned} \tag{2.4}$$

$\mathcal{L}$ has 9 minors of size 2×2, but there are only 6 monomials of degree 2 of a,b,c:

$$a^2, b^2, c^2, ab, ac, bc.$$

So there are 3 linear relations among the 2×2 minors. In terms of the Plücker coordinates, they are

$$z_{146}=-z_{245}, \quad z_{345}=z_{156}, \quad z_{346}=-z_{256}. \tag{2.5}$$

The monomials $a^2, b^2, c^2, ab, ac, bc$ are not algebraically independent; they satisfy the relation

$$\operatorname{rank}\begin{bmatrix} a^2 & ab & ac\\ ab & b^2 & bc\\ ac & bc & c^2 \end{bmatrix} = \operatorname{rank}\begin{bmatrix} a\\ b\\ c \end{bmatrix}\begin{bmatrix} a & b & c \end{bmatrix} \le 1; \tag{2.6}$$

i.e., all the 2×2 minors of (2.6) are zero, which induce 6 quadratic relations among the 2×2 minors of $\mathcal{L}$:

$$\begin{aligned} &z_{346}^2+z_{246}z_{356}=0, && z_{146}^2+z_{246}z_{145}=0,\\ &(z_{246}+z_{345})^2-z_{356}z_{145}=0, && z_{246}(z_{246}+z_{345})-z_{346}z_{146}=0,\\ &z_{346}(z_{246}+z_{345})+z_{356}z_{146}=0, && z_{146}(z_{246}+z_{345})+z_{145}z_{346}=0. \end{aligned} \tag{2.7}$$

Every point in $\mathrm{Grass}(3,\mathbb{K}^6)$ defined by $z_{123}=0$ and (2.4), (2.5), (2.7) is indeed a limit point of (2.3); therefore $\bar{\mathcal{L}}$ is defined by (2.4), (2.5), and (2.7) in $\mathrm{Grass}(3,\mathbb{K}^6)\subset\mathbb{P}^{19}$.

Note that every element in $\bar{\mathcal{L}}$ can be simply represented by a subspace of the form $\operatorname{rowsp}[F_1\ F_2]$, where the $m\times m$ matrix F_1 is not necessarily invertible. Row span $[F_1\ F_2]$ describes an element of $\mathcal{L}$ if and only if F_1 is invertible. Note that a characteristic equation is even defined if F_1 is singular unless the polynomial in (2.1) is the zero polynomial.

Let $f_i,\ i=0,\dots,N$, be the Plücker coordinates of $\operatorname{rowsp}[F_1\ F_2]$. In terms of the Plücker coordinates the characteristic equation can then be written as

$$\det\begin{bmatrix} F_1 & F_2\\ -sH+G & sI-A \end{bmatrix} = \sum_{i=0}^{N} f_i p_i(s), \tag{2.8}$$

where the $p_i(s)$ is the cofactor of f_i in the determinant (2.8).

Let $\mathcal{Z}\subset\mathbb{P}^N$ be the linear subspace defined by

$$\mathcal{Z}=\left\{ z\in\mathbb{P}^N \,|\, \sum_{i=0}^{N} p_i(s) z_i = 0 \right\}. \tag{2.9}$$

Following [14, 15, 18] we identify a closed loop characteristic polynomial $\varphi(s)$ with a point in $\mathbb{P}^n$. In analogy to the situation of the static pole placement problem considered in [2, 18] (compare also with [15, section 5]) one has a well-defined characteristic map

$$\begin{array}{rccc} \chi: & \bar{\mathcal{L}} - \mathcal{Z} & \longrightarrow & \mathbb{P}^n, \\ & \operatorname{rowsp}[F_1 \ F_2] & \longmapsto & \displaystyle\sum_{i=0}^{N} f_i p_i(s). \end{array} \tag{2.10}$$

It will turn out that surjectiveness of the map χ will imply the generic pole assignability of system (A, G, H) in the class of compensators $\mathcal{L}$.

Recall the notion of degree of a variety [10, Chapter I, section 7] and the notion of a central projection (see [17, Chapter I, section 4]). The geometric properties of the map χ are as follows.

THEOREM 2.3. *The map χ defines a central projection. In particular, if $\mathcal{Z} \cap \bar{\mathcal{L}} = \emptyset$ and $\dim \mathcal{L} = n$, then χ is surjective, and there are $\deg \bar{\mathcal{L}}$ many preimages (counted with multiplicity) for each point in $\mathbb{P}^n$, where $\deg \bar{\mathcal{L}}$ is the degree of the projective variety $\bar{\mathcal{L}}$ in $\mathbb{P}^N$.*

The proof for this theorem is identical to the one given in [14, 18]. In the algebraic geometry literature (see, e.g., [9, 13, 17]) χ is sometimes referred to as a *projection of $\mathcal{L}$ from the center $\mathcal{Z}$ to $\mathbb{P}^n$*, and $\mathcal{Z} \cap \bar{\mathcal{L}}$ is sometimes referred to as the *base locus*. Of course the interesting part of the theorem occurs when $\mathcal{Z} \cap \bar{\mathcal{L}} = \emptyset$ since in this situation very specific information on the number of solutions is provided. If $\mathcal{Z} \cap \bar{\mathcal{L}} = \emptyset$ and $\dim \mathcal{L} = n$, then one says that χ describes a *finite morphism* from the projective variety $\bar{\mathcal{L}}$ onto the projective space $\mathbb{P}^n$.

In analogy to the situation of the static pole placement problem [2, 18] and the dynamic pole placement problem [15] we introduce a definition for this important situation.

DEFINITION 2.4. *A particular system (A, G, H) is called* $\mathcal{L}$-nondegenerate *if $\mathcal{Z} \cap \bar{\mathcal{L}} = \emptyset$. A system which is not $\mathcal{L}$-nondegenerate will be called $\mathcal{L}$-degenerate.*

In general it will always happen that certain systems A, G, H are $\mathcal{L}$-degenerate. We first make a definition.

DEFINITION 2.5. *Let X be an arbitrary (affine or projective) variety. A subset $S \subset X$ is called a* generic set *of X if it contains a nonempty Zariski open set of X.*

The next theorem shows that if the dimension of $\mathcal{L}$ is not too large, then the set of systems A, G, H which are $\mathcal{L}$-degenerate are contained in a proper algebraic subset when viewed as a subset in the vector space $\mathbb{K}^{(n^2+2mn)}$.

LEMMA 2.6. *Assume the base field $\mathbb{K}$ is algebraically closed. If $\dim \mathcal{L} > n$, then every system A, G, H is $\mathcal{L}$-degenerate. If $\dim \mathcal{L} \leq n$, then a generic set of systems A, G, H is $\mathcal{L}$-nondegenerate.*

Proof. If $\dim \mathcal{L} > n$, then $\mathcal{Z} \cap \bar{\mathcal{L}}$ is nonempty by the (projective) dimension theorem (see, e.g., [10, Chapter I, Theorem 7.2]) and the fact that $\dim \mathcal{Z} \geq N - n - 1$.

Assume now that $\dim \mathcal{L} \leq n$. Consider

$$\det \begin{bmatrix} I_m & F \\ -sH+G & sE-A \end{bmatrix}, \tag{2.11}$$

and identify the set of matrices E, A, G, H with the vector space $\mathbb{K}^{2n(m+n)}$. In analogy to the proof of [15, Lemma 5.3] we compute the dimension of the coincidence set

$$\mathcal{S} := \left\{ (F_1, F_2;\ E, A, G, H) \in \bar{\mathcal{L}} \times \mathbb{K}^{2n(m+n)} \mid \det \begin{bmatrix} F_1 & F_2 \\ -sH+G & sE-A \end{bmatrix} = 0 \right\}.$$

Using the same arguments as in [15] one computes

$$\dim \mathcal{S} = \dim \bar{\mathcal{L}} + 2n(m+n) - n - 1.$$

Since $\bar{\mathcal{L}}$ is projective the projection onto the second factor (namely, $\mathbb{K}^{2n(m+n)}$) is an algebraic set by the main theorem of elimination theory (see, e.g., [13]). This projection can result in an algebraic set of dimension at most $\dim \mathcal{S} < 2n(m+n)$. So for generic matrices E, A, G, H, we have $\det E \neq 0$ and

$$\det \begin{bmatrix} F_1 & F_2 \\ -sH+G & sE-A \end{bmatrix} \neq 0$$

for all $[F_1 \quad F_2]$ in $\bar{\mathcal{L}}$. For such matrices $\{E, A, G, H\}$, the systems $\hat{A} = E^{-1}A, \hat{G} = E^{-1}G, \hat{H} = E^{-1}H$ are $\mathcal{L}$-nondegenerate, and the claim therefore follows. □

We are now in a position to state one of the main theorems of this paper.

THEOREM 2.7. *Assume the base field $\mathbb{K}$ is algebraically closed. Let $\mathcal{L} \subset Mat_{m\times n}$ be a fixed subspace. Then the map χ introduced in* (2.10) *is surjective for a generic set of matrices A, G, H if and only if $\dim \mathcal{L} \geq n$. If $\dim \mathcal{L} = n$, then for a generic set of matrices A, G, H the intersection $\mathcal{Z} \cap \bar{\mathcal{L}} = \emptyset$ and there are $\deg \bar{\mathcal{L}}$ many preimages (counted with multiplicity) for each point in $\mathbb{P}^n$.*

Proof. If $\dim \mathcal{L} < n$, then a simple dimension argument shows that χ cannot be surjective. The result for $\dim \mathcal{L} = n$ follows from Lemma 2.6 and Theorem 2.3. If $\dim \mathcal{L} > n$, we can always choose a subspace of $\mathcal{L}$ of dimension n. It follows that χ is surjective as soon as $\dim \mathcal{L} \geq n$. □

Theorem 2.7 gives a partial answer to Problem 1.1. The following result makes this clear.

THEOREM 2.8. *Let $\mathcal{L} \subset Mat_{m\times n}$ be a fixed subspace, and assume that $\mathbb{K}$ is algebraically closed. If $\dim \mathcal{L} < n$, then for almost all monic polynomials $\varphi(s) \in \mathbb{K}[s]$ of degree n there does not exist a $F \in \mathcal{L}$ such that* (1.1) *holds true.*

If $\dim \mathcal{L} \geq n$, then for a generic set of matrices A, G, H and a generic set of monic polynomials $\varphi(s)$ of degree n Problem 1.1 *has a solution.*

Finally, if $\dim \mathcal{L} = n$, then for a generic set of matrices A, G, H and a generic set of polynomials the number of solutions is always finite, and when counted with multiplicities there are exactly $\deg \bar{\mathcal{L}}$ solutions.

Proof. The only statement which does not immediately follow from Theorem 2.7 is the claim about the number of solutions in the critical dimension. We therefore assume that $\dim \mathcal{L} = n$. Consider the characteristic map χ introduced in (2.10). According to Theorem 2.7, χ is a finite morphism of degree $\bar{\mathcal{L}}$. For a generic set of polynomials $\varphi(s) \in \mathbb{K}^n \subset \mathbb{P}^n$ the inverse image $\chi^{-1}(\varphi(s))$ contains $\deg\left(\bar{\mathcal{L}}\right)$ different solutions, and all these solutions are contained in $\mathcal{L} \subset \bar{\mathcal{L}}$. □

Theorems 2.7 and 2.8 assume that the field is algebraically closed. For general fields it is often possible to deduce some results by considering the corresponding question over the algebraic closure. The following result is of this sort.

COROLLARY 2.9. *If the degree of the variety $\bar{\mathcal{L}}$ defined over the complex numbers $\mathbb{C}$ is odd and if $\dim \mathcal{L} \geq n$, then χ is also surjective over the real numbers $\mathbb{R}$ for a generic set of real matrices A, G, H.*

Proof. We will view a triple of complex matrices A, G, H as a point in $\mathbb{C}^{n(2m+n)}$. If $d := \dim \mathcal{L} = n$, then χ represents a finite morphism for a generic set of complex matrices A, G, H by Theorem 2.7. Since the subset of real matrices inside $\mathbb{C}^{n(2m+n)}$ is not contained in any algebraic set, χ is even a finite morphism for a generic set of real matrices A, G, H. Over the complex numbers the inverse image $\chi^{-1}(y) \subset \bar{\mathcal{L}}$

represents a finite set of complex conjugate points for every real point $y \in \mathbb{P}^n$; in particular, $\chi^{-1}(y)$ contains a real point for a generic set of real matrices A, G, H.

If $d < n$ we follow the reasoning in the proof of [14, Theorem 2.14]: for a generic set of real matrices A, G, H one has

$$\dim \mathcal{Z} \cap \bar{\mathcal{L}} = \dim \mathcal{L} - n - 1 = d - n - 1.$$

If this dimension formula holds choose a subspace $H \subset \mathbb{P}^N$ having codimension $d-n$ inside $\mathbb{P}^N$ and having the property that

$$\bar{\mathcal{L}} \cap \mathcal{Z} \cap H = \emptyset. \tag{2.12}$$

Such a subspace H exists by [13, Corollary (2.29)]. Let $\pi_1 : \bar{\mathcal{L}} \to \mathbb{P}^d$ be the central projection with center $\mathcal{Z} \cap H$, and let $\pi_2 : \mathbb{P}^d - \pi_1(\mathcal{Z}) \to \mathbb{P}^n$ be the central projection with center $\pi_1(\mathcal{Z})$. Then π_1 is a finite morphism which is surjective over $\mathbb{C}$. π_2 is a linear map, it is surjective as well, and

$$\chi = \pi_2 \circ \pi_1.$$

If the degree of $\bar{\mathcal{L}}$ is odd, then π_1 is surjective over the reals, and the claim follows. □

3. The degree of $\bar{\mathcal{L}}$ in some special situations. From Theorems 2.7 and 2.8 it became clear that $\deg \bar{\mathcal{L}}$ is equal to the number of solutions for Problem 1.1 in the critical dimension, at least generically. In this and in the next section we will compute $\deg \bar{\mathcal{L}}$ in many situations.

For the rest of the paper we will assume that $\mathbb{K}$ is an algebraically closed field of characteristic zero. We will show in a moment that the compactification $\bar{\mathcal{L}}$ is, in many cases, isomorphic to the product of some Schubert varieties. This will allow us to compute the degree of $\bar{\mathcal{L}} \subset \mathbb{P}^N$ in these cases.

For the convenience of the reader we summarize the basic notions. More details can be found in [12, 16] and [6, Chapter 14].

Consider a flag of linear subspaces

$$\mathcal{F} : \ \{0\} \subset V_1 \subset V_2 \subset \cdots \subset V_{m+n} = \mathbb{K}^{m+n},$$

where we assume that $\dim V_q = q$ for $q = 1, \ldots, m+n$. Let $\nu = (\nu_1, \ldots, \nu_m)$ be an ordered index set satisfying

$$1 \le \nu_1 < \cdots < \nu_m \le m+n.$$

With respect to the flag $\mathcal{F}$ one defines the Schubert variety

$$S(\nu_1, \ldots, \nu_m) := \{W \in \mathrm{Grass}(m, \mathbb{K}^{m+n}) \mid \dim(W \bigcap V_{\nu_k}) \ge k \ \text{ for } k = 1, \ldots, m\} \tag{3.1}$$

and the Schubert cell

$$C(\nu_1, \ldots, \nu_m) := \{W \in S(\nu_1, \ldots, \nu_m) \mid \dim(W \bigcap V_{\nu_k - 1}) = k - 1 \text{ for } k = 1, \ldots, m\}. \tag{3.2}$$

The closure of the Schubert cell $C(\nu_1, \ldots, \nu_m)$ inside the variety $\mathrm{Grass}(m, \mathbb{K}^{m+n}) \subset \mathbb{P}^N$ is equal to the Schubert variety $S(\nu_1, \ldots, \nu_m)$. By definition, $S(\nu_1, \ldots, \nu_m)$ is

a projective variety. There is a well-known formula for the degree of a Schubert variety [12, Chapter XIV, section 6, equation (7)]:

$$\deg S(\nu_1,\ldots,\nu_k) = \left(\sum_i(\nu_i - i)\right)!\frac{\prod_{j>i}(\nu_j-\nu_i)}{\prod_i(\nu_i-1)!}.$$

Let $\mathcal{B} := \{v_1,\ldots,v_{m+n}\} \subset \mathbb{K}^{m+n}$ be a basis which is compatible with the flag $\mathcal{F}$. In other words this basis has the property that $V_i = \operatorname{span}(v_1,\ldots,v_i)$. With respect to the basis $\mathcal{B}$ one can represent the Schubert cell $C(\nu_1,\ldots,\nu_m)$ as the set of all m-dimensional subspaces in $\mathbb{K}^{m+n}$ which are the rowspaces of a matrix of the form

$$\begin{bmatrix} * & \cdots & * & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ * & \cdots & * & 0 & * & \cdots & * & 1 & \cdots & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ * & \cdots & * & 0 & * & \cdots & * & 0 & \cdots & * & \cdots & * & 1 & 0 & \cdots & 0 \end{bmatrix}, \tag{3.3}$$

where the 1's are in the columns $\nu_1,\ldots,\nu_m$.

The cell $C(\nu_1,\ldots,\nu_m)$ is isomorphic to $\mathbb{K}^d$, where $d = \sum_{i=0}^m(\nu_i - i)$. In particular, the cell $C(\nu_1,\ldots,\nu_m)$ is isomorphic to every subspace $\mathcal{L} \subset Mat_{m\times n}$ having dimension $\dim\mathcal{L} = d$. In general it is not true that the closures $S(\nu_1,\ldots,\nu_m) \subset \mathbb{P}^N$ and $\bar{\mathcal{L}} \subset \mathbb{P}^N$ are isomorphic. This happens, however, in the following situation.

Let $E_{i,j}$ be the $m \times n$ matrix whose i,j-entry is 1 and all the other entries are 0. Let $\mu = (\mu_1,\ldots,\mu_m)$ be an ordered index set satisfying

$$0 \le \mu_1 \le \cdots \le \mu_m \le n.$$

Definition 3.1. *$\mathcal{L} \subset Mat_{m\times n}$ is called a* lower left filled linear space *of type μ if $\mathcal{L}$ is spanned by the matrices*

$$E_{i,j} \quad \textit{for} \quad j \le \mu_i,\ i = 1,\ldots,m.$$

Lemma 3.2. *If $\mathcal{L} \subset Mat_{m\times n}$ is a lower left filled linear space of type μ, then $\bar{\mathcal{L}}$ is isomorphic to the Schubert variety $S(\mu_1+1,\mu_2+2,\ldots,\mu_m+m)$.*

Proof. Let $\nu_i := \mu_i + 1$, $i = 1,\ldots,m$. There is a fixed $(m+n)\times(m+n)$ permutation matrix P such that the set

$$\{[I_m \;\; F]\,P \mid F \in \mathcal{L}\} \subset Mat_{m\times(m+n)}$$

is equal to the cell $C(\nu_1,\ldots,\nu_m)$ described in (3.3). The linear transformation $P \in Gl_{m+n}$ extends to a linear transformation in $\mathbb{P}(\wedge^m\mathbb{K}^{m+n}) = \mathbb{P}^N$, and this linear transformation maps $\bar{\mathcal{L}}$ isomorphically onto $S(\nu_1,\ldots,\nu_m)$. ∎

The proof of the lemma shows in particular that permutations of the columns inside $Mat_{m\times n}$ result in isomorphic compactifications. The following lemma shows that a broader range of transformations do not change the topological properties of the compactification.

Lemma 3.3. *Assume there are subspaces $\mathcal{L}_1,\mathcal{L}_2 \subset Mat_{m\times n}$. If there are linear transformations $S \in Gl_m$ and $T \in Gl_n$ such that $\mathcal{L}_2 = S\mathcal{L}_1T^{-1}$, then there exists an automorphism of $\mathbb{P}^N$ which maps the compactification $\overline{\mathcal{L}_1}$ isomorphically onto the compactification $\overline{\mathcal{L}_2}$.*

Proof.

$$\begin{bmatrix} I_m & S\mathcal{L}_1T^{-1} \end{bmatrix} = S\begin{bmatrix} I_m & \mathcal{L}_1 \end{bmatrix}\begin{bmatrix} S^{-1} & 0 \\ 0 & T^{-1} \end{bmatrix}.$$

The matrix to the right, an element of GL_{m+n}, induces a linear transformation on the projective space $\mathbb{P}(\wedge^m \mathbb{K}^{m+n}) = \mathbb{P}^N$ which maps $\overline{\mathcal{L}_1}$ onto $\overline{\mathcal{L}_2}$. □

THEOREM 3.4. *Assume there are linear transformations $S \in Gl_m$ and $T \in Gl_n$ such that*

$$S\mathcal{L}T^{-1} = \begin{pmatrix} \mathcal{L}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathcal{L}_k \end{pmatrix},$$

where each $\mathcal{L}_l, l = 1, \ldots, k$, is the space of $m_l \times n_l$ lower left filled matrices of type μ^l:

$$0 \le \mu_1^l \le \cdots \le \mu_{m_l}^l \le n_l.$$

Then $\bar{\mathcal{L}}$ is isomorphic to the product of Schubert varieties

$$S(\mu_1^1 + 1, \mu_2^1 + 2, \ldots, \mu_{m_1}^1 + m_1) \times \cdots \times S(\mu_1^k + 1, \mu_2^k + 2, \ldots, \mu_{m_k}^k + m_k)$$

and

$$\deg \bar{\mathcal{L}} = \left(\sum_{i,l} \mu_i^l \right)! \frac{\prod_{i,l_r > l_s} \left(\mu_i^{l_r} + l_r - \mu_i^{l_s} - l_s \right)}{\prod_{i,l} (\mu_i^l + l - 1)!}.$$

Proof. The closure of $[I_{m_l} \;\; \mathcal{L}_l]$ in the Grassmann variety $\mathrm{Grass}(m_l, \mathbb{K}^{m_l+n_l})$ is the Schubert variety $S(\mu_1^l + 1, \ldots, \mu_{m_l}^l + m_l)$, and $\bar{\mathcal{L}}$ is a product of Schubert varieties.

The degree formula of a product of projective varieties under the Segre embedding [19, Proposition 2.1] is given by

$$\deg Z_1 \times \cdots \times Z_k = \frac{(\sum_i \dim Z_i)!}{\prod_i (\dim Z_i)!} \prod_i \deg Z_i.$$

Combining these formulas gives the result. □

COROLLARY 3.5. *When $\mu_1^1 = \cdots = \mu_{m_1}^1 = n_1$ and $\mu_i^l = 0$ for $l > 1$, then the compactification $\bar{\mathcal{L}} = \mathrm{Grass}(m_1, \mathbb{K}^{m_1+n_1})$ and its degree is*

$$\frac{(m_1 n_1)! 1! 2! \cdots (m_1 - 1)!}{n_1! (n_1 + 1)! \cdots (n_1 + m_1 - 1)!}.$$

Using Lemma 3.3 and Corollary 3.5 we can deduce Theorem 1.4, the result of Brockett and Byrnes. For this assume that $H = 0$, $G = I$, and $\mathcal{L} = \{BFC \mid F \in Mat_{m\times p}\}$. Without loss of generality we can assume that B, C have full rank, $\operatorname{rank} B = m$, and $\operatorname{rank} C = p$. (Theorem 1.4 assumes genericity!) There are invertible matrices S, T such that $SB = \left[\begin{smallmatrix} I_m \\ 0 \end{smallmatrix}\right]$ and $CT^{-1} = \left[\begin{smallmatrix} I_p \\ 0 \end{smallmatrix}\right]$. It follows that

$$S\mathcal{L}T^{-1} = \left\{ \begin{bmatrix} F & 0 \\ 0 & 0 \end{bmatrix} \in Mat_{n\times n} \mid F \in Mat_{m\times p} \right\}.$$

According to Lemma 3.3 and Corollary 3.5 the compactification is isomorphic to the Grassmannian $\mathrm{Grass}(m, \mathbb{K}^{m+p})$ as predicted by Theorem 1.4. In order to fully prove

Theorem 1.4 it remains to be shown that for a generic set of matrices $A \in Mat_{n\times n}$ the system is $\mathcal{L}$-nondegenerate as soon as $n = mp$.

COROLLARY 3.6. *When $m_l = n_l = 1$ and $\mu_1^l = 1$ for all l, then $\bar{\mathcal{L}} = \prod_{i=1}^n \mathbb{P}^1 = \mathbb{P}^1 \times \cdots \times \mathbb{P}^1$ and its degree is*

$$n!.$$

Corollary 3.6 covers a result first studied by Friedland [4, 5]. Indeed the subspaces $\mathcal{L} \subset Mat_{n\times n}$ correspond in this case exactly to the set of diagonal matrices. By Theorem 2.7 we know that for a generic set of matrices A, G, H the characteristic map χ is a finite morphism of mapping degree $n!$. Friedland [4, 5] and Byrnes and Wang [3] did show that the set of all matrices of the form $A, I_n, 0$ belongs to this generic set. We therefore have the following result.

THEOREM 3.7 (see [3, 4, 5]). *Let $\mathcal{L} \subset Mat_{n\times n}$ be the set of all diagonal matrices defined over an algebraically closed field $\mathbb{K}$. If $A \in Mat_{n\times n}$ is an arbitrary matrix and $\varphi \in \mathbb{K}[s]$ is an arbitrary monic polynomial of degree n, then there are exactly $n!$ diagonal matrices $F \in \mathcal{L}$ (when counted with multiplicity) such that the matrix $A + F$ has characteristic polynomial $\varphi(s)$.*

4. The degree of $\bar{\mathcal{L}}$ in the generic situation. In the previous section we computed the degree of the variety $\bar{\mathcal{L}}$ in many special cases. The set of all subspaces $\mathcal{L} \subset Mat_{m\times n}$ having the property that $\dim \mathcal{L} = d$ can be identified with the Grassmannian variety $\mathrm{Grass}(d, \mathbb{K}^{mn})$. The degree in our concern attains its maximal value on a Zariski open subset of $\mathrm{Grass}(d, \mathbb{K}^{mn})$. This largest possible degree is sometimes referred to as the *generic degree*. In other words, the generic degree is obtained by algebraic perturbation of the subvariety $\mathcal{L}$ in the ambient space, $\mathbb{K}^{mn}$. In this section we determine an upper bound of the generic degree in the case when $d = n$ and prove that this upper is reached when $m = n < 5$. Let $\mathbb{K}$ be an algebraically closed field.

LEMMA 4.1. *Let $H_1, \ldots, H_k$, $k \le n$, be hypersurfaces in $\mathbb{P}^n$ of degrees $d_1, \ldots, d_k$, respectively, and let $Z_1, \ldots, Z_m$ be the irreducible components of $\bigcap_{i=1}^k H_i$ (not necessarily the same dimensions). Then*

$$\sum_{j=1}^m \deg Z_j \le \prod_{i=1}^k d_i.$$

Proof. It is sufficient to prove that for any l-dimensional projective variety Z of degree d, $l > 0$, and for any hypersurface H of degree q, the sum of the degrees of the irreducible components of $Z \cap H$ is at most dq.

If $Z \subset H$, then $Z \cap H = Z$ and $\deg Z = d \le dq$. On the other hand, if $Z \not\subset H$, then each irreducible component of $Z \cap H$ has dimension $l - 1$ (see the proof of [10, Chapter 1, Proposition 7.1]), and by Bézout's theorem [9, Theorem 18.4, p. 228], the sum of the degrees of the irreducible components of $Z \cap H$ is dq counted with multiplicity. ∎

LEMMA 4.2. *Let $p_i(x) \in \mathbb{K}[x_1, \ldots, x_n]$ be polynomials of degree d_i, $i = 1, \ldots, k$, $k \le n$. Then the number of irreducible components of*

$$\{x \in \mathbb{K}^n \mid p_i(x) = 0, i = 1, \ldots, k\}$$

is at most $\prod_{i=1}^k d_i$.

Proof. Let $\hat{p}_i(z)$ be the homogenization of $p(x)$, i.e.,

$$\hat{p}_i(z) = z_0^{d_i} p(z_1/z_0, \ldots, z_n/z_0).$$

Then $\hat{p}_i(z)$ defines a hypersurface of degree d_i in $\mathbb{P}^n$. Lemma 4.1 implies that the number of irreducible components of

$$\{z \in \mathbb{P}^n | \hat{p}_i(z) = 0, i = 1, \ldots, k\}$$

is at most $\prod_{i=1}^{k} d_i$. So is the number of irreducible components of

$$\{x \in \mathbb{K}^n \mid p_i(x) = 0, i = 1, \ldots, k\}. \qquad \square$$

THEOREM 4.3. *Let $\mathbb{K}$ be an algebraically closed field. The generic degree of $\bar{\mathcal{L}} \subset \mathbb{P}(\wedge^m \mathbb{K}^{m+n}) = \mathbb{P}^N$ is at most $\min(m,n)^n$. When $m = n$ the sharper bound $n(n-1)^{n-1}$ even holds.*

Proof. For a generic $(N-n)$-dimensional projective subspace $H \subset \mathbb{P}^N$, the intersection $H \cap \bar{\mathcal{L}}$ contains exactly $\deg \bar{\mathcal{L}}$ many points, and all of them are in $\mathcal{L}$. Let

$$\sum_{\nu} z_\nu (e_{\nu_1} \wedge \cdots \wedge e_{\nu_n}) \tag{4.1}$$

be the homogeneous coordinates of the points in $\wedge^m \mathbb{K}^{m+n}$, where $\{e_i\}$ is the standard basis of $\mathbb{K}^{m+n}$. Then each $(N-n)$-dimensional projective subspace H is defined by n linear equations in $\{z_\nu\}$. Let $\mathcal{L} = \text{span}\,\{F_1, \ldots, F_n\}$, and define

$$F(x) = x_1 F_1 + \cdots + x_n F_n. \tag{4.2}$$

Consider the Plücker coordinates z_ν for the row space of $[I_m,\ F(x)]$. There are now two cases. When $m \le n$, then the coordinates z_ν have degree at most m when viewed as polynomials of $\mathbb{K}[x_1, \ldots, x_n]$. Intersecting with the subspace H results in n polynomial equations in n variables having degree at most m. By the previous lemma the number of solutions is at most m^n.

When $m \ge n$, then the z_ν have degree at most n, and we get the upper bound n^n. When $m = n$ we get the sharpened upper bound as follows: In order to describe the generic plane H we can always choose a set of equations such that at most one of them contains the last Plücker coordinate $z_{(n+1,\ldots,2n)}$ which is equal to $\det F(x)$, a polynomial of degree n at most. All other polynomials z_ν have degree $n-1$ at most. Therefore $\deg \bar{\mathcal{L}}$ equals the number of solutions of one polynomial equation of degree at most n and $n-1$ of polynomial equations of degrees at most $n-1$. By Lemma 4.2, if the number of solutions is finite, then it is at most $n(n-1)^{n-1}$, and hence $\deg \bar{\mathcal{L}}$ is at most $n(n-1)^{n-1}$ when $m = n$. $\square$

The reader may wonder how good the bound of the degree in Theorem 4.3 is. It is readily shown that the bound is sharp when $\min(m,n) = 1$. In the following two pages we show that the degree is $n(n-1)^{n-1}$ for $m = n < 5$.

Note that the Plücker coordinate $z_\nu = z_{(\nu_1,\ldots,\nu_n)}$ defined by (4.1) is the full size minor consisting of the $\nu_1, \ldots, \nu_n$ columns of $[F_1,\ F_2] \in \text{Grass}(n, 2n)$. Define

$$|\nu| = \sum_{i=1}^{n} (\nu_i - i)$$

and partial order $\nu \le \mu$ if $\nu_i \le \mu_i$ for all i. Under the standard flag (i.e., the flag spanned by the ordered standard basis), the Schubert variety (3.1) is defined by

$$S(\nu) := S(\nu_1, \ldots, \nu_i) = \{z \in \text{Grass}(n, \mathbb{K}^{2n}) \mid z_\mu = 0 \text{ for } \mu \not\le \nu\},$$

and the Schubert cell (3.2) is defined by

$$C(\nu) := C(\nu_1, \ldots, \nu_i) = \{z \in \mathrm{Grass}(n, \mathbb{K}^{2n}) \mid z_\mu = 0 \text{ for } \mu \not\leq \nu \text{ and } z_\nu \neq 0\},$$

and $\dim S(\nu) = \dim C(\nu) = |\nu|$.

LEMMA 4.4. *Let*

$$f_k(z) = \sum_{|\nu|=k} z_\nu,$$

and let Z_k be the algebraic subset of $\mathrm{Grass}(n, \mathbb{K}^{2n})$ *defined by $f_{k+1}(z) = 0, f_{k+2}(z) = 0, \ldots, f_{n^2}(z) = 0$. Then*

$$Z_k = \bigcup_{|\nu|=k} S(\nu).$$

Proof. Clearly $\bigcup_{|\nu|=k} S(\nu) \subset Z_k$. Let $z \in Z_k$. If $z \notin \bigcup_{|\nu|=k} S(\nu)$, then z must be in a Schubert cell $C(\mu)$ for some μ with $|\mu| > k$. Then it implies that $z_\nu = 0$ for all ν such that $|\nu| = |\mu|$ and $\nu \neq \mu$. From one of the defining equations, $f_{|\mu|}(z) = 0$, we derive a contradiction: $z_\mu = 0$. Therefore $z \in \bigcup_{|\nu|=k} S(\nu)$ and $Z_k = \bigcup_{|\nu|=k} S(\nu)$. ∎

We call a coordinate z_ν type i if exactly i of the indices $\{\nu_1, \ldots, \nu_n\}$ are in the set $\{n+1, \ldots, 2n\}$. For example, $z_{(n+1,\ldots,2n)}$ is type n, and $z_{(1,3,\ldots,n,n+1)}$ is type 1. Let

$$\widehat{Mat_{n\times n}} = \{[I_n,\ F] \mid F \in Mat_{n\times n}\}.$$

Then $\widehat{Mat_{n\times n}}$ is an open set of $\mathrm{Grass}(n, \mathbb{K}^{2n})$ which is isomorphic to $Mat_{n\times n}$. In $\widehat{Mat_{n\times n}}$, a type k coordinate is a homogeneous polynomial of degree k of the entries of F.

LEMMA 4.5. *Let $k > 0$ and Z be a k-dimensional subvariety of* $\mathrm{Grass}(n, \mathbb{K}^{2n})$ *such that it is not completely contained in the hypersurface*

$$H_0 := \{z \in \mathrm{Grass}(n, \mathbb{K}^{2n}) | z_{(1,2,\ldots,n)} = 0\}.$$

Then each irreducible component of $Z \cap H$ has dimension $k-1$ for generic hypersurfaces H defined by linear equations of type 1 *coordinates.*

Proof. For any H, either $Z \subset H$ or each irreducible component of $Z \cap H$ has dimension $k-1$ (see the proof of [10, Chapter 1, Proposition 7.1]). Therefore we need only to show that $Z \not\subset H$ for generic H's. Since the set of all such H's form a Zariski open set, we need only to show that it is nonempty. Since $\dim Z > 0$ and $Z \not\subset H_0$, we can always find a point $[I_n,\ F] \in Z$ with $F \neq 0$. Let z_ν be the type 1 Plücker coordinate corresponding to a nonzero entry of F. Then Z is not contained in the hypersurface defined by $z_\nu = 0$. ∎

LEMMA 4.6 (see [5]). *Let $p_1(x), \ldots, p_n(x)$ be polynomials on $\mathbb{K}^n$ of degrees $d_1, \ldots, d_n$, respectively, and let $p_i^h(x)$ be the homogeneous part of $p_i(x)$ of the highest degree (degree d_i). If $p_1^h(x) = 0, \ldots, p_n^h(x) = 0$ have only zero solutions, then the system of polynomial equations*

$$\begin{aligned} p_1(x) &= b_1, \\ &\vdots \\ p_n(x) &= b_n \end{aligned}$$

has $\prod_{i=1}^n d_i$ solutions counted with multiplicity for any $(b_1, \ldots, b_n) \in \mathbb{K}^n$.

Proof. Let $(1, x_1, \ldots, x_n) = (1, z_1/z_0, \ldots, z_n/z_0)$. Then $\mathbb{K}^n$ can be considered as an open subset of $\mathbb{P}^n$ defined by $z_0 \neq 0$. Let $h_i(z)$ be the homogenization of $p_i(x) - b_i$; i.e., $h_i(z) = z_0^{d_i}(p_i(z_1/z_0, \ldots, z_n/z_0) - b_i)$. $h_i(z)$ defines a hypersurface H_i of degree d_i in $\mathbb{P}^n$. Let H_0 be the hyperplane in $\mathbb{P}^n$ defined by $z_0 = 0$. The condition implies that the system of equations $z_0 = 0$ and $h_i(z) = 0$, for $i = 1, \ldots, n$, has only zero solution $z_i = 0$, $i = 0, \ldots, n$; i.e., in $\mathbb{P}^n$, $\cap_{i=0}^n H_i = \emptyset$. By the projective dimension theorem, $\dim \cap_{i=1}^n H_i = 0$, and by Bézout's theorem, $\cap_{i=1}^n H_i$ contains exactly $\prod_{i=1}^n d_i$ points counted with multiplicity and again by the given condition, all of them are in $\mathbb{K}^n$. ∎

THEOREM 4.7. *The generic degree of $\bar{\mathcal{L}}$ is $n(n-1)^{n-1}$ for $m = n < 5$.*

Proof. By Theorem 4.3, we need only to find one $\mathcal{L}$ such that $\deg \bar{\mathcal{L}} = n(n-1)^{n-1}$. From Lemma 4.4 we know that

$$\{z \in \mathrm{Grass}(n, \mathbb{K}^{2n}) \mid f_{n^2}(z) = 0, f_{n^2-1}(z) = 0, \ldots, f_{n^2-n+1}(z) = 0\} = \bigcup_{|\nu|=n^2-n} S(\nu).$$

By repeatedly using Lemma 4.5 we can find $n^2 - n$ linear equations

$$l_1(z) = 0, \ldots, l_{n^2-n}(z) = 0$$

of type 1 coordinates such that each irreducible component of

$$Z := \left\{ z \in \bigcup_{|\nu|=n^2-n} S(\nu) \mid l_1(z) = 0, \ldots, l_{n^2-n}(z) = 0 \right\}$$

either is contained completely in H_0 or has dimension 0.

Note that for $n < 5$, $f_{n^2-1}(z) = 0, \ldots, f_{n^2-n+1}(z) = 0$ are linear equations of type $n-1$ coordinates, and in $\widehat{Mat_{n\times n}} = \{[I_n,\ F]\}$ they are homogeneous equations of degree $n-1$ of the entries of F. Therefore if $[I_n,\ F]$ is in Z, then $[I_n,\ tF]$ are also in Z for all t. Since $\dim Z \cap \widehat{Mat_{n\times n}} = 0$, we must have

$$\dim Z \cap \widehat{Mat_{n\times n}} = \{[I_n,\ 0]\}. \tag{4.3}$$

On $\widehat{Mat_{n\times n}}$, $l_1(z) = 0, \ldots, l_{n^2-n}(z) = 0$ are linear equations of the entries of matrices $F \in Mat_{n\times n}$. Let $\mathcal{L}$ be the n-dimensional linear subspace of $Mat_{n\times n}$ defined by these linear equations, and let $F(x)$ be defined as in (4.2). Then in terms of x, the equation $0 = f_{n^2}(z) = \det F(x)$ is a homogeneous equation of degree n, and the equations $f_{n^2-1}(z) = 0, \ldots, f_{n^2-n+1}(z) = 0$ are homogeneous equations of degree $n-1$, and (4.3) implies that the system of these equations has only zero solution. Therefore, by Lemma 4.6, the system $f_{n^2}(z) = b_1, \ldots, f_{n^2-n+1}(z) = b_n$ has $n(n-1)^{n-1}$ solutions counted with multiplicity in $\mathcal{L}$, which means that the linear system of the Plücker coordinates

$$\begin{aligned} f_{n^2}(z) &= b_1 z_{(1,2,\ldots,n)}, \\ &\vdots \\ f_{n^2-n+1}(z) &= b_n z_{(1,2,\ldots,n)} \end{aligned}$$

has $n(n-1)^{n-1}$ solutions in $\bar{\mathcal{L}}$ counted with multiplicity, i.e., $\deg \bar{\mathcal{L}} = n(n-1)^{n-1}$. ∎

Theorem 2.8 showed the existence of solutions over an algebraically closed field when $d \geq n$. In the critical dimension ($d = n$) we already know that the number of

solutions is finite generically. Theorem 4.3 gives the following upper bound for the number of solutions in Problem 1.1.

THEOREM 4.8. *Let $\mathbb{K}$ be an arbitrary field. Let $\mathcal{L} \subset Mat_{m\times n}$ be a subspace of dimension n, let A, G, H be a "generic set of matrices," and let $\varphi(s) \in \mathbb{K}^n$ be an arbitrary monic polynomial $\varphi(s) \in \mathbb{K}[s]$ of degree n. Then there exist at most $\min(m,n)^n$ different feedback laws $F \in \mathcal{L}$ such that (1.1) holds. If, in addition, $m = n$, then the number of solutions is bounded by $n(n-1)^{n-1}$.*

Proof. Consider the problem over the algebraic closure $\bar{\mathbb{K}}$ of $\mathbb{K}$. By Theorem 2.7 there are for every monic polynomial $\varphi(s)$ of degree n exactly $\deg \bar{\mathcal{L}}$ many feedback laws $[F_1\ F_2] \in \bar{\mathcal{L}}$ (when counted with multiplicity) such that

$$\det \begin{bmatrix} F_1 & F_2 \\ -sH+G & sI-A \end{bmatrix} = \varphi(s).$$

Therefore, there are at most $\deg \bar{\mathcal{L}}$ many feedback laws $F \in \mathcal{L}$ whose coefficients are in the base field $\mathbb{K}$. $\deg \bar{\mathcal{L}}$ is always less than the generic degree. By Theorem 4.3 the generic degree is at most $\min(m,n)^n$ (respectively, $n(n-1)^{n-1}$ when $m = n$). ∎

Acknowledgment. We started to work on this project in 1998. Meeyoung was very enthusiastic about this research since it brought together nontrivial ideas from algebraic geometry and systems theory. She hoped that it would be the start of a longer research project. In June 2001, Meeyoung died in a tragic drowning accident in the Mediterranean Sea at the young age of 35. Her passing leaves everybody who knew her saddened, and we feel a deep sense of loss.

REFERENCES

[1] J. A. BALL, I. GOHBERG, L. RODMAN, AND T. SHALOM, *On the eigenvalues of matrices with given upper triangular part*, Integral Equations Operator Theory, 13 (1990), pp. 488–497.
[2] R. W. BROCKETT AND C. I. BYRNES, *Multivariable Nyquist criteria, root loci and pole placement: A geometric viewpoint*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 271–284.
[3] C. I. BYRNES AND X. WANG, *The additive inverse eigenvalue problem for Lie perturbations*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 113–117.
[4] S. FRIEDLAND, *Matrices with prescribed off-diagonal elements*, Israel J. Math., 11 (1972), pp. 184–189.
[5] S. FRIEDLAND,*Inverse eigenvalue problems*, Linear Algebra Appl., 17 (1977), pp. 15–51.
[6] W. FULTON, *Intersection Theory*, Ergeb. Math. Grenzgeb. (3) 2, Springer-Verlag, Berlin, Heidelberg, New York, 1984.
[7] I. GOHBERG, M. A. KAASHOEK, AND F. VAN SCHAGEN, *Partially Specified Matrices and Operators: Classification, Completion, Applications*, Birkhäuser, Boston, Basel, Berlin, 1995.
[8] P. GRIFFITHS AND J. HARRIS, *Principles of Algebraic Geometry*, John Wiley and Sons, New York, 1978.
[9] J. HARRIS, *Algebraic Geometry, A First Course*, Grad. Texts in Math., Springer-Verlag, New York, Berlin, 1992.
[10] R. HARTSHORNE, *Algebraic Geometry*, Springer-Verlag, Berlin, 1977.
[11] W. HELTON, J. ROSENTHAL, AND X. WANG, *Matrix extensions and eigenvalue completions, the generic case*, Trans. Amer. Math. Soc., 349 (1997), pp. 3401–3408.
[12] W. V. D. HODGE AND D. PEDOE, *Methods of Algebraic Geometry*, Vol. 2, Cambridge University Press, Cambridge, UK, 1952.
[13] D. MUMFORD, *Algebraic Geometry* I: *Complex Projective Varieties*, Springer-Verlag, Berlin, New York, 1976.
[14] M. S. RAVI, J. ROSENTHAL, AND X. WANG, *Dynamic pole assignment and Schubert calculus*, SIAM J. Control Optim., 34 (1996), pp. 813–832.
[15] J. ROSENTHAL, *On dynamic feedback compensation and compactification of systems*, SIAM J. Control Optim., 32 (1994), pp. 279–296.

[16] J. Rosenthal and X. Wang, *Inverse eigenvalue problems for multivariable linear systems*, in Systems and Control in the Twenty-First Century, C. I. Byrnes, B. N. Datta, D. Gilliam, and C. F. Martin, eds., Birkhäuser, Boston, Basel, Berlin, 1997, pp. 289–311.
[17] I. R. Shafarevich, *Basic Algebraic Geometry*, Springer-Verlag, Berlin, 1977.
[18] X. Wang, *Pole placement by static output feedback*, J. Math. Systems Estim. Control, 2 (1992), pp. 205–218.
[19] X. Wang, *Decentralized pole assignment and product Grassmannians*, SIAM J. Control Optim., 32 (1994), pp. 855–875.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 2094–2115

CONVERGENCE PROPERTIES OF POLICY ITERATION*

MANUEL S. SANTOS† AND JOHN RUST‡

Abstract. This paper analyzes asymptotic convergence properties of policy iteration in a class of stationary, infinite-horizon Markovian decision problems that arise in optimal growth theory. These problems have continuous state and control variables and must therefore be discretized in order to compute an approximate solution. The discretization may render inapplicable known convergence results for policy iteration such as those of Puterman and Brumelle [*Math. Oper. Res.*, 4 (1979), pp. 60–69]. Under certain regularity conditions, we prove that for piecewise linear interpolation, policy iteration converges quadratically. Also, under more general conditions we establish that convergence is superlinear. We show how the constants involved in these convergence orders depend on the grid size of the discretization. These theoretical results are illustrated with numerical experiments that compare the performance of policy iteration and the method of successive approximations.

Key words. policy iteration, method of successive approximations, quadratic and superlinear convergence, complexity, computational cost

AMS subject classifications. 49M15, 65K05, 90C30, 93B40

DOI. 10.1137/S0363012902399824

1. Introduction.

The goal of this paper is to provide new insights into the convergence properties of *policy iteration,* an algorithm developed by Bellman (1955, 1957) and Howard (1960) for solving stationary, infinite-horizon Markovian dynamic programming (MDP) problems. Policy iteration has some rough similarities to the simplex algorithm of linear programming (LP). Just as the simplex algorithm generates a sequence of improving trial solutions to the LP problem (along with their associated costs), policy iteration generates an improving sequence of decision rules to the MDP problem (along with their associated value functions). Also, similarly to the simplex algorithm, policy iteration has been found to converge to the optimal solution in a remarkably small number of iterations. Typically fewer than 10 to 20 policy iterations are required to find the optimal solution. But analogously to LP, where the number of possible vertices increases exponentially fast as the number of variables M and constraints N increases, the number of possible decision rules of an MDP problem with M states and N actions in each state is N^M, which also grows exponentially fast in M. Klee and Minty (1972) have constructed worst-case families of LP problems where the simplex algorithm visits a large and exponentially increasing number of vertices before converging to the optimal solution. Are there also "worst-case" families of MDP problems where policy iteration visits an exponentially increasing number of trial decision rules before converging to the optimal solution?

Although we do not answer this question in the present paper, we consider an example due to Tsitsiklis (2000) of a family of finite state MDP problems with M states and two actions in each state where roughly M policy iteration steps are required to find the optimal solution. While this example may not represent the "worst case," it is a fairly pessimistic result that suggests that policy iteration will not be an

*Received by the editors March 12, 2002; accepted for publication (in revised form) August 9, 2003; published electronically February 18, 2004.
http://www.siam.org/journals/sicon/42-6/39982.html
†Department of Economics, Arizona State University, Tempe, AZ 85287 (Manuel.Santos@asu.edu).
‡Department of Economics, University of Maryland, College Park, MD 20742 (jrust@gemini.econ.umd.edu).

efficient method for solving all types of MDP problems. The reason is that each policy iteration step is expensive: it involves (among other computations) a solution of a linear system with M equations in M unknowns at a total cost of $O(M^3)$ arithmetic operations using standard linear equation solvers.

This paper was motivated by a desire to characterize those families of MDP problems for which only a small number of policy iteration steps are required to find the solution. Puterman and Brumelle (1979) were among the first to analyze the convergence properties of policy iteration for MDP problems with continuous state and action spaces. They showed that policy iteration is mathematically equivalent to *Newton's method*, and so under certain regularity conditions the policy iteration algorithm displays a quadratic order of convergence. The shortcoming of Puterman and Brumelle's abstract infinite-dimensional space analysis is that the generalized version of policy iteration that they describe is not actually computationally feasible in problems where there are an infinite number of states and actions. Their analysis assumed that the exact value functions are computed at each policy evaluation step, and this requires solving an infinite-dimensional system of linear equations. Furthermore, they impose a Lipschitz order condition which is not easily verifiable in their framework.

We analyze a computationally feasible version of policy iteration for a class of optimal growth problems with a continuum of states and decisions. We establish sufficient conditions under which the policy iteration solution for discretized versions of this problem exhibits a quadratic rate of convergence and a global operative estimate with a convergence rate equal to 1.5. We study how the constants involved in the convergence orders depend on the grid size of the discretization. Also, under more general conditions we show that convergence is superlinear. Thus, desirable convergence properties that Puterman and Brumelle demonstrated for an abstract, idealized version of policy iteration also hold for a practical, computationally feasible version that can be applied to solve actual problems in economics and finance. To the best of our knowledge, this is the first numerical algorithm with such convergence properties.

Despite the scant theoretical work and evaluations of the performance of policy iteration in economic models, computational economists (e.g., see Chapter 12 of Judd (1998)) have been well aware of this numerical procedure. Rust (1987) applies policy iteration to solve a durable good replacement problem. A recent paper by Benitez-Silva et al. (2001) offers an extensive numerical evaluation of policy iteration in general economic models. Unfortunately, there is little guidance about the conditions under which one should use policy iteration as opposed to a variety of alternative algorithms including the *method of successive approximations.* This latter method can be shown to be globally convergent since the value function to an MDP problem is the fixed point to a contraction mapping. Section 6 contains a computational illustration that compares successive approximations and our computationally feasible version of policy iteration for an optimal growth problem whose solution can also be computed analytically.

The paper is structured as follows. In section 2 we present some background on the policy iteration algorithm, reviewing known results in the finite and infinite state cases, respectively. In section 3, we focus our analysis on a class of multidimensional optimal growth problems. Then, in section 4 we describe our computationally feasible version of the policy iteration algorithm using bilinear interpolation. In section 5 we state and prove the main results on the convergence properties of our algorithm. Section 6 provides a comparison of the performance of both policy iteration and the method of successive approximations in the case of an optimal growth problem that

admits an analytic solution. This enables us to evaluate the uniform approximation errors of the two algorithms as well as compare their relative computational speed. The computational results are consistent with theoretical predictions. We conclude in section 7 with a discussion of our main findings.

2. Background. As noted in the introduction, policy iteration is a commonly used algorithm for solving stationary, infinite-horizon MDP problems. The MDP problem is mathematically equivalent to computing the fixed point V^* to *Bellman's equation*

$$V^* = \Gamma(V^*), \tag{2.1}$$

where Γ is the *Bellman operator* given by

$$\Gamma(V)(s) \equiv \max_{a \in A(s)} \left[u(s,a) + \beta \int V(s') p(ds'|s,a) \right], \tag{2.2}$$

and $\beta \in (0,1)$ is the discount factor, $u(s,a)$ represents the utility or payoff earned in state s when action a is taken, and $p(ds'|s,a)$ is a conditional probability distribution for next period's state s' given the current state s and action a. As is well known (see, e.g., Blackwell (1965)), under mild regularity conditions the Bellman operator is a contraction mapping, and hence its fixed point V^*, the *value function,* is unique. The solution to Bellman's equation is of considerable interest because it has been shown that from V^* we can compute the corresponding *optimal decision rule or policy function* α^* given by

$$\alpha^*(s) \equiv \arg\max_{a \in A(s)} \left[u(s,a) + \beta \int V^*(s') p(ds'|s,a) \right]. \tag{2.3}$$

The policy iteration algorithm computes V^* via a sequence of trial value functions $\{V_n\}$ and decision rules $\{\alpha_n\}$ under an alternating sequence of *policy improvement* and *policy evaluation* steps. The policy improvement step computes an improved policy α_n implied by the previous value function V_n:

$$\alpha_n(s) = \arg\max_{a \in A(s)} \left[u(s,a) + \beta \int V_n(s') p(ds'|s,a) \right] \quad \text{(policy improvement)}. \tag{2.4}$$

The policy evaluation step computes a new value function V_{n+1} implied by policy α_n:

$$V_{n+1}(s) = \left[u(s,\alpha_n(s)) + \beta \int V_{n+1}(s') p(ds'|s,\alpha_n(s)) \right] \quad \text{(policy evaluation)}. \tag{2.5}$$

Policy iteration continues until $V_{n+1} = V_n$, $\alpha_{n+1} = \alpha_n$, or the difference between two successive value functions or decision rules is less than a prescribed solution tolerance. Under fairly general conditions, policy iteration can be shown to generate a monotonically improving sequence of trial value functions, $V_{n+1} \geq V_n$. In the case of MDP problems where both the state space S and action sets $A(s)$ are finite, there is only a finite number of possible policies bounded by N^M, where M is the number of states and N is the maximum number of possible actions in the constraint sets $A(s)$. This, together with the fact that policy iteration is monotonic (and thus cannot cycle) implies that the policy iteration algorithm will converge to the true fixed point in a finite number of steps (assuming all arithmetic operations are carried

out exactly, without any rounding error). Although the number of potential policies is vast (for example a stopping problem with two choices and 1000 states has 2^{1000} or about 1.07×10^{301} possible policies), it has been observed in most computational examples that policy iteration converges in a remarkably small number of iterations, typically less than 20. Furthermore, the number of policy iteration steps appears to be independent of the number of states and decisions.

The most commonly used alternative to policy iteration is the method of successive approximations

$$V_{n+1} = \Gamma(V_n). \tag{2.6}$$

As is well known, this algorithm is globally convergent (an implication of the fact that Γ is a contraction mapping), but it converges at a geometric rate. The error after $\widehat{T}$ successive approximations steps is $O(\beta^{\widehat{T}}/(1-\beta))$. Since each successive approximation step requires $O(M^2N)$ arithmetic operations (and thus each iteration is an order faster than policy iteration), as the discount factor $\beta \to 1$, the number of successive approximation steps required to obtain acceptable accuracy rises rapidly, calling for a huge number of computations. Hence, when β is close to 1 (as in the calibration of models with quarterly data or smaller time intervals), the total computational burden of doing a small number of more expensive policy iteration steps may be less than the total amount of work involved in doing a large number of less expensive successive approximation steps.

Of course, this advice will not hold if the number of policy iteration steps increases sufficiently rapidly with M. Unfortunately, the following counterexample (Tsitsiklis (2000)) shows that in general the number of policy iteration steps cannot be bounded by a constant that is independent of M. Consider an MDP problem with $M+1$ states $s \in \{0, 1, \ldots, M\}$. In each state there are two possible actions, $a \in \{-1, +1\}$, with the interpretation that $a = -1$ corresponds to "moving one state to the left" and $a = +1$ corresponds to "moving one state to the right." This can be mapped into a transition probability given by $p(ds'|s, a)$ that puts probability mass 1 on state $s' = s-1$ if $a = -1$ and probability mass 1 on state $s' = s+1$ if $a = +1$. States $s = 0$ and $s = M$ are zero–cost-absorbing states regardless of the action taken, so that we have

$$u(s, a) = 0 \quad \text{if } s = 0 \text{ or } s = M. \tag{2.7}$$

For states $s = 1, 2, \ldots, M-2$, the payoff equals -1 for moving left and -2 for moving right:

$$u(s, a) = \begin{cases} -1 & \text{if } a = -1, \\ -2 & \text{if } a = +1. \end{cases} \tag{2.8}$$

Finally, for state $s = M-1$ there is a reward of $2M$ for moving right and a reward of -1 for moving left:

$$u(s, a) = \begin{cases} -1 & \text{if } a = -1, \\ 2M & \text{if } a = +1. \end{cases} \tag{2.9}$$

Consider policy iteration starting from the initial value function $V_0(s) = 0$ for all s. Then it is easy to see that the optimal policy α_0 implied by this value function is

$\alpha_0(s) = -1$ for states $s = 1, \ldots, M-2$ and $\alpha_0(M-1) = +1$ (observe that the choice of policy is irrelevant at the absorbing states $s = 0$ and $s = M$). It is not difficult to check that when $\beta = 1$, the value V_1 associated with this policy α_0 is given by

$$V_1(s) = \begin{cases} 0 & \text{if } s = 0, M, \\ -s & \text{if } s \in \{1, 2, \ldots, M-2\}, \\ 2M & \text{if } s = M-1. \end{cases} \tag{2.10}$$

This is the value function generated by the first policy iteration step. For the second step, notice that optimal policy α_1 implied by V_1 is to go left at states $s = 1, 2, \ldots, M-3$, but to go right at states $s = M-2$ and $s = M-1$. This new policy α_1 differs from the initial policy α_0 in only one state, $s = M-2$, where it is now optimal to go right because the updated value V_1 assigns a higher value $V_1(M-1) = 2M$ to the penultimate state $s = M-1$. Moreover, after each succeeding policy iteration $n = 2, \ldots, M-1$, the updated decision rule α_n differs from the previous policy α_{n-1} in state $s = M-n-1$, flipping the optimal action from left to right. This process continues for $M-1$ policy iteration steps until the optimal policy $\alpha^*(s) = +1$ for all s; i.e., it is optimal to always move to the right. The optimal value function $V^*(s)$ for this policy is then given by

$$V^*(s) = \begin{cases} 0 & \text{if } s = 0, M, \\ 2(s+1) & \text{if } s \in \{1, 2, \ldots, M-1\}. \end{cases} \tag{2.11}$$

It should be stressed that the same type of result occurs for discounted problems with $\beta \in (0, 1)$ as long as β is sufficiently close to 1.

In summary, this counterexample provides a family of MDP problems with $M+1$ states in which a total of $M-1$ policy iteration steps are required to converge to the optimal solution starting from an initial guess $V_0 = 0$. Therefore, it will not be possible to provide a bound on the number of steps that is independent of the number of states M. But an alternative analysis of the convergence of policy iteration, initiated by Puterman and Brumelle (1979), does suggest that under additional conditions only a small number of policy iterations should be required to solve the MDP problem and this number should be essentially independent of the number of states M. Puterman and Brumelle were among the first to notice that policy iteration is mathematically equivalent to Newton's method for solving nonlinear functional equations. Then, building on some generalized results on the quadratic convergence of the Newton iteration with supports replacing derivatives, Puterman and Brumelle showed that for some constant L the iteration (2.4)–(2.5) satisfies the quadratic convergence bound

$$\|V_{n+1} - V^*\| \leq \frac{\beta L}{(1-\beta)} \|V_n - V^*\|^2, \tag{2.12}$$

where $\|V\|$ is a norm in the space of functions V. There are, however, two limitations to this result. For the case of finite state, finite action MDP problems, the quadratic convergence bound is not effective. First, (2.12) may hold only in a neighborhood of the fixed-point solution V^*. Consequently, the errors $\|V_n - V^*\|$ tend to be either very large until V_n is in a "domain of attraction" of V^*, in which case the error immediately falls to 0 after one further policy iteration step. In more technical terms, following the analysis below one can show that for the case of finite state, finite action MDP problems, constant L is equal to zero if there is no policy switch after a small

perturbation in the value function, but L becomes undefined (i.e., $L = \infty$) at points with multiple maxima in which a small perturbation in the value function leads to a sudden switch in the optimal action. Therefore, for MDP problems with a finite number of states and actions the Puterman and Brumelle results cannot be applied. The second limitation was noted in the introduction; namely, the abstract infinite-dimensional version of policy iteration in (2.4)–(2.5) is not computationally feasible for problems where there are infinite numbers of states. In these problems some sort of discretization procedure must be employed. And it seems crucial to understand how constant L will depend on the discretization procedure.

3. A reduced form model of economic growth. For expositional convenience and for our later computational purposes, we now consider a reduced-form version of our MDP problem (2.1)–(2.2) that encompasses most standard models of economic growth (cf. Stokey, Lucas with Prescott (1989)). As in most of the economics literature, we distinguish between endogenous and exogenous growth variables. This is commonly assumed in macroeconomic applications, where Bellman's equation is usually expressed in the following form:

$$V^*(k_0, z_0) = \max_{k_1} v(k_0, k_1, z_0) + \beta \int_Z V^*(k_1, z') Q(dz', z_0) \tag{3.1}$$
$$\text{subject to (s.t.) } (k_0, k_1, z_0) \in \Omega,$$
$$(k_0, z_0) \text{ fixed, } 0 < \beta < 1.$$

Here, k is a vector of endogenous state variables in a set K, which may include several types of capital stocks and measures of wealth. Also, z is a vector made up of stochastic exogenous variables such as some indices of productivity or market prices. This latter random vector lies in a set Z, and is governed by a probability law Q which is assumed to be weakly continuous. As is typical in economic growth theory (cf. Stokey, Lucas with Prescott (1989)), this formulation is written in reduced-form in the sense that the action or control variables are not explicitly laid out. It should be understood that for every (k_0, k_1, z_0) an optimal control has been selected from which the one-time payoff $v(k_0, k_1, z_0)$ can be defined.

The technological constraints of this economy are represented by a given feasible set $\Omega \subset K \times K \times Z$, which is the graph of a continuous correspondence $\Gamma : K \times Z \to K$. The intertemporal objective is characterized by a return function v and a given discount factor $0 < \beta < 1$. Then, $s = (k, z)$ is the vector of state variables lying in the set $S = K \times Z$. Let $(S, \mathbb{S})$ denote a measurable space.

As discussed in Stokey, Lucas with Prescott (1989) (see p. 240), the optimization problem of the preceding section is clearly more general. That is, with an appropriate choice of the action space and laws of motion for the state variables, our MDP problem (3.1) can be embedded in the framework of functional equations (2.1)–(2.2). The converse is not true. Our main interest, however, is to apply policy iteration to a standard macroeconomic setting. This optimization problem will become handy for our numerical experiments in section 6 of a neoclassical growth model with leisure. Moreover, it becomes transparent from our arguments below that our results can be extended to related MDP problems since the analysis centers on the following specific assumptions.

Assumption 1. The set $S = K \times Z \subset \mathbb{R}^\ell \times \mathbb{R}^m$ is compact, and $\mathbb{S}$ is the Borel σ-field. For each fixed z the set $\Omega_z = \{(k, k') \mid (k, k', z) \in \Omega\}$ is convex.

Assumption 2. The mapping $v : \Omega \longrightarrow \mathbb{R}$ is continuous. Also, there exists $\eta > 0$ such that for every z function $v(k, k', z) + \frac{\eta}{2}\|k'\|^2$ is concave in (k, k').

These assumptions are fairly standard. In Assumption 2, $\|\cdot\|$ denotes the Euclidean norm. The asserted uniform concavity of v will be needed only for some stronger versions of our results. Under these conditions the value function $V^*(k_0, z_0)$, given in (3.1), is well defined and jointly continuous in $K \times Z$. Moreover, for each fixed z_0 the mapping $V^*(\cdot, z_0)$ is concave. The optimal value $V^*(k_0, z_0)$ is attained at a unique point k_1 given by the policy function $k_1 = g(k_0, z_0)$ characterizing the set of optimal solutions.

4. Policy iteration in a space of piecewise linear interpolations. In what follows we shall be concerned with a policy iteration algorithm in a space of piecewise bilinear interpolations. Most of our results below apply to other interpolation schemes provided that (a) the operator is monotone, and (b) the operator has a fixed point. These two conditions may be quite restrictive for some approximation schemes such as polynomial and spline interpolations, but they hold true for several forms of piecewise linear interpolation.

We would like to remark that piecewise linear interpolation has certain computational advantages over the usual discretization procedure in which the domain and functional values are restricted to a fixed grid of prespecified points. Indeed, discretizations based on functional evaluations over a discrete set of points are not computationally efficient for smooth problems. Most calculations may become awkward under a discrete state space, and hence the corresponding algorithms are usually rather slow. For instance, the most powerful maximization procedures make use of the information provided by the values of the functional derivatives. These powerful techniques may still be applied under piecewise linear interpolation (cf. Santos (1999)). Therefore, piecewise linear interpolation preserves the aforementioned properties of monotonicity and existence of a fixed point for the operator, and at the same time allows for the use of efficient searching procedures over a continuous state space.

4.1. Formulation of the numerical algorithm. Let us assume that both K and Z are convex polyhedra. This does not entail much loss of generality in most economic applications. Let $\{S^j\}$ be a finite family of *simplices*[1] in K such that $\cup_j S^j = K$ and $\text{int}(S^i) \cap \text{int}(S^j) = \varnothing$ for every pair S^i, S^j. Also, let $\{D^i\}$ be a finite family of simplices in Z such that $\cup_i D^i = Z$ and $\text{int}(D^i) \cap \text{int}(D^j) = \varnothing$ for every pair D^i, D^j. Define the *grid size* or *mesh level* as

$$h = \max_{j,i} diam\left\{S^j, D^i\right\}.$$

Let (k^j, z^i) be a generic vertex of the triangulation. Then, every k and z can be represented as a convex combination of $\{k^j\}$ and $\{z^i\}$. More precisely, for $k \in S^j$ and $z \in D^i$ there is a unique set of nonnegative weights $\lambda_j(k)$ and $\varphi_i(z)$, with $\sum_j \lambda_j(k) = 1$ and $\sum_i \varphi_i(z) = 1$, such that

$$k = \sum_j \lambda_j(k) k^j \quad \text{and} \quad z = \sum_i \varphi_i(z) z^i \quad \text{for all } k^j \in S^j \text{ and } z^i \in D^i. \tag{4.1}$$

We next define a finite-dimensional space of numerical functions compatible with the simplex structure $\{S^j, D^i\}$. Each element V^h is determined by its nodal values

[1] A simplex S^j in $\mathbb{R}^\ell$ is the set of all convex combinations of $\ell + 1$ given points (cf. Rockafellar (1970)). Thus, a simplex in $\mathbb{R}^1$ is an interval, a simplex in $\mathbb{R}^2$ would be a triangle, and a simplex in $\mathbb{R}^3$ would be a tetrahedron.

$\{V^h(k^j, z^i)\}$ and is extended over the whole domain by the bilinear interpolation

$$V^h(k,z) = \sum_j \lambda_j(k)\left[\sum_i \varphi_i(z) V^h(k^j, z^i)\right].$$

Note that the interpolation is first effected in space Z and then in space K. This interpolation ordering is appropriate for carrying out the numerical integrations and maximizations outlined below. Let the function space

$$\mathcal{V}^h = \left\{ V^h : K \times Z \longrightarrow \mathbb{R} \;\middle|\; \begin{array}{l} V^h(k,z) = \sum_j \lambda_j(k)\left[\sum_i \varphi_i(z) V^h(k^j, z^i)\right] \\ \text{for } \lambda_j \text{ and } \varphi_i \text{ satisfying (4.1)} \end{array} \right\}. \tag{4.2}$$

It follows that $\mathcal{V}^h$ is a Banach space when equipped with the norm

$$\|V^h\| = \max_{(k^j, z^i) \in K \times Z} |V^h(k^j, z^i)| \quad \text{for } V^h \in \mathcal{V}^h. \tag{4.3}$$

We now consider the following algorithm for policy iteration in space $\mathcal{V}^h$.

(i) *Initial step*: Select an accuracy level ε and an initial guess V_0^h.

(ii) *Policy improvement step*: Find $k_1 = g_n^h(k_0^j, z_0^i)$ that solves

$$B^h(V_n^h)(k_0^j, z_0^i) \equiv -V_n^h(k_0^j, z_0^i) + \max_{k_1} v(k_0^j, k_1, z_0^i) + \beta \int_Z V_n^h(k_1, z') Q(dz', z_0^i) \tag{4.4}$$

for each vertex point (k_0^j, z_0^i).

(iii) *Policy evaluation step*: Find $V_{n+1}^h(k_0^j, z_0^i)$ that solves

$$V_{n+1}^h(k_0^j, z_0^i) = v(k_0^j, g_n^h(k_0^j, z_0^i), z_0^i) + \beta \int_Z V_{n+1}^h(g_n^h(k_0^j, z_0^i), z') Q(dz', z_0^i) \tag{4.5}$$

for each vertex point (k_0^j, z_0^i).

(iv) *End of iteration*: If $\|V_{n+1}^h - V_n^h\| \le \varepsilon$, stop; else, increment n by 1 and return to step (ii).

Observe that in step (ii) for a given V_n^h we first carry out the integration operation and then find the optimal policy g_n^h. And in step (iii) for a given g_n^h we search for a fixed-point solution V_{n+1}^h. Technically, a natural approach for solving (4.5) is to use numerical integration and then limit the search for the fixed point to a finite system of linear equations (cf. Dahlquist and Bjorck (1974) (pp. 396–397)). Regarding step (iv), the iteration process will stop once the term $\|V_{n+1}^h - V_n^h\|$ falls within a prespecified accuracy level, ε.

The solution to (4.5) can be written as

$$v_{g_n^h} = [I - \beta P_{g_n^h}] V_{n+1}^h, \tag{4.6}$$

where $v_{g_n^h}$ represents the utility implied by policy function g_n^h, that is, $v_{g_n^h}(k_0^j, z_0^i) = v(k_0^j, g_n^h(k_0^j, z_0^i), z_0^i)$ for each vertex point (k_0^j, z_0^i), and $P_{g_n^h}$ is the Markov (conditional expectation) operator defined by

$$\begin{aligned} P_{g_n^h}(V(k_0^j, z_0^i)) &= \int_Z V(g_n^h(k_0^j, z_0^i), z') Q(dz', z_0^i) \\ &= \sum_{j'} \lambda_{j'}(g_n^h(k_0^j, z_0^i)) \left[\int_Z \sum_{i'} \varphi_{i'}(z') V(k_0^{j'}, z_0^{i'}) Q(dz', z_0^i) \right]. \end{aligned} \tag{4.7}$$

It should be understood that in this latter expression $g_n^h(k_0^j, z_0^i)$ belongs to some simplex $S^{j'}$, and z' is integrated out over simplices of the form $D^{i'}$.

Observe that $P_{g_n^h}$ defines a linear operator in the space $\mathcal{V}^h$ of functions V^h that can be represented by its nodal values $\{V^h(k_0^j, z_0^i)\}$. Since this operator is positive and bounded and $\beta \in (0,1)$, it can be shown that the inverse operator $[I - \beta P_{g_n^h}]^{-1}$ exists and has the following series representation:

$$[I - \beta P_{g_n^h}]^{-1} = \sum_{t=0}^{\infty} \beta^t P_{g_n^h}^t. \tag{4.8}$$

The distance between two operators P_g and $P_{\widehat{g}}$ will be defined by

$$\|P_g - P_{\widehat{g}}\| = \max_{\{(k_0^j, z_0^i)\}} \left\{ \sum_{S^{j'}} \left[\sum_{j'} |\lambda_{j'}(g(k_0^j, z_0^i)) - \lambda_{j'}(\widehat{g}(k_0^j, z_0^i))| \right] \right\}. \tag{4.9}$$

The summation of these absolute differences goes over all weights $\lambda_{j'}$ and over all simplices $S^{j'}$, under the convention that $\lambda_{j'}(g(k_0^j, z_0^i)) = 0$ if $g(k_0^j, z_0^i)$ does not belong to $S^{j'}$.

If $\{V_n^h\}_{n \geq 1}$ is a sequence of functions generated by (4.4)–(4.5), one can readily check from these equations that such a sequence satisfies

$$V_{n+1}^h = V_n^h + [I - \beta P_{g_n^h}]^{-1} B^h(V_n^h). \tag{4.10}$$

Moreover, if in the policy improvement step the set of maximizers g_n^h is unique, then $-[I - \beta P_{g_n^h}]$ is the derivative of B^h at V_n^h when $P_{g_n^h}$ is considered as a linear operator in the finite-dimensional space $\mathcal{V}^h$. Hence, (4.10) implies that policy iteration is equivalent to Newton's method applied to operator B^h defined in (4.4). As is well known (e.g., Traub and Woźniakowski (1979)), Newton's method exhibits locally a quadratic rate of convergence provided that the functional derivative satisfies a certain regular Lipschitz condition. But the uniqueness of the set of maximizers seems fairly stringent for operator B^h. Indeed, function V_{n+1}^h is not necessarily concave, since it is obtained as the solution to (4.5). Therefore, there is no guarantee that the maximizer in (4.4) is unique, and consequently that B^h has a well-defined derivative. To circumvent these technicalities, Puterman and Brumelle (1979) apply an extension of Newton's method to policy iteration following a familiar procedure with supports replacing derivatives. Even though operator B^h may not have a well-defined derivative, the following general property follows from the above maximization step: If g_n^h is a selection of the correspondence of maximizers in (4.4), then it must be the case that $-[I - \beta P_{g_n^h}]$ is the support of operator B^h at V_n^h. More precisely, (4.4) implies that for any other function V^h in $\mathcal{V}^h$ the following condition must hold:

$$B^h(V^h) - B^h(V_n^h) \geq -[I - \beta P_{g_n^h}][V^h - V_n^h]. \tag{4.11}$$

Of course, one readily sees from (4.11) that if there is a unique set of maximizers g_n^h, then $-[I - \beta P_{g_n^h}]$ is the derivative of B^h at V_n^h.

4.2. Existence of a fixed point and monotonicity. For our later analysis, we need to establish the existence of a unique fixed point for our algorithm and the monotone convergence to such a solution. We begin with the following discretized

version of Bellman's equation.

$$V^h(k_0^j, z_0^i) = \max_{k_1} v(k_0^j, k_1, z_0^i) + \beta \int_Z V^h(k_1, z') Q(dz', z_0^i) \quad \text{s.t. } (k_0^j, k_1, z_0^i) \in \Omega \quad \text{for each vertex point } (k_0^j, z_0^i). \tag{4.12}$$

Note that this equation needs to be satisfied only at each vertex point (k_0^j, z_0^i).

LEMMA 4.1. *Equation* (4.12) *has a unique solution* V^h *in* $\mathcal{V}^h$.

Proof. The proof is standard. One just defines the discretized dynamic programming operator $V_{n+1}^h = T^h(V_n^h)$ given by

$$V_{n+1}^h(k_0^j, z_0^i) = \max_{k_1} v(k_0^j, k_1, z_0^i) + \beta \int_Z V_n^h(k_1, z') Q(dz', z_0^i) \quad \text{s.t. } (k_0^j, k_1, z_0^i) \in \Omega \quad \text{for each vertex point } (k_0^j, z_0^i). \tag{4.13}$$

One immediately sees that T^h is a contraction mapping in $\mathcal{V}^h$ with modulus $0 < \beta < 1$. By a well-known fixed-point theorem, T^h has a unique fixed point V^h in $\mathcal{V}^h$.

Notice that $V^h = T^h(V^h)$ implies that $B^h(V^h) = 0$. Therefore, the method of successive approximations as defined by (4.13) allows us to prove the existence of a fixed point for our algorithm as defined by (4.4)–(4.5). We next verify the monotone convergence of the algorithm to the fixed-point solution. □

LEMMA 4.2. *Assume that* $\{V_n^h\}_{n\geq 0}$ *is a sequence satisfying* (4.4)–(4.5). *Then* $V_{n+1}^h \geq V_n^h$ *for all* $n \geq 1$.

Proof. From (4.5), consider the following equation:

$$V_n^h = v_{g_{n-1}^h} + \beta P_{g_{n-1}^h} V_n^h. \tag{4.14}$$

That is, V_n^h is the value function under policy g_{n-1}^h. Now, call g_n^h the corresponding set of maximizers over the right-hand side of (4.4) under function V_n^h. Then,

$$v_{g_n^h} + \beta P_{g_n^h} V_n^h \geq V_n^h. \tag{4.15}$$

Moreover, a further application of this procedure for function V_n^h on the left-hand side of (4.15) yields

$$v_{g_n^h} + \beta P_{g_n^h} v_{g_n^h} + \beta^2 [P_{g_n^h}]^2 V_n^h \geq V_n^h.$$

Hence, after t iterations, we obtain

$$\sum_{s=0}^{t} \beta^s [P_{g_n^h}]^s v_{g_n^h} + \beta^{t+1} [P_{g_n^h}]^{t+1} V_n^h \geq V_n^h. \tag{4.16}$$

Now, letting $t \to \infty$, it follows from (4.6)–(4.8) that the left-hand side of (4.16) converges to V_{n+1}^h. Therefore, $V_{n+1}^h \geq V_n^h$. □

Remark 4.3. Assume that $\{V_n^h\}_{n\geq 1}$ is a sequence satisfying (4.4)–(4.5). Let T^h be the discretized dynamic programming operator as defined by (4.13). Then, from the previous method of proof one can readily establish that $V_{n+1}^h \geq T^h(V_n^h) \geq V_n^h$. In this sense, the policy iteration algorithm converges faster to the fixed-point solution than the method of successive approximations generated by operator T^h.

5. Convergence properties of the numerical algorithm. We now establish some convergence properties of our policy iteration algorithm. We begin with a global result for concave interpolations in which the convergence order is equal to 1.5 and the constant involved in the convergence order is relatively easy to estimate. Among other applications, this result may be useful in placing an upper bound on the number of policy iteration steps over a well-defined convergence region to reach a tolerance level ε. Then, the same strategy of proof used for this global convergence result will be applied to address further local convergence properties. Thus for concave interpolations we prove quadratic convergence of the algorithm, and for a more general setting of continuous functions we prove that convergence is superlinear. The constants involved in these convergence orders are shown to depend on the mesh level of the discretization.

5.1. Global convergence in a space of concave functions. For present purposes, we shall assume that either the fixed point $V^h(k,z)$ is a concave function in k or the sequence $\{V_n^h(k,z)\}_{n\geq 1}$ generated by policy iteration is a sequence of concave functions in k.

In what follows, for a real-valued function the norm $\|V_n^h\|$ is as defined in (4.3), and the distance between operators $\|P_g - P_{\widehat{g}}\|$ is as defined in (4.9). Also, for an ℓ-dimensional function $g = (g_1, \ldots, g_r, \ldots, g_\ell)$ let

$$\|g\| = \max_{0\leq r\leq \ell} \left| \max_{(k_0^j, z_0^i)} g_r(k_0^j, z_0^i) \right|. \tag{5.1}$$

The following simple result will be very useful.

LEMMA 5.1. *Let h be the grid size of triangulation $\{S^j, D^i\}$. Then, there exists a constant κ that depends on the grid configuration and the dimension ℓ such that $\|P_g - P_{\widehat{g}}\| \leq \frac{\kappa}{h}\|g - \widehat{g}\|$.*

As presently shown, constant κ depends on the uniformity of the grid; also, $\frac{\kappa}{h}$ converges to ∞ as h goes to 0. Therefore, the constants involved in our convergence results below will depend on the uniformity of the grid and on the mesh level, h, of the discretization. These constants get unbounded as h goes to zero.

Proof of Lemma 5.1. The proof becomes more transparent for the case $\ell = 1$. In this simple case, we can show that

$$\kappa = \frac{2h}{h_0}, \tag{5.2}$$

where $h_0 = \min_r\{k^{r+1} - k^r\}$ is the minimum distance over all pairs of adjacent grid points. Thus, for a uniform grid we obtain $\kappa = 2$.

To verify (5.2), consider two points $k_1 = g(k_0^j, z_0^i)$ and $\widehat{k}_1 = \widehat{g}(k_0^j, z_0^i)$. If k_1 and $\widehat{k}_1$ are contained in the same grid interval $[k^r, k^{r+1}]$, then $k = \lambda(k)k^r + (1-\lambda(k))k^{r+1}$ for $k = k_1, \widehat{k}_1$. Hence,

$$\begin{aligned} &|\lambda(k_1) - \lambda(\widehat{k}_1)| + |(1-\lambda(k_1)) - (1-\lambda(\widehat{k}_1))| \\ &\qquad = 2|\lambda(k_1) - \lambda(\widehat{k}_1)| \leq \frac{2|k_1 - \widehat{k}_1|}{k^{r+1} - k^r} \leq \frac{2|k_1 - \widehat{k}_1|}{h_0}. \end{aligned} \tag{5.3}$$

If $k_1 < \widehat{k}_1$ belong to different grid intervals, then $k^r < k_1 < k^{r+1} < \cdots < k^{r+n} < \widehat{k}_1 < k^{r+n+1}$ for some integers r and n. Notice that for $n > 1$ we get from (4.9) that

$\|P_g - P_{\widehat{g}}\| = 2$. Thus, for this case the result is trivially satisfied for κ in (5.2). If $n = 1$, then (4.9) amounts to

$$|\lambda(k_1)| + |(1 - \lambda(k_1)) - \lambda(\widehat{k}_1)| + |1 - \lambda(\widehat{k}_1)|. \tag{5.4}$$

By continuity, for $\widehat{k}_1 = k^{r+1}$ the bound in (5.3) is also valid for (5.4). From $\widehat{k}_1 = k^{r+1}$ this bound can be extended for all $\widehat{k}_1$ in the interval $[k^{r+1}, k^{r+2}]$, since $\lambda(\widehat{k}_1)$ has Lipschitz constant bounded by $1/h_0$. Consequently,

$$|\lambda(k_1)| + |(1 - \lambda(k_1)) - \lambda(\widehat{k}_1)| + |1 - \lambda(\widehat{k}_1)| \leq \frac{2|k_1 - \widehat{k}_1|}{h_0}.$$

Therefore, in all cases the stated result holds true for (k_0^j, z_0^i), and (k_0^j, z_0^i) is an arbitrarily chosen vertex point.

The proof in the multidimensional case is very similar. Let us assume that the domain K is subdivided into a family of simplices $\{S^j\}$ such that $\bigcup_j S^j = K$ and $\text{int}(S^i) \cap \text{int}(S^j) = \varnothing$ for every pair S^i, S^j. Then, every point k in S^j has a unique representation as a convex combination of the vertex points, $k = \sum_{j=1}^{\ell+1} \lambda_j k_0^j$. Moreover, every λ_j is a Lipschitz function on K, and we can find a uniform Lipschitz constant that applies for all λ_j. Therefore, given any pair of points $k_1 = g(k_0^j, z_0^i)$ and $\widehat{k}_1 = \widehat{g}(k_0^j, z_0^i)$, the proof proceeds in the same way as above. □

PROPOSITION 5.2. *Let $\{V_n^h(k,z)\}_{n\geq 1}$ be a sequence of functions generated by* (4.4)–(4.5)*, and assume that every function V_n^h is concave in k. Let $\{g_n^h\}_{n\geq 1}$ be the corresponding sequence of policy functions. Then, under Assumptions* 1 *and* 2 *there exists a constant L such that for any pair of functions V_n^h, V_{n+1}^h, it must hold that $\|P_{g_n^h} - P_{g_{n+1}^h}\| \leq L\|V_n^h - V_{n+1}^h\|^{1/2}$.*

Proof. First, the contraction property of operator T^h implies that

$$\left\|v_{g_n^h} + \beta P_{g_n^h} V_n^h - v_{g_{n+1}^h} - \beta P_{g_{n+1}^h} V_{n+1}^h\right\| \leq \beta \left\|V_n^h - V_{n+1}^h\right\|.$$

Moreover, by (4.3) and (4.7), we have $\|\beta P_{g_{n+1}^h} V_n^h - \beta P_{g_{n+1}^h} V_{n+1}^h\| \leq \beta\|V_n^h - V_{n+1}^h\|$. Hence, an application of the triangle inequality yields

$$\left\|v_{g_n^h} + \beta P_{g_n^h} V_n^h - v_{g_{n+1}^h} - \beta P_{g_{n+1}^h} V_n^h\right\| \leq 2\beta \left\|V_n^h - V_{n+1}^h\right\|. \tag{5.5}$$

Now, as is well known (e.g., see Lemma 3.2 of Santos (2000)) by the concavity of V_n^h in k, the convexity of Ω, and the postulated concavity of v in Assumption 2, we can assert from (5.5) that

$$\left\|g_n^h - g_{n+1}^h\right\| \leq 2\left(\frac{\beta}{\eta}\right)^{1/2} \left\|V_n^h - V_{n+1}^h\right\|^{1/2}. \tag{5.6}$$

Therefore, a straightforward application of Lemma 5.1 proves the result for $L = 2\frac{\kappa}{h}(\frac{\beta}{\eta})^{1/2}$. □

It should be stressed that in the preceding proof only one value function V_n^h needs to be concave. Hence, the following result is an easy consequence of the previous arguments.

COROLLARY 5.3. *Let V^h be the fixed point of the discretized Bellman equation* (4.12)*, and assume that $V^h(k,z)$ is a concave function in k. Let $\{V_n^h\}_{n\geq 1}$ be a sequence of (not necessarily concave) functions generated by* (4.4)–(4.5)*. Then, under*

Assumptions 1 *and* 2, *there exists a constant* L *such that*

$$\left\|P_{g^h} - P_{g_n^h}\right\| \leq L \left\|V^h - V_n^h\right\|^{1/2} \quad \textit{for all } n.$$

Remark 5.4. Observe that for $\ell = 1$ the fixed-point solution $V^h(k,z)$ is in fact a concave function in k. Indeed, for $\ell = 1$ the concavity of $V^h(k,z)$ can be established by the method of successive approximations.

Remark 5.5. Note that in Corollary 5.3 operator $P_{g_n^h}$ is not necessarily unique, since there may not be a unique set of maximizers g_n^h. But the result holds true for any $P_{g_n^h}$, and the distance between two optimal policies must be arbitrarily small for a large enough n.

All the basic ingredients are now in place to demonstrate that the algorithm displays a convergence rate equal to 1.5 (cf. Puterman and Brumelle (1979)).

THEOREM 5.6. *Assume that the conditions of Proposition* 5.2 *are satisfied. Then*

$$\left\|V_{n+1}^h - V_n^h\right\| \leq \frac{\beta L}{1-\beta} \left\|V_n^h - V_{n-1}^h\right\|^{1.5} \quad \textit{for all } n.$$

Proof. First, for any two functions V_n^h and V_{n-1}^h we must have (see (4.11))

$$B^h(V_{n-1}^h) - B^h(V_n^h) \geq -[I - \beta P_{g_n^h}][V_{n-1}^h - V_n^h]. \tag{5.7}$$

Moreover, a further application of (5.7) yields

$$\begin{aligned} B^h(V_n^h) - [I - \beta P_{g_n^h}][V_{n-1}^h - V_n^h] &\leq B^h(V_{n-1}^h) \\ &\leq B^h(V_n^h) - [I - \beta P_{g_{n-1}^h}][V_{n-1}^h - V_n^h]. \end{aligned}$$

Now, subtracting $B^h(V_n^h) - [I - \beta P_{g_n^h}][V_{n-1}^h - V_n^h]$ from each of the three terms, we obtain

$$\begin{aligned} 0 &\leq B^h(V_{n-1}^h) - B^h(V_n^h) + [I - \beta P_{g_n^h}][V_{n-1}^h - V_n^h] \\ &\leq ([I - \beta P_{g_n^h}] - [I - \beta P_{g_{n-1}^h}])[V_{n-1}^h - V_n^h]. \end{aligned} \tag{5.8}$$

Then, for any V_{n+1}^h and V_n^h satisfying (4.4)–(4.5) we must have

$$\begin{aligned} \left\|V_{n+1}^h - V_n^h\right\| &= \left\|[I - \beta P_{g_n^h}]^{-1} B^h(V_n^h)\right\| \\ &\leq \left\|[I - \beta P_{g_n^h}]^{-1}\right\| \left\|B^h(V_n^h)\right\| \\ &= \left\|[I - \beta P_{g_n^h}]^{-1}\right\| \left\|B^h(V_n^h) - B^h(V_{n-1}^h) + [I - \beta P_{g_{n-1}^h}][V_n^h - V_{n-1}^h]\right\| \\ &\leq \left\|[I - \beta P_{g_n^h}]^{-1}\right\| \left\|([I - \beta P_{g_n^h}] - [I - \beta P_{g_{n-1}^h}])[V_n^h - V_{n-1}^h]\right\|. \end{aligned} \tag{5.9}$$

Here, both equalities come from (4.10). Also, the first inequality follows from the definition of the norm, and the last inequality is a consequence of (5.8).

Therefore, from (5.9) we obtain

$$\left\|V_{n+1}^h - V_n^h\right\| \leq \left\|[I - \beta P_{g_n^h}]^{-1}\right\| \left\|[I - \beta P_{g_n^h}] - [I - \beta P_{g_{n-1}^h}]\right\| \left\|V_n^h - V_{n-1}^h\right\|. \tag{5.10}$$

Finally, Proposition 5.2 together with (5.10) implies that

$$\left\|V_{n+1}^h - V_n^h\right\| \leq \frac{\beta L}{1-\beta} \left\|V_n^h - V_{n-1}^h\right\|^{1.5}. \qquad \square$$

THEOREM 5.7. *Assume that the conditions of Corollary* 5.3 *are satisfied. Then,*

$$\left\|V^h - V^h_{n+1}\right\| \leq \frac{\beta L}{1-\beta}\left\|V^h - V^h_n\right\|^{1.5} \quad \textit{for all } n.$$

Proof. From the monotonicity of policy iteration we have $0 \leq V^h - V^h_{n+1}$. Then,

$$\begin{aligned} 0 \leq V^h - V^h_{n+1} &= V^h - V^h_n - [I - \beta P_{g^h_n}]^{-1} B^h(V^h_n) \\ &\leq [I - \beta P_{g^h_n}]^{-1}[I - \beta P_{g^h_n}][V^h - V^h_n] - [I - \beta P_{g^h_n}]^{-1}[I - \beta P_{g^h}][V^h - V^h_n] \\ &= [I - \beta P_{g^h_n}]^{-1}([I - \beta P_{g^h_n}] - [I - \beta P_{g^h}])[V^h - V^h_n]. \end{aligned}$$

Here, the first equality comes from (4.10); the inequality is a consequence of the maximization involved in operator B^h (cf. (4.11)) and the fact that $B^h(V^h) = 0$; and after a simple factorization we get the last equality.

Now, taking norms and applying Proposition 5.2 it follows that

$$\begin{aligned} \left\|V^h - V^h_{n+1}\right\| &\leq \left\|[I - \beta P_{g^h_n}]^{-1}\right\| \left\|[I - \beta P_{g^h_n}] - [I - \beta P_{g^h}]\right\| \left\|V^h - V^h_n\right\| \\ &\leq \frac{\beta L}{1-\beta}\left\|V^h - V^h_n\right\|^{1.5}. \qquad \square \end{aligned}$$

5.2. Local convergence properties. Suitable variations of the preceding arguments will now allow us to establish further convergence properties near the fixed-point solution V^h.

5.2.1. Quadratic convergence. To guarantee the quadratic convergence of policy iteration we need the following strengthening of Corollary 5.3.

PROPOSITION 5.8. *Let V^h be the fixed point of the discretized Bellman equation* (4.12). *Assume that $V^h(k,z)$ is a concave function in k. For each vertex point (k^j_0, z^i_0), assume that $g^h(k^j_0, z^i_0)$ is not a grid point in the family of simplices $\{S^j\}$. Let $\{V^h_n\}_{n\geq 1}$ be a sequence of functions generated by* (4.4)–(4.5). *Then, under Assumptions* 1 *and* 2, *there are constants $\widehat{L}$ and $\widehat{N}$ such that*

$$\left\|P_{g^h} - P_{g^h_n}\right\| \leq \widehat{L}\left\|V^h - V^h_n\right\| \quad \textit{for all } n \geq \widehat{N}.$$

After some obvious adjustments in the power estimates of the proof of Theorem 5.7, we now get the convergence result in Theorem 5.9.

THEOREM 5.9. *Assume that the conditions of Proposition* 5.8 *are satisfied. Then,*

$$\left\|V^h - V^h_{n+1}\right\| \leq \frac{\beta\widehat{L}}{1-\beta}\left\|V^h - V^h_n\right\|^2 \quad \textit{for all } n \geq \widehat{N}.$$

Remark 5.10. One may argue that Theorem 5.9 is a stronger result than Theorem 5.7. But Theorem 5.7 may be of interest for numerical applications[2] as constant L is easier to estimate. Also, note that constant $L = 2\frac{\kappa}{h}(\frac{\beta}{\eta})^{1/2}$ is $O(\frac{1}{h})$, whereas the proof of Proposition 5.8 below reflects that constant $\widehat{L}$ is $O(\frac{1}{h^2})$. Finally, the arguments

[2]For instance, we can define the region of convergence $R_\alpha = \{V \in \mathcal{V}^h | \frac{\beta L}{1-\beta}\|V - V^h\|^{0.5} \leq 1-\alpha\}$, where V^h is the fixed-point solution and $\alpha > 0$. Then, for all V in R_α we can find an upper bound on the number of policy iteration steps that are necessary to reach a certain tolerance level ε.

leading to the proof of Theorem 5.9 are heavily dependent on certain properties of piecewise linear approximations and the assumed interiority of the solution. In contrast, the arguments involved in the proof of Theorem 5.7 seem to be less specific.

Proof of Proposition 5.8. Let every function V^h in $\mathcal{V}^h$ be represented by a finite-dimensional vector u^h that lists all nodal values $\{V^h(k_0^j, z_0^i)\}$. Then, we may rewrite optimization problem (4.12) as follows:

$$\max_{k_1} \ \psi(k_0^j, z_0^i, k_1, u^h) = \max_{k_1} \ v(k_0^j, k_1, z_0^i) + \beta \int_Z V^h(k_1, z') Q(dz', z_0^i) \quad \text{s.t. } (k_0^j, k_1, z_0^i) \in \Omega. \tag{5.11}$$

In what follows, there is no restriction of generality to focus on a single vertex point (k_0^j, z_0^i). The mapping $\psi(k_0^j, z_0^i, \cdot, \cdot)$ has the following properties:

(i) The mapping $\psi(k_0^j, z_0^i, \cdot, \cdot)$ is continuous in (k_1, u).

(ii) For (k_0^j, z_0^i, u^h) function $\psi(k_0^j, z_0^i, k_1, u^h) + \frac{\eta}{2}\|k_1\|^2$ is concave in k_1.

(iii) Let $D_3\psi(k_0^j, z_0^i, k_1, u; v)$ be the directional derivative of function ψ at (k_0^j, z_0^i, k_1, u) with respect to k_1 in the direction v. Let $B_\delta(u^h) = \{u \mid \|u - u^h\| < \delta\}$. Then, for some small $\delta > 0$, and for all $\|v\| = 1$ and all k_1 sufficiently close to $k_1^h = g^h(k_0^j, z_0^i)$, there is a constant $H > 0$ such that $|D_3\psi(k_0^j, z_0^i, k_1, u; v) - D_3\psi(k_0^j, z_0^i, k_1, u^h; v)| \leq H\|u - u^h\|$ for every u in $B_\delta(u^h)$.

The continuity of $\psi(k_0^j, z_0^i, \cdot, \cdot)$ follows from standard arguments. In (ii) the curvature parameter η of Assumption 2 still applies as the fixed-point solution $V^h(k_1, z')$ is assumed to be concave in k_1. The Lipschitz property in (iii) comes from the fact that small changes in the nodal values lead to bounded variations in the slopes or directional derivatives of a piecewise linear function. In fact, for a given choice of norm $\|u - u^h\|$ constant H will depend on the form of the triangulation $\{S^j\}$, especially on $1/h_0$ (where h_0 is the minimum distance between two grid points).

Now, the proof of this proposition will result from some simple extensions of standard arguments[3] (e.g., Fleming and Rishel (1975) (p. 170) and Montrucchio (1987) (p. 263)). For fixed (k_0^j, z_0^i) let k_1^h be the unique maximum point in (5.11) and let $\widehat{k}_1$ be a maximum solution under $\psi(k_0^j, z_0^i, k_1, \widehat{u})$ for $\widehat{u}$ in $B_\delta(u^h)$.

By Assumption 2 and the presupposed concavity of $V^h(k_1, z')$ in k_1, we have

$$\psi(k_0^j, z_0^i, k_1^h, u^h) - \psi(k_0^j, z_0^i, \widehat{k}_1, u^h) \geq -D_3\psi(k_0^j, z_0^i, k_1^h, u^h; \widehat{k}_1 - k_1^h) + \frac{\eta}{2}\|\widehat{k}_1 - k_1^h\|^2. \tag{5.12}$$

Since the objective reaches the maximum value at k_1^h, the directional derivative in (5.12) is nonpositive. Moreover, concave and piecewise linear functions in $\mathbb{R}^\ell$ are absolutely continuous. Hence, we can apply the integral form of the mean-value theorem to the left-hand side of (5.12) so as to obtain

$$-\int_0^1 D_3\psi(\Phi_1(\lambda))d\lambda \geq \frac{\eta}{2}\|\widehat{k}_1 - k_1^h\|^2 \tag{5.13}$$

for $\Phi_1(\lambda) = (k_0^j, z_0^i, k_1^h + \lambda(\widehat{k}_1 - k_1^h), u^h; \widehat{k}_1 - k_1^h)$.

[3]The added difficulty in the proof is that concavity in k_1 is assumed only at point $u = u^h$; i.e., see property (ii) above.

Also, by definition, $\psi(k_0^j, z_0^i, \widehat{k}_1, \widehat{u}) - \psi(k_0^j, z_0^i, k_1^h, \widehat{u}) \geq 0$. Hence,

$$\int_0^1 D_3\psi(\Phi_2(\lambda))d\lambda \geq 0 \tag{5.14}$$

for $\Phi_2(\lambda) = (k_0^j, z_0^i, k_1^h + \lambda(\widehat{k}_1 - k_1^h), \widehat{u}; \widehat{k}_1 - k_1^h)$. Adding up inequalities (5.13)–(5.14), we get

$$-\int_0^1 [D_3\psi(\Phi_1(\lambda)) - D_3\psi(\Phi_2(\lambda))]d\lambda \geq \frac{\eta}{2}\|\widehat{k}_1 - k_1^h\|^2. \tag{5.15}$$

By (iii) above,

$$\int_0^1 |D_3\psi(\Phi_1(\lambda)) - D_3\psi(\Phi_2(\lambda))|d\lambda \leq H\|\widehat{u} - u^h\|\|\widehat{k}_1 - k_1^h\|. \tag{5.16}$$

Therefore, (5.15)–(5.16) implies

$$\|\widehat{k}_1 - k_1^h\| \leq \frac{2H}{\eta}\|\widehat{u} - u^h\|. \tag{5.17}$$

Finally, a straightforward application of Lemma 5.1 establishes Proposition 5.8 for $\widehat{L} = \frac{\kappa}{h}\frac{2H}{\eta}$. □

5.2.2. Superlinear convergence. In this part we lift the convexity and concavity conditions. Hence, we just assume that Ω is the graph of a continuous correspondence $\Gamma : K \times Z \to K$ defined on compact set $K \times Z$ and the reward function v is a continuous mapping on Ω.

Even though we are not able to guarantee quadratic convergence for the iteration scheme, we shall establish the intermediate property of superlinear convergence. This is a faster rate of convergence than that of the method of successive approximations, where by the contraction property of the dynamic programming operator one easily shows that

$$\|V^h - V_{n+1}^h\| \leq \beta\|V^h - V_n^h\| \tag{5.18}$$

for all $n \geq 1$, for every sequence $\{V_n^h\}_{n\geq 1}$ generated by (4.13).

PROPOSITION 5.11. *Let $\{V_n^h\}_{n\geq 1}$ be a sequence of functions generated by* (4.4)–(4.5) *that converge to the fixed point V^h. Then, for every $\epsilon > 0$ there is $\widehat{n}$ such that for each $P_{g_n^h}$ with $n \geq \widehat{n}$ there exists some P_{g^h} with the property that $\|P_{g_n^h} - P_{g^h}\| \leq \epsilon$.*

What this result states is that the correspondence of maximizers is upper-semicontinuous. Hence, for every $P_{g_n^h}$ one can find an arbitrarily close P_{g^h}, provided that n is sufficiently large. With Proposition 5.11 at hand, from the proof of Theorem 5.7 one can easily obtain the following result.

THEOREM 5.12. *Assume that the conditions of Proposition* 5.11 *are satisfied. Then,*

$$\limsup_{n\to\infty} \frac{\|V^h - V_{n+1}^h\|}{\|V^h - V_n^h\|} = 0. \tag{5.19}$$

6. Numerical experiments. In this section we report some simple numerical experiments which are intended to evaluate the performance of policy iteration and the method of successive approximations. The analysis centers on a one-sector deterministic growth model with leisure, and the main purpose is to evaluate the computing cost and approximation errors of the value and policy functions under each computational method.

Formally, our one-sector growth model is described by the following optimization problem:

$$
\begin{aligned}
V(k_0) = \max_{\{c_t, l_t, i_t\}} & \sum_{t=0}^{\infty} \beta^t [\lambda \log c_t + (1-\lambda)\log l_t] \\
\text{s.t. } & c_t + i_t = A k_t^{\alpha} (1-l_t)^{1-\alpha}, \\
& k_{t+1} = i_t + (1-\delta) k_t, \\
& k_t, c_t \geq 0,\ 0 \leq l_t \leq 1,\ k_0 \text{ given},\ t = 0, 1, 2, \ldots, \\
& 0 < \beta < 1,\ 0 < \lambda < 1,\ A > 0, \\
& 0 < \alpha < 1,\ 0 < \delta \leq 1.
\end{aligned} \tag{6.1}
$$

This is a familiar optimization problem in which c represents consumption, k represents the stock of physical capital, and l the fraction of time devoted to leisure activities. It is well known that for $\delta = 1$ the value function $V(k_0)$ takes the simple form $V(k_0) = B + C \log k_0$, where B is a certain constant and $C = \frac{\lambda\alpha}{(1-\alpha\beta)}$. Also, the policy function $k_{t+1} = \alpha\beta A k_t^{\alpha}(1-l_t)^{1-\alpha}$ with $l_t = \frac{(1-\lambda)(1-\alpha\beta)}{\lambda(1-\alpha)+(1-\lambda)(1-\alpha\beta)}$. Under these simple functional forms, there is a unique steady state $k^* = g(k^*)$, which is globally stable.

The existence of one state variable, k, and two controls, l and c, suggests that all numerical maximizations may be efficiently carried out with a unique decision variable. Let us then write the model in a more suitable form for our computations. Note that at time $t = 0$ the first-order conditions for our two control variables c and l are given by

$$
\frac{\lambda}{c_0} = \mu \qquad \text{and} \qquad \frac{(1-\lambda)(1-l_0)^{\alpha}}{l_0 A k_0^{\alpha}(1-\alpha)} = \mu,
$$

where μ is a Lagrange multiplier. After some simple rearrangements, we obtain

$$
c_0 = \frac{\lambda l_0 A k_0^{\alpha}(1-\alpha)}{(1-\lambda)(1-l_0)^{\alpha}}.
$$

Hence,

$$
k_1 = A k_0^{\alpha}(1-l_0)^{-\alpha}\left[(1-l_0) - \frac{\lambda l_0(1-\alpha)}{1-\lambda}\right].
$$

The iterative process $V_{n+1}^h = T^h(V_n^h)$ in (4.13) is then effected as follows:

$$
\begin{aligned}
V_{n+1}^h(k_0) = \max_{l_0} & \lambda \log\left(\frac{\lambda A k_0^{\alpha} l_0(1-\alpha)}{(1-\lambda)(1-l_0)^{\alpha}}\right) + (1-\lambda)\log l_0 \\
& + \beta V_n^h\left(A k_0^{\alpha}(1-l_0)^{-\alpha}\left[(1-l_0) - \frac{\lambda l_0(1-\alpha)}{1-\lambda}\right]\right).
\end{aligned} \tag{6.2}
$$

Although expression (6.2) may appear rather cumbersome, this form will prove appropriate for our computations as it involves maximization only in one single variable.

Our numerical exercises will focus on the following parameterization:

$$\beta = 0.95, \qquad \lambda = \frac{1}{3}, \qquad A = 10, \qquad \alpha = 0.34, \qquad \delta = 1.$$

For such values the stationary state is $k^* = 1.9696$.

We consider a uniform grid of points k^j with step size h. For the purposes of this exercise, the domain of possible capitals is restricted to the interval $[h, 10]$. We should remark that in this simple univariate case our interpolations will yield concave value functions $\{V_n^h\}_{n\geq 1}$ for the method of successive approximations, but the sequence of functions $\{V_n^h\}_{n\geq 1}$ may not be concave under policy iteration, since each V_n^h must solve the equation system (4.5) for a given g_{n-1}^h.

Our numerical computations were coded in standard C and run on a Silicon Graphics Octane 2 (with a dual processor, each component rated at 600 MHz), which in a double precision floating-point arithmetic allows for a 16-digit accuracy. All required numerical maximizations were effected by Brent's algorithm (cf. Press et al. (1992)) with an accuracy of 10^{-12}. Such a high precision should allow us to trace out the errors derived from other discretizations embedded in our algorithms. Also, for both policy iteration and the method of successive approximations the iteration process will stop once two consecutive value functions V_n^h and V_{n+1}^h satisfy the tolerance bound

$$\left\|V_n^h - V_{n+1}^h\right\| \leq \frac{1}{5}h^2. \tag{6.3}$$

The adequacy of this stopping rule for the method of successive approximations is discussed in Santos (1999). Roughly, constant $\frac{1}{5}h^2$ is selected so as to balance the approximation error from the use of a finite grid of points and the truncation error from stopping the iteration process in finite time.

For the method of successive approximations as specified in (6.2), we start each numerical exercise with a given grid size h and initial condition $V_0 \equiv 0$. The program then stops once condition (6.3) is satisfied. For each h, Table 6.1 reports the number of iterations, computing time, and the maximum observed errors in the last iteration for the value and policy functions.[4] For policy iteration, we follow the procedure specified in (4.4)–(4.5); for each h the iteration process starts with an initial condition $V_0 \equiv 0$, and it stops once condition (6.3) is satisfied. These findings are displayed in Table 6.2.

From the calculations reflected in Tables 6.1 and 6.2, we now discuss the computational cost and speed of convergence of these two algorithms.

(a) *Complexity.* As one can determine from Table 6.1, for $h = 10^{-1}$ the average time cost per iteration is roughly 0.017 seconds, and for $h = 10^{-2}$ the average time cost per iteration is roughly 0.17 seconds. Hence, for the method of successive approximations the average time cost per iteration grows linearly with the number of grid points. (This regular pattern is also observed for pairwise comparisons of other grid sizes.) To a certain extent, this result is to be expected since the major computational cost in each iteration is the number of maximizations, which grows linearly with the number of grid points. (Incidentally, this exercise shows that the cost of each maximization remains roughly invariant to the grid size.) In contrast, for policy iteration

[4]These values are defined, respectively, by $\|V - V_n^h\|$ and $\|g - g_n^h\|$, where V and g are the closed-form solutions for (6.1), and V_n^h and g_n^h are the computed value and policy functions corresponding to the last iteration, n, under a grid size h.

TABLE 6.1
Computational method: The method of successive approximations with linear interpolation.[a]

No. of vertex points	Mesh size h	Iterations	CPU time	Max. error in V	Max. error in g
100	1.00×10^{-1}	91	1.54	3.84×10^{-2}	5.44×10^{-2}
300	3.33×10^{-2}	128	5.42	3.44×10^{-3}	1.56×10^{-2}
1000	1.00×10^{-2}	181	30.17	3.68×10^{-4}	5.83×10^{-3}
3000	3.33×10^{-3}	223	94.58	3.42×10^{-5}	1.68×10^{-3}
10000	1.00×10^{-3}	271	354.27	3.36×10^{-6}	5.84×10^{-4}

[a]Parameter values: $\beta = 0.95$, $\lambda = \frac{1}{3}$, $A = 10$, $\alpha = 0.34$, and $\delta = 1$.

TABLE 6.2
Computational method: Policy iteration with linear interpolation.[b]

No. of vertex points	Mesh size h	Iterations	CPU time	Constant $\widehat{\mathcal{L}}^h$	Max. error in V	Max. error in g
100	1.00×10^{-1}	4	0.11	27.35	4.36×10^{-2}	6.136×10^{-2}
300	3.33×10^{-2}	5	2.54	1629.63	8.47×10^{-4}	2.264×10^{-2}
1000	1.00×10^{-2}	7	215.93	19816.63	8.51×10^{-6}	7.160×10^{-3}
3000	3.33×10^{-3}	10	16868.33	78308.85	1.35×10^{-6}	2.400×10^{-3}

[b]Parameter values: $\beta = 0.95$, $\lambda = \frac{1}{3}$, $A = 10$, $\alpha = 0.34$, and $\delta = 1$.

we can see from Table 6.2 that for $h = 10^{-1}$ the average time cost per iteration is roughly 0.03 seconds, whereas for $h = 10^{-2}$ the average time cost per iteration goes up to about 31 seconds. Hence, for policy iteration an increase in the number of grid points by a factor of 10 leads to an increase in the average time cost by over a factor of 10^3. Again, this result is to be expected since the most complicated step in policy iteration is (4.5), which involves a matrix inversion. This simple complexity analysis illustrates that policy iteration is faster for small grids, but it becomes relatively more costly for fine grids, unless further operational procedures are introduced for the matrix inversion required in (4.5). Under our present methods, it is extremely costly to go beyond grids of 3000 points for policy iteration, whereas we can carry out the method of successive approximations over grids of about 50000 points.

(b) *Convergence.* It is well known that the dynamic programming algorithm approaches the fixed-point solution at a linear rate, and this has been observed in many applications. In order to evaluate the quadratic convergence of policy iteration, we have computed the corresponding constant

$$\widehat{\mathcal{L}}_n^h = \frac{\left\|V_{\widehat{n}}^h - V_{n+1}^h\right\|}{\left\|V_{\widehat{n}}^h - V_n^h\right\|^2},$$

where $V_{\widehat{n}}^h$ is the value function of an arbitrarily high iteration, $\widehat{n}$, so that $V_{\widehat{n}}^h$ is a good approximation of the fixed point V^h (cf. Theorem 5.9). For each h, in Table 6.2 we report the max value, $\widehat{\mathcal{L}}^h = \max_n \widehat{\mathcal{L}}_n^h$. This constant takes on relatively high values, and it seems to grow as predicted by our analysis. Indeed, from the previous section (cf. Remark 5.10) we may conclude that our worst-case theoretical bounding constant $\widehat{\mathcal{L}} = \frac{\beta\widehat{L}}{1-\beta}$ is at least $O(\frac{1}{h^2})$, which appears to be in line with the observed estimates. The quadratic convergence near the fixed-point solution was further confirmed by a

TABLE 6.3
Computational method: The method of successive approximations with linear interpolation.[c]

No. of vertex points	Mesh size h	Iterations	CPU time	Max. error in V	Max. error in g
100	1.00×10^{-1}	460	6.20	2.09×10^{-1}	4.88×10^{-2}
300	3.33×10^{-2}	694	27.75	1.98×10^{-2}	1.63×10^{-2}
1000	1.00×10^{-2}	920	126.02	1.95×10^{-3}	5.58×10^{-3}
3000	3.33×10^{-3}	1126	470.65	1.95×10^{-4}	1.87×10^{-3}
10000	1.00×10^{-3}	1379	1748.29	1.93×10^{-5}	5.99×10^{-4}

[c]Parameter values: $\beta = 0.99$, $\lambda = \frac{1}{3}$, $A = 10$, $\alpha = 0.34$, and $\delta = 1$.

detailed analysis of the evolution of the errors $\|V_{\widehat{n}}^h - V_n^h\|$. For the sake of brevity, these results are not reported here.

It may seem paradoxical that the number of required iterations in Table 6.2 does not vary greatly with the grid size. But one could argue that stopping rule (6.3) is suitable for the method of successive approximations, which features a linear rate of convergence, but such a stopping rule is not so sensitive for algorithms displaying faster rates of convergence. The insensitivity on the number of policy steps was also observed as we varied the discount factor. Variations in β were reflected in changes in the above constant $\widehat{\mathcal{L}}^h$, but the number of required policy iterations was always below 20. For the method of successive approximations, however, the required number of iterations changes substantially with variations in the discount factor. For instance, in Table 6.3 the discount factor is increased from $\beta = 0.95$ to $\beta = 0.99$. Then, as compared to Table 6.1, the number of required iterations and the corresponding computational cost go up by a factor of 5.

As one can see, the bounding constant $\widehat{\mathcal{L}} = \frac{\beta \widehat{L}}{1-\beta}$ of Theorem 5.9 varies inversely with the discount factor β. A related type of dependence under (6.3) can be established for the bounding constant in the method of successive approximations (see (5.18)). But convergence is linear for the method of successive approximations, and hence changes in the bounding constant should necessarily be reflected in changes of the same magnitude in the number of iteration. Policy iteration, however, displays a faster convergence rate. Thus, for changes in the discount factor and corresponding bounding constant, the extra required iteration steps should be of a smaller order of magnitude. Therefore, quadratic convergence seems to be driving the insensitivity of the required number of policy iteration steps under this algorithm for changes in the grid size of the discretization and the discount factor. This does not mean that it is possible to bound the required number of policy iteration steps regardless of the discount factor or the mesh level of the discretization. Indeed, this paper shows that the theoretical constants involved in the orders of convergence get unbounded as the grid size of the discretization converges to zero.

7. Concluding remarks. This paper provides new convergence results on the policy iteration algorithm. Our work is motivated by the fact that this algorithm usually converges in a small number of steps (typically fewer than 20) even though the number of feasible policies that the policy iteration algorithm could evaluate is huge and increases exponentially fast with the number of states and possible actions. Unfortunately, in section 2 we presented an example of an MDP problem in which the number of iteration steps grows equally with the number of states, thus dispelling

the hope of proving a general result that only a small number of policy iteration steps will be necessary to solve an MDP problem.

We then focused on the observation that policy iteration is equivalent to Newton's method and thus ought to have quadratic rates of convergence. This equivalence holds even for MDP problems where the state and action spaces are no longer finite sets. This very rapid rate of convergence suggests that policy iteration could dominate the method of successive approximations, at least for discount factors β close to 1. A standard sufficient condition used to establish quadratic convergence is that the derivative of the nonlinear operator defining the system of equations to be solved by Newton's method (which in the case of policy iteration equals the identity minus the derivative of the Bellman operator) satisfies a Lipschitz condition. In the original work of Puterman and Brumelle (1979) this Lipschitz condition was assumed to hold globally, but they did not provide easily verifiable primitive conditions under which this Lipschitz bound could be satisfied. As a result, an open question remains: Under what conditions and for which classes of MDP problems might we obtain quadratic convergence rates for policy iteration?

This paper attempts to address this question by considering a relatively narrow class of dynamic models arising in economic growth theory. This class involves continuous state and control variables, and thus some sort of discretization procedure must be used to implement policy iteration in practical situations. The key Lipschitz condition that guarantees quadratic convergence is shown to hold for piecewise linear interpolation under a concavity assumption at the optimal interior solution, and it may be extended to other approximation schemes. But we also demonstrate that a weaker general bound does hold that involves the *square root* of the maximum absolute difference in the value function from the fixed-point solution. To establish that this weaker bound holds globally, we have imposed a concavity condition on both the return function and the fixed-point solution which are defined on a convex domain of state and control variables. Under this concavity assumption, we were able to prove a global result for the policy iteration algorithm of superlinear convergence with an exponent equal to 1.5. Furthermore, the best bounding constant on the errors of our algorithm is inversely proportional to the grid size h of the discretization of the state space. These results suggest the following two pessimistic conclusions about policy iteration: (1) As h goes to zero, not only will the number of states in the approximate MDP problem tend to infinity, but the theoretical convergence bounds (and possibly the number of policy iteration steps) will also tend to infinity as well. (2) Policy iteration will not converge globally at a rapid quadratic rate; quadratic convergence was validated under the concavity assumption and piecewise interpolation, but the general rate of convergence is expected to be superlinear.

Section 6 reported the results of some numerical experiments in which these theoretical results were verified. We corroborated the quadratic convergence of the algorithm, and that the bounding constants evolved as predicted by our analysis. As a result of the computational cost involved in finding the solution in each policy iteration step, our numerical experiments illustrated that for very fine grids policy iteration is substantially slower than simple successive approximations. Thus, our results provide a rather pessimistic perspective on the usefulness of policy iteration for solving large-scale MDP problems. Fine discretizations with large numbers of states are required to approximate the solutions accurately in most applications in economics and engineering. Other studies (see Benitez-Silva et al. (2001)) have suggested that an alternative type of policy iteration known as *parametric policy iteration* can be more effective for solving such large continuous-state MDP problems. Relatively little is

known, however, about the convergence properties of this latter algorithm. In fact, the results of this paper suggest that there is still much to be learned about the convergence properties of the standard policy iteration algorithm.

REFERENCES

R. Bellman (1955), *Functional equations in the theory of dynamic programming.* V. *Positivity and quasi-linearity*, Proc. Nat. Acad. Sci. USA, 41, pp. 743–746.

R. Bellman (1957), *Dynamic Programming*, Princeton University Press, Princeton, NJ.

H. Benitez-Silva, G. Hall, G. Hitch, G. Pauletto, and J. Rust (2001), *A Comparison of Discrete and Parametric Approximation Methods for Continuous-state Dynamic Programming Problems*, manuscript, Yale University, New Haven, CT.

D. Blackwell (1965), *Discounted dynamic programming*, Ann. Math. Statist., 36, pp. 226–235.

G. Dahlquist and A. Bjorck (1974), *Numerical Methods*, MIT Press, Cambridge, MA.

W. H. Fleming and R. W. Rishel (1975), *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York.

R. Howard (1960), *Dynamic Programming and Markov Processes*, MIT Press, Cambridge, MA.

K. L. Judd (1998), *Numerical Methods in Economics*, MIT Press, Cambridge, MA.

V. Klee and G. J. Minty (1972), *How good is the simplex algorithm?*, in Inequalities III, O. Shisha, ed., Academic Press, New York, pp. 159–175.

L. Montrucchio (1987), *Lipschitz continuous policy functions for strongly concave optimization problems*, J. Math. Econom., 16, pp. 259–273.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992), *Numerical Recipes in Fortran* 77*: The Art of Scientific Computing*, Cambridge University Press, Cambridge, UK.

M. L. Puterman and S. L. Brumelle (1979), *On the convergence of policy iteration in stationary dynamic programming*, Math. Oper. Res., 4, pp. 60–69.

R. T. Rockafellar (1970), *Convex Analysis*, Princeton University Press, Princeton, NJ.

J. Rust (1987), *Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher*, Econometrica, 55, pp. 999–1033.

M. S. Santos (1999), *Numerical solution of dynamic economic models*, in Handbook of Macroeconomics, J. B. Taylor and M. Woodford, eds., Elsevier Science, Amsterdam, The Netherlands, pp. 311–386.

M. S. Santos (2000), *Accuracy of numerical solutions using the Euler equation residuals*, Econometrica, 68, pp. 1377–1402.

N. L. Stokey, R. E. Lucas with E. C. Prescott (1989), *Recursive Methods in Economic Dynamics*, Harvard University Press, Cambridge, MA.

J. N. Tsitsiklis (2000), *private communication*, MIT, Cambridge, MA.

J. F. Traub and H. Woźniakowski (1979), *Convergence and complexity of Newton iteration for operator equations*, J. Assoc. Comput. Mach., 26, pp. 250–258.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 2116–2129

POLE ASSIGNMENT FOR A VIBRATING SYSTEM WITH AERODYNAMIC EFFECT*

J. N. WANG†, S. H. CHOU‡, Y. C. CHEN§, AND W. W. LIN§

Abstract. This paper deals with a pole assignment problem by single-input state feedback control arising from a one-dimensional vibrating system with aerodynamic effect. On the practical side, we derive explicit formulae for the required controlling force terms, which can reassign part of the spectrum to the desired values while leaving the remaining spectrum unchanged. On the mathematical side, unlike the classical Sturm–Liouville problem, our eigenvalue problem is associated with a cubic pencil with unbounded operators as coefficients and has many interesting new features, one of which is that a new controllability condition appears. This condition together with the known controllability condition in the quadratic case are necessary and sufficient. This sheds light on the adjustment of the model parameters. We also analyze the spectrum of the associated noncompact operator and in particular show that the discrete spectrums of controlled and uncontrolled systems lie outside a closed interval on the negative real axis.

Key words. vibrating system, aerodynamic effect, state feedback control, pole assignment

AMS subject classifications. 93B55, 93B52

DOI. 10.1137/S0363012902411568

1. Introduction.

Consider a vibrating system whose displacement $u = u(x,t)$ is governed by the initial boundary value problem

$$\begin{aligned} &\partial_x[p_1(x)\partial_x u + p_2(x)\mathcal{W}(\partial_x u)] - q(x)\partial_t^2 u = 0, \qquad 0 < x < L,\ t > 0, \\ &u(0,t) = u(L,t) = 0 \end{aligned} \tag{1.1}$$

with proper initial conditions, where $p_1(x)$, $p_2(x)$, and $q(x)$ are real-valued positive functions, and $\mathcal{W}$ is an integral operator defined by

$$\mathcal{W}v(x,t) = \rho v + \rho \int_0^t e^{\omega(t-s)} v(x,s)ds \tag{1.2}$$

for any complex-valued function $v : [0,L] \times [0,\infty) \to \mathbb{C}$, where $\rho \neq 0$ and $\omega \neq 0$ are real constants. The function $\mathcal{W}(\partial_x u)$ is called the Wagner lift-growth buildup function accounting for some dynamic effect. If $\mathcal{W}$ is zero, we recover the familiar lateral vibrating string or longitudinal vibrating rod case, depending on the boundary conditions. The more general nonzero case, i.e., the vibrating system with aerodynamic effect incorporated, has its origin in a dynamic loads analysis system (DYLOFLEX) [1].

Applying (1.2) to (1.1) with $v = \partial_x u$, multiplying (1.1) by $e^{-\omega t}$, and then differentiating with respect to t, we obtain the third order differential system

$$\begin{aligned} &\partial_x(\alpha(x)\partial_x u) + \partial_x(\beta(x)\partial_x\partial_t u) + \omega q(x)\partial_t^2 u - q(x)\partial_t^3 u = 0, \\ &u(0,t) = u(L,t) = 0, \end{aligned} \tag{1.3}$$

*Received by the editors July 18, 2002; accepted for publication (in revised form) August 30, 2003; published electronically February 18, 2004.

http://www.siam.org/journals/sicon/42-6/41156.html

†Department of Mathematics, National Taiwan University, Taipei, Taiwan (jnwang@math.ntu.edu.tw). This author was supported by NSC, Taiwan.

‡Department of Mathematics and Statistics, Bowling Green State University Bowling Green, OH 43403 (chou@bgnet.bgsu.edu). This author was supported by NCTS, Taiwan, and in part by NSF under DMS-0074259.

§Department of Mathematics, National Tsing Hua University, Hsinchu, 30043, Taiwan (g893260@oz.nthu.edu.tw, wwlin@math.nthu.edu.tw). The fourth author was supported by NSC, Taiwan.

where

$$\begin{aligned} \alpha(x) &= (1-\omega)\rho p_2(x) - \omega p_1(x), \\ \beta(x) &= p_1(x) + \rho p_2(x). \end{aligned} \tag{1.4}$$

To look for vibration modes, we substitute the form $u(x,t) = \phi(x)e^{\lambda t}, \lambda \in \mathbb{C}$, into (1.3) and obtain the eigenvalue problem

$$\begin{aligned} &\mathcal{L}(x, D, \lambda)\phi := (\alpha(x)\phi')' + \lambda(\beta(x)\phi')' + \lambda^2\omega q(x)\phi - \lambda^3 q(x)\phi = 0, \\ &\phi(0) = \phi(L) = 0, \end{aligned} \tag{1.5}$$

where $D = ' = \frac{d}{dx}$. Unlike the classical Sturm–Liouville eigenvalue problem, this problem is cubic in λ. To the best of our knowledge we have not seen such a formulation before. The main purpose of this paper is to study the pole assignment associated with this problem, using the state feedback control function $b(x)$. Namely, we look at the controlled system

$$\begin{aligned} &\partial_x[p_1(x)\partial_x v + p_2(x)\mathcal{W}(\partial_x v)] - q(x)\partial_t^2 v = b(x)w(t), \qquad 0 < x < L,\ t > 0, \\ &v(0,t) = v(L,t) = 0, \end{aligned} \tag{1.6}$$

where the feedback control force $w(t)$ has the form

$$w(t) = \int_0^L [f_1(x)\partial_t v(x,t) + f_2(x)v(x,t) + g_1(x)\tilde{\mathcal{W}}(\partial_x v)(x,t) + g_2(x)\tilde{\mathcal{W}}(v)(x,t)]dx$$

with

$$\tilde{\mathcal{W}}(v)(x,t) = \rho \int_0^t e^{\omega(t-s)} v(x,s)ds. \tag{1.7}$$

Note that here the first term ρv in the definition (1.2) has been absorbed into the first two terms on the right side of (1.7). Now substitution of $v(x,t) = \psi(x)e^{\lambda t}$ as done previously yields the eigenvalue problem associated with the controlled problem:

$$\begin{aligned} \mathcal{L}_c(x, D, \lambda)\psi &:= (\alpha(x)\psi')' + \lambda(\beta(x)\psi')' + \lambda^2\omega q(x)\psi - \lambda^3 q(x)\psi \\ &\quad - b\int_0^L \lambda(\lambda f_1 + f_2)\psi - \omega(\lambda f_1 + f_2)\psi + \rho g_1\psi' + \rho g_2\psi dx = 0, \\ \psi(0,t) &= \psi(L,t) = 0. \end{aligned} \tag{1.8}$$

From now on, we write $\mathcal{L}$ or $\mathcal{L}(\lambda)$ for $\mathcal{L}(x, D, \lambda)$ when no confusion can arise. Similar notation is also used for $\mathcal{L}_c(x, D, \lambda)$. Before investigating the pole assignment problem for (1.1), we first analyze its spectrum structure. Let the function $\alpha(x)$ in (1.4) be continuously differentiable and $\alpha(x) \neq 0$ for all x in the interval $[0, L]$. For definiteness, we assume that α is positive throughout the whole interval. The function $\beta = \beta(x) > 0$ is continuously differentiable. Then it is shown using the analytic Fredholm theorem [13] (cf. Theorem 3.3 and Remark 3.1) that the operator pencil $\mathcal{L}$ has only discrete spectrum in $\mathbb{C} \setminus E := E^c$, where

$$E := \left[-\max_{0\le x\le L} \frac{\alpha(x)}{\beta(x)}, -\min_{0\le x\le L} \frac{\alpha(x)}{\beta(x)} \right]. \tag{1.9}$$

Since the interval E lies in the negative real axis, we are mainly concerned with relocating discrete spectrum or poles of (1.5) in $E^c \setminus (-\infty, 0)$. More precisely, let the discrete spectrum of $\mathcal{L}$ in E^c be $\{\lambda_1, \dots, \lambda_\ell, \lambda_{\ell+1}, \dots\}$ and the first ℓ poles $\{\lambda_1, \dots, \lambda_\ell\}$ be distinct, closed under complex conjugation, and furthermore $\{\lambda_1, \dots, \lambda_\ell\} \cap \{\lambda_{\ell+1}, \dots\} = \emptyset$. Our goal is to replace $\{\lambda_1, \dots, \lambda_\ell\}$ by $\{\mu_1, \dots, \mu_\ell\}$, which is a conjugate set of distinct complex values in E^c, with $\{\mu_1, \dots, \mu_\ell\} \cap \{\lambda_1, \lambda_2, \dots\} = \emptyset$. It turns out under suitable *controllability conditions* that we can find explicitly the functions f_1, f_2, g_1, and g_2 so that $\{\lambda_1, \dots, \lambda_\ell\}$ are replaced by $\{\mu_1, \dots, \mu_\ell\}$ and other poles $\{\lambda_{\ell+1}, \lambda_{\ell+2}, \dots\}$ remain unchanged. In Theorem 4.1, the reader can find the following formulae for the above functions with ϕ_j being the eigenfunction:

$$f_1(x) = q \sum_{j=1}^{\ell} \xi_j \phi_j,$$

$$f_2(x) = q \sum_{j=1}^{\ell} \xi_j \lambda_j \phi_j,$$

$$g_1(x) = \left(\frac{\beta}{\rho}\right) \sum_{j=1}^{\ell} \xi_j \phi_j',$$

$$g_2(x) = \left(\frac{q}{\rho}\right) \sum_{j=1}^{\ell} \xi_j \lambda_j^2 \phi_j,$$

where

$$\xi_j = \frac{\lambda_j - \mu_j}{\int_0^L b\phi_j dx} \prod_{r=1,\, r\neq j}^{\ell} \frac{\lambda_j - \mu_r}{\lambda_j - \lambda_r}$$

and

$$\int_0^L b\phi_j \neq 0 \tag{1.10}$$

for $j = 1, 2, \dots, \ell$. Note that (1.10) can be seen as a continuous version of the usual controllability condition in the matrix (discrete) case (see, for example, [3]). However, unlike the matrix case, (1.10) alone does not guarantee that λ_j is a controllable mode. As a matter of fact, for the continuous case, we need to define an additional controllability condition

$$(3\lambda_j^2 - 2\omega\lambda_j) \int_0^L q(x)\phi_j(x)\phi_j(x)dx + \int_0^L \beta(x)\phi_j'(x)\phi_j'(x)dx \neq 0 \tag{1.11}$$

for any $1 \leq j \leq \ell$. With the help of the two controllability conditions (1.10) and (1.11), we show in Theorem 4.2 that outside E no extra discrete spectrum of $\mathcal{L}_c$ are generated except those poled. That is, the discrete spectrum of $\mathcal{L}_c$ in E^c is precisely described by $\{\mu_1, \dots, \mu_\ell, \lambda_{\ell+1}, \dots\}$. To further refine the answer to the pole assignment problem for (1.1), we also prove that the essential spectrum of $\mathcal{L}_c$ is identical to that of $\mathcal{L}$, i.e., the essential spectrum of $\mathcal{L}$ does not change in the course of state feedback control. This property is, roughly, due to the fact that $\mathcal{L}_c$ is a compact perturbation of $\mathcal{L}$. Finally, in order for the controlled system to be realizable, the functions f_1, f_2, g_1, and g_2 need to be real, which is shown in Theorem 4.4.

The pole assignment problem, which is concerned with assigning all eigenvalues to desired locations, is a well known and important problem in control theory. It has been extensively studied for the linear system with state or output feedback control. We refer to [16] for the detailed description of this problem and to [2] for the state-of-the-art numerical methods. Our problem here is a variant of the pole assignment problem called the *partial* pole assignment problem, which is concerned with assigning some eigenvalues to desired positions and keeping all other eigenvalues unchanged. The partial pole assignment is more practical than the pole assignment problem, especially for distributed parameter models where we encounter infinitely many eigenvalues. The most studied distributed parameter model in this respect is the vibrating system (see, for example, [10]). The partial pole assignment problem with state feedback control for the usual vibrating distributed parameter system has been considered in [5], [6], [7], [12]. It should be noted that systems in the matrix formulation can be treated as approximations of distributed parameter systems by finite-difference or finite-element methods. In this setting, the problem becomes finite-dimensional.

This paper is partly motivated by the article [12], which considered the partial pole assignment for the vibrating rod without aerodynamic effect ($\mathcal{W} = 0$ in (1.1)). The resulting equation is a standard Sturm–Liouville type. Therefore, the spectrum consists of only the discrete spectrum (i.e., eigenvalues). In [12], an explicit solution to the partial pole assignment problem with suitable state feedback was constructed and the conditions under which this solution is unique were determined. Before [12], by the similar state feedback law, Russell [15] considered the (full) pole assignment for a class of hyperbolic distributed parameter control systems. Another motivation of this paper is the partial pole assignment problem for a discrete version of (1.1) [11]. In [11], we dealt with a cubic matrix pencil rather than an operator pencil. The similar pole assignment problem for the quadratic matrix pencil was studied in [3], [4].

In the presence of aerodynamic effect, the equation in (1.1) is a Volterra integro-differential type. This type of equation also arises in modelling phenomena involving viscoelasticity. Our pole assignment method can be used to stabilize the vibrating system with aerodynamic effect by suitable state feedback control. From the perspective of the stabilization by state feedback control, some related results for Volterra integrodifferential equations were obtained in [8], where the authors considered exponential stabilization of an abstract linear Volterra integrodifferential equation in a Hilbert space

$$u'' = -E_1 A u(t) + E_2 \int_0^t k(t-s) A u(s) ds + f(t)$$

by a state feedback control given as

$$f(t) = -C_0 u(t) - C_1 u'(t),$$

where A is a positive semidefinite self-adjoint unbounded operator, E_1, E_2 are positive constants, and C_0, C_1 are bounded linear operators of finite rank. Here k is a nonnegative, convex, and exponentially decreasing function with finite value at 0. Besides considering a simpler system and having different viewpoints from ours, [8] used techniques from semigroup theory. The similar problem was also considered in [9] with a slightly different system.

2. Elementary properties of the eigensystem. In this section, we will derive some general properties of the eigenstructure of (1.5). In mathematical formalism,

the cubic eigenvalue problem is to find a complex number λ and a complex function ϕ such that

$$\mathcal{L}(\lambda)\phi = 0, \tag{2.1}$$

$$\phi(0) = \phi(L) = 0. \tag{2.2}$$

Assume for the time being that the eigenpairs (2.1) and (2.2) exist and that the eigenmodes are C^1 functions.

THEOREM 2.1 (dimension of eigenspace). *Let λ be an eigenvalue of* (1.5). *Then the dimension of the eigenspace corresponding to λ is one, provided* $(\alpha(x)+\lambda\beta(x)) \neq 0$ *identically.*

Proof. Let $\{\lambda, \phi\}$ and $\{\lambda, \psi\}$ be eigenpairs of the eigenvalue problem (1.5), where ϕ and ψ are smooth. Then we have

$$\mathcal{L}\phi = (\alpha(x)\phi')' + \lambda(\beta(x)\phi')' + \lambda^2\omega q(x)\phi - \lambda^3 q(x)\phi = 0, \tag{2.3}$$

$$\mathcal{L}\psi = (\alpha(x)\psi')' + \lambda(\beta(x)\psi')' + \lambda^2\omega q(x)\psi - \lambda^3 q(x)\psi = 0. \tag{2.4}$$

Subtracting $\phi\mathcal{L}\psi$ from $\psi\mathcal{L}\phi$, we get

$$(\alpha(x)\phi')'\psi + \lambda(\beta(x)\phi')'\psi - (\alpha(x)\psi')'\phi - \lambda(\beta(x)\psi')'\phi = 0,$$

which implies

$$\frac{d}{dx}((\alpha(x) + \lambda\beta(x))(\phi'\psi - \phi\psi')) = 0.$$

Using the boundary conditions, we have

$$(\alpha(x) + \lambda\beta(x))(\phi'\psi - \phi\psi') = 0, \qquad 0 < x < L.$$

Noting that $(\alpha(x) + \lambda\beta(x)) \neq 0$ and ϕ, and that ψ are continuously differentiable, we conclude that the Wronskian

$$\phi'\psi - \phi\psi' = 0,$$

and hence the functions ϕ and ψ are dependent. This implies that the dimension of the eigenspace corresponding to λ is one. ∎

THEOREM 2.2 (orthogonality relation). *Given two complex functions f and g, we define*

$$\langle f, g\rangle := \int_0^L f(x)g(x)dx. \tag{2.5}$$

Let $\{\lambda_m, \phi_m\}$ and $\{\lambda_n, \phi_n\}$ be two distinct eigenpairs, i.e., $\lambda_m \neq \lambda_n$. Then we have the following relation:

$$((\lambda_m^2 + \lambda_m\lambda_n + \lambda_n^2) - \omega(\lambda_m + \lambda_n))\langle q\phi_m, \phi_n\rangle + \langle \beta\phi_m', \phi_n{}'\rangle = 0. \tag{2.6}$$

Proof. Note that

$$\begin{aligned}
\mathcal{L}(\lambda_m)\phi_m &= (\alpha(x)\phi_m')' + \lambda_m(\beta(x)\phi_m')' + \lambda_m^2\omega q(x)\phi_m - \lambda_m^3 q(x)\phi_m = 0,\\
\mathcal{L}(\lambda_n)\phi_n &= (\alpha(x)\phi_n')' + \lambda_n(\beta(x)\phi_n')' + \lambda_n^2\omega q(x)\phi_n - \lambda_n^3 q(x)\phi_n = 0,\\
\phi_m(0) &= \phi_n(0) = \phi_m(L) = \phi_n(L) = 0.
\end{aligned}$$

Then from the fact that

$$\phi_n \mathcal{L}(\lambda_m)\phi_m - \phi_m \mathcal{L}(\lambda_n)\phi_n = 0,$$

we can deduce

$$\begin{aligned}(\alpha(x)\phi_m')'\phi_n + \lambda_m(\beta(x)\phi_m')'\phi_n - (\alpha(x)\phi_n')'\phi_m - \lambda_n(\beta(x)\phi_n')'\phi_m \\ +(\lambda_m^2 - \lambda_n^2)\omega q(x)\phi_m\phi_n - (\lambda_m^3 - \lambda_n^3)q(x)\phi_m\phi_n = 0,\end{aligned} \tag{2.7}$$

or equivalently

$$\begin{aligned}\frac{d}{dx}(\alpha(x)(\phi_m'\phi_n - \phi_n'\phi_m)) + \frac{d}{dx}(\beta(x)(\lambda_m\phi_m'\phi_n - \lambda_n\phi_n'\phi_m)) \\ -(\lambda_m - \lambda_n)\beta(x)\phi_m'\phi_n' + (\lambda_m^2 - \lambda_n^2)\omega q(x)\phi_m\phi_n \\ -(\lambda_m^3 - \lambda_n^3)q(x)\phi_m\phi_n = 0.\end{aligned} \tag{2.8}$$

Now integrating over the interval $[0, L]$ leads to

$$\begin{aligned}\int_0^L \frac{d}{dx}(\alpha(x)(\phi_m'\phi_n - \phi_n'\phi_m))dx + \int_0^L \frac{d}{dx}(\beta(x)(\lambda_m\phi_m'\phi_n - \lambda_n\phi_n'\phi_m))dx \\ - \int_0^L (\lambda_m - \lambda_n)\beta(x)\phi_m'\phi_n'dx + \int_0^L (\lambda_m^2 - \lambda_n^2)\omega q(x)\phi_m\phi_n dx \\ - \int_0^L (\lambda_m^3 - \lambda_n^3)q(x)\phi_m\phi_n dx = 0,\end{aligned} \tag{2.9}$$

which, upon using the boundary conditions and cancelling the nonzero common factor $\lambda_m - \lambda_n$, gives

$$((\lambda_m^2 + \lambda_m\lambda_n + \lambda_n^2) - \omega(\lambda_m + \lambda_n)) \int_0^L q(x)\phi_n\phi_m dx + \int_0^L \beta(x)\phi_m'\phi_n' dx = 0. \qquad \square$$

Note that the product in (2.5) is not really an inner product. Nevertheless we can think of (2.6) as an orthogonality relation for the eigenvalue system (1.5). From the proof we can view it as a generalization of Green's identity in the Sturm–Liouville eigenvalue problem to the present cubic eigenvalue problem.

3. Analysis of the spectrum of $\mathcal{L}$. Let us write down the operator pencil associated with (1.5):

$$\mathcal{L}(x, D, \lambda) = P_0(x, D) + \lambda P_1(x, D) + \lambda^2 P_2(x, D) + \lambda^3 P_3(x, D),$$

where

$$\begin{aligned}P_0(x, D)\phi &= (\alpha(x)\phi')', \\ P_1(x, D)\phi &= (\beta(x)\phi')', \\ P_2(x, D)\phi &= \omega q(x)\phi, \\ P_3(x, D)\phi &= -q(x)\phi.\end{aligned}$$

Recall that $\alpha(x)$ and $\beta(x)$ are positive C^1 functions on the closure of $\Omega := (0, L)$. In this section we want to analyze the spectrum structure of $\mathcal{L}$ as an unbounded operator in $L^2(\Omega)$ with domain $\mathrm{Dom}(\mathcal{L}) := H^2(\Omega) \cap H_0^1(\Omega) \subset L^2(\Omega)$ and range $L^2(\Omega)$. We need a few definitions first.

DEFINITION 3.1. *We say that $\lambda \in r(\mathcal{L})$, the resolvent set of $\mathcal{L}$, if $\mathcal{L}^{-1}(\lambda)$ exists and is bounded on $L^2(\Omega)$.*

DEFINITION 3.2. *The set $\sigma(\mathcal{L}) = \mathbb{C} \setminus r(\mathcal{L})$ is called the spectrum of $\mathcal{L}$. To further classify the spectrum of $\mathcal{L}$, we define $\lambda_0 \in \sigma_{disc}(\mathcal{L})$, the discrete spectrum (eigenvalues) of $\mathcal{L}$, if λ_0 is an isolated point of $\sigma(\mathcal{L})$ and the Laurent expansion of $\mathcal{L}^{-1}$ near λ_0 can be written as*

$$\mathcal{L}^{-1}(\lambda) = \sum_{j=-k}^{\infty} A_j(\lambda - \lambda_0)^j, \tag{3.1}$$

where $0 \le k < \infty$ and the coefficients $A_{-k}, \ldots, A_{-1}$ are all finite rank operators. The essential spectrum of $\mathcal{L}$, denoted by $\sigma_{ess}(\mathcal{L})$, is defined by $\sigma_{ess}(\mathcal{L}) = \sigma(\mathcal{L}) \setminus \sigma_{disc}(\mathcal{L})$. In other words, $\mathcal{L}^{-1}$ is a meromorphic function in $\mathbb{C} \setminus \sigma_{ess}(\mathcal{L})$ and the coefficients of the negative terms in the Laurent expansion of $\mathcal{L}^{-1}$ at the point in $\sigma_{disc}(\mathcal{L})$ are finite rank operators.

Remark 3.1. If we consider the operator pencil $B = Q - \lambda$, where Q is a closed operator, then the definitions of $\sigma_{\rm disc}(B)$ and $\sigma_{\rm ess}(B)$ in Definition 3.2 agree with the usual ones defined, for example, in [14, p. 13, p. 108]. Moreover, any isolated point of $\sigma(B)$ is in the discrete spectrum. We can also see that if $\lambda_0 \in \sigma_{\rm disc}(\mathcal{L})$ then the kernel of $\mathcal{L}(\lambda_0)$ is nontrivial. In fact, with the help of Theorem 2.1, the kernel of $\mathcal{L}(\lambda_0)$ is one-dimensional whenever $\lambda_0 \in E^c$, where E is as in (1.9).

To understand the structure of $\sigma(\mathcal{L})$, let us examine $\sigma(\mathcal{A})$, where

$$\mathcal{A} = P_0 + \lambda P_1.$$

Intuitively, $\sigma(\mathcal{L})$ and $\sigma(\mathcal{A})$ have similar structures except discrete points, since $\mathcal{L}$, in some sense, can be treated as $\mathcal{A}$ plus compact perturbations. Note that $P_1^{-1} : L^2(\Omega) \to H^2(\Omega)$ is bounded. Since P_0 is self-adjoint, we have that $P_0P_1^{-1} : L^2(\Omega) \to L^2(\Omega)$ is a closed operator. Also, it is readily seen that $\sigma_{ess}(\mathcal{A}) = \sigma_{ess}(P_0P_1^{-1}+\lambda)$ and $\sigma_{disc}(\mathcal{A}) = \sigma_{disc}(P_0P_1^{-1} + \lambda)$. For consistency, here we have used the unconventional notation $\sigma_{ess}(P_0P_1^{-1}+\lambda)$ and $\sigma_{disc}(P_0P_1^{-1}+\lambda)$ to represent, respectively, the essential and discrete spectrum of $P_0P_1^{-1}$. In view of the assumptions on $\alpha(x)$ and $\beta(x)$ and the standard elliptic regularity theorem, we can see that $\sigma(\mathcal{A}) \subseteq E$, where E is as in (1.5). Clearly, E is an interval in the negative real axis. We are now ready to give a description of $\sigma(\mathcal{L})$.

THEOREM 3.3. *Let $\alpha(x), \beta(x)$ satisfy the assumptions as stated and $q(x) \in L^\infty(\Omega)$. Then the operator pencil $\mathcal{L}$ has only discrete spectrum in $\mathbb{C} \setminus \sigma(\mathcal{A})$.*

Proof. If $\lambda \notin \sigma(\mathcal{A})$, then $\mathcal{L}^{-1}$ exists if and only if $(I + (\lambda^2P_2 + \lambda^3P_3)\mathcal{A}^{-1})^{-1}$ exists. Observe that

$$(\lambda^2P_2 + \lambda^3P_3)\mathcal{A}^{-1} = (\lambda^2P_2 + \lambda^3P_3)P_1^{-1}(P_0P_1^{-1} + \lambda)^{-1}$$

is compact for all $\lambda \notin \sigma(\mathcal{A})$ and is an analytic operator-valued function of λ in $\mathbb{C}\setminus\sigma(\mathcal{A})$. In addition, we can check that $(I + (\lambda^2P_2 + \lambda^3P_3)\mathcal{A}^{-1})^{-1}$ exists at $0 \in \mathbb{C} \setminus \sigma(\mathcal{A})$. Now by the analytic Fredholm theorem [13, p. 201], we conclude that there exists a set of discrete points S in $\mathbb{C} \setminus \sigma(\mathcal{A})$ such that $(I + (\lambda^2P_2 + \lambda^3P_3)\mathcal{A}^{-1})^{-1}$ exists in $\mathbb{C} \setminus (\sigma(\mathcal{A}) \cup S)$. Moreover, $(I + (\lambda^2P_2 + \lambda^3P_3)\mathcal{A}^{-1})^{-1}$ is a meromorphic of λ in $\mathbb{C} \setminus \sigma(\mathcal{A})$ and the coefficients of the negative terms in the Laurent expansion of $(I + (\lambda^2P_2 + \lambda^3P_3)\mathcal{A}^{-1})^{-1}$ at $\lambda_0 \in S$ are finite rank operators. In other words, points in S belong to $\sigma_{disc}(\mathcal{L})$. ∎

Remark 3.2. It follows from Theorem 3.3 that $\mathcal{L}$ has only discrete spectrum in E^c.

Next, we want to discuss the essential spectrum of $\mathcal{L}$. It turns out that $\sigma_{ess}(\mathcal{L})$ and $\sigma_{ess}(\mathcal{A})$ are equal.

THEOREM 3.4.

$$\sigma_{ess}(\mathcal{L}) = \sigma_{ess}(\mathcal{A}).$$

Proof. Notice that the operator

$$\Phi(\lambda) := (\mathcal{L} - \mathcal{A})\ \mathcal{A}^{-1} = (\lambda^2 P_2 + \lambda^3 P_3) P_1^{-1} (P_0 P_1^{-1} + \lambda)^{-1}$$

is compact and analytic in $\mathbb{C} \setminus \sigma(\mathcal{A})$. Now if $\lambda \in \mathbb{C} \setminus \sigma(\mathcal{A})$, we can see that $\mathcal{L}^{-1}$ exists if and only if $(I + \Phi(\lambda))^{-1}$ exists. Obviously, $0 \in \mathbb{C} \setminus \sigma_{ess}(\mathcal{A})$ and $(I + \Phi(0))^{-1} = I$ exists. Recall that $(P_0 P_1^{-1} + \lambda)^{-1}$ is a meromorphic function in $\mathbb{C} \setminus \sigma_{ess}(\mathcal{A})$ with finite rank residues at points in $\sigma_{disc}(\mathcal{A})$. Hence, the operator $\Phi(\lambda)$ is compact for all $\lambda \in \mathbb{C} \setminus \sigma_{\rm ess}(\mathcal{A})$ and meromorphic in $\mathbb{C} \setminus \sigma_{ess}(\mathcal{A})$ with finite rank residues at points in $\sigma_{disc}(\mathcal{A})$. Now, by the meromorphic Fredholm theorem [14, p. 107], there exists a set of discrete points S' such that $(I + \Phi(\lambda))^{-1}$ is meromorphic in $\mathbb{C} \setminus \sigma_{ess}(\mathcal{A})$ with finite rank residues at points in S'. In other words, $\mathcal{L}$ has only discrete spectrum in $\mathbb{C} \setminus \sigma_{ess}(\mathcal{A})$. Therefore, we conclude that $\sigma_{ess}(\mathcal{L}) \subseteq \sigma_{ess}(\mathcal{A})$. Conversely, by exchanging the roles of $\mathcal{L}$ and $\mathcal{A}$ and going over the same argument, we can show that $\sigma_{ess}(\mathcal{A}) \subseteq \sigma_{ess}(\mathcal{L})$. ∎

4. Pole assignment. In this section we focus mainly on the method for replacing some particular poles of $\mathcal{L}$ in E^c with the prescribed poles while keeping the others in E^c unchanged. In other words, the whole concept is to determine a control force required to do such a job. We have shown that $\mathcal{L}$ has only discrete spectrum in E^c and the kernel of $\mathcal{L}$ at this discrete spectrum is one-dimensional. Therefore, we will call (λ, ϕ) an eigenpair of $\mathcal{L}$ whenever $\mathcal{L}(\lambda)\phi = 0$ for $\lambda \in E^c$. Suppose that $b(x)$ is a real control function, and $w(t)$ is a control force being applied to (1.1). Let $v(x,t)$ be the response of the controlled system. For our study, we take the control force $w(t)$ of the form

$$w(t) = \int_0^L [f_1(x)\partial_t v(x,t) + f_2(x)v(x,t) + g_1(x)\tilde{W}(\partial_x v)(x,t) + g_2(x)\tilde{W}(v)(x,t)]dx,$$

where

$$\tilde{W}(v)(x,t) = \rho \int_0^t e^{\omega(t-s)} v(x,s)ds. \tag{4.1}$$

(We use $\tilde{W}$, since the first term ρ in the function W has been absorbed into other terms.) Thus, the controlled system is expressed as

$$\begin{aligned} &\partial_x[p_1(x)\partial_x v + p_2(x)W(\partial_x v)] - q(x)\partial_t^2 v \\ &\qquad = b(x) \int_0^L [f_1(x)v + f_2(x)\partial_t v + g_1(x)\tilde{W}(\partial_x v) + g_2(x)\tilde{W}(v)]dx, \\ &v(0,t) = v(L,t) = 0. \end{aligned} \tag{4.2}$$

Simplifying (4.2) as done previously for the free system and letting $v(x,t) = \psi(x)e^{\lambda t}$ yields

$$\begin{aligned}\mathcal{L}_c(\lambda)\psi := {} & (\alpha(x)\psi')' + \lambda(\beta(x)\psi')' + \lambda^2\omega q(x)\psi - \lambda^3 q(x)\psi \\ & -b\int_0^L [\lambda(\lambda f_1 + f_2)\psi - \omega(\lambda f_1 + f_2)\psi + \rho g_1\psi' + \rho g_2\psi]dx = 0, \\ (4.3)\quad & \psi(0,t) = \psi(L,t) = 0.\end{aligned}$$

Here we restate that $\alpha(x)$ and $\beta(x)$ belong to $C^1(\bar{\Omega})$, and $q(x)$ is in $L^\infty(\Omega)$.

Before we begin, we would like to do some simple adjustments and arrangements. Let $\{\lambda_1, \lambda_2, \dots\} = \sigma_{disc}(\mathcal{L}) \setminus E$ with associated eigenvectors $\{\phi_1, \phi_2, \dots\}$ and let the first ℓ discrete spectrum $\{\lambda_1, \dots, \lambda_\ell\}$ satisfy $\{\lambda_1, \dots, \lambda_\ell\} \cap \{\lambda_{\ell+1}, \lambda_{\ell+2}, \dots\} = \emptyset$. Now we will show how to obtain the control force $w(t)$, which assigns poles $\{\lambda_1, \dots, \lambda_\ell\}$ of $\mathcal{L}$ to prescribed values (still lying in E^c) while leaving the other poles in E^c unchanged. More precisely, let $\{\mu_1, \dots, \mu_\ell\}$ be in E^c with $\{\mu_1, \dots, \mu_\ell\} \cap \{\lambda_1, \lambda_2, \dots\} = \emptyset$. Then we wish to find $f_1(x)$, $f_2(x)$, $g_1(x)$ and $g_2(x)$ in (4.3) such that $\{\mu_1, \dots, \mu_\ell, \lambda_{\ell+1}, \dots\} = \sigma_{disc}(\mathcal{L}_c) \setminus E$.

THEOREM 4.1. *Let $b(x) \in L^2(\Omega)$ satisfy*

$$(4.4)\qquad \int_0^L b\phi_j dx = \langle b, \phi_j \rangle \neq 0 \qquad \text{for } j = 1, 2, \dots, \ell.$$

Assume that $\{\lambda_1, \dots, \lambda_\ell\}$ has distinct elements and so does $\{\mu_1, \dots, \mu_\ell\}$. Let the following functions be defined:

$$\begin{aligned} f_1(x) &:= q\sum_{j=1}^{\ell} \xi_j \phi_j, \\ f_2(x) &:= q\sum_{j=1}^{\ell} \xi_j \lambda_j \phi_j, \\ g_1(x) &:= \left(\frac{\beta}{\rho}\right)\sum_{j=1}^{\ell} \xi_j \phi_j', \\ (4.5)\qquad g_2(x) &:= \left(\frac{q}{\rho}\right)\sum_{j=1}^{\ell} \xi_j \lambda_j^2 \phi_j, \end{aligned}$$

where

$$(4.6)\qquad \xi_j = \frac{\lambda_j - \mu_j}{\langle b, \phi_j \rangle} \prod_{r=1,\, r\neq j}^{\ell} \frac{\lambda_j - \mu_r}{\lambda_j - \lambda_r}, \qquad j = 1, 2, \dots, \ell.$$

Then we have that

(i) *$\sigma_{ess}(\mathcal{L}_c) = \sigma_{ess}(\mathcal{L})$ and $\mathcal{L}_c$ has only discrete spectrum in E^c,*
(ii) *$\{\mu_1, \dots, \mu_\ell, \lambda_{\ell+1}, \dots\} \subset \sigma_{disc}(\mathcal{L}_c) \setminus E$.*

Proof. (i) Since $\alpha(x), \beta(x) \in C^1(\bar{\Omega})$ and $q(x) \in L^\infty(\Omega)$, we get from the elliptic regularity theorem that $\phi_j \in H^2(\Omega) \cap H_0^1(\Omega)$. Using the integration by parts in the

integral term $\int_0^L \rho g_1 \psi' dx$ of (4.3), we obtain that

$$\begin{aligned} & b\int_0^L [\lambda(\lambda f_1 + f_2)\psi - \omega(\lambda f_1 + f_2)\psi + \rho g_1 \psi' + \rho g_2 \psi] dx \\ &= b\int_0^L \{\lambda(\lambda f_1 + f_2) - \omega(\lambda f_1 + f_2)\psi - \rho g_1' + \rho g_2\}\psi dx \\ &:= G(\lambda)\psi(x). \end{aligned}$$

It is clear that $G(\lambda)$ is an integral operator depending on λ analytically. In view of (4.5) and the fact that $b \in L^2(\Omega)$, we have that $G(\lambda)$ is a Hilbert–Schmidt operator on $L^2(\Omega)$ and hence compact. Going over the proofs of Theorems 3.3 and 3.4 once again, we immediately conclude that $\mathcal{L}_c$ has only discrete spectrum in E^c and $\sigma_{ess}(\mathcal{L}_c) = \sigma_{ess}(\mathcal{A}) = \sigma_{ess}(\mathcal{L})$.

(ii) We first show that with (4.5), for each $k \geq \ell + 1$, the eigenpair $\{\lambda_k, \phi_k\}$ of $\mathcal{L}$ satisfies $\mathcal{L}_c(\lambda_k)\phi_k = 0$. The idea is simply to show that the control term in (4.3) is zero when λ is replaced by any $\lambda_k, k \geq \ell + 1$. Clearly,

$$\begin{aligned} & (\alpha\phi_k')' + \lambda_k(\beta\phi_k')' + \lambda_k^2\omega q\phi_k - \lambda_k^3 q\phi_k \\ &\qquad -b\int_0^L \lambda_k(\lambda_k f_1 + f_2)\phi_k - \omega(\lambda_k f_1 + f_2)\phi_k + \rho g_1 \phi_k' + \rho g_2 \phi_k \, dx \\ &= -b\int_0^L \lambda_k(\lambda_k f_1 + f_2)\phi_k - \omega(\lambda_k f_1 + f_2)\phi_k + \rho g_1 \phi_k' + \rho g_2 \phi_k \, dx \\ &= -b\sum_{i=1}^{\ell} \xi_i \{((\lambda_k^2 + \lambda_k\lambda_i + \lambda_i^2) - \omega(\lambda_k + \lambda_i))\langle q\phi_k, \phi_i\rangle + \langle \beta\phi_k', \phi_i{}'\rangle\}. \end{aligned}$$

By the orthogonality relation (2.6) we then conclude that $\mathcal{L}_c(\lambda_k)\phi_k = 0$, i.e., $\lambda_k \in \sigma_{disc}(\mathcal{L}_c)$.

Now suppose that $\{\mu_k, \chi_k\}$ is the solution of the system

$$\begin{aligned} & (\alpha\chi_k')' + \mu_k(\beta\chi_k')' + \mu_k^2\omega q\chi_k - \mu_k^3 q\chi_k = b, \\ & \chi_k(0) = \chi_k(L) \end{aligned} \tag{4.7}$$

for $1 \leq k \leq \ell$. The existence of the solution is guaranteed by the fact that $\mu_k \in r(\mathcal{L})$ for all k and $b \in L^2(\Omega)$. We will now show that $\mathcal{L}_c(\mu_k)\chi_k = 0$ for all $1 \leq k \leq \ell$. First we consider the following:

$$\begin{cases} (\alpha\chi_k')' + \mu_k(\beta\chi_k')' + \mu_k^2\omega q\chi_k - \mu_k^3 q\chi_k = b, \\ \chi_k(0) = \chi_k(L) = 0, \qquad 1 \leq k \leq \ell, \end{cases} \tag{4.8}$$

and

$$\begin{cases} (\alpha\phi_j')' + \lambda_j(\beta\phi_j')' + \lambda_j^2\omega q\phi_j - \lambda_j^3 q\phi_j = 0, \\ \phi_j(0) = \phi_j(L) = 0, \qquad 1 \leq j \leq \ell. \end{cases} \tag{4.9}$$

Multiplying (4.8) by ϕ_j and subtracting from it the quantity (4.9) times χ_k, we get, on the one hand,

$$\begin{aligned} & (\lambda_j - \mu_k)\langle \beta\chi_k', \phi_j{}'\rangle + (\mu_k^2 - \lambda_j^2)\omega\langle q\chi_k, \phi_j\rangle \\ & \qquad -(\mu_k^3 - \lambda_j^3)\langle q\chi_k, \phi_j\rangle = \langle b, \phi_j\rangle. \end{aligned} \tag{4.10}$$

On the other hand, let $\mu \in \mathbb{C}$ and $\psi : [0, L] \to \mathbb{C}$, and define

$$\begin{aligned} F(\mu, \psi) &:= \int_0^L \mu(\mu f_1 + f_2)\psi - \omega(\mu f_1 + f_2)\psi + \rho g_1 \psi' + \rho g_2 \psi dx \\ &= \sum_{j=1}^{\ell} \xi_j [\langle \beta\psi', \phi_j{}' \rangle - (\mu + \lambda_j)\omega\langle q\psi, \phi_j\rangle \\ &\qquad +(\lambda_j^2 + \mu\lambda_j + \mu^2)\langle q\psi, \phi_j\rangle]. \end{aligned} \tag{4.11}$$

Here in the notation we suppress the dependence of F on ϕ_j's.

By writing

$$\begin{aligned} \xi_j &= \left(\frac{\lambda_j - \mu_j}{\langle b, \phi_j\rangle} \prod_{r=1,\, r\neq j}^{\ell} \frac{\lambda_j - \mu_r}{\lambda_j - \lambda_r} \right) \\ &= (\lambda_j - \mu_k) \left(\frac{\prod_{r=1, r\neq k}^{\ell} (\lambda_j - \mu_r)}{\prod_{r=1, r\neq j}^{\ell} (\lambda_j - \lambda_r)} \right) \left(\frac{1}{\langle b, \phi_j\rangle} \right), \end{aligned} \tag{4.12}$$

we obtain

$$\begin{aligned} F(\mu_k, \psi) = \sum_{j=1}^{\ell} &\left(\frac{\prod_{r=1, r\neq k}^{\ell} (\lambda_j - \mu_r)}{\prod_{r=1, r\neq j}^{\ell} (\lambda_j - \lambda_r)} \right) \left(\frac{1}{\langle b, \phi_j\rangle} \right) \\ &\times [(\lambda_j - \mu_k)\langle \beta\psi', \phi_j{}'\rangle + (\mu_k^2 - \lambda_j^2)\omega\langle q\psi, \phi_j\rangle \\ &-(\lambda_j^3 - \mu_k^3)\langle q\psi, \phi_j\rangle]. \end{aligned} \tag{4.13}$$

Let us check the validity of the equation

$$(\alpha\chi_k')' + \mu_k(\beta\chi_k')' + \mu_k^2\omega q\chi_k - \mu_k^3 q\chi_k - bF(\mu_k, \chi_k) = 0. \tag{4.14}$$

Using (4.8), (4.10), and (4.12), we see that the left-hand side of (4.14) is

$$b - b\left\{ \sum_{j=1}^{\ell} \frac{\prod_{r=1, r\neq k}^{\ell} (\lambda_j - \mu_r)}{\prod_{r=1, r\neq j}^{\ell} (\lambda_j - \lambda_r)} \right\} = 0, \quad 1 \le k \le \ell,$$

where we used an identity from [3]: for any $1 \le k \le \ell$,

$$\sum_{j=1}^{\ell} \frac{\prod_{r=1, r\neq k}^{\ell} (\lambda_j - \mu_r)}{\prod_{r=1, r\neq j}^{\ell} (\lambda_j - \lambda_r)} = 1.$$

This completes the proof of (ii). □

Next we want to show that the discrete spectrum of $\mathcal{L}_c$ in E^c is precisely given by $\{\mu_1, \dots, \mu_\ell, \lambda_{\ell+1}, \dots\}$. That is, no new extra discrete spectrum occurs in E^c except those prescribed.

THEOREM 4.2. *Let f_1, f_2, g_1, and g_2 be defined as in* (4.5) *and* (4.6), *and the assumptions of Theorem* 4.1 *hold. Moreover, assume that (λ_k, ϕ_k) satisfies*

$$(3\lambda_k^2 - 2\omega\lambda_k)\langle q\phi_k, \phi_k\rangle + \langle \beta\phi_k', \phi_k'\rangle \neq 0 \tag{4.15}$$

for $1 \le k \le \ell$. *Then the discrete spectrum of* $\mathcal{L}_c$ *in* E^c *is precisely given by* $\{\mu_1, \dots, \mu_\ell, \lambda_{\ell+1}, \dots\}$, *i.e.,*

$$\{\mu_1, \dots, \mu_\ell, \lambda_{\ell+1}, \dots\} = \sigma_{disc}(\mathcal{L}_c) \setminus E.$$

Proof. The previous theorem shows that

$$\{\mu_1, \dots, \mu_\ell, \lambda_{\ell+1}, \dots\} \subset \sigma_{disc}(\mathcal{L}_c) \setminus E.$$

Here, we only need to prove the opposite inclusion.

We first claim that $\{\lambda_1, \dots, \lambda_\ell\} \notin \sigma_{disc}(\mathcal{L}_c) \setminus E$. Equivalently, those poled will not show up again. To begin, we show that $\mathcal{L}_c(\lambda_k)\phi_k \neq 0$ for $1 \le k \le \ell$. In fact, since $\{\lambda_k, \phi_k\}$ is an eigenpair of $\mathcal{L}$, we have that

$$\begin{aligned}\mathcal{L}_c(\lambda_k)\phi_k &= \mathcal{L}(\lambda_k)\phi_k - b(x)F(\lambda_k, \phi_k)\\ &= -b(x)F(\lambda_k, \phi_k).\end{aligned}$$

By virtue of (4.11) and the orthogonality relation, we can see that

$$F(\lambda_k, \phi_k) = \xi_k\{(3\lambda_k^2 - 2\omega\lambda_k)\langle q\phi_k, \phi_k\rangle + \langle \beta\phi_k', \phi_k'\rangle\}.$$

It is clear that $\xi_k \neq 0$ and, therefore, $F(\lambda_k, \phi_k) \neq 0$ by the condition (4.15).

Next, we claim $\mathcal{L}_c(\lambda_k)\psi \neq 0$ for $1 \le k \le \ell$ if ψ and ϕ_k are linearly independent. In fact, since the dimension of the eigenspace of $\mathcal{L}$ is one, we have $\mathcal{L}(\lambda_k)\psi \neq 0$. Suppose, on the contrary, $0 = \mathcal{L}_c(\lambda_k)\psi = \mathcal{L}(\lambda_k)\psi - b(x)F(\lambda_k, \psi)$. On the one hand,

$$\begin{aligned}0 &= \langle \psi, \mathcal{L}(\lambda_k)\phi_k\rangle = \langle \mathcal{L}(\lambda_k)\psi, \phi_k\rangle\\ &= \langle bF(\lambda_k, \psi), \phi\rangle = F(\lambda_k, \psi)\langle b, \phi_k\rangle.\end{aligned}$$

But in view of (4.4), this leads to $F(\lambda_k, \psi) = 0$, and hence $0 = \mathcal{L}(\lambda_k)\psi$, a contradiction.

Finally we show that if $\mu \notin \{\lambda_1, \lambda_2, \dots\} \cup \{\mu_1, \dots, \mu_k\}$ and $\mu \notin E$, then $\mu \notin \sigma_{disc}(\mathcal{L}_c) \setminus E$. We will use a contradiction. Suppose that $\{\mu, \phi\}$ is an eigenpair of $\mathcal{L}_c$. Thus

$$\begin{aligned}0 &= \mathcal{L}_c(\mu)\phi\\ &= \mathcal{L}(\mu)\phi - b(x)C,\end{aligned}$$

where the constant $C = F(\mu, \phi)$. Note that $C \neq 0$; otherwise $\mu \in \sigma_{disc}(\mathcal{L}) \setminus E$. By normalization, we can find $\{\mu, \psi\}$, where $\psi = \frac{1}{C}\phi$, so that

$$\mathcal{L}(\mu)\psi = b.$$

Now we can compute $F(\mu, \psi)$ as we did for $F(\mu_k, \psi)$ going from (4.8) to (4.13), and hence

$$\begin{aligned}0 &= \mathcal{L}_c(\mu)\psi\\ &= \mathcal{L}(\mu)\psi - b(x)F(\mu, \psi)\\ &= b - bF(\mu, \psi)\\ &= b\left[1 - \sum_{j=1}^{\ell} \frac{1}{\lambda_j - \mu} \frac{\Pi_{r=1}^{\ell}(\lambda_j - \mu_r)}{\Pi_{r=1, r\neq j}^{\ell}(\lambda_j - \lambda_r)}\right]\\ &\neq 0,\end{aligned}$$

which is a contradiction since the bracketed term can be transformed into a polynomial of degree ℓ in μ and $\mu_1, \ldots, \mu_\ell$ have been used up as the ℓ roots. This completes the proof. ∎

In view of the above, conditions (4.4) and (4.15) can be legitimately called the controllability conditions. We now show that if either one of (4.4) and (4.15) is violated, then the mode λ_j cannot be relocated by our designed control force, where $1 \le j \le \ell$.

THEOREM 4.3. *For any $1 \le j \le \ell$, let (λ_j, ϕ_j) satisfy either $\langle b, \phi_j \rangle = 0$ or $(3\lambda_j^2 - 2\omega\lambda_j)\langle q\phi_j, \phi_j \rangle + \langle \beta\phi_j', \phi_j' \rangle = 0$. Then $\lambda_j \in \sigma_{disc}(\mathcal{L}_c)$, with f_1, f_2, g_1, and g_2 of $\mathcal{L}_c$ being given in* (4.5) *and* (4.6).

Proof. From the proof of Theorem 4.2, we immediately see that $\mathcal{L}_c(\lambda_j)\phi_j = 0$ if $(3\lambda_j^2 - 2\omega\lambda_j)\langle q\phi_j, \phi_j \rangle + \langle \beta\phi_j', \phi_j' \rangle = 0$. Now we assume that $\langle b, \phi_j \rangle = 0$. In view of the form of $\mathcal{L}_c$, we have $\langle \mathcal{L}_c(\lambda_j)\psi, \phi_j \rangle = 0$ for any $\psi \in H^2(\Omega) \cap H_0^1(\Omega)$, which implies that $\mathcal{L}_c(\lambda_j)$ is not invertible. For, if $\mathcal{L}_c(\lambda_j)^{-1}$ exists, then we can find a $\psi_j \in H^2(\Omega) \cap H_0^1(\Omega)$ such that $\mathcal{L}_c(\lambda_j)\psi_j = \bar{\phi}_j$ and therefore $\langle \mathcal{L}_c(\lambda_j)\psi, \phi_j \rangle \neq 0$. Since $\mathcal{L}_c$ has only discrete spectrum in E^c, λ_j must be in $\sigma_{disc}(\mathcal{L}_c)$. ∎

Finally, in order for the control to be realizable, we need to show that f_1, f_2, g_1, and g_2 are real functions.

THEOREM 4.4. *Assume that both sets $\{\lambda_1, \ldots, \lambda_\ell\}$ and $\{\mu_1, \ldots, \mu_\ell\}$ are closed under complex conjugation. Then the functions f_1, f_2, g_1, and g_2 are real functions.*

Proof. Rewriting ξ_j yields

$$\begin{aligned}\xi_j &= \frac{\lambda_j - \mu_j}{\langle b, \phi_j \rangle} \prod_{r=1,\, r\neq j}^{\ell} \frac{\lambda_j - \mu_r}{\lambda_j - \lambda_r} \\ &= \frac{\prod_{r=1}^{\ell}(\lambda_j - \mu_r)}{\langle b, \phi_j \rangle \prod_{r=1,\, r\neq j}^{\ell}(\lambda_j - \lambda_r)}.\end{aligned}$$

Therefore, if λ_j is real and its associated eigenfunction ϕ_j is also real, then $\bar{\xi}_j = \xi_j$, i.e., ξ_j is real. Now assume that λ_j and λ_{j+1} are a conjugate pair and their eigenfunctions satisfy $\phi_j = \bar{\phi}_j$. Then we can see that

$$\begin{aligned}\xi_{j+1} &= \frac{\prod_{r=1}^{\ell}(\lambda_{j+1} - \mu_r)}{\langle b, \phi_{j+1} \rangle \prod_{r=1,\, r\neq j+1}^{\ell}(\lambda_{j+1} - \lambda_r)} \\ &= \frac{\prod_{r=1}^{\ell}(\bar{\lambda}_j - \mu_r)}{\langle b, \bar{\phi}_j \rangle \prod_{r=1,\, r\neq j+1}^{\ell}(\bar{\lambda}_j - \lambda_r)} \\ &= \frac{\prod_{r=1}^{\ell}(\bar{\lambda}_j - \mu_r)}{\langle b, \bar{\phi}_j \rangle \prod_{r=1,\, r\neq j}^{\ell}(\bar{\lambda}_j - \bar{\lambda}_r)} \\ &= \bar{\xi}_j.\end{aligned}$$

Thus, for f_1, we have that

$$\bar{f} = q \sum_{j=1}^{\ell} \bar{\xi}_j \bar{\phi}_j = q \sum_{j=1}^{\ell} \xi_j \phi_j = f_1,$$

i.e., f_1 is real. Similarly, we can prove that f_2, g_1, and g_2 are real. ∎

REFERENCES

[1] R. Miller, R. Kroll, and R. Clemmons, *Dynamic Loads Analysis System (DYLOFLEX) Summary*, Tech. Rep. NASA Contractor Report 2846-1, NASA, 1979.

[2] B. N. Datta, *Numerical Methods for Linear Control Systems Design and Analysis*, Academic Press, New York, to appear.

[3] B. N. Datta, S. Elhay, and Y. M. Ram, *Orthogonality and partial pole assignment of the symmetric definite quadratic pencil*, Linear Algebra Appl., 257 (1997), pp. 29–48.

[4] B. N. Datta, S. Elhay, Y. M. Ram, and D. R. Sarkissian, *Partial eigenstructure assignment for the quadratic pencil*, J. Sound Vibration, 230 (2000), pp. 101–110.

[5] B. N. Datta, Y. M. Ram, and R. D. Sarkissian, *Spectrum modification for gyroscopic systems*, ZAMM Z. Angew. Math. Mech., 82 (2002), pp. 191–200.

[6] B. N. Datta and R. D. Sarkissian, *Feedback control in distributed parameter gyroscopic systems: A solution of the partial eigenvalue assignment problem*, Mechanical Systems and Signal Processing, special issue on vibration control, 16 (2001), pp. 3–17.

[7] B. N. Datta and R. D. Sarkissian, *A computational method for feedback control in distributed parameter systems*, in Proceedings of the 8th IEEE International Conference on Methods and Models in Robotics and Automation, 2002, pp. 139–144.

[8] W. Desch and R. K. Miller, *Exponential stabilization of Volterra integro-differential equations in Hilbert space*, J. Differential Equations, 70 (1987), pp. 366–389.

[9] W. Desch and R. K. Miller, *Exponential stabilization of Volterra integral equations with singular kernels*, J. Integral Equations Appl., 1 (1988), pp. 397–433.

[10] D. J. Inman, *Vibration: With Control, Measurement, and Stability*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[11] W. W. Lin and J. N. Wang, *Robust partial pole assignment for the vibrating system with aerodynamic effect*, Numer. Linear Algebra Appl., to appear.

[12] Y. M. Ram, *Pole assignment for the vibrating rod*, Quart. J. Mech. Appl. Math., 51 (1998), pp. 461–476.

[13] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* I. *Functional Analysis*, Academic Press, New York, 1973.

[14] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* IV. *Analysis of Operators*, Academic Press, New York, 1978.

[15] D. L. Russell, *Canonical forms and spectral determination for a class of hyperbolic distributed parameter control systems*, J. Math. Anal. Appl., 62 (1978), pp. 186–225.

[16] W. Wonham, *Murray Linear Multivariable Control: A Geometric Approach*, 3rd ed., Appl. Math. 10, Springer-Verlag, New York, 1985.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 2130–2144

WIENER–HOPF FACTORIZATION INDICES AND INFINITE STRUCTURE OF RATIONAL MATRICES*

A. AMPARAN†, S. MARCAIDA†, AND I. ZABALLA†

Abstract. The relationship between the infinite structure and the Wiener–Hopf factorization indices of polynomial matrix representations of controllable systems is investigated. On the way, local Wiener–Hopf factorization indices must be defined. In addition, the finite and infinite structure of a rational matrix is characterized given its Wiener–Hopf factorization indices.

Key words. matrix polynomial, rational matrix, Wiener–Hopf factorization indices, finite invariant factors, infinite structure, majorization

AMS subject classifications. 15A23, 93B55, 15A45

DOI. 10.1137/S0363012902402591

1. Introduction.

Given a linear time invariant system of the form

$$\dot{x}(t) = Ax(t) + Bu(t), \quad A,\ n \times n, \quad B,\ n \times m \tag{1.1}$$

we can associate with it a nonsingular polynomial matrix as follows: Define a nonsingular $m \times m$ polynomial matrix $D(s)$ to be a polynomial matrix representation of system (1.1) if $D(s)$ is the denominator of any right coprime factorization of the transfer function $(sI_n - A)^{-1}B$; i.e., $(sI_n - A)^{-1}B = N(s)D(s)^{-1}$ for some $n \times m$ polynomial matrix such that $D(s), N(s)$ are relatively right coprime. Then it was proven in [16] that given two matrix pairs (A_1, B_1) and (A_2, B_2) with polynomial matrix representations $D_1(s)$ and $D_2(s)$, respectively, there is an invertible matrix T such that $(T^{-1}A_1T, T^{-1}B_1) = (A_2, B_2)$ if and only if there is a unimodular matrix $U(s)$ such that $D_1(s)U(s) = D_2(s)$ (notice that no mention is made of the corresponding numerators). In other words, there is a one-to-one correspondence between the set of orbits of matrix pairs under similarity and the set of orbits of nonsingular polynomial matrices under right equivalence. This fact has some important consequences. Let us mention just three of them:

(i) Since the Hermite normal form of any nonsingular polynomial matrix is a canonical representative of each orbit for the right equivalence, we can associate each class of system similar matrix pairs to an upper right proper polynomial matrix with the identity as leading coefficient: the Hermite normal form of any polynomial matrix representation of the system.

(ii) As any nonsingular polynomial matrix is right equivalent to a column proper matrix (see [15]), and the column degrees of this matrix are uniquely determined, these degrees are invariant under system similarity (these degrees are the controllability indices of the system). This is a well-known fact that can be derived, of course, from other approaches (see, for example, [9, 11]).

(iii) All polynomial matrix representations of system similar matrix pairs have the same invariant factors, and these are (apart from invariant factors equal to 1)

*Received by the editors February 12, 2002; accepted for publication (in revised form) June 20, 2003; published electronically February 18, 2004. Partially supported by the Dirección General de Investigación Científica y Técnica, Proyecto de Investigación PB97-0599-CO3-01.

http://www.siam.org/journals/sicon/42-6/40259.html

†Departamento de Matemática Aplicada y EIO, Universidad del País Vasco, Apdo. Correos 644, Bilbao 48080, Spain (mepamlaa@lg.ehu.es, mepmabes@lg.ehu.es, izaballa@picasso.lc.ehu.es).

the invariant factors of the state matrix in any pair in the similarity class (see, for example, [9]).

Feedback equivalence of matrix pairs is also related to an equivalence relationship between the corresponding polynomial matrix representations. In order to expose it, we need to introduce some notation and terminology. First, since we are going to deal only with algebraic properties of systems, we will assume that the underlying field, $\mathbb{F}$, is arbitrary. Let $\mathbb{F}[s]$ and $\mathbb{F}(s)$ be the ring of polynomials and the field of rational functions with coefficients in $\mathbb{F}$. Let $\mathbb{F}_{pr}(s)$ denote the subset of rational functions of $\mathbb{F}(s)$ whose denominators have a degree higher than or equal to that of the numerators. This is a Euclidean ring called the ring of proper rational functions. As usual, a polynomial matrix $U(s) \in \mathbb{F}[s]^{m\times m}$ is said to be unimodular if it is invertible in $\mathbb{F}[s]^{m\times m}$; i.e., its determinant is a constant nonzero polynomial. The unities in $\mathbb{F}_{pr}(s)^{m\times m}$ are called biproper or bicausal matrices; their determinants are biproper rational functions, i.e., rational functions whose denominators and numerators have equal degrees.

Two rational matrices $T_1(s), T_2(s) \in \mathbb{F}(s)^{m\times n}$ are said to be (left) Wiener–Hopf equivalent if there exist matrices $U(s) \in \mathbb{F}[s]^{n\times n}$, unimodular, and $B(s) \in \mathbb{F}_{pr}(s)^{m\times m}$, biproper, such that $T_2(s) = B(s)T_1(s)U(s)$. The right Wiener–Hopf equivalence is defined similarly by exchanging the roles of $B(s)$ and $U(s)$. It is well known (see, for example, [1, 4, 5]) that any $m \times n$ rational matrix $T(s)$ is (left) Wiener–Hopf equivalent to a diagonal matrix

$$\Delta = \begin{bmatrix} \operatorname{diag}(s^{k_1}, s^{k_2}, \ldots, s^{k_r}) & 0 \\ 0 & 0 \end{bmatrix},$$

where the integers $k_1 \geq \cdots \geq k_r$ are uniquely determined by $T(s)$ and are called the (left) Wiener–Hopf factorization indices of $T(s)$. In other words $T_1(s)$ and $T_2(s)$ are Wiener–Hopf equivalent if and only if they have the same Wiener–Hopf factorization indices. In particular, for a nonsingular polynomial matrix $P(s) \in \mathbb{F}[s]^{m\times m}$ its (left) Wiener–Hopf factorization indices are nonnegative integers. These indices are the column degrees of any column proper matrix right equivalent to $P(s)$ (see [4]).

As shown in [4, 16], given two controllable systems (A_1, B_1) and (A_2, B_2), if $D_1(s)$ and $D_2(s)$ are their respective polynomial matrix representations, then (A_1, B_1) and (A_2, B_2) are feedback equivalent if and only if $D_1(s)$ and $D_2(s)$ are left Wiener–Hopf equivalent. In other words, there are invertible matrices T and Q and a matrix R such that $(A_2, B_2) = (T^{-1}(A_1 + B_1R)T, T^{-1}B_1Q)$ if and only if there are matrices $U(s)$, unimodular, and $B(s)$, biproper, such that $D_2(s) = B(s)D_1(s)U(s)$. This is actually a result that follows from the fact that a proper precompensator, $C(s)$, can be implemented as a state feedback on a system with transfer function $T(s) = N(s)D(s)^{-1}$ if and only if $C(s)$ is biproper and $C(s)^{-1}D(s)$ is polynomial (see [7]).

As a conclusion we have that the controllability indices of (A, B) are the Wiener–Hopf factorization indices of any of its polynomial matrix representations.

We recall now that given a controllable system, (A, B), the famous Rosenbrock's theorem on pole placement [12] calls for the existence of a feedback gain F such that the state matrix of the closed-loop system, $A+BF$, lies in a prescribed similarity class. We can easily translate this theorem into a result about polynomial matrices. In fact, let (A, B) be a controllable pair and assume that there is a matrix F such that $A+BF$ lies in a prescribed similarity class. Then any polynomial matrix representation of $(A + BF, B)$ has Wiener–Hopf factorization indices of (A, B), and invariant factors of $A + BF$ (apart from invariant factors equal to 1). This implies the existence of a

polynomial matrix with prescribed Wiener–Hopf factorization indices and invariant factors. Conversely, if there is a nonsingular polynomial matrix, $D(s) \in \mathbb{F}[s]^{m\times m}$, with $k_1 \geq \cdots \geq k_m$ as Wiener–Hopf indices and $\alpha_1(s) \mid \cdots \mid \alpha_m(s)$ as invariant factors and $(A, B) \in \mathbb{F}^{n\times n} \times \mathbb{F}^{n\times m}$ is a controllable pair with $k_1 \geq \cdots \geq k_m$ as controllability indices, then there is a matrix $F \in \mathbb{F}^{m\times n}$ such that $A + BF$ has $\alpha_1(s), \ldots, \alpha_m(s)$ and $n - m$ polynomials equal to 1 as invariant factors. For if $\tilde{D}(s)$ is a polynomial matrix representation of (A, B) and (A_1, B_1) is a controllable pair such that $D(s)$ is a polynomial matrix representation of this pair (such a pair always exists), then $D(s)$ and $\tilde{D}(s)$ are Wiener–Hopf equivalent and so (A, B) and (A_1, B_1) are feedback equivalent. Since A_1 and $D(s)$ have the same nontrivial invariant factors, there is a matrix F such that $A + BF = TA_1T^{-1}$ has the desired invariant factors.

Thus Rosenbrock's theorem can be stated as follows.

THEOREM 1.1. *Let $k_1 \geq k_2 \geq \cdots \geq k_m$ and $\alpha_1(s) \mid \alpha_2(s) \mid \cdots \mid \alpha_m(s)$ be nonnegative integers and monic polynomials, respectively. Then there exists a nonsingular matrix $D(s) \in \mathbb{F}[s]^{m\times m}$ with $\alpha_1(s), \alpha_2(s), \ldots, \alpha_m(s)$ as invariant factors and $k_1, k_2, \ldots, k_m$ as Wiener–Hopf factorization indices if and only if the following relations hold:*

$$\sum_{i=1}^{j} k_i \leq \sum_{i=1}^{j} d(\alpha_{m-i+1}(s)), \quad j = 1, 2, \ldots, m-1, \tag{1.2}$$

$$\sum_{i=1}^{m} k_i = \sum_{i=1}^{m} d(\alpha_i(s)), \tag{1.3}$$

where $d(\cdot)$ stands for "degree of."

Rosenbrock's result relates the invariant factors (Smith structure) of a nonsingular polynomial matrix and its Wiener–Hopf factorization indices. In this paper we generalize Rosenbrock's result by studying the relationship between the finite and infinite Smith–McMillan structure of a nonsingular rational matrix and its Wiener–Hopf factorization indices (see [2] for a control interpretation of the infinite structure of a rational matrix). A partial result on the problem of the relationship between the Wiener–Hopf factorization indices and the Smith–McMillan structure at infinity of nonsingular polynomial matrices can be found in [5].

It should be noted that these problems admit different but equivalent formulations. In fact, we aim to completely characterize, for a given nonsingular rational matrix $T(s) \in \mathbb{F}(s)^{m\times m}$, the possible finite Smith–McMillan structure of matrices in the set

$$\{B(s)T(s) : B(s) \text{ biproper}\}$$

and the infinite Smith–McMillan structure of

$$\{T(s)U(s) : U(s) \text{ unimodular}\}.$$

The infinite structure of rational matrices has shown to be of interest in control theory (see, for example, [14]). A test for the solution of the well-known model matching problem is based on the equality of the infinite structure of several matrices; this structure appears significantly in the solution of the decoupling problem (see, for example, [8, 10]).

First, we recall the finite and infinite Smith–McMillan forms of a rational matrix: Two rational matrices $T_1(s), T_2(s) \in \mathbb{F}(s)^{p\times m}$ are said to be equivalent if there are unimodular matrices $U(s) \in \mathbb{F}[s]^{p\times p}$ and $V(s) \in \mathbb{F}[s]^{m\times m}$ such that $T_1(s) = U(s)T_2(s)V(s)$. And they are said to be equivalent at infinity if there exist biproper rational matrices $B_1(s) \in \mathbb{F}_{pr}(s)^{p\times p}$ and $B_2(s) \in \mathbb{F}_{pr}(s)^{m\times m}$ such that $T_1(s) = B_1(s)T_2(s)B_2(s)$.

Any rational matrix $T(s) \in \mathbb{F}(s)^{p\times m}$ is equivalent to a diagonal matrix

$$S(s) = \begin{bmatrix} \operatorname{diag}\left(\frac{\epsilon_1(s)}{\psi_1(s)}, \ldots, \frac{\epsilon_r(s)}{\psi_r(s)}\right) & 0 \\ 0 & 0 \end{bmatrix},$$

where $r = \operatorname{rank} T(s)$, $\epsilon_i(s), \psi_i(s) \in \mathbb{F}[s]$ are monic and coprime such that $\epsilon_1(s) \mid \epsilon_2(s) \mid \cdots \mid \epsilon_r(s)$ and $\psi_r(s) \mid \psi_{r-1}(s) \mid \cdots \mid \psi_1(s)$. The rational functions, $\frac{\epsilon_i(s)}{\psi_i(s)}$, $1 \le i \le r$, are said to form the finite structure of $T(s)$. The zeros of $T(s)$ are the roots, with their respective multiplicites, of $\epsilon_i(s)$, $1 \le i \le r$, and the poles of $T(s)$ are the roots of $\psi_i(s)$, $1 \le i \le r$, with their respective multiplicites. The matrix $S(s)$ is the Smith–McMillan form of $T(s)$ (see, for example, [12, 13]). The rational functions $\frac{\epsilon_1(s)}{\psi_1(s)}, \ldots, \frac{\epsilon_r(s)}{\psi_r(s)}$ will be called invariant rational functions of $T(s)$. (Recall that they are uniquely determined by $T(s)$.)

Similarly, any rational matrix is equivalent at infinity to a diagonal matrix [2, 13]

$$D(s) = \begin{bmatrix} \operatorname{diag}\left(s^{q_1}, s^{q_2}, \ldots, s^{q_r}\right) & 0 \\ 0 & 0 \end{bmatrix},$$

where $r = \operatorname{rank} T(s)$ and $q_1 \ge q_2 \ge \cdots \ge q_r$, $q_i \in \mathbb{Z}$. The rational functions $s^{q_1}, s^{q_2}, \ldots, s^{q_r}$ are called the invariant factors at infinity of $T(s)$. In particular, a nonsingular polynomial matrix $P(s) \in \mathbb{F}[s]^{m\times m}$ is equivalent at infinity to a diagonal matrix $D(s) = \operatorname{diag}(s^{q_1}, s^{q_2}, \ldots, s^{q_m})$, where $q_1 \ge q_2 \ge \cdots \ge q_m$ are (possibly negative) integers.

The paper is organized as follows: In section 2 we extend Rosenbrock's result to rational matrices with prescribed finite structure. In order to deal with the problem related to the infinite structure of rational matrices we work first with matrix polynomials. We will need the concept of *local Wiener–Hopf factorization indices*. This is not a new concept (see [5]) but since we want to work over an arbitrary field the definition in [5] cannot be applied. So we will adopt a different and more general approach. This is done in section 3. In section 4 we give some auxiliary results that will provide a necessary and sufficient condition for the existence of a nonsingular rational matrix with both Wiener–Hopf and infinite invariant factors prescribed. In the last section we deal with the problem of the existence of a rational matrix with the three types of invariants prescribed: Wiener–Hopf factorization indices and finite and infinite structures.

Remark. We finish this section with two remarks:

1. If $P(s) \in \mathbb{F}[s]^{m\times m}$, we can see from [13] that $\sum_{i=1}^{j} q_i$ is the greatest degree among the degrees of all minors of $P(s)$ of order j, $j \in \{1, 2, \ldots, m\}$. In particular, $q_1 = d(P(s))$; i.e., q_1 is the degree of the element of $P(s)$ of highest degree.

2. Conditions (1.2) and (1.3) will play an important role along the way. They are usually summarized as

$$(k_1, \ldots, k_m) \prec (d(\alpha_m(s)), \ldots d(\alpha_1(s)))$$

by using the majorization symbol, $\prec$, in the sense of Hardy, Littlewood, and Polya (see [6]).

2. Wiener–Hopf factorization indices and finite structure of rational matrices. Many problems about the structure of rational matrices can be deduced from the corresponding problems about polynomial matrices. This is the case when extending the Rosenbrock's polynomial version given in Theorem 1.1 to rational matrices. The idea is as follows: Let $T(s) \in \mathbb{F}(s)^{m\times m}$ be a rational matrix and let $d(s)$ be the monic least common denominator of all the elements $t_{ij}(s) = \frac{n_{ij}(s)}{d_{ij}(s)}$ of $T(s)$. Then $T(s)$ can be written as

$$T(s) = \frac{1}{d(s)} N(s), \tag{2.1}$$

where $N(s) \in \mathbb{F}[s]^{m\times m}$.

The following three lemmas examine the relationship between the finite and infinite Smith–McMillan form of a rational matrix $T(s)$ and the finite and infinite Smith–McMillan form of the corresponding polynomial matrix $N(s)$, as well as the relationship between the Wiener–Hopf factorization indices of $T(s)$ and $N(s)$. The proofs of the three lemmas are straightforward.

LEMMA 2.1. *Let $T(s) \in \mathbb{F}(s)^{m\times m}$ be a nonsingular rational matrix, $d(s)$ the least common multiple of its denominators, and $d(s)T(s) = N(s) \in \mathbb{F}[s]^{m\times m}$. Let $\frac{\epsilon_1(s)}{\psi_1(s)}, \ldots, \frac{\epsilon_m(s)}{\psi_m(s)}$ be irreducible rational functions, where $\epsilon_i(s), \psi_i(s) \in \mathbb{F}[s]$ are monic and coprime such that $\epsilon_i(s) \mid \epsilon_{i+1}(s)$, $i = 1, 2, \ldots, m-1$, while $\psi_{i+1}(s) \mid \psi_i(s)$, $i = 1, 2, \ldots, m-1$. Let $\sigma_i(s) = \frac{d(s)}{\psi_i(s)}$, $i = 1, \ldots, m$. Then $\frac{\epsilon_1(s)}{\psi_1(s)}, \ldots, \frac{\epsilon_m(s)}{\psi_m(s)}$ are the invariant functions of $T(s)$ if and only if $\epsilon_1(s)\sigma_1(s), \ldots, \epsilon_m(s)\sigma_m(s)$ are the invariant factors of $N(s)$.*

Notice that since $\psi_{i+1}(s) \mid \psi_i(s)$, we have that $\sigma_i(s) \mid \sigma_{i+1}(s)$, $i = 1, \ldots, m$; and then $\epsilon_1(s)\sigma_1(s) \mid \cdots \mid \epsilon_m(s)\sigma_m(s)$.

LEMMA 2.2. *With $T(s)$, $d(s)$, and $N(s)$ as in Lemma 2.1, let $d = d(d(s))$ and $q_1 \geq q_2 \geq \cdots \geq q_m$ integers. Then $s^{q_1}, s^{q_2}, \ldots, s^{q_m}$ are the invariant factors at infinity of $T(s)$ if and only if $s^{q_1+d}, s^{q_2+d}, \ldots, s^{q_m+d}$ are the invariant factors at infinity of $N(s)$.*

LEMMA 2.3. *With $T(s)$, $d(s)$, $N(s)$, and d as in Lemma 2.2, let $k_1 \geq k_2 \geq \cdots \geq k_m$ be integers. Then $k_1, k_2, \ldots, k_m$ are the Wiener–Hopf factorization indices of $T(s)$ if and only if $k_1 + d, k_2 + d, \ldots, k_m + d$ are the Wiener–Hopf factorization indices of $N(s)$.*

Remark. As we already said, the Wiener–Hopf factorization indices of polynomial matrices are nonnegative integers. In the case of rational matrices, these are not always positive. In particular, the following is true for nonsingular proper rational matrices:

1. If $T(s) \in \mathbb{F}(s)^{m\times m}_{pr}$, then $k_i \leq 0$, $i = 1, 2, \ldots, m$. In particular, if $T(s)$ is strictly proper, then $k_i < 0$, $i = 1, 2, \ldots, m$.

2. If $T(s) \in \mathbb{F}(s)^{m\times m}_{pr}$, then $q_i \leq 0$, $i = 1, 2, \ldots, m$ (see [13]). In particular, if $T(s)$ is strictly proper, then $q_i < 0$, $i = 1, 2, \ldots, m$.

With the help of Lemmas 2.1 and 2.3, we can give a necessary and sufficient condition for the existence of a nonsingular rational matrix with prescribed Wiener–Hopf factorization indices and finite structure.

THEOREM 2.4. *Let $k_1 \geq \cdots \geq k_m$ be integers and $\frac{\epsilon_1(s)}{\psi_1(s)}, \ldots, \frac{\epsilon_m(s)}{\psi_m(s)}$ irreducible rational functions, where $\epsilon_i(s), \psi_i(s) \in \mathbb{F}[s]$ are monic and coprime such that $\epsilon_i(s) \mid \epsilon_{i+1}(s)$, $i = 1, 2, \ldots, m-1$, and $\psi_{i+1}(s) \mid \psi_i(s)$, $i = 1, 2, \ldots, m-1$. Then there exists a nonsingular matrix $T(s) \in \mathbb{F}(s)^{m\times m}$ with $k_1, \ldots, k_m$ as Wiener–Hopf factorization*

indices and $\frac{\epsilon_1(s)}{\psi_1(s)}, \dots, \frac{\epsilon_m(s)}{\psi_m(s)}$ *as invariant rational functions if and only if*

$$(k_1, \dots, k_m) \prec (d(\epsilon_m(s)) - d(\psi_m(s)), \dots, d(\epsilon_1(s)) - d(\psi_1(s))). \tag{2.2}$$

Proof. Let $T(s)$, $d(s)$, $N(s)$, and d be as in Lemma 2.2. By using Lemmas 2.1 and 2.3 we know that if $\sigma_i(s) = \frac{d(s)}{\psi_i(s)}$, $i = 1, \dots, m$, then $\epsilon_1(s)\sigma_1(s) \mid \cdots \mid \epsilon_m(s)\sigma_m(s)$ are the invariant factors of $N(s)$ and $k_1 + d, \dots, k_m + d$ are its Wiener–Hopf factorization indices. Thus, by Theorem 1.1 we have that

$$(k_1 + d, \dots, k_m + d) \prec (d(\epsilon_m(s)\sigma_m(s)), \dots, d(\epsilon_1(s)\sigma_1(s))).$$

But $d(\sigma_i(s)) = d - d(\psi_i(s))$, $i = 1, \dots, m$. Thus,

$$(k_1 + d, \dots, k_m + d) \prec (d(\epsilon_m(s)) - d(\psi_m(s)) + d, \dots, d(\epsilon_1(s)) - d(\psi_1(s)) + d),$$

and condition (2.2) follows.

Conversely, as by definition of majorization $k_m \geq d(\epsilon_1(s)) - d(\psi_1(s))$, we have that $k_m + d(\psi_1(s)) \geq d(\epsilon_1(s)) \geq 0$. Thus, $k_1 + d(\psi_1(s)) \geq \cdots \geq k_m + d(\psi_1(s)) \geq 0$.

On the other hand, as $\psi_i(s)$ divides $\psi_1(s)$, $i = 1, \dots, m$, we have that $\frac{\epsilon_1(s)}{\psi_i(s)}\psi_1(s)$ are polynomials satisfying $\frac{\epsilon_1(s)}{\psi_1(s)}\psi_1(s) \mid \cdots \mid \frac{\epsilon_m(s)}{\psi_m(s)}\psi_1(s)$.

Now, from (2.2),

$$\begin{aligned}&(k_1 + d(\psi_1(s)), \dots, k_m + d(\psi_1(s)))\\ &\prec (d(\epsilon_m(s)) - d(\psi_m(s)) + d(\psi_1(s)), \dots, d(\epsilon_1(s)) - d(\psi_1(s)) + d(\psi_1(s))).\end{aligned} \tag{2.3}$$

Hence

$$(k_1 + d(\psi_1(s)), \dots, k_m + d(\psi_1(s))) \prec \left(d\left(\frac{\epsilon_m(s)}{\psi_m(s)}\psi_1(s)\right), \dots, d\left(\frac{\epsilon_1(s)}{\psi_1(s)}\psi_1(s)\right)\right). \tag{2.4}$$

By Theorem 1.1 there exists $N(s) \in \mathbb{F}[s]^{m\times m}$ with $k_1 + d(\psi_1(s)), \dots, k_m + d(\psi_1(s))$ as Wiener–Hopf factorization indices and $\frac{\epsilon_1(s)}{\psi_1(s)}\psi_1(s) \mid \cdots \mid \frac{\epsilon_m(s)}{\psi_m(s)}\psi_1(s)$ as invariant factors. By Lemmas 2.1 and 2.3, $T(s) = \frac{1}{\psi_1(s)}N(s)$ is the desired matrix. ∎

In the following two sections we study the existence of polynomial matrices with prescribed infinite structure and Wiener–Hopf factorization indices. Then we will extend the obtained results to rational matrices with the help of Lemmas 2.2 and 2.3.

3. Local Wiener–Hopf factorization indices of a polynomial matrix. We are led by the following idea to relate the infinite structure and the Wiener–Hopf factorization indices of a polynomial matrix: By a conformal transformation we can bring the point at infinity to any desired finite point, and then use the known properties (Rosenbrock's theorem) for the finite structure of the corresponding polynomial matrix in order to obtain information about those properties at infinity by reversing the transformation. Locally the invariant factors reduce to the elementary divisors, but we still need the concept of local Wiener–Hopf factorization indices. As we said in the introduction, this is a well-known concept when $\mathbb{F} = \mathbb{C}$, the field of the complex numbers (see, for example, [5]), but since we are working on an arbitrary field we will use a different approach.

We need some preliminary results.

PROPOSITION 3.1. *Let* $P(s) \in \mathbb{F}[s]^{m\times m}$ *be a nonsingular polynomial matrix and* $\pi(s) \in \mathbb{F}[s]$ *a monic irreducible polynomial. Then there exist matrices* $A(s) \in \mathbb{F}[s]^{m\times m}$ *and* $B(s) \in \mathbb{F}[s]^{m\times m}$ *such that*

(i) $P(s) = A(s)B(s)$,
(ii) *the invariant factors of* $A(s)$ *are powers of the polynomial* $\pi(s)$,
(iii) *the invariant factors of* $B(s)$ *are relatively prime with* $\pi(s)$.

Proof. Let $S(s) = \operatorname{diag}(\alpha_1(s), \dots, \alpha_m(s))$ be the Smith canonical form of $P(s)$, $\alpha_1(s) \mid \cdots \mid \alpha_m(s)$ being its invariant factors. So there exist unimodular matrices $U(s), V(s) \in \mathbb{F}[s]^{m\times m}$ such that

$$P(s) = U(s)\operatorname{diag}(\alpha_1(s), \dots, \alpha_m(s))V(s). \tag{3.1}$$

If $\pi(s)$ is not a divisor of $\det P(s)$, then $A(s) = I_m$ and $B(s) = P(s)$ satisfy the requirements. If $\pi(s) \mid \det P(s)$, then $\alpha_i(s) = \pi(s)^{d_i}\beta_i(s)$ with $d_m \geq \cdots \geq d_1 \geq 0$ and $\gcd(\beta_i(s), \pi(s)) = 1$, $i \in \{1, 2, \dots, m\}$. Thus,

$$P(s) = A(s)B(s) \tag{3.2}$$

with

$$A(s) = U(s)\operatorname{diag}(\pi(s)^{d_1}, \dots, \pi(s)^{d_m}),$$
$$B(s) = \operatorname{diag}(\beta_1(s), \dots, \beta_m(s))V(s),$$

and the proposition follows. ∎

PROPOSITION 3.2. *Let* $P(s) \in \mathbb{F}[s]^{m\times m}$ *be a nonsingular polynomial matrix and* $\pi(s) \in \mathbb{F}[s]$ *a monic irreducible polynomial. If there exist polynomial matrices* $A_1(s), B_1(s) \in \mathbb{F}[s]^{m\times m}$ *and* $A_2(s), B_2(s) \in \mathbb{F}[s]^{m\times m}$ *such that*
(i) $P(s) = A_1(s)B_1(s) = A_2(s)B_2(s)$,
(ii) *the invariant factors of* $A_1(s)$ *and* $A_2(s)$ *are powers of* $\pi(s)$,
(iii) *the invariant factors of* $B_1(s)$ *and* $B_2(s)$ *are relatively prime with* $\pi(s)$,
then $A_1(s)$ *and* $A_2(s)$ *are right equivalent.*

Proof. By using conditions (i), (ii), and (iii), we can see that $A_1(s)$ and $A_2(s)$ as well as $B_1(s)$ and $B_2(s)$ have, respectively, the same invariant factors: the first being powers of $\pi(s)$ and the second relatively prime with $\pi(s)$.

By using (i) we have that

$$A_1(s) = A_2(s)B_2(s)B_1(s)^{-1}. \tag{3.3}$$

Let $G(s) = B_2(s)B_1(s)^{-1}$. As $B_1(s)^{-1} = \frac{1}{\det B_1(s)}Adj(B_1(s))$, it follows that the matrix $(\det B_1(s))G(s)$ is polynomial. Let $D(s) \in \mathbb{F}[s]^{m\times m}$ be the Smith canonical form of $(\det B_1(s))G(s)$. So there exist unimodular matrices $U(s)$, $V(s) \in \mathbb{F}[s]^{m\times m}$ such that

$$D(s) = U(s)(\det B_1(s))G(s)V(s).$$

Therefore

$$U(s)G(s)V(s) = \frac{1}{\det B_1(s)}D(s).$$

On the other hand, if we divide $D(s)$ by $\det B_1(s)$ and cancel out common factors, we obtain the Smith–McMillan form of $G(s)$ (see [13]):

$$S_G(s) = U(s)G(s)V(s) = \frac{1}{\det B_1(s)}D(s) = \operatorname{diag}\left(\frac{\epsilon_1(s)}{\psi_1(s)}, \dots, \frac{\epsilon_m(s)}{\psi_m(s)}\right),$$

where $\epsilon_i(s), \psi_i(s)$ are monic and coprime polynomials such that $\epsilon_i(s)|\epsilon_{i+1}(s)$, while

$\psi_{i+1}(s)|\psi_i(s)$, $i = 1, 2, \ldots, m-1$. Furthermore, as $\psi_k(s)$ is a divisor of $\det B_1(s)$ for all k, we conclude that the polynomials $\psi_k(s)$ are relatively prime with the invariant factors of $A_2(s)$.

Note that since $\det(B_2(s)B_1(s)^{-1})$ is a constant, $\det(S_G(s)) = 1$. We claim that, in fact, $S_G(s) = I_m$. If this were not true, there would exist a k such that $\psi_k(s) \neq 1$ (because otherwise, since $\det(S_G(s)) = 1$, we would have that $\epsilon_i(s) = 1$ for all i and then $S_G(s) = I_m$).

Let $a_{ik}(s)\frac{\epsilon_k(s)}{\psi_k(s)}$ be an arbitrary element of the kth column of $A_2(s)U(s)^{-1}S_G(s)$, where $a_{ik}(s)$ is the element in the (i,k) position of $A_2(s)U(s)^{-1}$. As $A_1(s)V(s) = A_2(s)U(s)^{-1}S_G(s)$ is a polynomial matrix, $a_{ik}(s)\frac{\epsilon_k(s)}{\psi_k(s)}$ is a polynomial. But since $\gcd(\epsilon_k(s), \psi_k(s)) = 1$ we conclude that necessarily $\psi_k(s) \mid a_{ik}(s)$ for all i; i.e., $\psi_k(s)$ is a divisor of every entry in the kth column of $A_2(s)U(s)^{-1}$ and thus a divisor of $\det(A_2(s)U(s)^{-1})$. This would be a contradiction unless $\psi_k(s) = 1$ because this polynomial is a divisor of $\det B_1(s)$ and this one is relatively prime to $\det A_1(s)$, which is equal (up to the product by a constant) to $\det(A_2(s)U(s)^{-1})$. So $S_G(s) = I_m$ as claimed and $A_1(s) = A_2(s)U(s)^{-1}V(s)^{-1}$ as desired. □

As said in the introduction, the Wiener–Hopf factorization indices of a polynomial matrix $P(s)$ are the column degrees of any of its right equivalent column reduced matrices. In fact, there may be more than one column reduced matrix right equivalent to $P(s)$ but all of them have the same column degrees. Thus, by Propositions 3.1 and 3.2, the following definition makes sense.

DEFINITION 3.3. *Let $P(s) \in \mathbb{F}[s]^{m\times m}$ be a nonsingular polynomial matrix, and $\pi(s) \in \mathbb{F}[s]$ a monic irreducible polynomial. Let $A(s), B(s) \in \mathbb{F}[s]^{m\times m}$ be polynomial matrices such that*

(i) *$P(s) = A(s)B(s)$,*
(ii) *the invariant factors of $A(s)$ are powers of the polynomial $\pi(s)$,*
(iii) *the invariant factors of $B(s)$ are relatively prime with $\pi(s)$.*

Then the Wiener–Hopf factorization indices of $A(s)$ will be called local Wiener–Hopf factorization indices of $P(s)$ with respect to $\pi(s)$.

4. Wiener–Hopf factorization indices and infinite structure. In this section we first give the relationship between the Wiener–Hopf factorization indices of a nonsingular polynomial matrix and its invariant factors at infinity. We need some technical lemmas. The proof of the first one is straightforward.

LEMMA 4.1.

(a) *Let $P(s) \in \mathbb{F}[s]^{m\times m}$ and $q = d(P(s))$. Then $s^qP(1/s)$ is a polynomial matrix.*

(b) *Let $U(s) \in \mathbb{F}[s]^{m\times m}$ be a unimodular matrix. Then $\det(U(1/s)) = c \in \mathbb{F}, c \neq 0$ and $U(1/s)$ is a biproper matrix.*

(c) *Let $L(s) \in \mathbb{F}_{pr}(s)^{m\times m}$ be a biproper matrix and $m(s)$ the least common denominator of the elements of $L(1/s)$. Then $m(s)L(1/s)$ is a polynomial matrix, $\gcd(m(s), s) = 1$, and $\gcd(\det(m(s)L(1/s)), s) = 1$.*

(d) *If $P(s)$ is a column reduced matrix with column degrees $p_1, \ldots, p_m$ and $D(s) = \operatorname{diag}(s^{p_1}, \ldots, s^{p_m})$, then $P(1/s)D(1/s)^{-1}$ is a polynomial matrix.*

LEMMA 4.2. *Let $P(s) \in \mathbb{F}[s]^{m\times m}$ be a nonsingular polynomial matrix with $q = d(P(s))$ and let $k_1 \geq k_2 \geq \cdots \geq k_m$ be its Wiener–Hopf factorization indices. Then there exist both a nonnegative integer d and a polynomial, relatively prime with s, $m(s)$ such that $d + q - k_m, \ldots, d + q - k_1$ are the local Wiener–Hopf factorization indices of $m(s)s^{d+q}P(1/s)^T \in \mathbb{F}[s]^{m\times m}$ with respect to the polynomial s.*

Proof. As $k_1, k_2, \dots, k_m$ are the Wiener–Hopf factorization indices of $P(s)$, there exist matrices $U(s) \in \mathbb{F}[s]^{m\times m}$, unimodular, and $L(s) \in \mathbb{F}_{pr}(s)^{m\times m}$, biproper, such that

$$P(s) = L(s)\,\mathrm{diag}(s^{k_m}, \dots, s^{k_1})U(s). \tag{4.1}$$

Let $m(s)$ be the least common denominator of the elements of $L(1/s)$ and $d = d(U(s))$. Then, replacing s by $1/s$ and multiplying by $m(s)s^{d+q}$ in (4.1), we get

$$m(s)s^{d+q}P(1/s) = m(s)L(1/s)\,\mathrm{diag}(s^{d+q-k_m}, \dots, s^{d+q-k_1})U(1/s),$$

and transposing

$$m(s)s^{d+q}P(1/s)^T = U(1/s)^T\,\mathrm{diag}(s^{d+q-k_m}, \dots, s^{d+q-k_1})m(s)L(1/s)^T.$$

Now, let $B(s) = m(s)L(1/s)^T$ and $A(s) = U(1/s)^T\,\mathrm{diag}(s^{d+q-k_m}, \dots, s^{d+q-k_1})$. Then, using Lemma 4.1, the following hold.

(i) $m(s)s^{d+q}P(1/s)^T = A(s)B(s)$ is a polynomial matrix. $A(s)$ is also polynomial. In fact, on the one hand, $s^dU(1/s)^T$ is polynomial, and on the other hand, one can easily deduce from [15] that $q - k_i \geq 0$ for all $i = 1, \dots, m$.

(ii) The invariant factors of $A(s)$ are powers of s.

(iii) The invariant factors of $B(s)$ are relatively prime with s.

(iv) The Wiener–Hopf factorization indices of $A(s)$ are $d+q-k_m, \dots, d+q-k_1$, since $U(1/s)$ is a biproper matrix.

We conclude then, from Definition 3.3, that the local Wiener–Hopf factorization indices of $m(s)s^{d+q}P(1/s)^T$ are $d+q-k_m, \dots, d+q-k_1$ as desired. ∎

Recall now, from the first remark in this paper, that if $s^{q_1}, \dots, s^{q_m}$ are the invariant factors at infinity of $P(s)$, then $q_1 = d(P(s))$. So $s^{q_1}P(1/s)$ is a polynomial matrix (Lemma 4.1).

LEMMA 4.3. *Let $P(s) \in \mathbb{F}[s]^{m\times m}$ be a nonsingular polynomial matrix and $s^{q_1}, \dots, s^{q_m}$, with $q_1 \geq q_2 \geq \cdots \geq q_m$ its invariant factors at infinity. Then for any nonnegative integer d and any polynomial relatively prime with s, $m(s)$, the polynomials $s^{d+q_1-q_1} \mid \cdots \mid s^{d+q_1-q_m}$ are the finite elementary divisors of the polynomial matrix $m(s)s^{d+q_1}P(1/s)^T \in \mathbb{F}[s]^{m\times m}$ associated to s.*

Proof. Let $\alpha_1(s) \mid \cdots \mid \alpha_m(s)$ be the finite invariant factors of $s^{q_1}P(1/s)$. Then there exist unimodular matrices $U(s), V(s) \in \mathbb{F}[s]^{m\times m}$ such that

$$U(s)s^{q_1}P(1/s)V(s) = \mathrm{diag}(\alpha_1(s), \dots, \alpha_m(s)). \tag{4.2}$$

If we write $\alpha_i(s) = s^{a_i}\beta_i(s)$ with $\gcd(\beta_i(s), s) = 1$, $i = 1, \dots, m$, then $0 \leq a_1 \leq \cdots \leq a_m$. By changing s by $1/s$ in (4.2) we have

$$U(1/s)\frac{1}{s^{q_1}}P(s)V(1/s) = \mathrm{diag}\left(\frac{1}{s^{a_1}}\beta_1(1/s), \dots, \frac{1}{s^{a_m}}\beta_m(1/s)\right). \tag{4.3}$$

Notice that $\beta_i(1/s)$ is a biproper rational function because $\gcd(\beta_i(s), s)) = 1$. Thus, by multiplying (4.3) on the right by the biproper matrix

$$B(s) = \mathrm{diag}\left(\frac{1}{\beta_1(1/s)}, \dots, \frac{1}{\beta_m(1/s)}\right)$$

and setting $B_1(s) = U(1/s)$ and $B_2(s) = V(1/s)B(s)$, we have that

$$B_1(s)P(s)B_2(s) = \mathrm{diag}(s^{q_1-a_1}, \dots, s^{q_1-a_m}),$$

with $B_1(s)$ and $B_2(s)$ biproper matrices and $q_1 - a_1 \geq \cdots \geq q_1 - a_m$. This means that $s^{q_1-a_1}, s^{q_1-a_2}, \ldots, s^{q_1-a_m}$ are the invariant factors at infinity of $P(s)$. But, by hypothesis, these are $s^{q_1}, \ldots, s^{q_m}$. Therefore $a_i = q_1 - q_i$, $i = 1, \ldots, m$.

Now, let d be any nonnegative integer and $m(s)$ any polynomial relatively prime with s. By multiplying (4.2) by $m(s)s^d$ and transposing, we get

$$\begin{aligned} &V(s)^T(m(s)s^{d+q_1}P(1/s)^T)U(s)^T \\ &= \operatorname{diag}(s^{d+q_1-q_1}m(s)\beta_1(s), \ldots, s^{d+q_1-q_m}m(s)\beta_m(s)), \end{aligned}$$

where $\gcd(m(s)\beta_i(s), s) = 1$ for all $i = 1, \ldots, m$. This means that

$$s^{d+q_1-q_1}m(s)\beta_1(s), \ldots, s^{d+q_1-q_m}m(s)\beta_m(s)$$

are the invariant factors of $m(s)s^{d+q_1}P(1/s)^T$ and so $s^{d+q_1-q_1} \mid \cdots \mid s^{d+q_1-q_m}$ are its elementary divisors associated to the irreducible polynomial s. ∎

LEMMA 4.4. *Let $k_1 \geq \cdots \geq k_m$, $t_m \geq \cdots \geq t_1$ be nonnegative integers. Let $\pi(s) \in \mathbb{F}[s]$ be a monic irreducible polynomial. Then there exists a nonsingular matrix $P(s) \in \mathbb{F}[s]^{m\times m}$ with $k_1, \ldots, k_m$ as its local Wiener–Hopf factorization indices with respect to $\pi(s)$ and $\pi(s)^{t_1} \mid \cdots \mid \pi(s)^{t_m}$ as its finite elementary divisors associated to $\pi(s)$ if and only if*

$$(k_1, \ldots, k_m) \prec (t_m d(\pi(s)), \ldots, t_1 d(\pi(s))). \tag{4.4}$$

Proof. Assume that there is a nonsingular matrix $P(s) \in \mathbb{F}[s]^{m\times m}$ such that $\alpha_1(s) \mid \cdots \mid \alpha_m(s)$ are its invariant factors. We can write $\alpha_i(s) = \pi(s)^{t_i}\beta_i(s)$ with $\gcd(\beta_i(s), \pi(s)) = 1$, $i = 1, \ldots, m$. The Smith canonical form of $P(s)$ is $\operatorname{diag}(\alpha_1(s), \ldots, \alpha_m(s))$ and there exist unimodular matrices $U(s), V(s) \in \mathbb{F}[s]^{m\times m}$ such that $P(s) = A(s)B(s)$ with

$$A(s) = U(s)\operatorname{diag}(\pi(s)^{t_1}, \ldots, \pi(s)^{t_m})$$

and

$$B(s) = \operatorname{diag}(\beta_1(s), \ldots, \beta_m(s))V(s).$$

By Definition 3.3 the local Wiener–Hopf factorization indices of $P(s)$ with respect to $\pi(s)$, i.e., $k_1, \ldots, k_m$, are the Wiener–Hopf factorization indices of $A(s)$. Since $\pi(s)^{t_1} \mid \cdots \mid \pi(s)^{t_m}$ are the invariant factors of $A(s)$, condition (4.4) follows from Rosenbrock's theorem (Theorem 1.1).

Conversely, by Rosenbrock's theorem there exists a nonsingular matrix $A(s) \in \mathbb{F}[s]^{m\times m}$ with Wiener–Hopf factorization indices $k_1, \ldots, k_m$ and $\pi(s)^{t_1} \mid \cdots \mid \pi(s)^{t_m}$ as invariant factors. So it is sufficient to take $P(s) = A(s)$. ∎

Now we can give our main result.

THEOREM 4.5. *Let $k_1 \geq \cdots \geq k_m \geq 0$ and $q_1 \geq \cdots \geq q_m$ be integers. Then there exists a nonsingular matrix $P(s) \in \mathbb{F}[s]^{m\times m}$ with $k_1, \ldots, k_m$ as Wiener–Hopf factorization indices and $s^{q_1}, \ldots, s^{q_m}$ as invariant factors at infinity if and only if*

$$(k_1, \ldots, k_m) \prec (q_1, \ldots, q_m). \tag{4.5}$$

Proof. First, we are going to prove the necessity. We know that $q_1 = d(P(s))$. By using Lemma 4.2, there exist both a polynomial $m(s)$ such that $\gcd(m(s), s) = 1$ and

a nonnegative integer d such that $d+q_1-k_m,\dots,d+q_1-k_1$ are the local Wiener–Hopf factorization indices of $m(s)s^{d+q_1}P(1/s)^T \in \mathbb{F}[s]^{m\times m}$ with respect to s, and by Lemma 4.3, $s^{d+q_1-q_1} \mid \cdots \mid s^{d+q_1-q_m}$ are its finite elementary divisors associated to the irreducible polynomial s. Then, by Lemma 4.4 applied to the polynomial matrix $m(s)s^{d+q_1}P(1/s)^T$, we have that

$$(d+q_1-k_m,\dots,d+q_1-k_1) \prec (d+q_1-q_m,\dots,d+q_1-q_1).$$

From the definition of majorization we conclude that condition (4.5) holds.

Conversely, if condition (4.5) holds and bearing in mind that $q_1 \geq k_1 \geq k_i$, $i=1,\dots,m$, we have that

$$(q_1-k_m,\dots,q_1-k_1) \prec (q_1-q_m,\dots,q_1-q_1).$$

By Rosenbrock's theorem, there is a nonsingular matrix $A(s) \in \mathbb{F}[s]^{m\times m}$ with $q_1-k_m,\dots,q_1-k_1$ as Wiener–Hopf factorization indices and $s^{q_1-q_1} \mid \cdots \mid s^{q_1-q_m}$ as invariant factors. As postmultiplication by unimodular matrices does not change either the Wiener–Hopf factorization indices or the invariant factors, we can assume that $A(s)$ is column-reduced and the degree of its ith column is q_1-k_{m-i+1}. Thus we can write

$$A(s) = L(s)\,\mathrm{diag}(s^{q_1-k_m},\dots,s^{q_1-k_1}) \tag{4.6}$$

with $L(s)$ a biproper matrix. Furthermore, since $q_1-k_m \geq \cdots \geq q_1-k_1$, it follows that $d(A(s)) = q_1-k_m$.

By replacing s by $1/s$ and multiplying equation (4.6) by s^{q_1} we have that

$$s^{q_1}A(1/s) = L(1/s)\,\mathrm{diag}(s^{k_m},\dots,s^{k_1}),$$

where $s^{q_1}A(1/s)$ is a polynomial matrix (Lemma 4.1). Then

$$s^{q_1}A(1/s)^T = \mathrm{diag}(s^{k_m},\dots,s^{k_1})L(1/s)^T.$$

Recall now that, by Lemma 4.1, if $D(s) = \mathrm{diag}(s^{q_1-k_m},\dots,s^{q_1-k_1})$, then matrix $L(1/s) = A(1/s)D(1/s)^{-1}$ is polynomial. But since $L(s)$ is biproper, $\det L(1/s)$ is both a biproper rational function and a polynomial. Thus $\det L(1/s) = c \in \mathbb{F}, c \neq 0$ and $L(1/s)$ is unimodular. Hence the Wiener–Hopf factorization indices of $s^{q_1}A(1/s)^T$ are $k_1,\dots,k_m$.

On the other hand, as $s^{q_1-q_1} \mid \cdots \mid s^{q_1-q_m}$ are the invariant factors of $A(s)$, there exist unimodular matrices $U(s), V(s) \in \mathbb{F}[s]^{m\times m}$ such that

$$U(s)A(s)V(s) = \mathrm{diag}(s^{q_1-q_1},\dots,s^{q_1-q_m}). \tag{4.7}$$

Then, replacing s by $1/s$ and multiplying by s^{q_1}, we have that

$$U(1/s)s^{q_1}A(1/s)V(1/s) = \mathrm{diag}(s^{q_1},\dots,s^{q_m})$$

and

$$V(1/s)^T s^{q_1}A(1/s)^T U(1/s)^T = \mathrm{diag}(s^{q_1},\dots,s^{q_m}),$$

where $U(1/s)^T$ and $V(1/s)^T$ are biproper matrices (Lemma 4.1). Hence $s^{q_1},\dots,s^{q_m}$ are the invariant factors at infinity of $s^{q_1}A(1/s)^T$. Therefore if we write $P(s) = s^{q_1}A(1/s)^T$ the theorem follows. ∎

We finish this section by extending the previous result to nonsingular rational matrices.

THEOREM 4.6. *Let $k_1 \geq \cdots \geq k_m$ and $q_1 \geq \cdots \geq q_m$ be integers. Then there exists a nonsingular matrix $T(s) \in \mathbb{F}(s)^{m\times m}$ with $k_1, \ldots, k_m$ as Wiener–Hopf factorization indices and $s^{q_1}, \ldots, s^{q_m}$ as invariant factors at infinity if and only if*

$$(k_1, \ldots, k_m) \prec (q_1, \ldots, q_m). \tag{4.8}$$

Proof. Recall that if $d(s)$ is the monic least common denominator of all the elements of $T(s)$ and $d = d(d(s))$, then

$$T(s) = \frac{1}{d(s)} N(s),$$

where $N(s) \in \mathbb{F}[s]^{m\times m}$. By using Lemmas 2.2 and 2.3 we know that $k_1+d, \ldots, k_m+d$ are the Wiener–Hopf factorization indices of $N(s)$ and $s^{q_1+d}, \ldots, s^{q_m+d}$ are its invariant factors at infinity. Now, by Theorem 4.5 applied to the nonsingular polynomial matrix $N(s)$, we have that

$$(k_1 + d, \ldots, k_m + d) \prec (q_1 + d, \ldots, q_m + d),$$

and (4.8) follows.

Conversely, let d be any integer such that $d \geq |k_m|$. Then $k_1 + d, \ldots, k_m + d$ are nonnegative integers and $q_1 + d, \ldots, q_m + d$ are integers. From (4.8)

$$(k_1 + d, \ldots, k_m + d) \prec (q_1 + d, \ldots, q_m + d).$$

Thus, by Theorem 4.5 there exists $N(s) \in \mathbb{F}[s]^{m\times m}$ with $k_1+d, \ldots, k_m+d$ as Wiener–Hopf factorization indices and $s^{q_1+d}, \ldots, s^{q_m+d}$ as invariant factors at infinity.

Let $T(s) = \frac{1}{d(s)} N(s)$, where $d(s)$ is any polynomial of degree d. By applying Lemmas 2.2 and 2.3, we have that $T(s)$ is the desired matrix. □

As a consequence we get the following relationship between the Wiener–Hopf factorization indices and invariant factors at infinity of proper rational function matrices.

COROLLARY 4.7. *Let $k_1 \geq \cdots \geq k_m$ and $q_1 \geq \cdots \geq q_m$ be nonpositive integers. Then there exist a nonsingular matrix $T(s) \in \mathbb{F}(s)_{pr}^{m\times m}$ with $k_1, \ldots, k_m$ as Wiener–Hopf factorization indices and $s^{q_1}, \ldots, s^{q_m}$ as invariant factors at infinity if and only if*

$$(k_1, \ldots, k_m) \prec (q_1, \ldots, q_m).$$

5. Wiener–Hopf factorization indices and finite and infinite structures of rational matrices. Although prescribing three types of invariants usually results in a more difficult problem, this is not the case for the one that we are studying. The reason is that, for a given rational matrix $T(s) \in \mathbb{F}(s)^{m\times m}$, the Wiener–Hopf factorization indices of all matrices in the sets

$$\{B(s)T(s) : B(s) \text{ biproper}\} \text{ and } \{T(s)U(s) : U(s) \text{ unimodular}\}$$

are the same. The following theorem characterizes the finite and infinite structure of nonsingular rational matrices when their Wiener–Hopf factorization indices are prescribed.

THEOREM 5.1. *Let $k_1 \geq \cdots \geq k_m$, $q_1 \geq \cdots \geq q_m$ be integers and $\frac{\epsilon_1(s)}{\psi_1(s)}, \ldots, \frac{\epsilon_m(s)}{\psi_m(s)}$ irreducible rational functions, where $\epsilon_i(s), \psi_i(s) \in \mathbb{F}[s]$ are monic and coprime such that $\epsilon_i(s) \mid \epsilon_{i+1}(s)$, $i = 1, 2, \ldots, m-1$, and $\psi_{i+1}(s) \mid \psi_i(s)$, $i = 1, 2, \ldots, m-1$. Then there exists a nonsingular rational matrix $T(s) \in \mathbb{F}(s)^{m\times m}$ with $k_1, \ldots, k_m$ as Wiener–Hopf factorization indices, $s^{q_1}, \ldots, s^{q_m}$ as invariant factors at infinity, and $\frac{\epsilon_1(s)}{\psi_1(s)}, \ldots, \frac{\epsilon_m(s)}{\psi_m(s)}$ as invariant rational functions if and only if*

$$(k_1, \ldots, k_m) \prec (d(\epsilon_m(s)) - d(\psi_m(s)), \ldots, d(\epsilon_1(s)) - d(\psi_1(s))), \tag{5.1}$$

$$(k_1, \ldots, k_m) \prec (q_1, \ldots, q_m). \tag{5.2}$$

Proof. The necessity of (5.1) and (5.2) follows from Theorems 2.4 and 4.6. Conversely, by Theorem 2.4 there exists a rational matrix $T_1(s)$ whose Wiener–Hopf factorization indices are $k_1, \ldots, k_m$ and whose invariant rational functions are $\frac{\epsilon_1(s)}{\psi_1(s)}, \ldots, \frac{\epsilon_m(s)}{\psi_m(s)}$. And by Theorem 4.6 there exists a rational matrix $T_2(s)$ whose Wiener–Hopf factorization indices are $k_1, \ldots, k_m$ and whose invariant factors at infinity are $s^{q_1}, \ldots, s^{q_m}$. As $T_1(s)$ and $T_2(s)$ have the same Wiener–Hopf factorization indices, they are Wiener–Hopf equivalent. So there exist both a biproper matrix $B(s) \in \mathbb{F}_{pr}(s)^{m\times m}$ and a unimodular matrix $U(s) \in \mathbb{F}[s]^{m\times m}$ such that $T_1(s) = B(s)T_2(s)U(s)$. Thus $T(s) = B(s)T_2(s) = T_1(s)U(s)^{-1}$ has the same infinite structure as $T_2(s)$, the same finite structure as $T_1(s)$ and the Wiener–Hopf factorization indices of both $T_1(s)$ and $T_2(s)$. That is to say, $T(s)$ is the desired matrix. ∎

Remark. All throughout the paper we have been working with left Wiener–Hopf factorization indices. The left Wiener–Hopf factorization indices of a nonsingular polynomial matrix $P(s) \in \mathbb{F}[s]^{m\times m}$ are the column degrees of any column proper polynomial matrix right equivalent to $P(s)$. And these are, in turn, the controllability indices of any controllable pair (A, B) with $P(s)$ as polynomial matrix representation.

As said in the introduction, we also have right Wiener–Hopf indices and right Wiener–Hopf equivalence of rational matrices. This is the same as the left Wiener–Hopf equivalence but exchanging the roles of matrices $B(s)$ and $U(s)$. That is to say, $T_1(s), T_2(s)$ are right Wiener–Hopf equivalent if there are matrices $U(s)$, unimodular, and $B(s)$, biproper, such that $T_1(s) = U(s)T_2(s)B(s)$. In other words, $T_1(s)$ and $T_2(s)$ are left Wiener–Hopf equivalent if and only if $T_1(s)^T$ and $T_2(s)^T$ are right Wiener–Hopf equivalent. For a polynomial matrix, P(s), the right Wiener–Hopf factorization indices are the row degrees of any row proper polynomial matrix *left* equivalent to $P(s)$. And these are, again, the observability indices of any pair (B^T, A^T) such that $P(s)^T$ is a polynomial matrix representation of (A, B).

Since matrix transposition of a rational matrix does not change its finite or infinite structure, all results in this manuscript can be translated straightforwardly into results about the right Wiener–Hopf factorization indices of rational matrices just by transposition.

We finish with two results that relate the left and right Wiener–Hopf factorization indices. A partial result on this problem can be found in [3]. Actually, in that paper the relationship between several types of factorization indices was studied. In particular, a necessary and sufficient condition for two integer vectors to be the left and right Wiener–Hopf factorization indices of a rational matrix function is provided. This result is a consequence of our Theorem 5.3 and is presented as Corollary 5.4.

We consider again the sets

$$\{U(s)T(s) : U(s) \text{ unimodular}\} \text{ and } \{T(s)U(s) : U(s) \text{ unimodular}\}.$$

All matrices in these sets have the same finite structure and all the matrices in the sets

$$\{B(s)T(s) : B(s) \text{ biproper}\} \text{ and } \{T(s)B(s) : B(s) \text{ biproper}\}$$

have the same infinite structure. Ideas similar to those of Theorem 5.1 enable us to prove the following.

THEOREM 5.2. *Let $k_1 \geq \cdots \geq k_m$, $l_1 \geq \cdots \geq l_m$ be integers and $\frac{\epsilon_1(s)}{\psi_1(s)}, \ldots, \frac{\epsilon_m(s)}{\psi_m(s)}$ m irreducible rational functions, where $\epsilon_i(s), \psi_i(s) \in \mathbb{F}[s]$ are monic and coprime such that $\epsilon_i(s) \mid \epsilon_{i+1}(s)$, $i = 1, 2, \ldots, m-1$, and $\psi_{i+1}(s) \mid \psi_i(s)$, $i = 1, 2, \ldots, m-1$. Then there exists a nonsingular rational matrix $T(s) \in \mathbb{F}(s)^{m\times m}$ with $k_1, \ldots, k_m$ as left Wiener–Hopf factorization indices, $l_1, \ldots, l_m$ as right Wiener–Hopf factorization indices, and $\frac{\epsilon_1(s)}{\psi_1(s)}, \ldots, \frac{\epsilon_m(s)}{\psi_m(s)}$ as invariant rational functions if and only if*

$$(k_1, \ldots, k_m) \prec (d(\epsilon_m(s)) - d(\psi_m(s)), \ldots, d(\epsilon_1(s)) - d(\psi_1(s))), \tag{5.3}$$

$$(l_1, \ldots, l_m) \prec (d(\epsilon_m(s)) - d(\psi_m(s)), \ldots, d(\epsilon_1(s)) - d(\psi_1(s))). \tag{5.4}$$

Proof. The necessity of (5.3) follows from Theorem 2.4. As $T(s)^T$ has $l_1, \ldots, l_m$ as left Wiener–Hopf factorization indices and $\frac{\epsilon_1(s)}{\psi_1(s)}, \ldots, \frac{\epsilon_m(s)}{\psi_m(s)}$ as invariant rational functions, (5.4) holds by applying Theorem 2.4 to $T(s)^T$. Conversely, by Theorem 2.4 there exists a rational matrix $T_1(s)$ whose left Wiener–Hopf factorization indices are $k_1, \ldots, k_m$ and whose invariant rational funtions are $\frac{\epsilon_1(s)}{\psi_1(s)}, \ldots, \frac{\epsilon_m(s)}{\psi_m(s)}$, and there exists another rational matrix $T_2(s)$ whose left Wiener–Hopf factorization indices are $l_1, \ldots, l_m$ and whose invariant rational funtions are $\frac{\epsilon_1(s)}{\psi_1(s)}, \ldots, \frac{\epsilon_m(s)}{\psi_m(s)}$. As $T_1(s)$ and $T_2(s)^T$ have the same finite structure, there are unimodular matrices $U_1(s), U_2(s) \in \mathbb{F}[s]^{m\times m}$ such that $T_1(s) = U_1(s)T_2(s)^T U_2(s)$. Thus the desired matrix is $T(s) = U_1(s)T_2(s)^T = T_1(s)U_2(s)^{-1}$. ∎

Similarly we can prove the following.

THEOREM 5.3. *Let $k_1 \geq \cdots \geq k_m$, $l_1 \geq \cdots \geq l_m$ and $q_1 \geq \cdots \geq q_m$ be integers. Then there exists a nonsingular rational matrix $T(s) \in \mathbb{F}(s)^{m\times m}$ with $k_1, \ldots, k_m$ as left Wiener–Hopf factorization indices, $l_1, \ldots, l_m$ as right Wiener–Hopf factorization indices, and $s^{q_1}, \ldots, s^{q_m}$ as invariant factors at infinity if and only if*

$$(k_1, \ldots, k_m) \prec (q_1, \ldots, q_m), \tag{5.5}$$

$$(l_1, \ldots, l_m) \prec (q_1, \ldots, q_m). \tag{5.6}$$

COROLLARY 5.4. *Let $k_1 \geq \cdots \geq k_m$ and $l_1 \geq \cdots \geq l_m$ be integers. Then there exists a nonsingular rational matrix $T(s) \in \mathbb{F}(s)^{m\times m}$ with $k_1, \ldots, k_m$ as left Wiener–Hopf factorization indices and $l_1, \ldots, l_m$ as right Wiener–Hopf factorization indices if and only if*

$$\sum_{i=1}^{m} k_i = \sum_{i=1}^{m} l_i. \tag{5.7}$$

Proof. The necessity of (5.7) is evident. Let us prove the sufficiency. We use the notation $q^+ = \max(q, 0)$ and choose integers $q_1 \geq \cdots \geq q_m$ in the following way:

$$q_1 = \max\left(\sum_{i=1}^{m} {k_i}^+, \sum_{i=1}^{m} {l_i}^+\right), \quad q_2 = \cdots = q_{m-1} = 0, \quad q_m = \sum_{i=1}^{m} k_i - q_1. \tag{5.8}$$

We have that

$$(k_1, \dots, k_m) \prec (q_1, \dots, q_m), \tag{5.9}$$

$$(l_1, \dots, l_m) \prec (q_1, \dots, q_m). \tag{5.10}$$

The existence of $T(s)$ follows from Theorem 5.3. ∎

6. Conclusions. In this paper the relationship between the Wiener–Hopf factorization indices of nonsingular polynomial matrices and their finite and infinite structures has been investigated. The main result states that they must satisfy a Rosenbrock-like theorem. In order to get this result the concept of local Wiener–Hopf factorization indices with respect to any irreducible polynomial has been introduced. The main results were generalized for rational matrices and the existence of matrices with all kind of invariants (left and right Wiener–Hopf indices, and finite and infinite structure) has been studied.

REFERENCES

[1] K. Clancey and I. Gohberg, *Factorization of Matrix Functions and Singular Integral Operators*, Birkhäuser Verlag, Basel, Boston, Stuttgart, 1981.

[2] J. M. Dion and C. Commault, *Smith-McMillan factorization at infinity of rational matrix functions and their control interpretation*, Systems Control Lett., 1 (1982), pp. 312–320.

[3] I. Feldman and A. Markus, *On some properties of factorization indices*, Integral Equations Operator Theory, 30 (1998), pp. 326–337.

[4] P. Fuhrmann and J. C. Willems, *Factorization indices at infinity for rational matrix functions*, Integral Equations Operator Theory, 2/3 (1979), pp. 287–301.

[5] I. Gohberg, M. A. Kaashoek, and F. van Schagen, *Partially Specified Matrices and Operators: Classification, Completion, Applications*, Birkhäuser Verlag, Basel, 1995.

[6] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, UK, 1967.

[7] M. L. J. Hautus and M. Heymann, *Linear feedback—an algebraic approach*, SIAM J. Control Optim., 16 (1978), pp. 83–105.

[8] M. L. J. Hautus and M. Heymann, *Linear feedback decoupling: Transfer function analysis*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 823–832.

[9] T. Kailath, *Linear Systems*, Prentice-Hall, London, 1980.

[10] M. Malabre and V. Kučera, *Infinite structure and exact model matching problem: A geometric approach*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 226–268.

[11] V. M. Popov, *Invariant description of linear, time-invariant controllable systems*, SIAM J. Control, 10 (1972), pp. 252–264.

[12] H. H. Rosenbrock, *State-space and Multivariable Theory*, Thomas Nelson and Sons, London, 1970.

[13] A. I. G. Vardulakis, *Linear Multivariable Control*, John Wiley and Sons, New York, 1991.

[14] G. L. Verghese, *Infinite Frequency Behaviour in Generalized Dynamical Systems*, Ph.D. Thesis, Stanford University, 1978.

[15] W. A. Wolovich, *Linear Multivariable Systems*, Appl. Math. Sci. 11, Springer-Verlag, New York, 1974.

[16] I. Zaballa, *Controllability and Hermite indices of matrix pairs*, Internat. J. Control, 68 (1997), pp. 61–86.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 2145–2171

ASYMPTOTIC CONTROL OF PAIRS OF OSCILLATORS COUPLED BY A REPULSION, WITH NONISOLATED EQUILIBRIA II: THE SINGULAR CASE*

MARC-OLIVIER CZARNECKI†

Abstract. Let $\phi : H \to \mathbb{R}$ be a $\mathcal{C}^1$ function on a real Hilbert space H, and let $\gamma > 0$ be a positive damping parameter. For any (singular) repulsive potential $V : H \setminus \{0\} \to \mathbb{R}_+$, i.e., such that $\lim_{z\to 0} V(z) = +\infty$, and any control function $\varepsilon : \mathbb{R}_+ \to \mathbb{R}_+ \setminus \{0\}$ which tends to zero as $t \to +\infty$, we study the asymptotic behavior of the trajectories of the coupled dissipative system of nonlinear oscillators

$$(\mathrm{HBFC}^2_{sing}) \qquad \begin{cases} \ddot{x} + \gamma\dot{x} + \nabla\phi(x) + \varepsilon(t)\nabla V(x-y) = 0, \\ \ddot{y} + \gamma\dot{y} + \nabla\phi(y) - \varepsilon(t)\nabla V(x-y) = 0. \end{cases}$$

This system is the singular version of the regular (HBFC^2_{reg}) system studied in [A. Cabot and M.-O. Czarnecki, *SIAM J. Control Optim.*, 41 (2002), pp. 1254–1280], where the potential V is defined on the whole space H. The purpose of this paper is to obtain whenever possible the same existence and convergence results in the singular case as in the regular case considered by A. Cabot and M.-O. Czarnecki. This study is mainly motivated by a better convergence behavior of (HBFC^2_{sing}) with the same sharp condition on the control ε exhibited in [H. Attouch and M.-O. Czarnecki, *J. Differential Equations*, 179 (2002), pp. 278–310] and by A. Cabot and M.-O. Czarnecki. Precisely, when $H = \mathbb{R}$, and if ε is a "slow" control, i.e., $\int_0^{+\infty} \varepsilon(t)dt = +\infty$, then the trajectories x and y converge to extremal points of the set $S = \{\lambda \in \mathbb{R}, \nabla\phi(\lambda) = 0\}$ of the equilibria of ϕ. The awkward case in A. Cabot and M.-O. Czarnecki, where the trajectories may have the same limit, disappears. Of importance, from a physical point of view, we thus can consider actual, for example electromagnetic, repulsive potentials.

Key words. nonlinear oscillator, coupled system, slow control, heavy ball with friction, global optimization, singular potential

AMS subject classifications. Primary, 37N40, 34G20; Secondary, 34H05, 34D05, 34E10, 49K15, 70F99

DOI. 10.1137/S036301290342320X

1. Introduction.

Let H be a real Hilbert space, with scalar product and corresponding norm, respectively, denoted by $\langle ., . \rangle$ and $|\ .\ |$. Let $\phi : H \to \mathbb{R}$ be a given $\mathcal{C}^1$ real-valued function, called the potential function. An important problem is the search of the equilibria of the function ϕ (i.e., the solutions of the equation $\nabla\phi(x) = 0$, where $\nabla\phi$ is the gradient of ϕ), among which the minima (global or local) play a particular role in optimization, physics, economics, etc. To obtain equilibria or minima of the function ϕ, a powerful method is to follow the trajectories of an associated dissipative gradient-like dynamical system, possibly discretized for numerical applications. In many practical problems (for example, when minimizing a convex function which is not strictly convex), the function ϕ has nonisolated equilibria and one wants to choose a particular equilibrium or minimum, or also one may desire a full and global description of the set of equilibria or minima of ϕ.

As motivated in [6], a fruitful direction is to consider coupled dynamical systems, exchanging information, and thus having more ability of globally exploring the

*Received by the editors February 15, 2003; accepted for publication (in revised form) July 30, 2003; published electronically February 18, 2004.
http://www.siam.org/journals/sicon/42-6/42320.html
†Institut de Mathématiques et de Modélisation de Montpellier, CNRS UMR 5149, case courrier 051, Université Montpellier 2, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France (marco@math.univ-montp2.fr).

function ϕ than a single noncoupled system. In order to obtain a global dynamical approach of the set of the equilibria of ϕ, [6] studies the asymptotic behavior of the regular (HBFC^2_{reg}) system

$$(\mathrm{HBFC}^2_{reg}) \qquad \begin{cases} \ddot{x} + \gamma\dot{x} + \nabla\phi(x) + \varepsilon(t)\nabla V_{reg}(x-y) = 0, \\ \ddot{y} + \gamma\dot{y} + \nabla\phi(y) - \varepsilon(t)\nabla V_{reg}(x-y) = 0, \end{cases}$$

where $\gamma > 0$ is a positive damping parameter, $\varepsilon : \mathbb{R}_+ \to \mathbb{R}_+ \setminus \{0\}$ is a control function such that $\lim_{t\to+\infty} \varepsilon(t) = 0$, and $V_{reg} : H \to \mathbb{R}_+$ is a (regular) coupling potential function.

We refer the reader to [6] for the detailed reasons that lead us to consider the regular (HBFC^2_{reg}) system and also for more references on the subject. Let us simply mention that to obtain more exploration properties than a first order in time dynamical system, one considers second order in time dynamical systems, among which the heavy ball with friction system

$$(\mathrm{HBF}) \qquad \ddot{x}(t) + \gamma\dot{x}(t) + \nabla\phi(x(t)) = 0$$

enjoys most of the nice properties of the steepest descent method (see the recent papers of Alvarez [1], Attouch, Goudou, and Redont [4], Goudou [7], Haraux and Jendoubi [8], and Jendoubi [9]). By adding a Tikhonov-like asymptotic control term $\varepsilon(t)x(t)$, Attouch and Czarnecki [3] show that the system

$$(\mathrm{HBFC}) \qquad \ddot{x}(t) + \gamma\dot{x}(t) + \nabla\phi(x(t)) + \varepsilon(t)x(t) = 0$$

will select a specific equilibrium when ϕ is convex and $\varepsilon : \mathbb{R}_+ \to \mathbb{R}_+$ is a $\mathcal{C}^1$ control function which tends to zero slowly, i.e., such that $\int_0^{+\infty} \varepsilon(t)dt = +\infty$.

It was then natural to examine if the selection property of the control in (HBFC) could be adapted to a repulsive coupling of two systems, a case which is considered in [6]. Precisely, when $H = \mathbb{R}$ and if ε is a "slow" control, i.e., $\int_0^{+\infty} \varepsilon(t)dt = +\infty$, then the trajectories x and y of (HBFC^2_{reg}) converge to extremal points of the set $S = \{\lambda \in \mathbb{R}, \nabla\phi(\lambda) = 0\}$ of the equilibria of ϕ or have the same limit. However, in this last case, one does not obtain more information than with the not coupled (HBF) system. This is due to the lack of strength of the (regular) repulsive potential at the origin, which allows the two trajectories to asymptotically collapse.

In order to forbid the solutions to collapse, we consider in this paper the case of a singular potential, defined on $H \setminus \{0\}$, which, moreover, tends to $+\infty$ at 0. Since we want the potential to always be "active," it is essential to impose $\varepsilon(t) > 0$ for every t. Note that the (HBFC^2_{sing}) system has a similar mechanical interpretation as the (HBF) system with an extra repulsion force—deriving from the potential V—between the two "balls." Also, because of this natural physical aspect of the coupled system, it is important to consider singular interaction potentials that are the ground of the gravitational and electromagnetic theories. For example, in the case where $V(z) = 1/|z|^2$, it corresponds to the electric potential between two particles having a varying load $\varepsilon(t)$ of same sign.

We obtain the same types of properties for the (HBFC^2_{sing}) system as for the (HBFC^2_{reg}) system, mainly showing that the trajectories x and y of (HBFC^2_{sing}) satisfy

$$\lim_{t\to+\infty} \nabla\phi(x(t)) = \lim_{t\to+\infty} \nabla\phi(y(t)) = 0,$$

$$\lim_{t\to+\infty} \dot{x}(t) = \lim_{t\to+\infty} \dot{y}(t) = 0.$$

But the proofs in the singular case cannot be derived from the regular case and are much more involved. In the one-dimensional case, we show that the (HBFC^2_{sing}) system enjoys nicer properties than the (HBFC^2_{reg}) system, precisely the awkward case in [6], where the trajectories may have the same limit disappears.

One may wonder whether the singular case can be seen as an asymptotic limit of regular cases. This question is of importance for numerical applications, since a singular potential is likely to be numerically approximated by a regular potential. A first look seems to indicate a positive answer (section 7.2).

Obviously one wants to obtain more results in higher, possibly infinite, dimension. As in the regular case, the techniques that we use for the one-dimensional convergence results are specific to the dimension one. These convergence results do not immediately apply in higher dimension (Remark 2.7), at least with the same precision. But a numerical study (section 7.1) seems to indicate that the (HBFC^2_{sing}) system still enjoys good global exploration properties in higher dimension and opens perspective of generalizations to higher dimension.

The paper is organized as follows. In section 2.1, we precisely state the global existence results and general asymptotic properties (Proposition 2.1 and Theorem 2.2). In section 2.2, we precisely state the asymptotic convergence results in dimension one (Theorems 2.4 and 2.6, Corollaries 2.7 and 2.8). In section 3, we provide remarks on and counterexamples of our results. The results are proved in section 4 (global existence), section 5 (asymptotic properties), and section 6 (asymptotic convergence in dimension one). Finally, in section 7, we give numerical experiments in higher dimension and address the question of the singular case as a limit of regular cases.

2. Main results. We assume the following (rather standard) set of assumptions.

Hypothesis 1. Let H be a real Hilbert space and $\phi : H \to \mathbb{R}$ be a map of class $\mathcal{C}^1$ such that

$$(\mathcal{H}_\phi) \quad \begin{cases} \text{(i)} & \text{the map } \phi \text{ is bounded from below on } H; \\ \text{(ii)} & \text{the map } \nabla\phi \text{ is Lipschitz continuous on the bounded subsets of } H. \end{cases}$$

Let $V : H \setminus \{0\} \to \mathbb{R}_+$ be a map of class $\mathcal{C}^1$ such that

$$(\mathcal{H}_V) \quad \begin{cases} \text{(i)} & \text{the map } \nabla V \text{ is locally Lipschitz continuous on } H \setminus \{0\}; \\ \text{(ii)} & \lim_{z\to 0} V(z) = +\infty \quad \textit{(singularity assumption).} \end{cases}$$

Let $\varepsilon : [0, +\infty) \to \mathbb{R}_+ \setminus \{0\}$ be a function of class $\mathcal{C}^1$ such that

$$(\mathcal{H}_\varepsilon) \quad \begin{cases} \text{(i)} & \text{the function } \varepsilon \text{ is nonincreasing, i.e., } \dot\varepsilon(t) \le 0 \quad \forall t \in \mathbb{R}_+; \\ \text{(ii)} & \lim_{t\to+\infty} \varepsilon(t) = 0. \end{cases}$$

Let $\gamma > 0$, set $\Gamma = \{(x, y) \in H^2, \quad x - y \neq 0\}$, and $\Big((x_0, y_0), (\dot x_0, \dot y_0)\Big) \in \Gamma \times H^2$, the ($HBFC^2_{sing}$) system is defined as follows:

$$(\mathrm{HBFC}^2_{sing}) \quad \begin{cases} \ddot x + \gamma\dot x + \nabla\phi(x) + \varepsilon(t)\nabla V(x - y) = 0, \\ \ddot y + \gamma\dot y + \nabla\phi(y) - \varepsilon(t)\nabla V(x - y) = 0 \\ \Big(x(0), y(0), \dot x(0), \dot y(0)\Big) = \Big(x_0, y_0, \dot x_0, \dot y_0\Big). \end{cases}$$

Remark 2.1. For the sake of readability, we take 0 as initial time. All the results of the paper clearly hold by taking any other initial time $t_0 \in \mathbb{R}$ and making the corresponding adaptations in the statements.

2.1. Global properties. The following proposition ensures the global existence of the solutions of the (HBFC^2_{sing}) system.

PROPOSITION 2.1 (global existence). *Assume Hypothesis* 1. *Then*

(i) *there exists a unique maximal solution* $(x, y) : [0, +\infty) \to \Gamma$ *of* (HBFC^2_{sing}) *which is of class* $\mathcal{C}^2$*;*

(ii) $(\dot{x}, \dot{y}) \in L^\infty([0, +\infty); H \times H) \cap L^2([0, +\infty); H \times H)$, $(\phi(x), \phi(y)) \in L^\infty([0, +\infty); \mathbb{R} \times \mathbb{R})$, *and* $\varepsilon V(x - y) \in L^\infty([0, +\infty); \mathbb{R})$*;*

(iii) *the energy function* E, *defined by* $E(t) = \frac{1}{2}|\dot{x}(t)|^2 + \frac{1}{2}|\dot{y}(t)|^2 + \phi(x(t)) + \phi(y(t)) + \varepsilon(t)V(x(t) - y(t))$, *is nonincreasing.*

Proposition 2.1 is proved in section 4 and commented on in section 3.1.

The next result shows first global (and essential for the following results) convergence properties of the solutions of the (HBFC^2_{sing}) system. The conclusions hold as in the regular case [6, Theorem 2.1], with a stronger assumption on the potential V for the convergence of $\dot{x}$ and $\dot{y}$, but the proofs heavily differ because of the singularity assumption. Most of the difficulties specific to the singular case are concentrated in the proof of Theorem 2.2 given in section 5.

We shall consider two additional assumptions in the statement of Theorem 2.2. The first assumption is a limit condition on the potential ϕ, which covers a wide set of cases.

Hypothesis 2 (limit condition (LIM)). For every sequence $(z_n) \subset H$ such that $\lim_{n\to+\infty} |z_n| = +\infty$, there exists a subsequence $(z_{\varphi(n)})$ such that $\lim_{n\to+\infty} \phi(z_{\varphi(n)}) = +\infty$ or $\lim_{n\to+\infty} \nabla\phi(z_{\varphi(n)}) = 0$.

The limit condition (LIM) is discussed in detail in [6, section 3.3]. Let us just mention that it is equivalent to the following assertion:

$$\forall \alpha > 0, \quad \forall A \in \mathbb{R}, \text{ the set } \{z \in H, \phi(z) \leq A \text{ and } |\nabla\phi(z)| \geq \alpha\} \text{ is bounded,}$$

and it is clearly satisfied in the two following more simple cases: (c) the map ϕ is coercive, i.e., $\lim_{|z|\to+\infty} \phi(z) = +\infty$, and (d) $\lim_{|z|\to+\infty} \nabla\phi(z) = 0$.

The second assumption is a double assumption on the potential V, repulsion and radial symmetry (with an additional error term).

Hypothesis 3 (radial symmetry and repulsion ($\mathcal{H}_V$)(iii)). There exist a decreasing function $W : \mathbb{R}_+ \setminus \{0\} \to \mathbb{R}_+$ of class $\mathcal{C}^1$ and a bounded "error" function $e : H \setminus \{0\} \to \mathbb{R}_+$ of class $\mathcal{C}^1$ such that, for every $z \in H \setminus \{0\}$,

$$V(z) = W(|z|) + e(z)$$

and such that, on a neighborhood of 0, ∇e is bounded and $\langle \nabla e(z), z \rangle \leq 0$.

Note that the typical electric potential $V(z) = 1/|z|^2$ satisfies Hypothesis 3. The error term is important, for instance to allow computations errors in numerical applications.

THEOREM 2.2 (asymptotic properties). *Part* 1. *Assume Hypothesis* 1*, and that, for every* $r > 0$ *the map* ∇V *is bounded on the bounded subsets of* $H \setminus B(0, r)$*. Assume, moreover, Hypothesis* 2*, i.e., the map* ϕ *satisfies the limit condition* (LIM)*, or that the trajectory* (x, y) *is bounded. Then*

$$\lim_{t\to+\infty} \nabla\phi(x(t)) = \lim_{t\to+\infty} \nabla\phi(y(t)) = 0.$$

Hence if x_∞*(resp.,* y_∞*) is a cluster point of* x *(resp.,* y*), then,* x_∞ *(resp.,* y_∞*)* $\in S$*, with* $S = \{z \in H, \nabla\phi(z) = 0\}$.

Part 2. *Additionally assume Hypothesis* 3; *i.e., the potential* V *is a radial repulsion. Then*

$$\lim_{t\to+\infty} \dot{x}(t) = \lim_{t\to+\infty} \dot{y}(t) = 0.$$

Theorem 2.2 is proved in section 5 and commented on in section 3.2.

The following result specifies the behavior in the convex case.

COROLLARY 2.3. *Under the assumptions of Theorem* 2.2, *additionally assume that the map* ϕ *is convex. Then*

(v) *if the trajectory* x *(resp.,* y*) is bounded, then* $\lim_{t\to+\infty} \phi(x(t)) = \inf \phi$ *(resp.,* $\lim_{t\to+\infty} \phi(y(t)) = \inf \phi$*);*

(vi) *if an element* x_∞*(resp.,* y_∞*) is a weak cluster point of* x*(resp.,* y*), then* x_∞*(resp.,* y_∞*)* $\in S$.

We refer the reader to [6] for the classical proof of Corollary 2.3.

2.2. Convergence in the one-dimensional case. In this section, we explore the convergence and minimizing properties of the solutions of the (HBFC$^2_{sing}$) system, under the additional assumption that the space H is one-dimensional (i.e., $H = \mathbb{R}$). The singularity assumption implies in particular that the trajectories x and y may never cross. We show that this property "passes to the limit" with a slow control and thus improves the (slow control) results of [6], when the potential V is a repulsion, i.e., satisfies

$$(\mathcal{H}_V)(\text{iv}) \qquad \forall z \in \mathbb{R} \setminus \{0\}, \quad zV'(z) \le 0.$$

Precisely, the awkward case in [6], where the two trajectories x and y have the same limit, does not appear in the forthcoming slow control results (section 2.2.2). Finally, since $H = \mathbb{R}$, note that the limit condition (LIM) is satisfied if ϕ is convex and bounded from below.

2.2.1. Convergence of the trajectory. The next result shows the convergence of the solutions of the (HBFC$^2_{sing}$) system. Recall that S denotes the set of the equilibria of ϕ: $S = \{z \in \mathbb{R}, \phi'(z) = 0\}$. In order to give a unified presentation of our results, we let

$$\widehat{S} = \{\overline{z} \in \overline{\mathbb{R}}, \lim_{z\to\overline{z}} \phi'(z) = 0\}.$$

THEOREM 2.4 (convergence of the solutions). *Assume Hypothesis* 1, *with* $H = \mathbb{R}$.

Part 1. *Assume that the solution* (x, y) *of the* (HBFC$^2_{sing}$) *system is bounded. Then it converges:*

(i) *There exists* $(x_\infty, y_\infty) \in \widehat{S} \times \widehat{S}$ *such that* $\lim_{t\to+\infty} \Big(x(t), y(t)\Big) = \Big(x_\infty, y_\infty\Big)$.

If $y_0 < x_0$ *(resp.,* $x_0 < y_0$*), then* $y_\infty \le x_\infty$ *(resp.,* $x_\infty \le y_\infty$*). Moreover, if* $x_\infty \ne y_\infty$, *then*

$$\text{(ii)} \quad \lim_{t\to+\infty} \dot{x}(t) = \lim_{t\to+\infty} \dot{y}(t) = 0.$$

Part 2. *Assume Hypothesis* 2, *i.e., the map* ϕ *satisfies the limit condition* (LIM) *(for example, if* ϕ *is convex), or that the trajectory* (x, y) *is bounded,*[1] *together with*

$$(\mathcal{H}_V)(\text{iv}) \qquad \forall z \in \mathbb{R} \setminus \{0\}, \quad zV'(z) \le 0.$$

Then the conclusions (i) *and* (ii) *of Part* 1 *hold.*

[1]This last case, mentioned again only for the convergence of $\dot{x}$ and $\dot{y}$ when $x_\infty = y_\infty$, could be deduced from the previous one—the (LIM) assumption—by changing the map ϕ.

Theorem 2.4, Part 2 is similar to its regular version [6, Theorem 2.2]. Part 1 can be also stated in the regular case and thus improves [6, Theorem 2.2]. This is precisely done in section 7.3. Theorem 2.4 is commented on in section 3.3 and proved in sections 6.1 and 6.2. As in the regular case, when the map ϕ is convex, the trajectory minimizes ϕ. The following corollary states this precisely.

COROLLARY 2.5. *Under the assumptions of Theorem* 2.4*, Part* 2*, additionally assume that the map ϕ is convex. Then* $\lim_{t\to+\infty}(\phi(x(t))\,,\,\phi(y(t))) = (\inf\phi\,,\,\inf\phi)$.

Remark 2.2. Corollary 2.5 may not hold if the map ϕ is not assumed to be convex, even if it is quasi-convex. Adapt the counterexample in [6].

Proof of Corollary 2.5. Consider the different cases: (i) $x_\infty \in \mathbb{R}$ (hence x is bounded), (ii) $x_\infty = -\infty$ (hence $\phi' \geq 0$), (iii) $x_\infty = +\infty$ (hence $\phi' \leq 0$). □

2.2.2. Slow control. By contrast with a "fast" control ε, i.e., such that $\int_0^{+\infty}\varepsilon(t)\,dt < +\infty$, which merely acts as a perturbation (see [3] and [6]), a "slow" control ε, i.e., such that $\int_0^{+\infty}\varepsilon(t)\,dt = +\infty$, has a more actual effect. The results in this section show the convergence of the solution map (x, y) toward specific points of $\widehat{S}$ with a "slow" control ε. They are the exact singular counterpoint of the regular versions given in [6, section 2.2.2]. The difference between the singular and regular results is more obvious for Corollary 2.8 and [6, Corollary 2.4] and their consequences (Corollary 2.9 and [6, Corollary 2.5]). An awkward case appears in the regular version, namely [6, Corollary 2.4(iii)] $\lim_{t\to+\infty} x(t) = \lim_{t\to+\infty} y(t)$, and it is the existence of this case that originally motivated the study of the singular case. The other convergence results are more general, but the differences with their regular version are less spectacular. For every $\lambda \in \overline{\mathbb{R}}$, we denote by $P^+(\lambda)$ (resp., $P^-(\lambda)$) the following proposition:

$$\text{For every neighborhood } V(\lambda) \text{ of } \lambda \quad \exists \mu \in V(\lambda)\cap\mathbb{R}, \quad \phi'(\mu) > 0 \quad (\text{resp., } \phi'(\mu) < 0).$$

THEOREM 2.6 (slow parametrization).[2] *Under the assumptions of Theorem* 2.4*, Part* 2*, additionally assume that*

$$(\mathcal{H}_\varepsilon)(\text{iii}) \qquad \int_0^{+\infty}\varepsilon(t)\,dt = +\infty,$$

$$(\mathcal{H}_V)(\text{v}) \qquad \forall M > 0, \quad \inf_{0<|z|\leq M}|V'(z)| > 0.$$

Let $(x_\infty, y_\infty) \in \widehat{S}\times\widehat{S}$ *be the limits of the trajectories. Assume* $y_0 < x_0$ *Then the solution* (x, y) *of the* (HBFC^2_{sing}) *system satisfies the following properties:*

- *if* $y_\infty > -\infty$ *and* $x_\infty < +\infty$, *then we have* $P^-(y_\infty)$ *and* $P^+(x_\infty)$.
- *if* $y_\infty = -\infty$ *and* $x_\infty < +\infty$, *then we have* $P^+(y_\infty)$ *or* $P^+(x_\infty)$.
- *if* $y_\infty > -\infty$ *and* $x_\infty = +\infty$, *then we have* $P^-(y_\infty)$ *or* $P^-(x_\infty)$.

If $x_0 < y_0$*, then the corresponding assertions hold (by exchanging the letters* x *and* y*).*

Theorem 2.6 is proved in section 6.3.

Remark 2.3. Assumption $(\mathcal{H}_V)$(v) means that the potential V always remains active near the origin, never being locally constant. This assumption is clearly satisfied if the potential V is strictly convex on $\mathbb{R}_+$ and on $\mathbb{R}_-$.

[2] Theorem 2.6 applies for every initial condition, contrary to its regular version [6, Theorem 2.3], which assumes that the trajectories do not have the same limit. The same holds for Corollary 2.7.

In fact, Theorem 2.6 specifies the points of $\widehat{S}$, where the trajectories x and y converge, as precisely stated in the next result.

COROLLARY 2.7 (slow parametrization). *Under the assumptions of Theorem* 2.6*, let $I(\lambda)$ be the connected component of λ in $\widehat{S}$. If $y_0 < x_0$ and if one of the following conditions is satisfied:*

$$\begin{array}{ll} \text{(a)} & -\infty < y_\infty, \\ \text{(b)} & y_\infty = -\infty \text{ and } \exists \lambda \in \mathbb{R},\ \phi'((-\infty, \lambda]) \leq 0, \end{array}$$

then the trajectory x converges to an extremal point of $I(x_\infty)$, precisely

$$x_\infty \in \{\inf I(x_\infty), \sup I(x_\infty)\}.$$

If $x_0 < y_0$, then the corresponding assertions hold (by exchanging the letters x and y).

Corollary 2.7 is proved in section 6.4.

Remark 2.4. The conclusion of Corollary 2.7 may not be more precise in general. Precisely, one cannot replace the conclusion $x_\infty \in \{\inf I(x_\infty), \sup I(x_\infty)\}$ by $x_\infty = \sup I(x_\infty)$. Adapt the counterexample in [6].

Remark 2.5. The conclusion of Corollary 2.7 does not hold if $y_\infty = -\infty$ and there exists no $\lambda \in \mathbb{R}$ such that $\phi'((-\infty, \lambda]) \leq 0$. Adapt the counterexample in [6].

When the set of equilibria is an interval, the behavior of the trajectories is more precise.

COROLLARY 2.8 (slow parametrization with a connected set of equilibria). *Under the assumptions of Theorem* 2.6*, additionally assume that the set $\widehat{S}$ is an interval (in $\overline{\mathbb{R}}$) (for example, if ϕ is convex). Then the solution (x,y) of the* (HBFC$^2_{sing}$) *system satisfies one of the following cases:*

$$\begin{array}{lll} \text{(i)} & \lim_{t\to+\infty} \Big(x(t)\,,\, y(t) \Big) = \Big(\sup \widehat{S}\,,\, \inf \widehat{S} \Big) & \text{if } y_0 < x_0; \\ \text{(ii)} & \lim_{t\to+\infty} \Big(x(t)\,,\, y(t) \Big) = \Big(\inf \widehat{S}\,,\, \sup \widehat{S} \Big) & \text{if } x_0 < y_0. \end{array}$$

Corollary 2.8 is proved in section 6.5. As a (clear) consequence of Corollary 2.8, the limit of the difference of the trajectories maximizes the (possibly) infinite diameter of the set S.

COROLLARY 2.9. *Under the assumptions of Theorem* 2.6*, additionally assume that S is a nonempty interval and that $\widehat{S} = \mathrm{cl}_{\overline{\mathbb{R}}}(S)$. Then the solution (x, y) of the* (HBFC$^2_{sing}$) *system satisfies*

$$\lim_{t\to+\infty} |x(t) - y(t)| = \mathrm{diam}\,(S).$$

The proof of Corollary 2.9 is immediate.

Remark 2.6. Corollary 2.9 may not hold if we do not assume $\widehat{S} = \mathrm{cl}_{\overline{\mathbb{R}}}(S)$, even if S is a nonempty interval. Adapt the counterexample in [6].

Remark 2.7. The formulation of Theorem 2.6 is specific to the dimension one. This is not the case of Corollary 2.9, but Corollary 2.9 may not remain true in higher dimension. Adapt the counterexample in [6].

3. Remarks and counterexamples.

3.1. On Proposition 2.1.

Remark 3.1. Proposition 2.1 easily generalizes by replacing the domain $H \setminus \{0\}$ by any open set $\Omega \subset H$, replacing $(\mathcal{H}_V)$(ii) by the following: for all $\bar{z} \in$

$\mathrm{bd}\,\Omega, \lim_{z\in\Omega,\, z\to\bar{z}} V(z) = +\infty$, and letting $\Gamma = \{(x,y) \in H^2, \quad x - y \in \Omega\}$. We let the reader check that the proof in section 4 still holds. This may be of interest, on one side, if one wants to avoid points other than 0 and, on the other side, since it then covers and thus generalizes the regular case (by taking $\Omega = H$).

Remark 3.2. In Proposition 2.1, we consider only the case where $\varepsilon(t) > 0$ for every t, i.e., the case where the control is effective. The case where $\varepsilon(T) = 0$ for some T can be easily proved. But, first, since the function ε is assumed to be nonincreasing, $\varepsilon(t) = 0$ for every $t \geq T$, and there is no control on the system for $t \geq T$. Second, one cannot ensure that the difference of the trajectories $x - y$ remains in the domain $H \setminus \{0\}$. For our purpose of controlling the system, the case where $\varepsilon(t) > 0$ for every t is the only interesting one.

3.2. On Theorem 2.2.

Remark 3.3. Theorem 2.2, Part 1 may not hold if the trajectory (x, y) is not bounded and if the map ϕ does not satisfy the limit condition (LIM). Adapt the counterexample in [6].

Remark 3.4. If the trajectory (x, y) is bounded, then $\lim_{t\to+\infty} \nabla\phi(x(t)) + \nabla\phi(y(t)) = 0$ without assuming that ∇V is bounded on the bounded subsets of $H \setminus B(0, r)$ (see Lemma 5.2 below).

Remark 3.5. Theorem 2.2, Part 1 may not hold if one replaces the set $H \setminus \{0\}$ by any open set $\Omega \subset H$ and makes the corresponding adjustment for $(\mathcal{H}_V)$(ii). Let $H = \mathbb{R}$, $\Omega = \mathbb{R} \setminus [-2, 2]$, $\phi(x) = \frac{x^2}{2}$, and $V(z) = \frac{1}{|z-2|}$ for $z \notin [-2, 2]$. Then $x(t) = 1 + \frac{1}{\sqrt{t}}$, $y(t) = -x(t)$, are solutions to the corresponding (HBFC^2_{sing}) system with $\varepsilon(t) = 4t^{-1} - 2t^{-5/2} + 3t^{-7/2}$ and $\nabla\phi(x(t)) \to 1$ when $t \to +\infty$. In some sense, defining the singular potential V on more general open subsets Ω may create too many constraints since the trajectory (x, y) must stay in the corresponding set Γ.

Remark 3.6. Theorem 2.2, Part 2 holds under the following more general assumptions on the potential V:

(i) $\lim_{n\to+\infty} V(z_n) = +\infty \Rightarrow \lim_{n\to+\infty} z_n = 0$ for every sequence (z_n) in H. In view of the singularity assumption $(\mathcal{H}_V)$(ii), the previous implication would be an equivalence.

(ii) The potential V is repulsive near 0, precisely $\langle \nabla V(z), z\rangle \leq 0$ for every $z \in B(0, r) \setminus \{0\}$ for some $r > 0$.

(iii) For some function $\lambda : H \to (-\infty, 0)$, the function $z \mapsto \nabla V(z) - \lambda(z)z$ is bounded.

These assumptions are clearly technical and are mentioned for the record.

Remark 3.7. Under the assumptions of Theorem 2.2, it is easy to find examples of nonconvergent (nonbounded) trajectories of the (HBFC^2_{sing}) system. In the regular case [6, section 3.8], one immediately deduces examples of bounded and nonconvergent trajectories of the (HBFC^2_{reg}) system by taking $\varepsilon(t) = 0$ for every t, in which case (HBFC^2_{reg}) becomes a noncontrolled (HBF) system, with examples of bounded and nonconvergent trajectories of (HBF); see Redont [11] (also in [4]) and Jendoubi and Poláčik [10]. This cannot be done in the singular case, since we impose the condition $\varepsilon(t) > 0$ for every t. But our concern being the exploration of the equilibria of ϕ by the (HBFC^2_{sing}) system, the extensive study of the converging properties of (HBFC^2_{sing}), of high interest for itself, is beyond the scope of this paper.

3.3. On Theorem 2.4.

Remark 3.8. Theorem 2.4, Part 1, (i) may not be true if the trajectory (x, y) is not assumed to be bounded. It is possible to build an example such that $\lim_{t\to+\infty} x(t) =$

$+\infty$, $\liminf_{t\to+\infty} y(t) = -1$, $\limsup_{t\to+\infty} y(t) = 1$. Take

$$x(t) = 3\log(t) - \cos(\log(t)), \qquad y(t) = \cos(\log(t)),$$

and $\varepsilon(t) = 1/\sqrt{t}$. Then the maps $t \mapsto x(t)$ and $t \mapsto x(t) - y(t)$ are strictly increasing, hence invertible, and their reciprocal functions are locally Lipschitz continuous. Define $\phi : \mathbb{R} \to \mathbb{R}$ of class $\mathcal{C}^1$, with a Lipschitz continuous derivate, by $\phi|_{[-1,1]} = 0$ and $\ddot{x}(t) + \ddot{y}(t) + \dot{x}(t) + \dot{y}(t) + \phi'(x(t)) = 0$ (for $t \geq e$), i.e., $\phi'(3\log(t) - \cos(\log(t))) = -3/t + 3/t^2$. Then $\phi'(z) \sim -3\alpha(z)e^{-z/3}$ with $\alpha(z) \in [e^{-1/3}, e^{1/3}]$ when $z \to +\infty$; hence ϕ is bounded from below. Define $V : (0, +\infty) \to \mathbb{R}$ of class $\mathcal{C}^1$, with a Lipschitz continuous derivate, by $\ddot{y}(t) + \dot{y}(t) - \varepsilon(t)V'(x(t) - y(t)) = 0$ (for $t \geq e$), i.e., $V'(3\log(t) - 2\cos(\log(t))) = -(1/\sqrt{t})\sin(\log(t)) + (1/t\sqrt{t})(\sin(\log(t)) - \cos(\log(t)))$. Then $|V'(z)| \leq e^{-z/6+1/3} + o(e^{-z/6})$; hence V can be chosen positive. The map (x, y) is by construction a solution of the corresponding (HBFC^2_{sing}) system with $\gamma = 1$, and the assumptions of Theorem 2.4, Part 1 are satisfied except for the bound on the trajectory x.

Remark 3.9. Theorem 2.4, Part 1, (ii) may not be true if $x_\infty = y_\infty$. In $\mathbb{R}$, take $\phi(x) = 0$, and consider a function $f : \mathbb{R} \to \mathbb{R}_+$ of class $\mathcal{C}^1$ such that $\operatorname{supp} f = [-1, 1]$, $\int_{\mathbb{R}} f = 1$, $f'|_{[-1,0]} \geq 0$, $f'|_{[0,1]} \leq 0$, and $f(0) = 1$. Define x by $\dot{x}(t) = -e^{-t} - \sum_{n\in\mathbb{N}} \sqrt{1 + 1/n}\, f(2^n(t - n))$ and $\lim_{n\to+\infty} x(t) = 0$, i.e.,

$$x(t) = e^{-t} + \sum_{n\in\mathbb{N}} \sqrt{1 + \frac{1}{n}} \int_{u\geq t} f\left(2^n(u - n)\right) du.$$

Then $\dot{x}(t) < 0$; hence $t \mapsto x(t)$ is monotone, it strictly decreases to 0, and the reciprocal map $z \mapsto t(z)$ is defined on $(0, x(0)]$ and is locally Lipschitz continuous (since $\dot{x}$ is bounded from above by some negative real number on every compact set).[3] Assume that for every n, there is a unique solution of $2^n f'(2^n(t - n)) + f(2^n(t - n)) = 0$ on the interval $(n - 1/2^n, n - 1/2^n)$, which we denote $n + \alpha_n$, such that $\lim_{n\to+\infty} n^2 2^n \alpha_n = 0$. This is the case if we take $f(x) = (1 - x^2)^2$, and in this case $\alpha_n = 1/2^{2n+1} + o(1/2^{2n})$. Let us choose a nonincreasing control $\varepsilon : [0, +\infty) \to (0, +\infty)$ of class $\mathcal{C}^\infty$ such that

$$\varepsilon(t) = \frac{1}{n} \text{ for every } t \in [n + \alpha_n + \beta_n,\ n + 1 + \alpha_{n+1} - \beta_{n+1}]$$

for some decreasing sequence $(\beta_n)_{n\in\mathbb{N}}$ of small enough positive real numbers. Define $V : (0, +\infty) \to \mathbb{R}$ by

$$V'(2x(t)) = -(\ddot{x}(t) + \dot{x}(t))/\varepsilon(t).$$

Then V' is locally Lipschitz continuous. Let us verify that $\lim_{z\to 0} V(z) = +\infty$. We have $dV(2x(t))/dt = 2V'(2x(t))\dot{x}(t) = -2\dot{x}(t)(\ddot{x}(t) + \dot{x}(t))/\varepsilon(t)$. The map $\dot{x}$ is negative, and for every n and every $t \in [n - 1 + 1/2^{n-1}, n + 1 - 1/2^{n+1}]$,

$$\ddot{x}(t) + \dot{x}(t) = -\sqrt{1 + \frac{1}{n}} \left(2^n f'\left(2^n\left(t - n\right)\right) + f\left(2^n\left(t - n\right)\right)\right).$$

[3]This immediate statement is of importance in order to obtain a locally Lipschitz continuous map V'. Without this constraint, one could forget the term e^{-t} in the definition of x and $\dot{x}$ and obtain a $\mathcal{C}^1$ coupling potential V, but nonunique solutions appear.

Then $\ddot{x}(t)+\dot{x}(t)=0$ if $t\in[n+1/2^n, n+1-1/2^{n+1}]$, $\ddot{x}(t)+\dot{x}(t)>0$ if $t\in(n+\alpha_n, n+1/2^n)$, and $\ddot{x}(t)+\dot{x}(t)<0$ if $t\in(n+1-1/2^{n+1}, n+1+\alpha_{n+1})$. Hence $t\mapsto V(2x(t))$ increases on the interval $[n+\alpha_n, n+1/2^n]$, is constant on $[n+1/2^n, n+1-1/2^{n+1}]$, and decreases on $[n+1-1/2^{n+1}, n+1+\alpha_{n+1}]$. It is then sufficient to estimate $V(2x(n+1+\alpha_{n+1}))-V(2x(n+\alpha_n))$.

$$\begin{aligned}
V(2x(n+1+\alpha_{n+1}))-V(2x(n+\alpha_n)) &= \int_{n+\alpha_n}^{n+1+\alpha_{n+1}} -2\frac{\dot{x}(t)\ddot{x}(t)+\dot{x}(t)^2}{\varepsilon(t)}dt\\
&= \int_{n+\alpha_n}^{n+1+\alpha_{n+1}} -2\frac{\dot{x}(t)\ddot{x}(t)+\dot{x}(t)^2}{1/n}dt + o\left(\frac{1}{n}\right)^4\\
&= -2n\int_{n+\alpha_n}^{n+1/2^n} \dot{x}(t)\ddot{x}(t)+\dot{x}(t)^2dt - 2n\int_{n+1-1/2^{n+1}}^{n+1+\alpha_{n+1}} \dot{x}(t)\ddot{x}(t)\\
&\quad +\dot{x}(t)^2dt + o\left(\frac{1}{n}\right)\\
&= n\left(\dot{x}(n+\alpha_n)^2-\dot{x}(n+1+\alpha_{n+1})^2\right)+o\left(\frac{1}{n}\right)^5\\
&= n\left(1+\frac{1}{n}-1-\frac{1}{n+1}\right)+o\left(\frac{1}{n}\right)^6\\
&= \frac{1}{n+1}+o\left(\frac{1}{n}\right).
\end{aligned}$$

Hence $\lim_{z\to 0}V(z)=+\infty$, and by choosing $V(2x(0))$ large enough, V is positive. Defining V symmetrically on $\mathbb{R}_-\setminus\{0\}$, all the assumptions of Theorem 2.4, Part 1 are satisfied. The map $(x,-x)$ is by construction a solution of the corresponding (HBFC^2_{sing}) system with $\gamma=1$, $x_\infty=y_\infty=0$, and, clearly, the map $\dot{x}$ does not converge to 0.

Remark 3.10. The radial symmetry and repulsion assumption $(\mathcal{H}_V)$(iii), with $e=0$, implies assumption $(\mathcal{H}_V)$(iv), but the converse is not true. However, since $H=\mathbb{R}$, and since $x-y$ keeps a constant sign, assumption $(\mathcal{H}_V)$(iv) is sufficient to obtain Theorem 2.2, Part 2, i.e., $\lim_{t\to+\infty}\dot{x}(t)=\lim_{t\to+\infty}\dot{y}(t)=0$.

4. Global existence: Proof of Proposition 2.1. The proof, which we briefly recall, is quite standard and differs from the regular case only when showing that the trajectories are defined on the whole set $[0,+\infty)$. We let

$$\begin{aligned}
&X=(x,y)\in H^2, \quad \Phi(X)=\phi(x)+\phi(y), \quad U(X)=V(x-y),\\
&X_0=(x_0,y_0), \quad \dot{X}_0=(\dot{x}_0,\dot{y}_0).
\end{aligned}$$

The system (HBFC^2_{sing}) then reduces to

$$(\mathrm{HBFC}) \quad \ddot{X}+\gamma\dot{X}+\nabla\Phi(X)+\varepsilon(t)\nabla U(X)=0, \quad X(0)=X_0, \quad \dot{X}(0)=\dot{X}_0.$$

Proof of (i). The second order system (HBFC) can be written as a first order system in $\Gamma\times H^2$, $\dot{Y}=F(t,Y)$ with $Y(t)=(X(t),\dot{X}(t))$, and $F(t,u,v)=(v,-\gamma v-$

[4]The equality holds true by choosing β_n and β_{n+1} small enough.

[5]The equality holds true since $\dot{x}(n+1/2^n)=-e^{n+1/2^n}=o(1/n^2)$ and $\dot{x}(n+1+1/2^{n+1})=o(1/n^2)$.

[6]The function f attains its maximum at 0, $f'(0)=0$, and $\dot{x}(n+\alpha_n)=-e^{n+\alpha_n}-\sqrt{1+\frac{1}{n}}f(2^n\alpha_n)=o(1/n^2)-\sqrt{1+\frac{1}{n}}(1+o(2^n\alpha_n))=-\sqrt{1+\frac{1}{n}}+o(1/n^2)$.

$\nabla\Phi(u) - \varepsilon(t)\nabla U(u))$. For $Y_0 = (X_0, \dot{X}_0)$ given in $\Gamma \times H^2$, the Cauchy–Lipschitz theorem and Hypothesis 1 ensure the existence of a unique local solution to the problem $\dot{Y} = F(t, Y)$, $Y(0) = Y_0$. Let X denote the maximal solution defined on the interval $[0, T_{max})$ with $0 < T_{max} \leq +\infty$. Observe that equation (HBFC) and the regularity assumptions on $\phi, V,$ and ε automatically imply that the map X is $\mathcal{C}^2$ on the interval $[0, T_{max})$. We first show that the map $\dot{X}$ is bounded.

We can define along every trajectory of (HBFC) the energy function by

$$E(t) = \frac{1}{2}|\dot{X}(t)|^2 + \Phi(X(t)) + \varepsilon(t)U(X(t)).$$

By differentiation of $E(t)$, and in view of (HBFC), we obtain for every $t \in [0, T_{max})$

$$\begin{aligned} \dot{E}(t) &= \left\langle \dot{X}(t), \ddot{X}(t) + \nabla\Phi(X(t)) + \varepsilon(t)\nabla U(X(t)) \right\rangle + \dot{\varepsilon}(t)U(X(t)) \\ &= -\gamma|\dot{X}(t)|^2 + \dot{\varepsilon}(t)U(X(t)). \end{aligned} \tag{4.1}$$

Since $\dot{\varepsilon}(t) \leq 0$ (assumption $(\mathcal{H}_\varepsilon)$(i)) and $U \geq 0$, we have $\dot{E}(t) \leq 0$. Hence the function E is nonincreasing, which proves (iii), and for all $t \in [0, T_{max})$ $E(t) \leq E(0)$. Equivalently,

$$\frac{1}{2}|\dot{X}(t)|^2 + \Phi(X(t)) + \varepsilon(t)U(X(t)) \leq E(0). \tag{4.2}$$

Since Φ is bounded from below, $U(X(t)) \geq 0$, and $\varepsilon(t) \geq 0$, we obtain that

$$\sup_{t\in[0,T_{max})} |\dot{X}(t)| < +\infty \quad \text{and} \quad \sup_{t\in[0,T_{max})} \varepsilon(t)U(X(t)) < +\infty.$$

We now prove $T_{max} = +\infty$, noting that assumption $(\mathcal{H}_V)$(ii) is crucial in the singular case. Indeed, assume that $T_{max} < +\infty$. We have $|X(t) - X(t')| \leq ||\dot{X}||_\infty |t - t'|$, and since $T_{max} < +\infty$, then $\lim_{t\to T_{max}} X(t) := X_\infty$ exists and belongs to $\mathrm{cl}\,\Gamma$. Hence, X and $\dot{X}$ are bounded on $[0, T_{max})$. From assumption $(\mathcal{H}_\phi)$(ii)

$$\nabla\Phi(X) \quad \text{is bounded on } [0, T_{max}). \tag{4.3}$$

From (4.2), we obtain for every $t \in [0, T_{max})$, $\varepsilon(t)U(X(t)) \leq E(0) - \inf\Phi$. Hence, since the function ε is nonincreasing

$$\forall t \in [0, T_{max}), \qquad U(X(t)) \leq \frac{E(0) - \inf\Phi}{\varepsilon(T_{max})}.$$

Assume $X_\infty \notin \Gamma$. Then $X_\infty \in \mathrm{bd}\,\Gamma$ and assumption $(\mathcal{H}_V)$(ii) implies

$$\lim_{t\to T_{max}} U(X(t)) = +\infty,$$

which contradicts the previous inequality. Hence $X_\infty \in \Gamma$. Then the set $\{\nabla U(X(t)), t \in [0, T_{max})\} \cup \{\nabla U(X_\infty)\}$ is compact and, in particular,

$$\nabla U(X) \quad \text{is bounded on } [0, T_{max}). \tag{4.4}$$

From (HBFC), (4.3), and (4.4) the map $\ddot{X}$ is also bounded on the interval $[0, T_{max})$. Hence $\lim_{t\to T_{max}} \dot{X}(t) = \dot{X}_\infty$ exists. Applying again the local existence theorem with

initial data $(X_\infty, \dot{X}_\infty) \in \Gamma \times H^2$, we extend the maximal solution to a strictly larger interval, a contradiction. Hence $T_{max} = +\infty$, which completes the proof of (i). ∎

Proof of (ii). From (4.1), since $\dot{\varepsilon} \leq 0$ and $U \geq 0$, we derive, for every t,

$$\int_0^t |\dot{X}(s)|^2 ds \leq \frac{1}{\gamma}(E(0) - E(t)).$$

Since the function E is nonincreasing, for every t, $E(0) \geq E(t) \geq \Phi(X(t))$; hence the map Φ is bounded from above, hence bounded, and the function E is bounded from below. Hence $\lim_{t\to+\infty} E(t) = E_\infty$ for some $E_\infty \in \mathbb{R}$. Hence $\dot{X} \in L^2([0,+\infty); H^2)$. We proved above that $\dot{X} \in L^\infty([0,+\infty); H^2)$ and $\varepsilon U(X) \in L^\infty([0,+\infty); \mathbb{R})$. ∎

5. Asymptotic properties: Proof of Theorem 2.2. The proof of Theorem 2.2 goes in three steps. In section 5.1, we prove preliminary results. In section 5.2, we prove Part 1. Finally, we prove Part 2 in section 5.3.

5.1. Preliminary results. In this section, we first obtain results by adding the two equations of (HBFC$^2_{sing}$), with no influence from the interaction potential V.

LEMMA 5.1. *Under the assumptions of Theorem* 2.2*,*

(i) *the maps* $\nabla\phi(x)$ *and* $\nabla\phi(y)$ *are bounded;*

(ii) $\lim_{t\to+\infty} \dot{x}(t) + \dot{y}(t) = 0$*, and the map* $\ddot{x} + \ddot{y}$ *is bounded.*

Proof of Lemma 5.1. *Proof of* (i). If the trajectory (x, y) is bounded (case (a)), it is an immediate consequence of assumption $(\mathcal{H}_\phi)$(ii). Let us now consider case (b). From Proposition 2.1(ii), the map $\phi(x)$ is bounded. The limit condition (LIM) implies that the set

$$C = \{z \in H, \phi(z) \leq \|\phi(x)\|_\infty \text{ and } |\nabla\phi(z)| \geq 1\}$$

is bounded. Since the map $\nabla\phi$ is bounded on the bounded sets, it is bounded on C. If $x(t) \notin C$, since $\phi(x(t)) \leq \|\phi(x)\|_\infty$, we deduce $|\nabla\phi(x(t))| < 1$. Hence the map $\nabla\phi(x)$ is bounded, and so is the map $\nabla\phi(y)$.

Proof of (ii). By adding the two equations of (HBFC$^2_{sing}$), we obtain

$$\ddot{x}(t) + \ddot{y}(t) + \gamma(\dot{x}(t) + \dot{y}(t)) + \nabla\phi(x(t)) + \nabla\phi(y(t)) = 0. \tag{5.1}$$

Since, by (i), $\nabla\phi(x)$ and $\nabla\phi(y)$ are bounded and since by Proposition 2.1(ii), $\dot{x}$ and $\dot{y}$ are bounded, then from (5.1) the map $\ddot{x} + \ddot{y}$ is bounded. From Proposition 2.1(ii), the map $\dot{x} + \dot{y}$ belongs to $L^2([0,+\infty), H)$, and we deduce by a classical argument

$$\lim_{t\to+\infty} \dot{x}(t) + \dot{y}(t) = 0. \qquad ∎$$

LEMMA 5.2. *Under the assumptions of Theorem* 2.2*, let* $(t_n) \subset \mathbb{R}_+$ *be a sequence such that* $\lim_{n\to+\infty} t_n = +\infty$ *and* $(x(t_n), y(t_n))$ *is bounded. Then*

$$\lim_{n\to+\infty} \nabla\phi(x(t_n)) + \nabla\phi(y(t_n)) = 0.$$

Proof of Lemma 5.2. In view of Lemma 5.1(ii) and (5.1), it is equivalent to prove that $\lim_{n\to+\infty} \ddot{x}(t_n) + \ddot{y}(t_n) = 0$. Take a real number $\tau > 0$. Since the maps $\dot{x}$ and $\dot{y}$ are bounded, and since the sequence $(x(t_n), y(t_n))$ is bounded, the sets $\bigcup_{n\geq0} x([t_n - \tau, t_n + \tau])$ and $\bigcup_{n\geq0} y([t_n - \tau, t_n + \tau])$ are also bounded. Hence the map $\nabla\phi$ is K-Lipschitz continuous on the bounded set

$$\bigcup_{n\geq0} x([t_n - \tau, t_n + \tau]) \cup \bigcup_{n\geq0} y([t_n - \tau, t_n + \tau])$$

for some $K > 0$. Let $z = x + y$; then $\dot{z}$ and $\ddot{z}$ are also Lipschitz continuous on the bounded set $\bigcup_{n\geq 0} x([t_n-\tau, t_n+\tau]) \cup \bigcup_{n\geq 0} y([t_n-\tau, t_n+\tau])$. The proof now consists in a.e. differentiating (5.1). We obtain, for almost every t,

$$\dddot{z}(t) + \gamma\ddot{z}(t)(t) + \frac{d}{dt}\nabla\phi(x(.))(t) + \frac{d}{dt}\nabla\phi(y(.))(t) = 0.$$

But $|\frac{d}{dt}\nabla\phi(x(.))(t) + \frac{d}{dt}\nabla\phi(y(.))(t)| \leq K(|\dot{x}(t)| + |\dot{y}(t)|)$. From the above equation,

$$\text{for a.e. } t,\ \dddot{z}(t) + \gamma\ddot{z}(t)(t) \leq K\left(|\dot{x}(t)| + |\dot{y}(t)|\right). \tag{5.2}$$

Assuming without loss of generality that $t_n - \tau \geq 0$, we multiply (5.2) by $e^{\gamma t}$ and integrate between $t_n - \tau$ and t_n.

$$\ddot{z}(t_n) \leq \ddot{z}(t_n - \tau)e^{-\gamma\tau} + e^{-\gamma t_n}\int_{t_n-\tau}^{t_n} e^{\gamma s}K\left(|\dot{x}(s)| + |\dot{y}(s)|\right)\, ds.$$

By Cauchy–Schwarz inequality

$$\begin{aligned}|\ddot{z}(t_n)| &\leq |\ddot{z}(t_n - \tau)|e^{-\gamma\tau} + e^{-\gamma t_n}\left(\int_{t_n-\tau}^{t_n} e^{2\gamma s}\, ds\right)^{\frac{1}{2}}\left(\int_{t_n-\tau}^{t_n} K^2\left(|\dot{x}(s)| + |\dot{y}(s)|\right)^2\, ds\right)^{\frac{1}{2}}\\ &\leq |\ddot{z}(t_n - \tau)|e^{-\gamma\tau} + \frac{1}{\sqrt{2\gamma}}\left(\int_{t_n-\tau}^{t_n} K^2\left(|\dot{x}(s)| + |\dot{y}(s)|\right)^2\, ds\right)^{\frac{1}{2}}.\end{aligned}$$

Recalling that $\dot{x}$ and $\dot{y}$ belong to $L^2([0,+\infty); H)$, let us take the upper limit of each member when $n \to +\infty$:

$$\limsup_{n\to+\infty} |\ddot{x}(t_n) + \ddot{y}(t_n)(t_n)| \leq \|\ddot{x} + \ddot{y}\|_\infty e^{-\gamma\tau}.$$

At the limit when $\tau \to +\infty$, we obtain $\lim_{t\to+\infty} \ddot{x}(t_n) + \ddot{y}(t_n) = 0$. □

5.2. Proof of Theorem 2.2, Part 1. We distinguish the two cases (a) and (b).

5.2.1. Proof in case (a). The trajectory (x, y) is bounded, and the fact that $\lim_{t\to+\infty} \nabla\phi(x(t)) = \lim_{t\to+\infty} \nabla\phi(y(t)) = 0$ is an immediate consequence of the following lemma. □

LEMMA 5.3. *Under the assumptions of Theorem* 2.2*, let* $(t_n) \subset \mathbb{R}_+$ *be a sequence such that* $\lim_{n\to+\infty} t_n = +\infty$ *and* $(x(t_n), y(t_n))$ *is bounded. Then*

$$\lim_{n\to+\infty} \nabla\phi(x(t_n)) = \lim_{n\to+\infty} \nabla\phi(y(t_n)) = 0.$$

Contrary to its regular version [6, Lemma 4.1], the proof of Lemma 5.3 uses the coupled structure of the system. An important step is the proof that the map $t \mapsto \nabla V(x(t) - y(t))$ is bounded on a convenient set.

Proof of Lemma 5.3. Assume that it is not true. Then there exists $\alpha > 0$ and a subsequence of (t_n), still denoted by (t_n), such that $\lim_{n\to\infty} t_n = +\infty$ and $|\nabla\phi(x(t_n))| \geq 2\alpha$ or $|\nabla\phi(y(t_n))| \geq 2\alpha$ for every n. Without loss of generality, we may assume that, for every $n \geq 0$,

$$|\nabla\phi(x(t_n))| \geq 2\alpha. \tag{5.3}$$

By assumption, the map $\nabla\phi$ is K-Lipschitz continuous on the bounded set

$$C = \{z \in H, \quad \exists n \in \mathbb{N}, \quad |z - x(t_n)| \leq 1 \text{ or } |z - y(t_n)| \leq 1\}$$

for some $K > 0$. Let $\tau = \frac{\alpha}{K \max\{\|\dot{x}\|_\infty, \|\dot{y}\|_\infty\}}$ (assuming, without loss of generality, that $\max\{\|\dot{x}\|_\infty, \|\dot{y}\|_\infty\} \neq 0$) and let $t \in [t_n, t_n + \tau]$. Then

$$|x(t) - x(t_n)| \leq \|\dot{x}\|_\infty \tau \leq \frac{\alpha}{K} \quad \text{and} \quad |y(t) - y(t_n)| \leq \|\dot{y}\|_\infty \tau \leq \frac{\alpha}{K}.$$

Assuming without loss of generality that $\frac{\alpha}{K} \leq 1$, then from above $x(t) \in C$ and $y(t) \in C$. Consequently,

$$|\nabla\phi(x(t)) - \nabla\phi(x(t_n))| \leq K|x(t) - x(t_n)| \leq K\frac{\alpha}{K} = \alpha, \tag{5.4}$$

$$|\nabla\phi(y(t)) - \nabla\phi(y(t_n))| \leq K|y(t) - y(t_n)| \leq K\frac{\alpha}{K} = \alpha. \tag{5.5}$$

For every $t \in [t_n, t_n + \tau]$, let us now multiply the first equation of (HBFC^2_{sing}) by $e^{\gamma s}$ and integrate on the interval $[t_n, t]$

$$e^{\gamma t}\dot{x}(t) - e^{\gamma t_n}\dot{x}(t_n) + \int_{t_n}^{t} e^{\gamma s}\,\nabla\phi(x(s))\,ds + \int_{t_n}^{t} e^{\gamma s}\,\varepsilon(s)\nabla V(x(s) - y(s))\,ds = 0. \tag{5.6}$$

In view of (5.4), we have

$$\left|\int_{t_n}^{t} e^{\gamma s}\,\nabla\phi(x(s))\,ds - \int_{t_n}^{t} e^{\gamma s}\,\nabla\phi(x(t_n))\,ds\right| \leq \int_{t_n}^{t} e^{\gamma s}\alpha\,ds = \alpha\,\frac{e^{\gamma t} - e^{\gamma t_n}}{\gamma}.$$

On the other hand,

$$\left|\int_{t_n}^{t} e^{\gamma s}\,\nabla\phi(x(t_n))\,ds\right| = \frac{e^{\gamma t} - e^{\gamma t_n}}{\gamma}|\nabla\phi(x(t_n))| \geq 2\alpha\,\frac{e^{\gamma t} - e^{\gamma t_n}}{\gamma}.$$

Then, noticing, moreover, that $\nabla\phi$ is bounded by some $M > 0$ on the bounded set C,

$$\alpha\,\frac{e^{\gamma t} - e^{\gamma t_n}}{\gamma} \leq \left|\int_{t_n}^{t} e^{\gamma s}\,\nabla\phi(x(s))\,ds\right| \leq M\,\frac{e^{\gamma t} - e^{\gamma t_n}}{\gamma}. \tag{5.7}$$

Let us now prove that the map $t \mapsto \nabla V(x(t) - y(t))$ is bounded on the set $\bigcup_{n\in\mathbb{N}}[t_n, t_n + \tau]$. From Lemma 5.2, $\lim_{n\to+\infty} \nabla\phi(x(t_n)) + \nabla\phi(y(t_n)) = 0$; hence there exists $N \in \mathbb{N}$ such that, for every $n \geq N$,

$$|\nabla\phi(x(t_n)) + \nabla\phi(y(t_n))| \leq \alpha. \tag{5.8}$$

Writing

$$\begin{aligned} 2\,\nabla\phi(x(t_n)) = \quad & \nabla\phi(x(t)) - \nabla\phi(y(t)) \quad + \quad \nabla\phi(x(t_n)) - \nabla\phi(x(t)) \\ & + \nabla\phi(y(t)) - \nabla\phi(y(t_n)) \quad + \quad \nabla\phi(x(t_n)) + \nabla\phi(y(t_n)) \end{aligned}$$

and taking into account (5.3), (5.4), (5.5), and (5.8), we find $4\,\alpha \leq |\nabla\phi(x(t)) - \nabla\phi(y(t))| + 3\,\alpha$, i.e.,

$$|\nabla\phi(x(t)) - \nabla\phi(y(t))| \geq \alpha.$$

Since $\nabla\phi$ is K-Lipschitz continuous on the bounded set C, we have $|\nabla\phi(x(t)) - \nabla\phi(y(t))| \le K\,|x(t)-y(t)|$, which combined with the previous inequality yields

$$|x(t)-y(t)| \ge \frac{\alpha}{K}.$$

Since $(x(t_n), y(t_n))$ is bounded, and since the maps $\dot{x}$ and $\dot{y}$ are bounded, the above equation implies that the set

$$\left\{ x(t)-y(t), t \in \bigcup_{n\ge N} [t_n, t_n+\tau] \right\}$$

is a bounded subset of $H \setminus B(0, \frac{\alpha}{K})$, and the map ∇V is bounded on this set. Equivalently, the map $t \mapsto \nabla V(x(t)-y(t))$ is bounded on the set $\bigcup_{n\ge N}[t_n, t_n+\tau]$, hence on the set $\bigcup_{n\in\mathbb{N}}[t_n, t_n+\tau]$. Since $\lim_{t\to+\infty}\varepsilon(t)=0$, for n large enough and for every $t \in [t_n, t_n+\tau]$,

$$\left| \int_{t_n}^{t} e^{\gamma s}\,\varepsilon(s)\nabla V(x(s)-y(s))\,ds \right| \le \frac{\alpha}{2}\,\frac{e^{\gamma t}-e^{\gamma t_n}}{\gamma}. \tag{5.9}$$

In view of (5.6), (5.7), and (5.9), for every $t \in [t_n, t_n+\tau]$,

$$\frac{\alpha}{2}\,\frac{e^{\gamma t}-e^{\gamma t_n}}{\gamma} \le \left| e^{\gamma t}\dot{x}(t) - e^{\gamma t_n}\dot{x}(t_n)\right| \le \left(M+\frac{\alpha}{2}\right)\,\frac{e^{\gamma t}-e^{\gamma t_n}}{\gamma},$$

which can be rewritten as

$$\frac{\alpha}{2}\,\frac{e^{\gamma(t-t_n)}-1}{\gamma} \le |e^{\gamma(t-t_n)}\dot{x}(t) - \dot{x}(t_n)| \le \left(M+\frac{\alpha}{2}\right)\,\frac{e^{\gamma(t-t_n)}-1}{\gamma}.$$

Taking the following claim into account, we obtain, for some $\beta>0$ and n large enough, $\int_{t_n}^{t_n+\tau} e^{\gamma(t-t_n)}\,|\dot{x}(t)|\,dt \ge \beta$. We then deduce from Cauchy–Schwarz inequality

$$\frac{e^{2\gamma\tau}-1}{2\gamma}\int_{t_n}^{t_n+\tau} |\dot{x}(t)|^2\,dt \;\ge\; \beta^2,$$

and, finally, $\int_{t_n}^{t_n+\tau}|\dot{x}(t)|^2\,dt \;\ge\; 2\gamma\,\beta^2\,e^{-2\gamma\tau}$, which contradicts $\dot{x} \in L^2([0,+\infty), H)$. Hence $\lim_{n\to+\infty}\nabla\phi(x(t_n)) = \lim_{n\to+\infty}\nabla\phi(y(t_n)) = 0$. □

Claim 5.1. Let $g \in \mathcal{C}^0([0,\tau],\mathbb{R}_+)$ and $h \in \mathcal{C}^0([0,\tau],\mathbb{R}_+)$ such that $\int_0^\tau g(t)\,dt > 0$, $h(0)=0$. There exists $\beta>0$ such that, for every $f \in \mathcal{C}^0([0,\tau],H)$ which satisfies $g(t) \le |f(t)-f(0)| \le h(t)$ for every $t\in[0,\tau]$,

$$\int_0^\tau |f(t)|dt \ge \beta.$$

Proof of Claim 5.1. Since $h(0)=0$, we clearly have $\lim_{t\to0}\frac{\int_0^t h(s)\,ds}{t}=0$. Let $\theta \in]0,\tau[$ such that $\int_0^\theta h(s)\,ds \le \frac{\int_0^\tau g(t)dt}{4\tau}\theta$. Let $\beta = \min\{\frac{\int_0^\tau g(t)\,dt}{2}, \frac{\int_0^\tau g(t)dt}{4\tau}\theta\}$. First assume $|f(0)| \le \frac{\int_0^\tau g(t)dt}{2\tau}$. Then $|f(t)| \ge g(t) - |f(0)|$ implies $\int_0^\tau |f(t)|dt \ge \frac{\int_0^\tau g(t)\,dt}{2}$. Now assume $|f(0)| \ge \frac{\int_0^\tau g(t)dt}{2\tau}$. Then $|f(t)| \ge |f(0)| - h(t)$ implies $\int_0^\tau |f(t)|dt \ge \int_0^\theta |f(t)|dt \ge \frac{\int_0^\tau g(t)dt}{2\tau}\theta - \frac{\int_0^\tau g(t)dt}{4\tau}\theta = \frac{\int_0^\tau g(t)dt}{4\tau}\theta$. □

5.2.2. Proof in case (b). Let us argue by contradiction and assume that it is not true; i.e., there exists $\alpha > 0$ and a sequence $(t_n) \subset \mathbb{R}_+$ such that $\lim_{n\to+\infty} t_n = +\infty$ and $|\nabla\phi(x(t_n))| \geq \alpha$ or $|\nabla\phi(y(t_n))| \geq \alpha$. Without loss of generality, we may assume that $|\nabla\phi(x(t_n))| \geq \alpha$. Since the map $\phi(x)$ is bounded (Proposition 2.1(ii)) and since the map ϕ satisfies the limit condition (LIM), the sequence $(x(t_n))$ is bounded. The following lemma from [6] shows that the sequence $(y(t_n))$ is also bounded. From Lemma 5.3, $\lim_{n\to+\infty} \nabla\phi(x(t_n)) = 0$, a contradiction. Hence $\lim_{t\to+\infty} \nabla\phi(x(t)) = \lim_{t\to+\infty} \nabla\phi(y(t)) = 0$. □

LEMMA 5.4 (see [6, Lemma 4.2]). *Assume Hypothesis* 1 *and Hypothesis* 2*; i.e., the map ϕ satisfies the limit condition* (LIM). *Let $\alpha > 0$ and $(t_n) \subset \mathbb{R}_+$ be a sequence such that $\lim_{n\to+\infty} t_n = +\infty$, $(x(t_n))$ is bounded, and $|\nabla\phi(x(t_n))| \geq \alpha$. Then the sequence $(y(t_n))$ is also bounded.*

The proof of Lemma 5.4 is independent from the singularity assumption on V, and the proof given in [6] remains valid in the present case.

5.3. Proof of Theorem 2.2, Part 2. The proof of Part 2 heavily relies on the direct study of the map $x - y$. By taking the difference of the two equations of (HBFC^2_{sing}), we obtain

(5.10)
$$\ddot{x}(t) - \ddot{y}(t) + \gamma(\dot{x}(t) - \dot{y}(t)) + \nabla\phi(x(t)) - \nabla\phi(y(t)) + 2\varepsilon(t)\nabla V(x(t) - y(t)) = 0.$$

We then deduce that the function $h(t) := \frac{1}{2}|x(t) - y(t)|^2$ satisfies the following equation:

$$\begin{aligned}\ddot{h}(t) + \gamma\dot{h}(t) =& |\dot{x}(t) - \dot{y}(t)|^2 - \langle \nabla\phi(x(t)) - \nabla\phi(y(t)), x(t) - y(t)\rangle \\ &-2\varepsilon(t)\langle \nabla V(x(t) - y(t)), x(t) - y(t)\rangle.\end{aligned} \tag{5.11}$$

Recalling that $\lim_{t\to+\infty} \dot{x}(t) + \dot{y}(t) = 0$ (Lemma 5.1), it is sufficient to prove that $\lim_{t\to+\infty} \dot{x}(t) - \dot{y}(t) = 0$. Assume that it is not true. Then there is $\alpha > 0$ and a sequence $(t_n) \subset \mathbb{R}_+$ such that $\lim_{n\to+\infty} t_n = +\infty$ and

$$\lim_{t\to+\infty} |\dot{x}(t_n) - \dot{y}(t_n)|^2 = \alpha.$$

Since $|\dot{x}-\dot{y}|^2 \in L^1([0,+\infty);\mathbb{R}_+)$ (Proposition 2.1), then $\liminf_{t\to+\infty} |\dot{x}(t)-\dot{y}(t)|^2 = 0$. Hence there exists two sequences (a_n) and (b_n) in $\mathbb{R}_+$ such that

$$|\dot{x}(t) - \dot{y}(t)|^2 \geq \frac{\alpha}{2} \text{ for every } t \in [t_n - a_n, t_n + b_n], \tag{5.12}$$

$$|\dot{x}(t_n - a_n) - \dot{y}(t_n - a_n)|^2 = |\dot{x}(t_n + b_n) - \dot{y}(t_n + b_n)|^2 = \frac{\alpha}{2}. \tag{5.13}$$

Without loss of generality, we assume $|\dot{x}(t_n)-\dot{y}(t_n)|^2 = \alpha$ for every n. Since $|\dot{x}-\dot{y}|^2 \in L^1([0,+\infty);\mathbb{R}_+)$, we have $\lim_{n\to+\infty} a_n = \lim_{n\to+\infty} b_n = 0$. From Rolle's theorem, there are $t^a_n \in (t_n - a_n, t_n)$ and $t^b_n \in (t_n + b_n, t_n)$ such that

$$2\,\langle \ddot{x}(t^a_n) - \ddot{y}(t^a_n), \dot{x}(t^a_n) - \dot{y}(t^a_n)\rangle = \frac{d}{dt}|\dot{x} - \dot{y}|^2(t^a_n) = \frac{\alpha}{2a_n}, \tag{5.14}$$

$$2\,\big\langle \ddot{x}(t^b_n) - \ddot{y}(t^b_n), \dot{x}(t^b_n) - \dot{y}(t^b_n)\big\rangle = \frac{d}{dt}|\dot{x} - \dot{y}|^2(t^b_n) = -\frac{\alpha}{2b_n}. \tag{5.15}$$

Taking the scalar product of (5.10) with $\dot{x}(t) - \dot{y}(t)$, we deduce

$$\begin{aligned}&\langle \ddot{x}(t) - \ddot{y}(t), \dot{x}(t) - \dot{y}(t)\rangle + \gamma|\dot{x}(t) - \dot{y}(t)|^2 + \langle \nabla\phi(x(t)) - \nabla\phi(y(t)), \dot{x}(t) - \dot{y}(t)\rangle \\ &\qquad +2\varepsilon(t)\langle \nabla V(x(t) - y(t)), \dot{x}(t) - \dot{y}(t)\rangle = 0.\end{aligned} \tag{5.16}$$

Since the map $t \mapsto \gamma|\dot{x}(t)-\dot{y}(t)|^2+\langle\nabla\phi(x(t))-\nabla\phi(y(t)),\dot{x}(t)-\dot{y}(t)\rangle$ is bounded, and in view of (5.14) and (5.15), for n large enough, we have

$$\langle\nabla V(x(t_n^a)-y(t_n^a)),\dot{x}(t_n^a)-\dot{y}(t_n^a)\rangle \le -\frac{\alpha}{5a_n},$$
$$\big\langle\nabla V(x(t_n^b)-y(t_n^b)),\dot{x}(t_n^b)-\dot{y}(t_n^b)\big\rangle \ge \frac{\alpha}{5a_n}.$$

At this point, let us first assume for simplicity that the error function e is equal to zero. Note that in this case, the assumption that the function W is of class $\mathcal{C}^1$ is automatically satisfied since we assumed that the potential V is of class $\mathcal{C}^1$. From $(\mathcal{H}_V)$(iii), $\nabla V(z)=W'(|z|)\frac{z}{|z|}$. Hence $\langle\nabla V(x-y),\dot{x}-\dot{y}\rangle=\frac{W'(|x-y|)}{|x-y|}\langle x-y,\dot{x}-\dot{y}\rangle=\frac{W'(|x-y|)}{|x-y|}\dot{h}$, and we deduce

$$\dot{h}(t_n^a)>0 \quad\text{and}\quad \dot{h}(t_n^b)<0.$$

Let us now multiply (5.11) by $e^{\gamma t}$ and integrate between t_n^a and t_n^b:

$$\frac{1}{\gamma}\left(e^{\gamma t_n^b}\,\dot{h}(t_n^b)-e^{\gamma t_n^a}\,\dot{h}(t_n^a)\right)=\int_{t_n^a}^{t_n^b}|\dot{x}(t)-\dot{y}(t)|^2-\langle\nabla\phi(x(t))-\nabla\phi(y(t)),x(t)-y(t)\rangle$$
$$-2\varepsilon(t)\langle\nabla V(x(t)-y(t)),x(t)-y(t)\rangle dt.$$

From $(\mathcal{H}_V)$(iii), $\langle\nabla V(z),z\rangle=W'(|z|)|z|$; hence $\langle\nabla V(x(t)-y(t)),x(t)-y(t)\rangle\le 0$. In view of (5.12), we deduce

$$\frac{\alpha}{2}(t_n^b-t_n^a)\le\sup_{t\in[t_n^a,t_n^b]}|x(t)-y(t)|\sup_{t\in[t_n^a,t_n^b]}|\nabla\phi(x(t))-\nabla\phi(y(t))|(t_n^b-t_n^a).$$

From the following lemma, the sequence $(\sup_{t\in[t_n^a,t_n^b]}|x(t)-y(t)|)_{n\in\mathbb{N}}$ is bounded (remark that this is clear when the trajectories x and y are bounded). From Theorem 2.2, Part 1, we have $\lim_{t\to+\infty}\nabla\phi(x(t))=\lim_{t\to+\infty}\nabla\phi(y(t))=0$, and we obtain a contradiction in the previous equation, for n large enough. Hence $\lim_{t\to+\infty}\dot{x}(t)-\dot{y}(t)=0$, which implies $\lim_{t\to+\infty}\dot{x}(t)=\lim_{t\to+\infty}\dot{y}(t)=0$. □

LEMMA 5.5.

$$\lim_{n\to+\infty}\sup_{t\in[t_n^a,t_n^b]}|x(t)-y(t)|=0.$$

Proof of Lemma 5.5. It is a consequence on a direct estimation on the energy function $E(t)=\frac{1}{2}|\dot{x}(t)|^2+\frac{1}{2}|\dot{y}(t)|^2+\phi(x(t))+\phi(y(t))+\varepsilon(t)V(x(t)-y(t))$. Recall that it is decreasing (Proposition 2.1(iii)); hence $E(t_n-a_n)\ge E(t_n)$. Since $|\dot{x}|^2+|\dot{y}|^2=\frac{1}{2}|\dot{x}-\dot{y}|^2+\frac{1}{2}|\dot{x}+\dot{y}|^2$, and since $\lim_{t\to+\infty}\dot{x}(t)+\dot{y}(t)=0$ (Lemma 5.1), for n large enough

$$\frac{1}{2}\,|\dot{x}(t_n-a_n)|^2+\frac{1}{2}\,|\dot{y}(t_n-a_n)|^2\le\frac{\alpha}{3},$$
$$\frac{1}{2}\,|\dot{x}(t_n)|^2+\frac{1}{2}\,|\dot{y}(t_n)|^2\ge\frac{\alpha}{2}.$$

Thus the inequality $E(t_n-a_n)\ge E(t_n)$ implies

$$\varepsilon(t_n-a_n)V(x(t_n-a_n)-y(t_n-a_n))\ge\frac{\alpha}{6}+\phi(x(t_n))+\phi(y(t_n))-\phi(x(t_n-a_n))$$
$$-\phi(y(t_n-a_n))+\varepsilon(t_n)V(x(t_n)-y(t_n)).$$

But $\frac{d}{dt}\phi(x(t)) + \frac{d}{dt}\phi(y(t)) = \langle \nabla\phi(x(t)), \dot{x}(t)\rangle + \langle \nabla\phi(y(t)), \dot{y}(t)\rangle$, the maps $\dot{x}$ and $\dot{y}$ are bounded, $\lim_{t\to+\infty} \nabla\phi(x(t)) = \lim_{t\to+\infty} \nabla\phi(y(t)) = 0$, and $\lim_{n\to+\infty} a_n = 0$; hence for n large enough

$$\varepsilon(t_n - a_n)V(x(t_n - a_n) - y(t_n - a_n)) \geq \frac{\alpha}{7}.$$

Since $\lim_{t\to+\infty} \varepsilon(t) = 0$, we have $\lim_{t\to+\infty} V(x(t_n - a_n) - y(t_n - a_n)) = +\infty$. Assumption $(\mathcal{H}_V)$(iii) clearly implies that the potential V is bounded on $H \setminus B(0, r)$, for every $r > 0$, and we deduce that $\lim_{n\to+\infty} |x(t_n - a_n) - y(t_n - a_n)| = 0$. Since $\lim_{n\to+\infty} a_n = \lim_{n\to+\infty} b_n = 0$, since the map $\dot{x} - \dot{y}$ is bounded, we deduce that $\lim_{n\to+\infty} \sup_{t\in[t_n-a_n,t_n+b_n]} |x(t) - y(t)| = 0$. Since $[t_n^a, t_n^b] \subset [t_n - a_n, t_n + b_n]$, this ends the proof of Lemma 5.5. □

We now proceed to the proof of Theorem 2.2, Part 2 in the general case. First notice that Lemma 5.5 remains valid. From $(\mathcal{H}_V)$(iii), $\nabla V(z) = W'(|z|)\frac{z}{|z|} + \nabla e(z)$. Hence $\langle \nabla V(x-y), \dot{x}-\dot{y}\rangle = \frac{W'(|x-y|)}{|x-y|}\langle x-y, \dot{x}-\dot{y}\rangle + \langle \nabla e(x-y), \dot{x}-\dot{y}\rangle = \frac{W'(|x-y|)}{|x-y|}\dot{h} + \langle \nabla e(x-y), \dot{x}-\dot{y}\rangle$. Since the gradient ∇e is bounded near 0, in view of Lemma 5.5, and since the map $\dot{x} - \dot{y}$ is bounded, we deduce that the term $\langle \nabla e(x-y), \dot{x}-\dot{y}\rangle$ is bounded on the interval $[t_n^a, t_n^b]$. We thus deduce again $\dot{h}(t_n^a) > 0$ and $\dot{h}(t_n^b) < 0$. Since $\langle \nabla V(z), z\rangle = W'(|z|)|z| + \langle \nabla e(z), z\rangle$, which is negative near 0, the remainder of the proof holds. □

5.3.1. Toward an alternate proof of Part 2. In this section, we show how a more classical proof can lead to the same result without the strong repulsion assumption on the potential V but, of course, with an additional assumption, namely, that the two trajectories do not collapse.

PROPOSITION 5.6. *Under the assumptions of Theorem* 2.2, *Part* 1, *for every* $\mu > \lambda > 0$,[7]

$$\lim_{t\to+\infty, |x(t)-y(t)|\in[\lambda,\mu]} \dot{x}(t) = \lim_{t\to+\infty, |x(t)-y(t)|\in[\lambda,\mu]} \dot{y}(t) = 0. \tag{5.17}$$

Assume, moreover, that (x, y) *is bounded, that the maps* $\phi \circ x$ *and* $\phi \circ y$ *converge,*[8] *and that* $x - y \not\to 0$. *Then*

$$\lim_{t\to+\infty} \dot{x}(t) = \lim_{t\to+\infty} \dot{y}(t) = 0.$$

Proof of Proposition 5.6. Assume that (5.17) is not true. Then there exist $\alpha > 0$ and a sequence $(t_n) \subset \mathbb{R}_+$ such that $\lim_{n\to+\infty} t_n = +\infty$ and $|\dot{x}(t_n)| \geq \alpha$, $|x(t_n) - y(t_n)| \in [\lambda, \mu]$. Let $\tau \in (0, \frac{\lambda}{2(\|\dot{x}\|_\infty + \|\dot{y}\|_\infty)}]$. Then for every $t \in [t_n, t_n + \tau]$, $|x(t) - y(t)| \in [\lambda/2, \mu + \lambda/2]$. Hence the map $t \mapsto \nabla V(x(t) - y(t))$ is bounded on the set $\bigcup_{n\in\mathbb{N}}[t_n, t_n + \tau]$. Hence the map $\ddot{x}$ is bounded on the set $\bigcup_{n\in\mathbb{N}}[t_n, t_n + \tau]$. By a classical argument, we obtain a contradiction with $\dot{x} \in L^2([0, +\infty))$.

Let us now prove the second part. Since $x-y \not\to 0$ there exist $\alpha > 0$ and a sequence $(t_n) \subset \mathbb{R}_+$ such that $\lim_{n\to+\infty} t_n = +\infty$ and $|x(t_n) - y(t_n)| \in [\alpha, \max|x - y|]$. From part (a), $\dot{x}(t_n) \to 0$ and $\dot{y}(t_n) \to 0$. Since $U(x(t_n), y(t_n)) = V(x(t_n) - y(t_n))$ is bounded, $E(t_n) \to \lim_{t\to+\infty} \phi(x(t)) + \phi(y(t))$. Since $\frac{1}{2}|\dot{x}(t)|^2 + \frac{1}{2}|\dot{y}(t)|^2 \leq E(t) - \phi(x(t)) - \phi(y(t))$, we deduce the result. □

[7] Of course we assume that the limits have a meaning, i.e., $\{t \geq T| \quad |x(t) - y(t)| \in [\lambda, \mu]\} \neq \emptyset$ for every T.

[8] This is here the case if ϕ is convex.

6. Proof of the one-dimensional convergence results. In this section, we consider the one-dimensional and singular case, i.e., $H = \mathbb{R}$. Note that V' is clearly bounded on the bounded subsets of $\mathbb{R} \setminus \{0\}$; hence Theorem 2.2, Part 1 applies. In particular, it implies

(6.1)
$$\phi'(\lambda) = 0 \text{ for every } \lambda \in \left[\liminf_{t\to+\infty} x(t), \limsup_{t\to+\infty} x(t)\right] \cup \left[\liminf_{t\to+\infty} y(t), \limsup_{t\to+\infty} y(t)\right]$$

such that $\lambda \in \mathbb{R}$. For the sake of readability, we prove the second part of Theorem 2.4 before the first part.

6.1. Proof of Theorem 2.4, Part 2. Without loss of generality, we assume $y_0 < x_0$. From Proposition 2.1(i), the trajectory (x, y) satisfies $y(t) \neq x(t)$ for every t; hence, since x and y are continuous,

$$y(t) < x(t).$$

Hence $V'(x(t) - y(t)) \leq 0$ from the repulsion assumption $(\mathcal{H}_V)$(iv). From the two equations of (HBFC^2_{sing}), we deduce

$$\ddot{x} + \gamma\dot{x} + \phi'(x) \geq 0 \quad \text{and} \quad \ddot{y} + \gamma\dot{y} + \phi'(y) \leq 0. \tag{6.2}$$

6.1.1. Proof of (ii). The proof in the general case remains valid in dimension one, noticing that $x - y$ has a constant sign. But we give here a short and (much) simpler proof. Since the maps $\dot{x}$ and $\dot{y}$ are bounded (Proposition 2.1(ii)) and the maps $\phi'(x)$ and $\phi'(y)$ are bounded, we deduce from (6.2) that the map $\ddot{x}$ (resp., $\ddot{y}$), is bounded from below (resp., above). Since $\dot{x} \in L^2([0, +\infty))$ and $\dot{y} \in L^2([0, +\infty))$, we deduce by a classical argument

$$\lim_{t\to+\infty} \dot{x}(t) = 0 \quad \text{and} \quad \lim_{t\to+\infty} \dot{y}(t) = 0. \qquad \square$$

6.1.2. Proof of (i). In view of (6.2) and (6.1), the proof of the convergence is identical to the proof of [6, Lemma 5.2] in the regular case. We now briefly recall the argument of [6]. Assume that the map x does not converge, i.e., $\liminf_{t\to+\infty} x(t) < \limsup_{t\to+\infty} x(t)$. There exist two sequences (t_n) and (t'_n) in $\mathbb{R}_+$ such that $t_n \leq t'_n$, $\lim_{n\to+\infty} t_n = \lim_{n\to+\infty} t'_n = +\infty$, and

$$\begin{aligned} x([t_n, t'_n]) &\subset \left[\liminf_{t\to+\infty} x(t), \limsup_{t\to+\infty} x(t)\right], \\ \lim_{n\to+\infty} x(t_n) &= \limsup_{t\to+\infty} x(t), \\ \lim_{n\to+\infty} x(t'_n) &= \liminf_{t\to+\infty} x(t). \end{aligned}$$

Indeed, take a sequence (τ_n) such that $\lim_{n\to+\infty} \tau_n = +\infty$ $\lim_{n\to+\infty} x(\tau_n) = \limsup_{t\to+\infty} x(t)$, and, for n large enough, let

$$t'_n = \sup\left\{u \geq \tau_n,\ x([\tau_n, u]) \geq \liminf_{t\to+\infty} x(t) + \frac{1}{n}\right\}, \tag{6.3}$$

$$t_n = \inf\left\{u \in [\tau_n, t'_n],\ x([u, T_n]) \leq \limsup_{t\to+\infty} x(t)\right\}. \tag{6.4}$$

Then, for every $u \in [t_n, t_n']$, $\phi'(x(u)) = 0$; hence, in view of (6.2), $\ddot{x}(u) + \gamma\dot{x}(u) \geq 0$. We integrate the previous inequality on $[t_n, t_n']$ to find

$$x(t_n') \geq x(t_n) - \frac{1}{\gamma}(\dot{x}(t_n') - \dot{x}(t_n)).$$

Since $\lim_{t\to+\infty} \dot{x}(t) = 0$ (assertion (ii) above) and passing to the limit when $n \to +\infty$, we obtain $\liminf_{t\to+\infty} x(t) \geq \limsup_{t\to+\infty} x(t)$, a contradiction. Hence there exists $(x_\infty, y_\infty) \in \overline{\mathbb{R}} \times \overline{\mathbb{R}}$ such that $\lim_{t\to+\infty} x(t) = x_\infty$ and $\lim_{t\to+\infty} y(t) = y_\infty$. Since $\lim_{t\to+\infty}(\phi'(x(t)), \phi'(y(t))) = (0,0)$ (Theorem 2.2, Part 1), then $(x_\infty, y_\infty) \in \widehat{S} \times \widehat{S}$, which ends the proof of Theorem 2.4, Part 2. □

6.2. Proof of Theorem 2.4, Part 1. The trajectory (x, y) is now assumed to be bounded. We distinguish the two cases, $x - y \not\to 0$ and $x - y \to 0$, and again assume $y_0 < x_0$.

6.2.1. Proof in the case $x - y \not\to 0$. In view of (6.1), the map ϕ is constant on each of the intervals $[\liminf_{t\to+\infty} x(t), \limsup_{t\to+\infty} x(t)]$ and $[\liminf_{t\to+\infty} y(t), \limsup_{t\to+\infty} y(t)]$. By continuity of the trajectory and ϕ, the maps $\phi \circ x$ and $\phi \circ y$ converge. Recalling that (x, y) is bounded, we obtain from Proposition 5.6

$$\lim_{t\to+\infty} \dot{x}(t) = \lim_{t\to+\infty} \dot{y}(t) = 0, \tag{6.5}$$

which proves (ii). We now prove the convergence of the map $x - y$.

LEMMA 6.1. *The map $x - y$ converges.*

Proof of Lemma 6.1. Assume that the map $x - y$ does not converge, i.e., $\liminf_{t\to+\infty} x(t) - y(t) < \limsup_{t\to+\infty} x(t) - y(t)$. Since the map V' is continuous, there exists some $z \in \big(\liminf_{t\to+\infty} x(t) - y(t), \limsup_{t\to+\infty} x(t) - y(t)\big)$ such that V' keeps a constant sign on a neighborhood of z (write $[a, b] = V'^{-1}(\{0\}) \cup V'^{-1}((0, +\infty)) \cup V'^{-1}((-\infty, 0))$). Without loss of generality, we assume that it is nonpositive. Consider $\alpha > 0$ such that

$$[z - 4\alpha, z + 4\alpha] \subset \left[\liminf_{t\to+\infty} x(t) - y(t), \limsup_{t\to+\infty} x(t) - y(t)\right], \tag{6.6}$$

$$V'(z') \leq 0 \text{ for every } z' \in [z - 4\alpha, z + 4\alpha]. \tag{6.7}$$

Consider $T \geq 0$ such that, for every $t \geq T$,

$$x(t) \in \left[\liminf_{s\to+\infty} x(s) - \alpha, \limsup_{s\to+\infty} x(s) + \alpha\right], \tag{6.8}$$

$$y(t) \in \left[\liminf_{s\to+\infty} y(s) - \alpha, \limsup_{s\to+\infty} y(s) + \alpha\right],$$

$$|\dot{x}(t)| \leq \gamma\alpha, \tag{6.9}$$

$$|\dot{y}(t)| \leq \gamma\alpha.$$

Consider $t' \geq t \geq T$ such that

$$(x - y)\,([t, t']) \subset [z - 4\alpha, z + 4\alpha], \tag{6.10}$$

$$x(t) - y(t) = z + 4\alpha, \tag{6.11}$$

$$x(t') - y(t') = z - 4\alpha. \tag{6.12}$$

Let us now show that

$$x(t') \geq x(t) - 3\alpha. \tag{6.13}$$

In view of (6.6), (6.7), and (6.10), for every $u \in [t, t']$, we have $V'(x(u) - y(u)) \leq 0$; hence $\ddot{x}(u) + \gamma\dot{x}(u) + \phi'(x(u)) \geq 0$.

First assume that $x([t, t']) \subset \left[\liminf_{s\to+\infty} x(s), \limsup_{s\to+\infty} x(s)\right]$. Then, for every $u \in [t, t']$, $\phi'(x(u)) = 0$; hence $\ddot{x}(u) + \gamma\dot{x}(u) \geq 0$. Integrating on $[t, t']$, we deduce $x(t') \geq x(t) + \frac{1}{\gamma}(\dot{x}(t) - \dot{x}(t'))$. In view of (6.9), $x(t') \geq x(t) - 2\alpha$.

Now assume that $x(u) > \limsup_{s\to+\infty} x(s)$ for some $u \in [t, t']$. If we have $x(t') \geq \limsup_{s\to+\infty} x(s)$, then $x(t') \geq x(t) - \alpha$ from (6.8). On the other hand, if $x(t') < \limsup_{s\to+\infty} x(s)$, then there exists $\tau \in [t, t']$ such that $x(\tau) = \limsup_{s\to+\infty} x(s)$ and $x([\tau, t']) \subset \left[\liminf_{s\to+\infty} x(s), \limsup_{s\to+\infty} x(s)\right]$, assuming without loss of generality that $2\alpha < \limsup_{s\to+\infty} x(s) - \liminf_{s\to+\infty} x(s)$, hence that $x([t, t'])$ cannot meet both $\limsup_{s\to+\infty} x(s)$ and $\liminf_{s\to+\infty} x(s)$. Then again $x(t') \geq x(\tau) - 2\alpha = \limsup_{s\to+\infty} x(s) - 2\alpha \geq x(t) - 3\alpha$ in view of (6.8).

Finally, assume that $x(u) < \liminf_{s\to+\infty} x(s)$ for some $u \in [t, t']$. If $x(t) \leq \liminf_{s\to+\infty} x(s)$, then $x(t') \geq x(t) - \alpha$ from (6.8). If $x(t) > \liminf_{s\to+\infty} x(s)$, then $x([t, \tau]) \subset \left[\liminf_{s\to+\infty} x(s), \limsup_{s\to+\infty} x(s)\right]$ and $x(\tau) = \liminf_{s\to+\infty} x(s)$ for some $\tau \in [t, t']$. Again $x(\tau) \geq x(t) - 2\alpha$, and, in view of (6.8), $x(t') \geq \liminf_{s\to+\infty} x(s) - \alpha = x(\tau) - \alpha \geq x(t) - 3\alpha$.

In the same way

$$y(t') \leq y(t) + 3\alpha.$$

In view of (6.13), $x(t') - y(t') \geq x(t) - y(t) - 6\alpha$. From (6.11) and (6.12), we deduce $z - 4\alpha \geq z + 4\alpha - 6\alpha$, a contradiction with $\alpha > 0$. Hence the map $x - y$ converges and its limit z_∞ is finite since (x, y) is bounded. □

Let us now prove that the map x converges. Assume that it is not true, i.e., $\liminf_{t\to+\infty} x(t) < \limsup_{t\to+\infty} x(t)$. The remainder of the proof is an adaptation of the proof of Theorem 2.4, Part 2, (i). Note that $\liminf_{t\to+\infty} y(t) = \liminf_{t\to+\infty} x(t) - z_\infty$ and $\limsup_{t\to+\infty} y(t) = \limsup_{t\to+\infty} x(t) - z_\infty$. Replacing x by $\min\{x, y + z_\infty\}$ (resp., $\max\{x, y + z_\infty\}$ in (6.3) (resp., (6.4)), we obtain two sequences (t_n) and (t'_n) in $\mathbb{R}_+$ such that $t_n \leq t'_n$, $\lim_{n\to+\infty} t_n = \lim_{n\to+\infty} t'_n = +\infty$, and

$$\begin{aligned}
x([t_n, t'_n]) &\subset \left[\liminf_{t\to+\infty} x(t), \limsup_{t\to+\infty} x(t)\right], \\
y([t_n, t'_n]) &\subset \left[\liminf_{t\to+\infty} y(t), \limsup_{t\to+\infty} y(t)\right], \\
\lim_{n\to+\infty} x(t_n) &= \lim_{n\to+\infty} y(t_n) + z_\infty = \limsup_{t\to+\infty} x(t), \\
\lim_{n\to+\infty} x(t'_n) &= \lim_{n\to+\infty} y(t'_n) + z_\infty = \liminf_{t\to+\infty} x(t).
\end{aligned}$$

Then, in view of Theorem 2.2, Part 1, $\phi'(x(u)) = \phi'(y(u)) = 0$ for every $u \in [t_n, t'_n]$. Recall (5.1):

$$\ddot{x}(u) + \ddot{y}(u) + \gamma(\dot{x}(u) + \dot{y}(u)) + \phi'(x(u)) + \phi'(y(u)) = 0;$$

hence

$$\ddot{x}(u) + \ddot{y}(u) + \gamma(\dot{x}(u) + \dot{y}(u)) = 0.$$

Integrating the previous equation on $[t_n, t'_n]$, we obtain

$$x(t'_n) + y(t'_n) + \gamma(\dot{x}(t'_n) + \dot{y}(t'_n)) = x(t_n) + y(t_n) + \gamma(\dot{x}(t_n) + \dot{y}(t_n)).$$

We have $\lim_{n\to+\infty} \dot{x}(t'_n) + \dot{y}(t'_n) = \lim_{n\to+\infty} \dot{x}(t_n) + \dot{y}(t_n) = 0$, from above or from Lemma 5.1. Taking the limit in the above equation when $n \to +\infty$, we deduce $\liminf_{t\to+\infty} x(t) = \limsup_{t\to+\infty} x(t)$, a contradiction. Hence the map x converges, and the very final argument in section 6.1.2 permits us to end the proof. □

6.2.2. Proof in the case $x - y \to 0$. In that case, we prove only the convergence of the trajectory. We may not have $\lim_{t\to+\infty} \dot{x}(t) = \lim_{t\to+\infty} \dot{y}(t) = 0$, but, in view of Lemma 5.1, $\lim_{t\to+\infty} \dot{x}(t) + \dot{y}(t) = 0$. Then the above proof remains valid, with $z_\infty = 0$. □

6.3. Proof of Theorem 2.6. The proofs of Theorem 2.6 and Corollary 2.7 are almost identical to their regular version. This is not surprising if we recall the assumption $\lim_{t\to+\infty} x(t) \neq \lim_{t\to+\infty} y(t)$ in the regular case, which thus makes the regular case analogous to a singular case (see section 7.2). Without loss of generality, we assume $y_0 < x_0$; hence $y(t) < x(t)$ for every t. Let us first consider the case where $y_\infty > -\infty$ and $x_\infty < +\infty$. Without loss of generality, we prove only the assertion $P^+(x_\infty)$. Let us argue by contradiction and assume that there exists a neighborhood $V(x_\infty)$ of x_∞ such that $\phi'_{|V(x_\infty)} \leq 0$. There exists $t_0 \geq 0$ such that, for all $t \geq t_0$, $x(t) \in V(x_\infty)$; hence

$$\forall t \geq t_0, \quad \phi'(x(t)) \leq 0. \tag{6.14}$$

Since, by assumption, $y_\infty > -\infty$ and $x_\infty < +\infty$, there exists $M > 0$ such that, for all $t \geq t_0$, $0 < x(t) - y(t) \leq M$. From $(\mathcal{H}_V)$(iv) and $(\mathcal{H}_V)$(v), there exists $\eta > 0$ such that

$$\forall t \geq t_0, \quad V'(x(t) - y(t)) \leq -\inf_{0<|z|\leq M} |V'(z)| = -\eta. \tag{6.15}$$

By using the first equation of (HBFC^2_{sing}), (6.14), and (6.15), we find

$$\forall t \geq t_0, \ \ddot{x}(t) + \gamma\dot{x}(t) \geq \eta\varepsilon(t).$$

Integrating the above differential inequality between t_0 and t, we find $\dot{x}(t) - \dot{x}(t_0) + \gamma(x(t) - x(t_0)) \geq \eta \int_{t_0}^{t} \varepsilon(u)du$. But the trajectory x is convergent, hence bounded, and $\int_{t_0}^{+\infty} \varepsilon(u)du = +\infty$. Hence $\lim_{t\to+\infty} \dot{x}(t) = +\infty$, a contradiction.

Let us now assume that $y_\infty = -\infty$ and $x_\infty < +\infty$. Let us argue by contradiction and assume that we have neither $P^+(y_\infty)$ nor $P^+(x_\infty)$. Then there exists a neighborhood $V(x_\infty)$ (resp., $V(y_\infty)$) of x_∞ (resp., y_∞) such that $\phi'_{|V(x_\infty)} \leq 0$ (resp., $\phi'_{|V(y_\infty)} \leq 0$). Consequently, there exists $t_0 \geq 0$ such that, for all $t \geq t_0$,

$$\phi'(x(t)) \leq 0 \quad \text{and} \quad \phi'(y(t)) \leq 0.$$

By adding the two equations of (HBFC^2_{sing}), we deduce

$$\forall t \geq t_0, \ \ddot{x}(t) + \ddot{y}(t) + \gamma(\dot{x}(t) + \dot{y}(t)) \geq 0.$$

A direct computation shows that $x + y$ is bounded from below, a contradiction with $y_\infty = -\infty$. The proof of the last case $y_\infty > -\infty$ and $x_\infty = +\infty$ goes along the same lines. □

6.4. Proof of Corollary 2.7. Let $I(x_\infty)$ be the connected component of x_∞ in the set $\widehat{S}$. Let us argue by contradiction and assume that $x_\infty \notin \{\inf I(x_\infty), \sup I(x_\infty)\}$. In particular, we have $x_\infty < +\infty$. If condition (a) is satisfied, we obtain, by Theorem 2.6, the assertions $P^-(y_\infty)$ and $P^+(x_\infty)$. If condition (b) is satisfied, the assertion $P^+(y_\infty)$ is false and then, by Theorem 2.6, we obtain the assertion $P^+(x_\infty)$. In both cases, we have $P^+(x_\infty)$, a contradiction with $x_\infty \in \operatorname{int} I(x_\infty)$. □

6.5. Proof of Corollary 2.8. Since $\widehat{S}$ is connected, $I(x_\infty) = I(y_\infty) = \widehat{S}$. From Corollary 2.7, $x_\infty \in \{\inf \widehat{S}, \sup \widehat{S}\}$ and $y_\infty \in \{\inf \widehat{S}, \sup \widehat{S}\}$. If the set $\widehat{S}$ is reduced to a singleton, there is nothing more to prove. So let us assume that $\inf \widehat{S} < \sup \widehat{S}$. Let us argue by contradiction and assume that $(x_\infty, y_\infty) = (\inf \widehat{S}, \inf \widehat{S})$ (resp., $(x_\infty, y_\infty) = (\sup \widehat{S}, \sup \widehat{S})$). From Theorem 2.6 we deduce that the assertion $P^+(\inf \widehat{S})$ (resp., $P^-(\sup \widehat{S})$) holds. On the other hand, it is immediate to verify that $\phi' \le 0$ (resp., $\phi' \ge 0$) in a neighborhood of $\inf \widehat{S}$ (resp., $\sup \widehat{S}$), which contradicts $P^+(\inf \widehat{S})$ (resp., $P^-(\sup \widehat{S})$). Hence, we have $(x_\infty, y_\infty) = (\inf \widehat{S}, \sup \widehat{S})$ or $(x_\infty, y_\infty) = (\sup \widehat{S}, \inf \widehat{S})$. □

7. Singular versus regular: Higher dimension and dimension one. From a physical point of view, it is clearly necessary to consider singular interaction potentials that naturally appear in various models as the classical gravitation and electromagnetism and also in molecular models. Even if models in dimension one do exist and are of interest in physics, this motivation is sufficient to examine at least the dimension three. In optimization, the advantage of a singular potential is a priori less obvious. In the next section, we indicate what we conjecture on the behavior of our coupled oscillators in higher dimension in both regular and singular cases. In the following section, we show that in dimension one the singular system can be seen as the limit case of regular systems.

7.1. Properties in higher dimension. We consider the coupled system

$$(\mathrm{HBFC}^p) \qquad \begin{cases} \ddot{x_1} + & \gamma \dot{x_1} + & \nabla\phi(x_1) + & \varepsilon(t)\frac{\partial U}{\partial x_1}(x_1, \dots, x_p) = & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \ddot{x_p} + & \gamma \dot{x_p} + & \nabla\phi(x_p) + & \varepsilon(t)\frac{\partial U}{\partial x_n}(x_1, \dots, x_p) = & 0 \end{cases}$$

with initial conditions ϕ and ε as in Hypothesis 1, Ω an open subset of H^p, $U : \Omega \to \mathbb{R}_+$ a potential such that $\operatorname{div} U = 0$, $\lim_{x \to \operatorname{bd}\Omega} U(x) = +\infty$. The existence part corresponding to Proposition 2.1 clearly holds.

For $p = 2$ and $\dim H = 1$, Corollary 2.9 shows that the corresponding system will (asymptotically) maximize the diameter $\operatorname{diam}(S)$. We know (Remark 2.7) that the statement of Corollary 2.9 does not remain true in dimension greater than 2, but we conjecture that under suitable assumptions on ϕ and V, say, for example, ϕ convex, the trajectory $(x_1, \dots, x_p)$, the solution of (HBFC^p), will converge to some $(x_{1,\infty}, \dots, x_{p,\infty})$ in Ω that shall minimize (at least locally) the interaction potential U on the solution set. The whole difference between the singular and regular version will reside in Ω, equal to the whole space H^p in the regular case. Assuming that the limit $(x_{1,\infty}, \dots, x_{p,\infty})$ will belong to Ω helps to avoid a priori uninteresting limits (as $x_{1,\infty} = \dots = x_{p,\infty}$).

7.1.1. Numerical experiments. Figure 1 illustrates our conjecture for $N = \dim H = 2$, $p = 4$, $\phi(a, b) = \max\{5(\sqrt{a^2/4 + b^2} - 1)^2, 0\}$, $\varepsilon(t) = 1/\log(t + 2)$, $\gamma = 2$. Other tests for different potentials ϕ, V, different controls ε, and mainly different

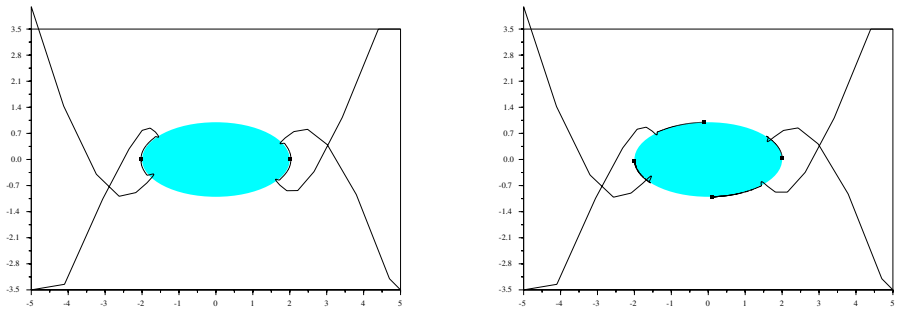


FIG. 1.

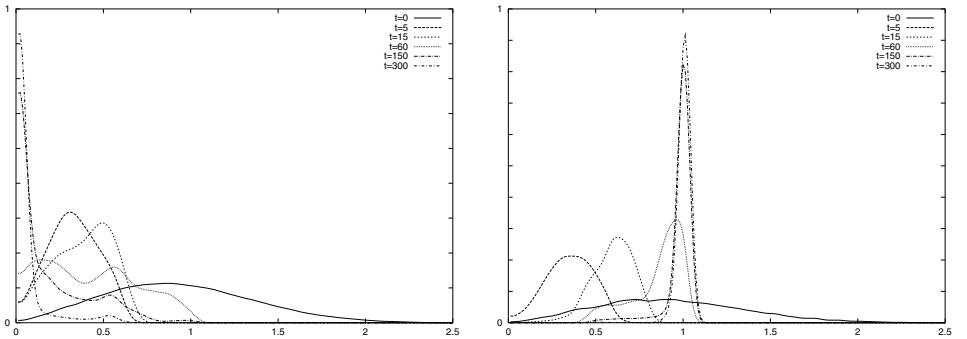


FIG. 2.

numbers of *systems* ($p = 3, 4, 5$ up to now) and different dimensions ($N = 2, 3$ up to now) show the same behavior. We give below the trajectories of the corresponding (HBFC^4_{reg}) system, with $V(z) = \exp(-|z|^2/50)$ and $U(x_1, \ldots, x_4) = \sum_{i<j} V(x_i - x_j)$, on the left, and of the corresponding (HBFC^4_{sing}) system, with $V(z) = 1/|z|^2$ and $U(x_1, \ldots, x_4) = \sum_{i<j} V(x_i - x_j)$, on the right.[9]

The numerical experiment that we now describe highlights the nice asymptotic properties of the (HBFC^p_{sing}) system. Taking the previous systems, Figure 2 illustrates the convergence of the area $\mathcal{A}rea(\mathrm{co}\{x_1(t), \ldots, x_4(t)\})$, up to a normalization factor[10] both in the regular case (on the left) and in the singular case (on the right). It precisely consists of evaluating the distribution of the function $\mathcal{A}rea(\mathrm{co}\{x_1(t), \ldots, x_4(t)\})$. For initial data $(x_1(0), \ldots, x_4(0), \dot{x}_1(0), \ldots, \dot{0}_4(t))$ in the set $([-5, 5] \times [-5, 5])^4 \times \{(0, 0)\}^4$ and for different times t, we compute the function $\mathcal{A}rea(\mathrm{co}\{x_1(t), \ldots, x_4(t)\})$ and, for a given number $n > 0$, the proportion of points which belong to an interval $[\frac{k}{n}, \frac{k+1}{n})$ for $n \in \mathbf{Z}$. The experiments are computed on a grid of $11^8 = 214\,358\,881$ points, and we limit the representation at $t = 300$ for a matter of readability. They indicate the likeliness of our conjecture and of the good asymptotic behavior of the (HBFC^p_{sing}) system.

[9] In both figures, the initial conditions are $x_1(0) = (5, -3.5)$, $\dot{x}_1(0) = (-1, 1)$, $x_2(0) = (-5, 3.5)$, $\dot{x}_2(0) = (0, 2)$, $x_3(0) = (-5, -3.5)$, $\dot{x}_3(0) = (3, 0.5)$, $x_4(0) = (5, 3.5)$, $\dot{x}_4(0) = (-2, 0)$.

[10] Precisely, the normalization factor is the area of the largest polygon included in S with (at most) four extremal points.

7.2. The singular case as a limit of regular cases in dimension one. When the potential V is defined on the whole line $\mathbb{R}$ (regular case), the convergence results are not as precise as in the singular case. Namely, the two trajectories x and y may have the same limit. Precisely, we recall the following result from [6].

THEOREM 7.1 (slow parametrization in the regular case [6]). *Under the assumptions of Corollary* 2.8, *let* $V : \mathbb{R} \to \mathbb{R}$ *be a (regular) repulsion potential, i.e., a map of class* $\mathcal{C}^1$ *such that*

$$(\mathcal{H}_{V_{reg}}) \qquad \begin{cases} \text{the map } \nabla V \text{ is locally Lipschitz continuous on } \mathbb{R} \\ \forall z \in \mathbb{R} \setminus \{0\}, \quad zV'(z) < 0. \end{cases}$$

Then there is a unique maximal solution $(x, y) : [0, +\infty) \to \mathbb{R} \times \mathbb{R}$ *of the* (HBFC^2_{reg}) *system*

$$(7.1) \qquad (\mathrm{HBFC}^2_{reg}) \qquad \begin{cases} \ddot{x} + \gamma\dot{x} + \nabla\phi(x) + \varepsilon(t)\nabla V(x-y) = 0, \\ \ddot{y} + \gamma\dot{y} + \nabla\phi(y) - \varepsilon(t)\nabla V(x-y) = 0, \\ \Big(x(0), y(0), \dot{x}(0), \dot{y}(0)\Big) = \Big(x_0, y_0, \dot{x}_0, \dot{y}_0\Big). \end{cases}$$

Moreover, the solution (x, y) *satisfies one of the following cases:*

(i) $\lim_{t\to+\infty} \Big(x(t)\,,\, y(t) \Big) = \Big(\sup \widehat{S}\,,\, \inf \widehat{S} \Big)$;

(ii) $\lim_{t\to+\infty} \Big(x(t)\,,\, y(t) \Big) = \Big(\inf \widehat{S}\,,\, \sup \widehat{S} \Big)$;

(iii) *there exists* $x_\infty \in \widehat{S}$ *such that* $\lim_{t\to+\infty} x(t) = \lim_{t\to+\infty} y(t) = x_\infty$.

It is then a natural question to know if the singular case can be seen as a limit of regular cases or if it corresponds to a singularity. If the (singular) repulsion potential $V : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ is a limit of (regular) potentials $V_n : \mathbb{R} \to \mathbb{R}$, how does the solution (x_n, y_n) of the corresponding $(\mathrm{HBFC}^2_{reg}(n))$ systems behave when $n \to +\infty$? More than a theoretical interest, the importance of this problem for numerical applications comes from an obvious remark: in numerical applications, a singular potential would be approximated by regular potentials. The general study of these questions is beyond the scope of this paper, but we give an indication of a positive answer with the following result, where the potential V_n corresponds to a truncated V.

PROPOSITION 7.2. *Under the assumptions of Corollary* 2.8, *let* (V_n) *be a sequence of regular repulsion potentials, i.e., for every* n, $V_n : \mathbb{R} \to \mathbb{R}_+$ *is a map of class* $\mathcal{C}^1$ *satisfying* $(\mathcal{H}_{V_{reg}})$, *and let* (α_n) *be a sequence in* $\mathbb{R}_+ \setminus \{0\}$ *such that* $\lim_{n\to+\infty} \alpha_n = 0$. *Assume that, for every* n,

$$V_n(x) = V(x) \text{ for every } x \notin (-\alpha_n, \alpha_n).$$

Assume that $\widehat{S}$ *is not reduced to a singleton. Then, for every initial condition* $(x_0, y_0, \dot{x}_0, \dot{y}_0) \in \mathbb{R}^4$, *such that* $x_0 \neq y_0$, *there is an integer* N *such that, for every* $n \geq N$, *the solution* (x_n, y_n) *of*

$$(7.2) \qquad (\mathrm{HBFC}^2_{reg}(n)) \qquad \begin{cases} \ddot{x} + \gamma\dot{x} + \nabla\phi(x) + \varepsilon(t)\nabla V_n(x-y) = 0, \\ \ddot{y} + \gamma\dot{y} + \nabla\phi(y) - \varepsilon(t)\nabla V_n(x-y) = 0, \\ \Big(x(0), y(0), \dot{x}(0), \dot{y}(0)\Big) = \Big(x_0, y_0, \dot{x}_0, \dot{y}_0\Big) \end{cases}$$

satisfies

$$(x_n, y_n) = (x, y),$$

where (x, y) *is the solution of the* (HBFC$^2_{sing}$) *system corresponding to* V*, with initial condition* $(x_0, y_0, \dot{x}_0, \dot{y}_0)$*. Hence*

$$\lim_{t\to+\infty} \Big(x_n(t)\,,\, y_n(t) \Big) = \Big(\sup \widehat{S}\,,\, \inf \widehat{S} \Big) \quad \text{if} \quad y_0 < x_0;$$

$$\lim_{t\to+\infty} \Big(x_n(t)\,,\, y_n(t) \Big) = \Big(\inf \widehat{S}\,,\, \sup \widehat{S} \Big) \quad \text{if} \quad x_0 < y_0.$$

Proof of Proposition 7.2. Let (x, y) be the solution of the (HBFC$^2_{sing}$) system corresponding to V, with initial condition $(x_0, y_0, \dot{x}_0, \dot{y}_0)$. Without loss of generality, assume that $x_0 > y_0$. From Corollary 2.8, $\lim_{t\to+\infty}(\, x(t)\,,\, y(t)\,) = (\, \sup \widehat{S}\,,\, \inf \widehat{S}\,)$. Since $x(t) > y(t)$ for every t, we obtain that $\inf_{t\geq 0}(x(t) - y(t)) > 0$. Let $N \in \mathbb{N}$ such that, for every $n \geq N$, $\alpha_n < \inf_{t\geq 0}(x(t) - y(t))$. Then (x, y) is a solution of the (HBFC$^2_{reg}(n)$) system, with initial condition $(x_0, y_0, \dot{x}_0, \dot{y}_0)$; hence $(x_n, y_n) = (x, y)$. □

7.3. Remark on the convergence in the one-dimensional and regular case. As we mentioned after stating Theorem 2.4, we can improve the one-dimensional and regular convergence result [6, Theorem 2.2] when the trajectory is assumed to be bounded. The following theorem states this precisely.

THEOREM 7.3 (convergence of the solutions, regular case). *Assume that* ϕ *and* V *are maps from* $\mathbb{R}$ *to* $\mathbb{R}$ *of class* $\mathcal{C}^1$ *and that their gradients are locally Lipschitz continuous. Assume that the solution* (x, y) *of the* (HBFC$^2_{reg}$) *system is bounded. Then it converges; precisely,*

(i) *there exists* $(\, x_\infty, y_\infty\,) \in \widehat{S} \times \widehat{S}$ *such that* $\lim_{t\to+\infty} \Big(x(t)\,,\, y(t) \Big) = \Big(x_\infty\,,\, y_\infty \Big)$.

We let the reader check that the proof in the singular case can be easily adapted (with a simplification since one already has $\lim_{t\to+\infty} \dot{x}(t) = \lim_{t\to+\infty} \dot{y}(t) = 0$ in the regular case).

Acknowledgments. Many thanks to Alexandre Cabot for his great help on the numerical experiments and valuable discussions. I'm grateful to an anonymous referee who raised the question of the possible nonconvergence of the trajectories of (HBFC$^2_{sing}$) and, consequently, pushed me to improve the one-dimensional convergence results (Theorem 2.4, Part 1, Remarks 3.8 and 3.9) in the nonrepulsive case.

REFERENCES

[1] F. ALVAREZ, *On the minimizing property of a second order dissipative system in Hilbert spaces*, SIAM J. Control Optim., 38 (2000), pp. 1102–1119.

[2] H. ATTOUCH AND R. COMINETTI, *A dynamical approach to convex minimization coupling approximation with the steepest descent method*, J. Differential Equations, 128 (1996), pp. 519–540.

[3] H. ATTOUCH AND M.-O. CZARNECKI, *Asymptotic control and stabilization of nonlinear oscillators with non isolated equilibria*, J. Differential Equations, 179 (2002), pp. 278–310.

[4] H. ATTOUCH, X. GOUDOU, AND P. REDONT, *The heavy ball with friction method.* I. *The continuous dynamical system*, Commun. Contemp. Math., 2 (2000), pp. 1–34.

[5] R.E. BRUCK, *Asymptotic convergence of nonlinear contraction semigroups in Hilbert space*, J. Funct. Anal., 18 (1975), pp. 15–26.

[6] A. Cabot and M.-O. Czarnecki, *Asymptotic control of pairs of oscillators coupled by a repulsion, with nonisolated equilibria* I: *The regular case*, SIAM J. Control Optim., 41 (2002), pp. 1254–1280.

[7] X. Goudou, *Genericity of the Convergence Towards a Local Minimum of the Heavy Ball Method*, manuscript.

[8] A. Haraux and M.A. Jendoubi, *Convergence of bounded weak solutions of the wave equation with dissipation and analytic nonlinearity*, Calc. Var. Partial Differential Equations, 9 (1999), pp. 95–124.

[9] M.A. Jendoubi, *Convergence of global and bounded solutions of the wave equation with linear dissipation and analytic nonlinearity*, J. Differential Equations, 144 (1998), pp. 302–312.

[10] M.A. Jendoubi and P. Poláčik, *Nonstabilizing solutions of semilinear hyperbolic and elliptic equations with damping*, Proc. Roy. Soc. Edinburgh Sect. A, 133 (2003), pp. 1137–1153.

[11] P. Redont, *Equation de la boule pesante avec frottement: exemple de solution non convergente*, Prépublication 99, Département de Mathématiques, Université de Montpellier II, http://www.math.univ-montp2.fr.

[12] A.N. Tikhonov and V. Ya. Arsenine, *Méthodes de résolution de problèmes mal posés*, MIR, Moscow, 1976.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 2172–2199

ROBUST POINT STABILIZATION OF UNDERACTUATED MECHANICAL SYSTEMS VIA THE EXTENDED CHAINED FORM*

DAVID A. LIZÁRRAGA†, NNAEDOZIE P. I. ANEKE‡, AND HENK NIJMEIJER§

Abstract. This paper addresses point stabilization for the extended chained form (ECF), a control system that may be used to model a number of mechanical underactuated systems. A control law is proposed, based on well-known hybrid open-loop/feedback techniques, which exponentially stabilizes the origin of a dynamic extension of the ECF and ensures a degree of robustness to additive disturbance terms that may represent, for instance, model uncertainties. Numerical simulations are included to illustrate the performance of the presented stabilizers.

Key words. extended chained form, second-order chained form, point stabilization, hybrid feedback, underactuated manipulator, surface vessel

AMS subject classifications. 93D21, 93C10

DOI. 10.1137/S0363012902405571

1. Introduction.

The study of mechanical control systems with fewer actuators than degrees of freedom constitutes a stimulating and active subject of research. Examples of such systems include underactuated manipulators [21], underactuated (surface and underwater) maritime vehicles [30, 5], underactuated spacecraft [18], and mechanical systems with internal degrees of freedom subject to virtual holonomic constraints [15, 24]. Besides the study of properties such as accessibility and controllability, the research efforts have focused mainly on problems such as open-loop steering from one configuration to another, trajectory tracking, and stabilization to an equilibrium point (or configuration). For underactuated mechanical systems, the latter problem is especially challenging since such systems typically do not meet Brockett's necessary condition for stabilization to a point by continuous, pure-state feedback [3]. As a consequence, solutions usually involve elaborate control techniques, such as time-varying feedback or hybrid control. In this paper we are particularly interested in stabilization to a point.

A valuable tool when addressing control problems is the possibility of transforming the system dynamics, via coordinate change and feedback, into a "canonical" control system with a simpler, more tractable structure. Among such canonical representations, the *extended chained form* (ECF)

$$\begin{aligned} \ddot{x}_1 &= u_1, \\ \ddot{x}_2 &= u_2, \\ \ddot{x}_3 &= u_1 x_2 \end{aligned} \tag{1}$$

*Received by the editors April 15, 2002; accepted for publication (in revised form) July 17, 2003; published electronically April 7, 2004. This work was carried out while the first two authors were at the Department of Mechanical Engineering, Technische Universiteit Eindhoven, as postdoctoral fellow and Ph.D. student, respectively.

http://www.siam.org/journals/sicon/42-6/40557.html

†Departamento de Matemáticas Aplicadas y Sistemas Computacionales, Instituto Potosino de Investigación Científica y Tecnológica, Apartado Postal 3-74 Tangamanga, San Luis Potosí, S.L.P., México (D.Lizarraga@ipicyt.edu.mx).

‡Alternative Powertrains, Vehicle Electronics and Controls Group, Ford Forschungszentrum Aachen GmbH, Suesterfeldstrasse 200, D-52072 Aachen, Germany (naneke@ford.com).

§Department of Mechanical Engineering, Technische Universiteit Eindhoven, Den Dolech 2, Postbus 513, 5600 MB Eindhoven, The Netherlands (H.Nijmeijer@tue.nl).

plays, for some underactuated mechanical systems, a role similar to the one played by the *chained form* for driftless nonholonomic systems (cf. [19, 26]). By slight abuse of nomenclature we are calling the particular system (1), with six state variables and two inputs, the ECF, although more general extensions to the chained form can be envisaged and have been considered. System (1) has also been termed *second-order chained form*. However, no definitive unifying notation seems to exist as yet for the family of "chained systems." In [11], for instance, a two-input control system is introduced which is referred to as an *n-dimensional, high-order generalized chained system*. On the other hand, chained systems having more than two inputs have also been studied under the denomination *multi-input chained systems*, e.g., in [29]. Finally, the reader should be aware that in some references—but not in the present paper—the term *extended chained form* refers to a driftless chained system, as introduced in [19], with integrators added in cascade to each of its inputs, cf., e.g., [31].

The ECF made its appearance in the context of underactuated mechanical systems in [4], where it was shown that the dynamic model of a simplified underwater vehicle is feedback-equivalent to two interconnected ECFs. In [8], the model of a planar PP$\overline{\text{R}}$ manipulator was directly transformed into the ECF (PP$\overline{\text{R}}$ denotes a manipulator with two *prismatic* joints and one *revolute* joint at its most distal end; the bar above "R" designates an *unactuated* or *passive* joint). Among the two-input, three-DOF systems that are feedback-equivalent to the ECF one finds the planar, vertical take-off and landing (VTOL) system in the absence of gravity [25], a simplified underwater vehicle [22], the planar, serial-drive RR$\overline{\text{R}}$ manipulator [32], and the planar, parallel-drive RRR manipulators with any two joints actuated. Two additional examples are multibody systems possessing an unactuated, internal DOF which is required, by design, to satisfy a virtual holonomic constraint, namely the rigid body with internal DOF in [15] and the dynamics of the spring-coupled, third link of a planar PPR manipulator in [24]. It is worth noting that the transformations involved in these examples allow one to map generic equilibrium configurations of the mechanical system to *the origin* of the ECF, thereby reducing stabilization of any such configuration to stabilization of the latter point.

In view of these results, considerable emphasis has been given to the design of controllers for the ECF and some of its generalizations. For instance, a time-varying controller, updated in terms of the state only at isolated time-instants, was developed in [4] to achieve a "discrete-time" version of $\mathcal{K}$-exponential stability for the origin of two interconnected ECFs. Tracking controllers were proposed in [8] which, associated with carefully selected state trajectories (cf. also [32]), exponentially drive the state of the ECF towards the origin. In [11], discontinuous controllers were introduced to *almost-exponentially* stabilize the origin of two-input, generalized, n-dimensional chained form systems, including the ECF. More recently, the authors of [6] pointed out conditions for two-input systems with drift to be feedback-linearizable by non-smooth (and eventually discontinuous) state and input transformations. Once such a transformation is applied, linear controllers can be used to drive the system state exponentially to the origin, provided the initial conditions belong to a set where the new coordinates are well defined. In [1] a time-varying, continuous, homogeneous control-law was introduced which, to date and to the extent of our knowledge, is the only one capable of ensuring Lyapunov-stability as well as exponential convergence (indeed $\mathcal{K}$-exponential stability) for the origin of the ECF.

In this paper we propose controllers that are both *stabilizing* and *robust*—in appropriately defined senses—based on a well-known *hybrid open-loop/feedback* approach (also known as *iterative state steering*). Essentially, this goes along the lines

of *discrete-time* control of *continuous-time* systems: at a given sample instant t_k the state $x(t_k)$ is sensed, and an input function $t \mapsto u_k(t)$ is computed and used to drive the system until the next sample instant t_{k+1}. Within the interval (t_k, t_{k+1}) the input may change with t, but it is *independent* of the instantaneous value of the state $x(t)$. The input u_k is designed so that, at the end of the interval, the state $x(t_{k+1})$ is "closer" to the origin than it was at the beginning. This control algorithm is iterated indefinitely and, under appropriate assumptions, it leads to a robustly stable equilibrium point. Let us remark that the use of iterated control is not new and that important results have been reported in the literature. One example is [4], mentioned above, where iterated controls were developed, but where no robustness study was carried out. A hybrid control combining sampled-time control with continuous-time, linear feedback was proposed in [20] to stabilize chained form systems, with applications to wheeled mobile robots. Among the earliest references addressing the *robustness* of time-varying, iterative control in the framework of nonholonomic systems one finds [2], where control laws are developed for the three- and four-dimensional chained forms. These feedback laws render the origin exponentially stable (in the discrete-time sense) and this stability property is preserved in the presence of additive disturbance vector fields. The authors of [14] consider a large class of systems, possibly with drift, under iterative state steering control. Although no algorithm is presented to construct any such controller—it is assumed that one is known beforehand—conditions are pointed out for discrete-time stability of the origin and robustness to the presence of additive disturbance vector fields. A drawback of the reported conditions for robustness is that some of them are stated in terms of *the flow* of the disturbance vector field(s), thus limiting the class of disturbances for which robustness can be assessed in practice. For driftless systems, a powerful approach was presented in [17], where a constructive algorithm is given to design stabilizers for *any* driftless, analytic, controllable system. The controllers thus obtained guarantee local exponential stability of the origin for a *dynamic extension* of the original system, and the stability is robust to additive disturbance vector fields. Our controller design and methodology share similarities with [4] and [14], although the stability and robustness analysis is inspired by [17]. The presence of a drift term, however, makes the analysis—and the eventual generalization of the present approach to a larger class of systems—more difficult. As a consequence, our result is merely applicable to a class of systems which can be represented as a (perturbed) ECF.

This paper is organized as follows. Section 2 contains definitions of stability and robustness, as used in the present context, as well as a statement of the robust stabilization problem. In section 3, a feedback law is introduced and then shown to be a robust stabilizer in the specific sense considered here. Section 4 contains two simulation examples. Some concluding remarks are given in section 5. Finally, in the appendix, notational conventions are fixed and some technical lemmas are stated or proved.

2. Preliminaries and definition of the problem. Prior to stating the problem, let us precisely define the notions of stability and robustness used in this context. To this end consider the ECF, regarded as the *nominal system*, rewritten in the form

$$\dot{x} = b_0(x) + u_1 b_1(x) + u_2 b_2(x),$$

with

$$b_0(x) = x_2 \frac{\partial}{\partial x_1} + x_4 \frac{\partial}{\partial x_3} + x_6 \frac{\partial}{\partial x_5}, \quad b_1(x) = \frac{\partial}{\partial x_2} + x_3 \frac{\partial}{\partial x_6}, \quad b_2(x) = \frac{\partial}{\partial x_4}. \tag{2}$$

As a result of model errors, such as parameter uncertainties, disturbance vector fields may be present in the system to be actually controlled, and one way to model this is by considering the *perturbed system*

$$\dot{x} = b_0(x) + h_0(x,\varepsilon) + \sum_{i=1}^{2} u_i(b_i(x) + h_i(x,\varepsilon)), \tag{3}$$

where $\boldsymbol{h} = (h_0, h_1, h_2)$ is a 3-tuple of real-analytic mappings $h_i : U \times E \to \mathbb{R}^6$, and $E \subset \mathbb{R}$ is an interval containing 0. $\boldsymbol{h}$, referred to in what follows as a *disturbance*, is assumed to satisfy $h_0(0,\varepsilon) = 0$ for every $\varepsilon \in E$, so that $(x,u) = (0,0)$ is an equilibrium point for the perturbed system. The interpretation of ε is that of an additional parameter quantifying the "magnitude" of the perturbation. For ease of reference we denote by $\mathcal{D}^3$ the set of all disturbances $\boldsymbol{h} = (h_0, h_1, h_2)$, each defined on a set $U \times E$ (E may thus depend on the choice of $\boldsymbol{h}$). In what follows we also write $h_i^\varepsilon(x) = h_i(x,\varepsilon)$.

Essentially, these disturbances are intended to represent two kinds of error terms, namely, those that do not depend on ε, which may typically encompass "high-order" terms neglected when the model is derived, and those that result from inaccuracies—quantified by ε—in the knowledge of the physical dimensions involved in the model (cf. also Remark 2(i) after Proposition 3.1). Obviously, however, not all disturbances may be modeled by additive vector fields as in (3). Phenomena such as neglected modes, nonsmooth effects (e.g., friction) or measurement noise would require different representations. Therefore, the notion of robustness one can aim at by considering such disturbances bears some limitations.

Before we proceed, let us recall the notion of exponential stability for continuous-time systems. Let $0 \in U \subset \mathbb{R}^n$, with U open, and consider the system

$$\dot{z} = f(z,t), \quad f(0,\,\cdot\,) = 0, \quad f : U \times \mathbb{R} \to \mathbb{R}^n. \tag{4}$$

The mapping $(z,t) \mapsto f(z,t)$ is assumed to be continuous in z and piecewise continuous in t. The origin $z = 0$ is *locally exponentially stable* for (4) if there exist $K > 0$, $\gamma > 0$ and a neighborhood $V \subset U$ of 0 such that, for every $(z_0, t_0) \in V \times \mathbb{R}$, a solution $z(\cdot)$ satisfying $z(t_0) = z_0$ is defined on $[t_0, \infty)$ and also satisfies

$$\|z(t)\| \le K\|z_0\|e^{-\gamma(t-t_0)} \tag{5}$$

for all $t \ge t_0$.

Now suppose that a continuous, time-varying (T-periodic) feedback law $\alpha : U \times \mathbb{R} \to \mathbb{R}^2$ is given. As mentioned in the introduction, one intends to act on the perturbed system (3) by periodically iterating this control law in the hope that such process stabilizes the system exponentially to a point (the origin, say, without loss of generality). Nevertheless, according to the definition of (local) exponential stability, the iteration of such a control law cannot, in general, achieve that goal since the origin may even fail to be an equilibrium. Indeed, the state of the system may reach the origin at some time $t_0 \in (kT, (k+1)T)$, which need not coincide with any of the sampling instants. Since the control operates in "open-loop" between samples, it may continue acting on the system, thus causing the state to leave the origin again. In such a case, inequality (5)—which is required to hold for every choice of "initial data" $(z_0, t_0) \in V \times \mathbb{R}$—would not hold for $(0, t_0)$ and any selection of $K > 0, \gamma > 0$. One way to remedy this issue is to consider stability in the *discrete-time*

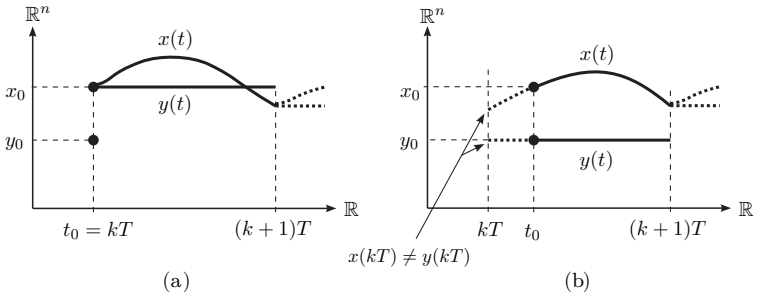


FIG. 1. *Initial conditions for system* (6): (a) *If* $t_0 \mod T = 0$, *both* $x(\cdot)$ *and* $y(\cdot)$ *are initialized to* x_0*;* (b) *If* $t_0 \mod T \neq 0$, $x(\cdot)$ *and* $y(\cdot)$ *are initialized to* x_0 *and* y_0, *respectively. Note that in the latter case the solutions are in general* not reversible in time, *since extending* $x(t)$ *and* $y(t)$ *for* $t \in [kT, t_0)$, *using the dynamics* (6), *may lead to the condition* $x(kT) \neq y(kT)$.

sense and concentrate only on the sequence of state values at the sampling instants, $(z(kT))_{k\in\mathbb{N}}$. However, since one is dealing with a continuous-time system (3), we adopt the alternative approach proposed in [17], where local exponential stability is considered *for a dynamic extension* of the perturbed system (3). More precisely, in order to cope with the case when $t_0 \mod T \neq 0$ (so t_0 does not equal any sampling instant) we adjoin a signal $t \mapsto y(t)$, which coincides with the state $x(kT)$ at the update instants indexed by $k \in \{\lfloor t_0/T \rfloor + 1, \lfloor t_0/T \rfloor + 2, \dots\}$, and then consider the *dynamically extended perturbed system*

$$\left\{ \begin{array}{rcl} \dot{x} & = & b_0(x) + h_0(x,\varepsilon) + \sum_{i=1}^{2} \alpha_i(y,t)(b_i(x) + h_i(x,\varepsilon)), \\ \dot{y} & = & \sum_{k=\lfloor t_0/T \rfloor + 1}^{\infty} \delta(t - kT)x(t), \end{array} \right. \tag{6}$$

under the proviso that its "initial condition" be defined, given any $(x_0, y_0) \in \mathbb{R}^6 \times \mathbb{R}^6$, by setting $(x(t_0), y(t_0))$ equal to (x_0, x_0) if $t_0 \mod T = 0$, or equal to (x_0, y_0) otherwise. (The symbol $\delta(t - kT)$ in (6) represents Dirac's delta "function" and satisfies $\int_{-\infty}^{\infty} \delta(t - kT) f(t) dt = f(kT)$ for any mapping $f : \mathbb{R} \to \mathbb{R}^n$.)

The meaning of the initial conditions for system (6) is illustrated in Figure 1. Clearly, the *first* sample instant *after* the initial time t_0 occurs at $t = (\lfloor t_0/T \rfloor + 1)T$ or, using the notation in the figure, at $t = (k+1)T$. This explains the initial value for the index k in the second summation of (6). Note also that the trajectories initialized in this way are defined for forward time ($t \geq t_0$), but they may fail to be reversible in time. In other words, when $t_0 \mod T \neq 0$, the solution $(x(\cdot), y(\cdot))$ may be prolonged to the interval $[kT, t_0)$ by using the dynamics (6); however, $x(kT)$ may differ from $y(kT)$.

Remark 1. It is worth pointing out that the dynamic extension in (6) is a technical artifice merely used to establish the proofs in a precise setting. In particular, the extension does not have to be "implemented," nor does it restrain the way the control signals are actually applied to system (1), or the set of allowable initial conditions for the latter.

The problem of robust stabilization may now be formulated as follows.

Problem 1 (robust stabilization of the extended chained form). Design a control law $\alpha : U \times \mathbb{R} \to \mathbb{R}^2$ which ensures that, for every disturbance $\boldsymbol{h}$ in a given set $\mathcal{A} \subset \mathcal{D}^3$, there is a constant $\varepsilon_0 > 0$ such that the origin $(x, y) = (0, 0)$ of system (6) is locally

exponentially stable whenever $\varepsilon \in E$ and $|\varepsilon| \leq \varepsilon_0$.

3. Robust stabilizers for the extended chained form. In this section we derive a solution to Problem 1 for the ECF system (1). The solution is obtained in two main steps: first the feedback law α is *designed* to have certain properties; then, in the slightly more involved second step, a stability/robustness *analysis* is carried out to guarantee that α indeed solves the problem. For more details on the notation used in this and the ensuing sections, the reader may consult section 6.1 in the appendix.

3.1. Design of the feedback law. Fix $T > 0$ and set $\omega = 2\pi/T$. Our goal is to design a feedback law $\alpha \in C^0(\mathbb{R}^6 \times \mathbb{R}; \mathbb{R}^2)$, T-periodic in its second argument, such that the solution $x(\cdot)$ to the controlled ECF

$$\dot{x} = b_0(x) + \sum_{i=1}^{2} \alpha_i(x_0, t) b_i(x), \quad x(0) = x_0 \in \mathbb{R}^6, \tag{7}$$

with b_0, b_1, b_2 given in (2), satisfies

$$x(T) = Ax_0 + \mathrm{o}(\|x_0\|), \tag{8}$$

where $A \in \mathbb{R}^{6\times 6}$ a *discrete–time-stable* matrix, i.e., a matrix whose spectrum is contained in $\{z \in \mathbb{C} : |z| < 1\}$. We propose the following controller structure:

$$\alpha_1(x,t) = a_1 x_1 + a_2 x_2 + G\rho(x)\cos(\omega t), \tag{9}$$

$$\alpha_2(x,t) = a_3 x_3 + a_4 x_4 - \frac{2\omega^2}{G}\frac{1}{\rho(x)}(a_5 x_5 + a_6 x_6)\cos(\omega t), \tag{10}$$

where the vector of control gains $a \in \mathbb{R}^6$ is determined below, $G > 0$, and ρ is given[1] by $\rho(x) = (\sum_{i=1}^{6} |x_i|^{\frac{2}{r_i}})^{\frac{1}{2}}$, with $r = (1,1,1,1,2,2)$. We set $\alpha(0,\cdot) = 0$. By virtue of the definition of ρ, one easily shows that $\alpha(x,t) \to 0$ whenever $x \to 0$, uniformly for $t \in \mathbb{R}$, so that α *is continuous on* $\mathbb{R}^6 \times \mathbb{R}$.

Now, the closed-loop system can be explicitly integrated thanks to the simple structure of the ECF and the fact that $u(t) = \alpha(x_0, t)$ is independent of $x(t)$ on the interval $(0, T)$. After some calculations, one verifies the solution $x(\cdot)$ is of the form

$$x(T) = Ax_0 + w(x_0), \tag{11}$$

where A is a block-diagonal matrix $A = \mathrm{diag}(A_1, A_2, A_3)$ with blocks defined by

$$A_i = \begin{pmatrix} 1 + \frac{1}{2}T^2 a_{2i-1} & T + \frac{1}{2}T^2 a_{2i} \\ T a_{2i-1} & 1 + T a_{2i} \end{pmatrix}, \quad i = 1,2,3. \tag{12}$$

The spectrum of A is the union of the spectra of the A_i, each of which can be made equal to $\{k_{i1}, k_{i2}\} \subset \{z \in \mathbb{C} : |z| < 1\}$—thus making A a discrete–time-stable matrix—by setting

$$a_{2i-1} = \frac{k_{i1} + k_{i2} - k_{i1}k_{i2} - 1}{T^2} \quad \text{and} \quad a_{2i} = \frac{k_{i1} + k_{i2} + k_{i1}k_{i2} - 3}{2T}, \quad i = 1,2,3. \tag{13}$$

[1] In the language of homogeneity, ρ is a homogeneous norm with respect to a dilation of weight r. In this paper, however, no further use is made of this terminology or the associated results, and the interested reader is referred to, e.g., [7, 9] for more detailed discussions on that subject.

Of course, a_{2i-1} and a_{2i} must be real, for which it suffices to choose k_{i1}, k_{i2} to be complex conjugate. On the other hand, it is readily checked that the function $w = (w_1, \dots, w_6) : \mathbb{R}^6 \to \mathbb{R}^6$ in (11) is given by $w_1 = \cdots = w_4 = 0$ and

$$(w_5, w_6)(x_0) = \rho(x_0)L(x_0) + (\rho(x_0))^{-1}P(x_0) + Q(x_0),$$

where $L : \mathbb{R}^6 \to \mathbb{R}^2$ is linear and $P, Q : \mathbb{R}^6 \to \mathbb{R}^2$ are quadratic. Since $\rho(x_0) = \mathrm{O}(\|x_0\|^{\frac{1}{2}})$, it follows that $w(x_0) = \mathrm{O}(\|x_0\|^{\frac{3}{2}})$ and hence $w(x_0) = \mathrm{o}(\|x_0\|)$, so the solution $x(T)$ has the form (8). Since A is discrete–time-stable, there exists a symmetric, positive-definite matrix $P \in \mathbb{R}^{6\times 6}$ and a real number $\tau \in [0, 1)$ such that $\|Ax_0\|_P \leq \tau\|x_0\|_P$ for every $x_0 \in \mathbb{R}^6$, with $\|x\|_P = x^T Px$ denoting the norm of x induced by P. This means that, locally around the origin, the mapping which assigns $x(T)$ to x_0 is a contraction in the norm $\|\cdot\|_P$.

3.2. Some links between the proposed controller and other approaches. The remarkably simple structure of the control law (9)–(10) shares common traits with the one in [1]. In particular, both involve terms that are linear in the state components governed by second-order chains of integrators, namely, $x_1, \dots, x_4$ in the notation of the present paper. In addition, both of them use normalization by ρ—multiplication of some terms by $1/\rho$—in order to adjust the "degree of homogeneity" of the control law α (see [1] for further details and definitions). The important difference, however, lies in the way the control signals are calculated and applied, to wit, iterative state steering vs. feedback. As a matter of fact, this difference is instrumental in establishing robustness.

Interestingly, the *frequency* ω of the time-varying terms in the control law (9)–(10) does not have to be large. In fact, that frequency may be taken arbitrarily small (i.e., the period between samples may be arbitrarily long) without qualitatively altering the nature of the result. This is in opposition with the control laws in [1] or, more generally, with previous results based on averaging of "highly oscillatory" systems, e.g., [28, 16].

Furthermore, in contrast with the control laws in [8], which provide tracking controllers that steer the state asymptotically towards the origin by following an appropriately designed trajectory, the computation of (9)–(10) does not require the use of any such trajectory.

It is also interesting to note that, while our approach and that of [4] exhibit similarities (e.g., both are intended to be implemented as hybrid open-loop/feedback) the control expressions (9)–(10) are less involved than the ones in [4], which make use of time-varying gains determined by the solutions of an exogenous system. Moreover, even though robustness is not explicitly addressed in [4], it seems difficult to assess whether those control laws ensure robustness in the sense considered in this paper or not. In particular, the result in [13], which allows us to ascertain nonrobustness of [1], does not apply in that case.

On the other hand, the work reported in [14], where stability is considered in the discrete-time sense, may be used to ascertain robustness of our controllers with respect to disturbances of a particularly simple nature. It is not clear, however, how a larger class of disturbances (such as the one considered in our main result; cf. Proposition 3.1 below) can be encompassed by the same methodology. In fact, the strongest result in [14] holds when disturbances are simple enough that adding them to the closed-loop system results in a vector field whose flow can be explicitly computed. Since our stability/robustness analysis uses a Chen–Fliess series expansion to scrutinize the

terms that add up to the flow, in a very loose sense it may be regarded as a refinement, for the special case of system (1) controlled by (9)–(10), of the results in [14].

To close this paragraph, let us add that our approach yields control laws that are globally defined on $\mathbb{R}^6 \times \mathbb{R}$; hence they are nonsingular on the whole domain of validity of the coordinate chart containing the point to be stabilized. A slightly different situation occurs for the control laws of [11] and [6], where singularities may appear near the target point due to the nature of the control laws and to the nature of the coordinate transformations, respectively.

3.3. Stability and robustness analysis. In this section we present our main result, Proposition 3.1, which characterizes the stability and robustness properties of the feedback law (9)–(10) applied to the ECF.

PROPOSITION 3.1. *The control law α defined in* (9)–(10) *is a local exponential stabilizer for the origin of system* (6)*, robust to disturbances in* $\mathcal{A} = \{(h_0^\varepsilon, h_1^\varepsilon, h_2^\varepsilon) \in \mathcal{D}^3 : \mathrm{Ord}(h_0^\varepsilon) \geq 1, \mathrm{Ord}(h_0^0) \geq 2 \text{ and } \mathrm{Ord}(h_i^0) \geq 1, i = 1, 2\}$.

Remark 2. (i) In view of the definition of $\mathcal{A}$, for $\boldsymbol{h} \in \mathcal{A}$ one can write $h_i(x, \varepsilon) = w_i^\varepsilon(x) + h_i^0(x)$, with $w_i^0(\,\cdot\,) = 0$, $h_0^\varepsilon(x) = \mathrm{O}(\|x\|^2)$, and $h_j^0(x) = \mathrm{O}(\|x\|^1)$, $(i = 1, 2, 3,$ $j = 1, 2)$. Hence each disturbance vector field can be thought of as consisting of two parts, one containing only "high-order" terms in x and the other one vanishing identically when $\varepsilon = 0$. The terms corresponding to these two parts may have different origins. For instance, $w_i^\varepsilon(x)$ may arise from uncertainty in the knowledge of the physical parameters; if ε is a quantitative measure of the uncertainty, then these terms should vanish when ε equals zero. On the other hand, $h_i^0(x)$ may include high-order terms truncated from a series expansion of the system's nominal model, and these terms do not necessarily vanish when $\varepsilon = 0$.

(ii) A measure of the extent to which robustness is ensured by a feedback law α lies in the nature of the set $\mathcal{A}$. Roughly stated, the larger this set is, the more sources of disturbances α can tolerate. In this respect, the control law in [1] is *not* robust to disturbances taken from $\mathcal{A}$; thus the origin may be destabilized by the addition of disturbances in $\mathcal{A}$ regardless of how small their magnitude is (i.e., for arbitrarily small $|\varepsilon| > 0$). This lack of robustness, which can be checked by using the results in [13], is illustrated through numerical simulation in the examples in section 4.

The proof of Proposition 3.1 shares the same basic structure as that of Theorem 1 in [17], and some other technical facts are easy modifications of proofs in [27] and [10]. For the sake of conciseness, we prove only those claims particular to our solution and explicitly refer the reader to the appropriate references when necessary.

Proof of Proposition 3.1. Let us fix a disturbance $\boldsymbol{h} \in \mathcal{A}$ defined on an open set $U \times E \subset \mathbb{R}^n \times \mathbb{R}$. It must be shown that there is $\varepsilon_0 > 0$ such that the origin of (6) is locally exponentially stable when $\varepsilon \in [-\varepsilon_0, \varepsilon_0] \cap E$. The proof is divided into two main steps corresponding to the following two claims.

Claim 1. For every compact interval $E' \subset E$ there is a compact neighborhood $U' \subset U$ of 0 such that if $x_0 \in U'$ and $\varepsilon \in E'$, the solution $t \mapsto x(t) = \pi(t, 0, x_0, \varepsilon)$ to

$$\dot{x} = b_0(x) + h_0^\varepsilon(x) + \sum_{i=1}^{2} \alpha_i(x_0, t)(b_i(x) + h_i^\varepsilon(x)), \quad x(0) = x_0, \tag{14}$$

satisfies

$$x(T) = Ax_0 + \lambda(\varepsilon, x_0) + \mu(\varepsilon, x_0) + \mathrm{o}(\|x_0\|),$$

where the mappings λ, μ (which need not be uniquely defined) are such that

$$\frac{\|\lambda(\varepsilon, x_0)\|}{\|x_0\|} \to 0 \quad \text{as} \quad \varepsilon \to 0, \quad \text{uniformly for } x_0 \in U' \setminus \{0\}, \tag{15}$$

$$\frac{\|\mu(\varepsilon, x_0)\|}{\|x_0\|} \to 0 \quad \text{as} \quad x_0 \to 0, \quad \text{uniformly for } \varepsilon \in E'. \tag{16}$$

Claim 2. [17, Theorem 1]. There exists a nonempty interval $E_0 \subset E$ containing 0 such that, for every $\varepsilon \in E_0$, the origin of system (6) is locally exponentially stable. The proof that Claim 1 implies Claim 2 can be found in [17, Theorem 1]; here we proceed with the proof of Claim 1. The first step consists in showing that the system's solution at time T can be represented by means of a Chen–Fliess series expansion and, to this end, the following lemma is instrumental.[2]

LEMMA 3.2. *Let M be a real-analytic manifold and let $\overline{x} \in M$. Assume that the following hold:* (1) *$f_0, \dots, f_m$ are real-analytic vector fields on M, with $f_0(\overline{x}) = 0$;* (2) *$\phi : M \to \mathbb{R}$ is real-analytic; and* (3) *$\alpha \in C^0(M \times \mathbb{R}; \mathbb{R}^m)$ is such that $\alpha(\overline{x}, \cdot\,) = 0$ and $\alpha(x, \cdot\,)$ is bounded for every $x \in M$. Then, given $T > 0$, there is a neighborhood $\in K$ of $\overline{x}$ such that, for $x_0 \in K$ and $t_0 \in \mathbb{R}$, the solution $t \mapsto \pi(t, t_0, x_0)$ to $\dot{x} = f_0(x) + \sum_{i=1}^m \alpha_i(x_0, t) f_i(x)$, $x(t_0) = x_0$ is defined for $t \in [t_0, t_0 + T]$, and the Chen–Fliess series $\mathrm{Ser}_{\phi, \boldsymbol{f}, \alpha}(t, t_0, x_0) = \sum_I f_I \phi(x_0) \int_{t_0}^t \alpha_I(x_0)$ converges to $\phi(\pi(t, t_0, x_0))$, absolutely and uniformly for $(x_0, t_0) \in K \times \mathbb{R}$ and $\in t[t_0, t_0 + T]$.*

Proof. The proof of Lemma 3.2 is given in the appendix. □

Let $E' \subset E$ be any compact interval containing 0. Define real-analytic vector fields g_0, g_1, g_2 on $U \times E$ and a feedback law $\overline{\alpha} \in C^0(U \times E \times \mathbb{R}; \mathbb{R}^m)$ by setting $g_i(x, \varepsilon) = b_i(x) + h_i^\varepsilon(x)$ and $\overline{\alpha}_i(x, \varepsilon, t) = \alpha_i(x, t)$. It is clear that $g_0(0, \varepsilon) = 0$ for $\varepsilon \in E$, and that $\boldsymbol{g} = (g_0, g_1, g_2)$ and $\overline{\alpha}$ satisfy the assumptions of Lemma 3.2. Hence, for every $\varepsilon \in E'$ there is an open neighborhood V_ε of $(0, \varepsilon) \in U \times E$ for which the conclusion of that lemma holds. But $(V_\varepsilon)_{\varepsilon \in E'}$ is an open cover for the compact set $\{0\} \times E'$; thus one can extract from it a finite, open subcover. This implies the existence of a neighborhood $U' \subset U$ of the origin with the property that, for any $\varepsilon \in E'$, the solution $t \mapsto x(t) = \pi(t, 0, x_0, \varepsilon)$ to system (14), issued from any point $x_0 \in U'$ at $t = 0$, is defined on $[0, T]$, and the corresponding Chen–Fliess series

$$S(x_0, \varepsilon, t) = \mathrm{Ser}_{\mathrm{id}, \boldsymbol{b} + \boldsymbol{h}^\varepsilon, \alpha}(t, 0, x_0) = \sum_I (b + h^\varepsilon)_I \mathrm{id}(x_0, \varepsilon) \int_0^t \alpha_I(x_0) \tag{17}$$

converges to $\pi(t, 0, x_0, \varepsilon)$ absolutely, uniformly for $(x_0, \varepsilon, t) \in U' \times E' \times [0, T]$. (Here we use the notation $(b + h^\varepsilon)_I \mathrm{id}(x_0) = (b_{i_1} + h_{i_1}^\varepsilon) \cdots (b_{i_r} + h_{i_r}^\varepsilon) \mathrm{id}(x_0)$, given a multi-index $I = (i_1, \dots, i_r)$.) Note that the terms $(x_0, \varepsilon) \mapsto (b + h^\varepsilon)_I \mathrm{id}(x_0)$ involved in the series (17) represent real-analytic, first-order differential operators iterated on the function id; hence these terms are real-analytic as well. We may therefore use (17) to express the solution at $t = T$ in order to prove that it satisfies Claim 1.

[2] In [27, Lemma 4.2], conditions are given for the Chen–Fliess series to converge for every t in a *sufficiently short* interval $[0, \tau]$. In the present case, however, one requires the value of the solution at the end of the interval $[0, T]$, with T fixed beforehand. When the system is driftless, the interval $[0, \tau]$ of validity of the series expansion can be made arbitrarily long by imposing small enough bounds on the control inputs $\|u(\cdot)\|$ (cf. [17, Prop. 1] and the remarks that follow it). Nevertheless, the system here contains a drift term, so the convergence results in [27] cannot be applied without modification. This motivates the role of Lemma 3.2, which states conditions for convergence of the series for arbitrarily large times and initial conditions near an equilibrium point.

Set $w_i^\varepsilon(x) = h_i^\varepsilon(x) - h_i^0(x)$ so that $h_i^\varepsilon = w_i^\varepsilon + h_i^0$, $i = 0, 1, 2$. Obviously, each $(x, \varepsilon) \mapsto w_i^\varepsilon(x)$ is real-analytic and vanishes when $\varepsilon = 0$. For convenience define the sets of vector fields $\mathcal{B} = \{b_0, b_1, b_2\}$, $\mathcal{W} = \{w_0^\varepsilon, w_1^\varepsilon, w_2^\varepsilon\}$, and $\mathcal{H} = \{h_0^0, h_1^0, h_2^0\}$. Considering that each of the iterated differential operators $(b + h^\varepsilon)_I$ in (17) can be written as $(b+w^\varepsilon+h^0)_I$, it is easy to check that, since $S(x_0, \varepsilon, T)$ converges absolutely, the series can be rearranged as $S(x_0, \varepsilon, T) = \sum_{i=1}^5 S_i(x_0, \varepsilon, T)$, where $S_1, \ldots, S_5$ are absolutely convergent series defined by

$$\begin{aligned}
S_1(x_0, \varepsilon, T) &= x_0 + \sum_{1 \le |I|} b_I \mathrm{id}(x_0) \textstyle\int_0^T \alpha_I(x_0), \\
S_2(x_0, \varepsilon, T) &= \sum_{1 \le |I|} X_I \mathrm{id}(x_0, \varepsilon) \textstyle\int_0^T \alpha_I(x_0), \\
S_3(x_0, \varepsilon, T) &= \sum_{1 \le |I| \le 2} Y_I \mathrm{id}(x_0, \varepsilon) \textstyle\int_0^T \alpha_I(x_0), \\
S_4(x_0, \varepsilon, T) &= \sum_{3 \le |I|} Y_I \mathrm{id}(x_0, \varepsilon) \textstyle\int_0^T \alpha_I(x_0), \\
S_5(x_0, \varepsilon, T) &= \sum_{1 \le |I|} Z_I \mathrm{id}(x_0, \varepsilon) \textstyle\int_0^T \alpha_I(x_0),
\end{aligned}$$

and, for $I = (i_1, \ldots, i_r)$, the iterated differential operators X_I, Y_I, Z_I satisfy the following:

1. For $j = 1, \ldots, r$, X_{i_j} and Y_{i_j} belong to $\mathcal{B} \cup \mathcal{W} \cup \mathcal{H}$, whereas Z_{i_j} belongs to $\mathcal{B} \cup \mathcal{H}$.
2. *At least one* of the X_{i_j} and *at least one* of the Y_{i_j} are contained in $\mathcal{W}$.
3. *None* of the X_{i_j} belongs to $\{b_0, w_0^\varepsilon, h_0^0\}$.
4. *At least one* of the Y_{i_j} belongs to $\{b_0, w_0^\varepsilon, h_0^0\}$.
5. *At least one* of the Z_{i_j} is contained in $\mathcal{H}$.

It follows from the first property that all of the Z_I are independent of ε and, from the second, that $X_I \mathrm{id}(x_0, 0) = Y_I \mathrm{id}(x_0, 0) = 0$ for every $x_0 \in U$. In what follows, S_1 through S_5 are analyzed separately in order to show that their sum has the form announced in Claim 1. Let us first present Lemma 3.3, which gathers some simple facts to be used below.

LEMMA 3.3. *Under the assumptions of Proposition* 3.1 *the following hold:*

(i) *For every compact neighborhood $U' \subset U$ of the origin there exists $K > 0$ such that $\|\alpha(x_0, t)\| \le K \|x_0\|^{\frac{1}{2}}$ for $(x_0, t) \in U' \times \mathbb{R}$.*

(ii) *Let $r \in \{1, 2\}$. For any* nonzero *multi-index $I \in \{0, 1, 2\}^r$, the iterated integral $\int_0^T \alpha_I$ satisfies $\int_0^T \alpha_I(x_0) = \mathrm{O}(\|x_0\|)$, and for any multi-index $I \in \{1, 2\}^r$ it satisfies $\int_0^T \alpha_I(x_0) = \mathrm{O}(\|x_0\|^2)$.*

(iii) *Say that $k_0 = 0$ and $k_1 = k_2 = 1/2$. Then for any multi-index $I = (i_1, \ldots, i_r) \in \{0, 1, 2\}^r$, $r > 0$, one has $\mathrm{Ord}(\int_0^T \alpha_I) \ge \sum_{j=1}^r k_{i_j}$.*

(iv) *For every multi-index I, $x_0 \mapsto \int_0^T \alpha_I(x_0)$ is continuous.*

(v) *For $i = 1, 2$ the following hold:*
   (i) $\mathrm{Ord}(b_0) = 0$, $\mathrm{Ord}(b_i) = -1$,
   (ii) $\mathrm{Ord}(h_0^\varepsilon) = \mathrm{Ord}(w_0^\varepsilon) \ge 1$, $\mathrm{Ord}(h_i^\varepsilon) = \mathrm{Ord}(w_i^\varepsilon) \ge -1$,
   (iii) $\mathrm{Ord}(h_0^0) \ge 2$, $\mathrm{Ord}(h_i^0) \ge 1$.

(vi) *If $\phi \in C^\infty(U; \mathbb{R})$ and $k \ge 1$, then $\mathrm{Ord}(b_0^k \phi) \ge k$. (Here $b_0^0 \phi = \phi$ and $b_0^j \phi = b_0(b_0^{j-1} \phi)$, $j \ge 1$.)*

Proof. Given in the appendix. ∎

The first sum S_1 converges to the solution of the nominal system (7) controlled by $u = \alpha(x_0, t)$; thus

$$S_1(x_0, \varepsilon, T) = x_0 + \sum_{1 \le |I|} b_I \mathrm{id}(x_0) \int_0^T \alpha_I(x_0) = Ax_0 + \mathrm{o}(\|x_0\|). \tag{18}$$

Let us now prove that S_2–S_4 can be written in terms of functions satisfying properties analogous to (15)–(16), while S_5 converges to an $\mathrm{o}(\|x\|)$ function. The following lemma is crucial to attaining this goal.

LEMMA 3.4. *Let $U \times E \subset \mathbb{R}^n \times \mathbb{R}$ be an open neighborhood of $(0,0)$ and assume that, for every I in a countable set $\mathcal{I}$, $a_I : U \times E \to \mathbb{R}^n$ is real-analytic and vanishes at $U \times \{0\}$ and $b_I : U \to \mathbb{R}$ is continuous. Assume further that $\sum_{I \in \mathcal{I}} a_I(x, \varepsilon) b_I(x)$ converges to $f(x, \varepsilon)$, absolutely and uniformly for $(x, \varepsilon) \in U \times E$. Then there is a compact neighborhood $U' \subset U$ of 0 such that*

(1) *f satisfies (15) (with $\lambda = f$) if $\mathcal{I}$ is finite and any of the following conditions holds:*
   (i) *$a_I(x, \varepsilon) = \mathrm{O}(\|x\|)$ for every $I \in \mathcal{I}$,*
   (ii) *$b_I(x) = \mathrm{O}(\|x\|)$ for every $I \in \mathcal{I}$.*

(2) *f satisfies (16) (with $\mu = f$) if any of the following conditions holds:*
   (i) *$a_I(x, \varepsilon) = \mathrm{o}(\|x\|)$ for every $I \in \mathcal{I}$,*
   (ii) *there is $c > 0$ such that $b_I(x) = \mathrm{O}(\|x\|^{1+c})$ for every $I \in \mathcal{I}$,*
   (iii) *$a_I(x, \varepsilon) = \mathrm{O}(\|x\|)$ and there is $d > 0$ such that $b_I(x, \varepsilon) = \mathrm{O}(\|x\|^d)$ for every $I \in \mathcal{I}$.*

Proof. The proof of Lemma 3.4 is given in the appendix. ∎

Consider the sum S_2. If $1 \le |I| \le 2$, Lemma 3.3(ii) yields $\mathrm{Ord}(\int_0^T \alpha_I) \ge 1$. On the other hand, since the X_I's do not involve any drift term (i.e., none of the indices in I equals zero), for the terms such that $|I| \ge 3$ one invokes Lemma 3.3(iii) to conclude that $\int_0^T \alpha_I(x_0) = \mathrm{O}(\|x_0\|^{1+c})$ with $c = 1/2$. Thus, by setting $S_2(x_0, \varepsilon, T) = \lambda_2(x_0, \varepsilon) + \mu_2(x_0, \varepsilon)$,

$$\lambda_2(x_0, \varepsilon) = \sum_{1 \le |I| \le 2} X_I \mathrm{id}(x_0, \varepsilon) \int_0^T \alpha_I(x_0) \quad \text{and}$$

$$\mu_2(x_0, \varepsilon) = \sum_{|I| \ge 3} X_I \mathrm{id}(x_0, \varepsilon) \int_0^T \alpha_I(x_0);$$

the first is a sum of finitely many terms, and the second is the limit of an absolutely convergent series. By virtue of Lemma 3.4(1)(ii) and Lemma 3.4(2)(ii), λ_2 and μ_2 satisfy properties analogous to (15) and (16), respectively.

Let us turn to S_3. If $I \in \{(0), (0,0)\}$, then, since $\mathrm{Ord}(b_0) = 0$, $\mathrm{Ord}(w_0^\varepsilon) \ge 1$, and $\mathrm{Ord}(h_0^0) \ge 2$, one has $\mathrm{Ord}(Y_I \mathrm{id}) \ge 1$ by virtue of Lemma 6.1(v). If $I \notin \{(0), (0,0)\}$, then Lemma 3.3(ii) implies $\mathrm{Ord}(\int_0^T \alpha_I) = 2$. The number of multi-indices I with $1 \le |I| \le 2$ being finite, one concludes by successive application of points (1)(i) and (1)(ii) of Lemma 3.4 that λ_3 defined by

$$S_3(x_0, \varepsilon, T) = \sum_{1 \le |I| \le 2} Y_I \mathrm{id}(x_0, \varepsilon) \int_0^T \alpha_I(x_0) = \lambda_3(x_0, \varepsilon) \tag{19}$$

satisfies (15) with $\lambda = \lambda_3$.

Now let us turn to S_4 and consider two cases according to the values of the multi-indices I.

Case (i) ($|I| \geq 3$ and I involves three or more *nonzero* indices). Lemma 3.3(iii) implies that $\int_0^T \alpha_I(x_0) = \mathrm{O}(\|x\|^{1+c})$ with $c = 1/2$. Thus the sum of the terms for which the multi-index I involves three or more nonzero indices converges to a function $(x_0, \varepsilon) \mapsto \mu_{4i}(x_0, \varepsilon)$ which, by virtue of Lemma 3.4(2)(ii), satisfies (16) with $\mu = \mu_{4i}$.

Case (ii) ($|I| \geq 3$ and I involves two or less *nonzero* indices). Consider the following four subcases:

- *Subcase* (a) ($|I| \geq 3$ and $I = (0, \dots, 0)$). By the definition of Y_I, w_0^ε appears at least once in Y_I; it follows that $\mathrm{Ord}(Y_I \mathrm{id}) \geq 2$ as a consequence of Lemma 3.3(v) and Lemma 6.1(v). Thus in this subcase $Y_I \mathrm{id}(x, \varepsilon) = \mathrm{o}(\|x\|)$, so the sum of these terms converges to a function $(x, \varepsilon) \mapsto \mu_{4a}(x, \varepsilon)$ which, by Lemma 3.4(2)(i), satisfies (16) with $\mu = \mu_{4a}$.
- *Subcase* (b) ($r = |I| \geq 3$, $I = (0, i_2, \dots, i_r)$ and one or two indices are *nonzero*). Using again Lemma 3.3(v) and Lemma 6.1(v), one deduces that $\mathrm{Ord}(Y_I \mathrm{id}) \geq 1$. Also, by virtue of Lemma 3.3(iii), $\mathrm{Ord}(\int_0^T \alpha_I) \geq \frac{1}{2}$. Thus the sum of terms in this subcase converges to a function $(x, \varepsilon) \mapsto \mu_{4b}(x, \varepsilon)$ which, in view of Lemma 3.4(2)(iii), satisfies (16) with $\mu = \mu_{4b}$.
- *Subcase* (c) ($|I| \geq 3$, $I = (i_1, 0, \dots, 0)$, $i_1 \neq 0$). It is clear that $\mathrm{Ord}(\int_0^T \alpha_I) = \frac{1}{2}$ as a consequence of Lemma 3.3(iii). Also, if neither w_0^ε nor h_0^0 is involved in Y_I, then Lemma 3.3(vi) implies $\mathrm{Ord}(Y_I \mathrm{id}) \geq -1 + 2 = 1$. If, on the contrary, any of w_0^ε or h_0^0 is involved at least once in Y_I, then $\mathrm{Ord}(Y_I \mathrm{id}) \geq -1 + \sum_{j=2}^r \mathrm{Ord}(Y_{i_j}) + \mathrm{Ord}(\mathrm{id}) \geq 1$ since, under that condition, one has $\sum_{j=2}^r \mathrm{Ord}(Y_{i_j}) \geq 1$ in view of Lemma 3.3(v) and Lemma 6.1. Therefore the sum of these terms converges to a function $(x, \varepsilon) \mapsto \mu_{4c}(x, \varepsilon)$ which, by Lemma 3.4(2)(iii), satisfies (16) with $\mu = \mu_{4c}$.
- *Subcase* (d) ($r = |I| \geq 3$, $I = (i_1, \dots, i_r)$, $i_1 \neq 0$ and *exactly one* of $i_2, \dots, i_r$ is *nonzero*). Let $\mathcal{I}$ denote the set of multi-indices corresponding to this subcase. One has $\mathrm{Ord}(\int_0^T \alpha_I) \geq 1$, since exactly two indices in I are nonzero. Assume that the nonzero indices are i_1 and i_j, $2 \leq j \leq r$, so both $\mathrm{Ord}(Y_{i_1})$ and $\mathrm{Ord}(Y_{i_j})$ are ≥ -1. Setting $\omega_1 = \sum_{k=2}^{j-1} \mathrm{Ord}(Y_{i_k})$ and $\omega_2 = \sum_{k=j+1}^{r} \mathrm{Ord}(Y_{i_k})$, one gets $\omega_1 \geq 0$ and $\omega_2 \geq 0$. For those terms with $|I| \leq 7$, Lemma 6.1(v) implies that $\mathrm{Ord}(Y_I \mathrm{id}) \geq 0$. For the terms with $|I| \geq 8$, on the other hand, either b_0 appears three times consecutively in Y_I, or it does not. In the former case, if the iterated differential operator $Y_{i_2} \cdots Y_{i_{j-1}}$ involves the three successive b_0's, then Lemma 3.3(vi) yields $\mathrm{Ord}(Y_I \mathrm{id}) \geq -1 + \max\{1, 3\} = 2$. If the three successive b_0's are involved in $Y_{i_{j+1}} \cdots Y_{i_r}$, the same lemma yields $\mathrm{Ord}(Y_{i_j} \cdots Y_{i_r} \mathrm{id}) \geq -1 + \max\{1, 3\} = 2$. Thus $\mathrm{Ord}(Y_{i_2} \cdots Y_{i_r} \mathrm{id}) \geq \omega_1 + 2$ and $\mathrm{Ord}(Y_I \mathrm{id}) \geq -1 + \max\{1, 2\} = 1$.

Consider now the case when Y_I does not involve three consecutive b_0's. In this case,

$$\begin{aligned} \mathrm{Ord}(Y_{i_j} \cdots Y_{i_r} \mathrm{id}) &\geq -1 + \max\{1, \omega_2 + 1\} = \omega_2, \\ \mathrm{Ord}(Y_{i_2} \cdots Y_{i_r} \mathrm{id}) &\geq \omega_1 + \max\{1, \omega_2\}; \end{aligned}$$

thus

$$\begin{aligned} \mathrm{Ord}(Y_{i_1} \cdots Y_{i_r} \mathrm{id}) &\geq -1 + \max\{1, \omega_1 + \max\{1, \omega_2\}\} \\ &= \max\{0, \max\{\omega_1, \omega_1 + \omega_2 - 1\}\}. \end{aligned}$$

But since $|I| \geq 8$, at least two vector fields from $\{w_0^\varepsilon, h_0^0\}$ appear in Y_I, so $\omega_1 + \omega_2 \geq 2$,

$\max\{\omega_1, \omega_1 + \omega_2 - 1\} \geq 1$, and, consequently, $\mathrm{Ord}(Y_I\mathrm{id}) \geq 1$. Therefore, by setting

$$\lambda_4(x_0, \varepsilon) = \sum_{I \in \mathcal{I}, |I| \leq 7} Y_I\mathrm{id}(x_0, \varepsilon) \int_0^T \alpha_I(x_0) \quad \text{and}$$

$$\mu_{4d}(x_0, \varepsilon) = \sum_{I \in \mathcal{I}, |I| \geq 8} Y_I\mathrm{id}(x_0, \varepsilon) \int_0^T \alpha_I(x_0),$$

one sees that these mappings are well defined, the first being the sum of finitely many terms and the second being the limit of an absolutely convergent sequence. But then, with the help of points (1)(ii) and (2)(iii) of Lemma 3.4, one concludes that λ_4 and μ_{4d} satisfy (15) and (16) with $\lambda = \lambda_4$ and $\mu = \mu_{4d}$, respectively. Summarizing the results from Cases (i) and (ii) for S_4, one obtains

$$S_4(x_0, \varepsilon, T) = \sum_{3 \leq |I|} Y_I\mathrm{id}(x_0, \varepsilon) \int_0^T \alpha_I(x_0) = \lambda_4(x_0, \varepsilon) + \mu_{4i}(x_0, \varepsilon) + \mu_{4ii}(x_0, \varepsilon) \tag{20}$$

with $\mu_{4ii} = \mu_{4a} + \mu_{4b} + \mu_{4c} + \mu_{4d}$.

Finally, let us show that S_5 converges to a function f_5 such that $f_5(x) = \mathrm{o}(\|x\|)$. Consider three cases according to the value of I.

Case (i) (I involves three or more *nonzero* indices). From Lemma 3.3(iii), we see that $\mathrm{Ord}(\int_0^T \alpha) \geq 3/2$.

Case (ii) (I involves one or two *nonzero* indices). If $r = |I| \in \{1, 2\}$, then $I \in \{1, 2\}^r$ is nonzero, so Lemma 3.3(ii) implies that $\mathrm{Ord}(\int_0^T \alpha) = 2$, whereas $\mathrm{Ord}(Z_I\mathrm{id}) \geq 0$. Now suppose that $r = |I| \geq 3$. From Lemma 3.3(iii), $\mathrm{Ord}(\int_0^T \alpha_I) \geq 1/2$. If $i_1 = 0$, then $\mathrm{Ord}(Z_I\mathrm{id}) \geq 0 + \mathrm{Ord}(Z_{i_2} \cdots Z_{i_r}\mathrm{id}) \geq 0 + \max\{1, 0\} = 1$. Now let us consider the case where $i_1 \neq 0$. If $I = (i_1, 0, \ldots, 0)$, then, by definition of Z_I, either $Z_{i_1} \in \mathcal{H}$, in which case $\mathrm{Ord}(Z_{\mathrm{id}}) \geq \mathrm{Ord}(Z_{i_1}) + \max\{1, 2\} \geq 3$, or $Z_{i_j} \in \mathcal{H}$ for some $j \in \{2, \ldots, r\}$, in which case $\mathrm{Ord}(Z_{i_2} \cdots Z_{i_r}\mathrm{id}) \geq \sum_{j=2}^r \mathrm{Ord}(Z_{i_j}) + 1 \geq 2$, and so, $\mathrm{Ord}(Z_I\mathrm{id}) \geq -1 + \max\{1, 2\} = 1$. If $i_1 \neq 0$ and $i_j \neq 0$ for some $j \in \{2, \ldots, r\}$, then $\mathrm{Ord}(\int_0^T \alpha_I) \geq 1$. Moreover, either $Z_{i_1} \in \mathcal{H}$, in which case $\mathrm{Ord}(Z_I\mathrm{id}) \geq 1 + \max\{1, 0\} = 2$, or $Z_{i_1} \notin \mathcal{H}$. Suppose the latter is true and set $\omega_1 = \sum_{k=2}^{j-1} \mathrm{Ord}(Z_{i_k})$ and $\omega_2 = \sum_{k=j+1}^r \mathrm{Ord}(Z_{i_k})$, so that $\omega_1 \geq 0, \omega_2 \geq 0$ and

$$\mathrm{Ord}(Z_{i_j} \cdots Z_{i_r}\mathrm{id}) \geq \mathrm{Ord}(Z_{i_j}) + \max\{1, \omega_2 + 1\} = \mathrm{Ord}(Z_{i_j}) + \omega_2 + 1,$$
$$\mathrm{Ord}(Z_{i_2} \cdots Z_{i_r}\mathrm{id}) \geq \omega_1 + \max\{1, \mathrm{Ord}(Z_{i_j}) + \omega_2 + 1\}.$$

If $Z_{i_j} \in \mathcal{H}$, then $\mathrm{Ord}(Z_{i_j}) \geq 1$ and hence $\mathrm{Ord}(Z_I\mathrm{id}) \geq -1 + \max\{1, 2 + \omega_2\} \geq 1$. If $Z_{i_j} \notin \mathcal{H}$, then $Z_{i_k} = h_0^0$ for some $k \in \{2, \ldots, r\} \setminus \{j\}$. In that case $\omega_1 + \omega_2 \geq 2 = \mathrm{Ord}(h_0^0)$ and $\mathrm{Ord}(Z_{i_j}) \geq -1$; thus $\mathrm{Ord}(Z_I\mathrm{id}) \geq -1 + \omega_1 + \max\{1, \omega_2\} \geq 1$. Summarizing, every term pertaining to Case (ii) satisfies $\mathrm{Ord}(Z_I\mathrm{id} \cdot \int_0^T \alpha_I) \geq 3/2$.

Case (iii) ($I = (0, \ldots, 0)$, $|I| = r$). Since (a) $\mathrm{Ord}(Z_{i_j}) \geq 0$ for $j = 1, \ldots, r$; (b) at least one of the Z_{i_j} is equal to h_0^0; and (c) $\mathrm{Ord}(h_0^0) \geq 2$, one has $\mathrm{Ord}(Z_I\mathrm{id}) = \sum_{j=1}^r \mathrm{Ord}(Z_{i_j}) + 1 \geq 3$.

All terms corresponding to Cases (i)–(iii) satisfy $\mathrm{Ord}(Z_I\mathrm{id} \cdot \int_0^T \alpha_I) \geq 3/2$; that is, $Z_I\mathrm{id}(x_0) \int_0^T \alpha_I(x_0) = \mathrm{o}(\|x\|)$. Thus their sum converges to a function f_5 with the required property. Clearly, the sum of finitely many functions $f_1, \ldots, f_N$ satisfying (15) on compact sets $U_1, \ldots, U_N$ (resp., (16)) also satisfies (15) on $U' = \bigcap_{i=1}^N U_i$

(resp., (16)). Therefore $x(T)$ is as in Claim 1, and the proof of Proposition 3.1 is complete. □

In Proposition 3.1, the condition that the disturbances belong to $\mathcal{A}$ is *sufficient* but not necessary for stability and robustness. In particular, disturbances in $\mathcal{A}$ satisfy $h_0(x,\varepsilon) = \mathrm{O}(\|x\|)$ or, stated otherwise, each component of the drift disturbance satisfies $h_{0,i} = \mathrm{O}(\|x\|^2)$. This is somewhat conservative since in some cases the latter condition is not satisfied and yet the conclusion of the previous proposition seems to hold in simulations. Indeed, a refinement of that result *seems plausible*, although the proof would require surmounting some technical obstacles. We are thus led to formulate the following conjecture which, as we shall see in the examples in section 4, might be of interest when addressing the stabilization of systems whose models can be written as an ECF with additional terms. By viewing these terms as disturbances, one might successfully use the control laws (9)–(10), *without modification*, to stabilize some of those systems to a point. A drawback of the stated condition, however, is that testing it may be difficult in practice.

CONJECTURE 1. *Let $\mathcal{A}'$ be the subset of $\mathcal{D}^3$ defined by stipulating that $(h_0^\varepsilon, h_1^\varepsilon, h_2^\varepsilon)$ belongs to $\mathcal{A}'$ if and only if* (1) $\mathrm{Ord}(h_0^\varepsilon) \geq 0$, $\mathrm{Ord}(h_0^0) \geq 2$, $\mathrm{Ord}(h_i^0) \geq 1$, $i = 1, 2$, *and* (2) *for every $k \geq 2$, every $X \in \{b_1, b_2, h_1^\varepsilon, h_2^\varepsilon\}$, and every k-tuple $(Y_1, \ldots, Y_k) \in \{b_0, h_0^\varepsilon\}^k$ having at least one of the Y_i equal to h_0^ε, one has $\mathrm{Ord}(XY_1 \cdots Y_k \mathrm{id}) \geq 1$. Then the control law α defined in* (9)–(10) *is a local exponential stabilizer for* (6)*, robust to disturbances in $\mathcal{A}'$.*

Remark 3. If this conjecture holds true, its proof should essentially coincide with that of Proposition 3.1. The only differences would arise in arguments that explicitly appeal to the assumption $\mathrm{Ord}(h_0^\varepsilon) \geq 1$ (i.e., $\mathrm{Ord}(w_0^\varepsilon) \geq 1$), namely, Subcases (a), (c), and (d) of Case (ii) in the sum S_4. One should show that, by dropping that assumption, the terms pertaining to those subcases satisfy the required properties. For Subcase (c) one has $\mathrm{Ord}(\int_0^T \alpha_I) \geq 1/2$ and, since in this subcase every multi-index I is of the form $I = (i_1, 0, \ldots, 0)$ with $i_1 \neq 0$, the corresponding terms are of the form $Y_I \mathrm{id} = XZ_1 \cdots Z_k \mathrm{id}$ with $X \in \{b_i, h_i^0, w_i^\varepsilon : i = 1, 2\}$ and $(Z_1, \ldots, Z_k) \in \{b_0, h_0^0, w_0^\varepsilon\}^k$. By a simple induction argument one sees that the definition of $\mathcal{A}'$, in particular condition (2) in Conjecture 1, implies that all such terms satisfy $\mathrm{Ord}(Y_I \mathrm{id}) \geq 1$, so by virtue of Lemma 3.4(2)(iii) these terms have the required properties. On the other hand, each term of the series involved in Subcase (a) satisfies $\mathrm{Ord}(Y_I \mathrm{id}) \geq 1$ and $\mathrm{Ord}(\int_0^T \alpha_I) = 0$ since $\int_0^T \alpha_I(x_0) = \frac{T^{|I|}}{|I|!}$, whereas the terms of the series in Subcase (d) satisfy $\mathrm{Ord}(Y_I \mathrm{id}) \geq 0$ and $\mathrm{Ord}(\int_0^T \alpha_I) \geq 1$. To prove the conjecture, then, it would suffice to show that the (infinite) series in these two subcases converge to functions λ_{4a} and λ_{4d} satisfying (15).

4. Examples.

4.1. Underactuated manipulator. Consider the example of a $\mathrm{PP}\overline{\mathrm{R}}$ manipulator, depicted in Figure 2, with unactuated third joint, constrained to move on a horizontal plane. Considering the links and joints as rigid bodies and neglecting gravitational and frictional forces, this system can be modeled by

$$\begin{aligned} M_1\ddot{q}_1 - m_3 l \sin(q_3)\ddot{q}_3 - m_3 l \cos(q_3)\dot{q}_3^2 &= \tau_1, \\ M_2\ddot{q}_2 + m_3 l \cos(q_3)\ddot{q}_3 - m_3 l \sin(q_3)\dot{q}_3^2 &= \tau_2, \\ -M_3 l \sin(q_3)\ddot{q}_1 + M_3 l \cos(q_3)\ddot{q}_2 + J\ddot{q}_3 &= 0, \end{aligned} \tag{21}$$

where m_i, $i = 1, 2, 3$, is the mass of the ith link, $M_i = \sum_{j=i}^3 m_j$, J is the moment of inertia of the third link with respect to the axis of the third joint, and l is the

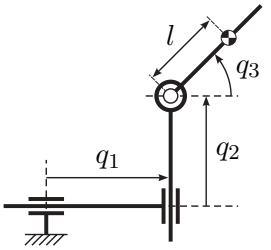


FIG. 2. *Schematic representation of the $PP\overline{R}$ manipulator.*

distance from the same axis to the center of mass of the third link. The input vector $\tau = (\tau_1, \tau_2)$ represents the forces applied in the q_1 and q_2 directions, respectively. The configuration manifold is $Q = \mathbb{R}^2 \times SS^1$, for which $q : Q \to \mathbb{R}^2 \times (-\pi, \pi)$ is a local coordinate system.

Given a target configuration $\overline{q} \in Q$, the dynamics can be transformed into the ECF, locally around $\overline{q}$, by using the coordinates of the third link's "center of percussion." A detailed description of the corresponding transformation can be found in [8]; for simplicity, however, in what follows we assume without loss of generality that the target configuration—the one that should be stabilized—is given by $q(\overline{q}) = (0, 0, 0) \in U$. After simple computations one verifies that, by setting $K = J/M_3/l$, the dynamic model (21) can be transformed into the ECF $\dot{x} = b_0(x) + u_1 b_1(x) + u_2 b_2(x)$ by means of the feedback transformation $x = \varphi(q, \dot{q})$, $u = A(q, \dot{q}) + B(q)[\tau_1 \;\; \tau_2]^T$, where

$$\varphi(q, \dot{q}) = \begin{pmatrix} q_1 + K(\cos(q_3) - 1) \\ \dot{q}_1 - K \sin(q_3)\dot{q}_3 \\ \tan(q_2) \\ (1 + \tan^2(q_3))\dot{q}_2 \\ q_2 + K \sin(q_3) \\ \dot{q}_2 + K \cos(q_3)\dot{q}_3 \end{pmatrix},$$

and

$$A(q, \dot{q}) = \frac{1}{\Delta(q)} \begin{pmatrix} \frac{(JK^2M_2 - J^2 - K^4M_1M_2 + JK^2M_1)\cos(q_3)\dot{q}_3^2}{K} \\ \frac{(2K^2M_1M_2 - 3JM_1\cos^2(q_3) - 2JM_2 + 3JM_2\cos^2(q_3))\sin(q_3)\dot{q}_3^2}{\cos^3(q_3)} \end{pmatrix},$$

$$B(q) = \frac{1}{\Delta(q)} \begin{pmatrix} (K^2M_2 - J)\cos^2(q_3) & (K^2M_1 - J)\cos(q_3)\sin(q_3) \\ \frac{KM_2\sin(q_3)}{\cos^2(q_3)} & -\frac{KM_1}{\cos(q_3)} \end{pmatrix},$$

$$\Delta(q) = K^2M_1M_2 - JM_1\cos^2(q_3) - JM_2\sin^2(q_3).$$

The control laws developed above can be iterated, after the system has been transformed into the ECF, in order to stabilize the origin $x = 0$. To this end, at each sample time $t_k = kT$ one uses the measurements of the state variables to calculate $x(t_k) = \varphi(q(t_k), \dot{q}(t_k))$, then the prescribed control law $u(t) = \alpha(x(t_k), t)$ is computed from (9)–(10). The actual force used to drive the system is obtained by using the inverse transformation $\tau(t) = [B(q)]^{-1}(u(t) - A(q, \dot{q}))$.

When the system parameters are not accurately known, which is most often the case, the functions φ, A, and B typically include additional terms. For the sake of illustration let us suppose that uncertainties are present in the values of the (cumulated) masses M_i, the position of the third link's center of mass l, and its inertia

moment J. This entails that only the erroneous values $\widetilde{M}_i = M_i + \nu_i$, $\widetilde{l} = l + \nu_4$, and $\widetilde{J} = J + \nu_5$, where $\nu = (\nu_1, \ldots, \nu_5) \in \mathbb{R}^5$ represents the parameter errors, are available to the controller. Note that, if one sets $\varepsilon = \|\nu\|^2$, the norm of the error tends to zero as $\varepsilon \to 0$. Ultimately, the effect of the inaccuracies results in disturbance vector fields $h = (h_0^\varepsilon, h_1^\varepsilon, h_2^\varepsilon)$ being added to the nominal ECF system, yielding a perturbed system in the form of (3). Using a computer algebra package, one readily verifies that for $i = 1, 2, 3$, the mappings $(x, \varepsilon) \mapsto h_i^\varepsilon(x)$ are analytic and have the following structures:

$$h_0^\varepsilon(x) = x_4^2(a_{2,0}^\varepsilon + \mathrm{O}(|x_3|^2))\frac{\partial}{\partial x_2} + x_4^2(a_{4,3}^\varepsilon x_3^3 + \mathrm{O}(|x_3|^5))\frac{\partial}{\partial x_4} + x_4^2(a_{6,1}^\varepsilon x_3 + \mathrm{O}(|x_3|^2))\frac{\partial}{\partial x_6},$$

$$h_1^\varepsilon(x) = (b_{2,0}^\varepsilon + \mathrm{O}(|x_3|^2))\frac{\partial}{\partial x_2} + (b_{4,1}^\varepsilon x_3 + \mathrm{O}(|x_3|^3))\frac{\partial}{\partial x_4} + (b_{6,1}^\varepsilon x_3 + \mathrm{O}(|x_3|^3))\frac{\partial}{\partial x_6},$$

$$h_2^\varepsilon(x) = (c_{2,1}^\varepsilon x_3 + \mathrm{O}(|x_3|^3))\frac{\partial}{\partial x_2} + (c_{4,0}^\varepsilon + \mathrm{O}(|x_3|^2))\frac{\partial}{\partial x_4} + (c_{6,0}^\varepsilon + \mathrm{O}(|x_3|^2))\frac{\partial}{\partial x_6},$$

where the symbols $a_{i,j}^\varepsilon$, $b_{i,j}^\varepsilon$, and $c_{i,j}^\varepsilon$ represent real numbers which vanish when $\varepsilon = 0$ but are nonzero for generic parameter and error values. This implies that $\mathrm{Ord}(h_0^\varepsilon) = 1$ and $\mathrm{Ord}(h_1^\varepsilon) = \mathrm{Ord}(h_2^\varepsilon) = -1$, so the assumptions in Proposition 3.1 are verified. As a result, the iterated application of the control laws (9)–(10) will ensure that the origin of the dynamically extended system (6) is locally exponentially stable, provided ε *is small enough*.

Now consider a PP$\overline{\mathrm{R}}$ manipulator whose *nominal*, physical dimensions are as follows. The three masses are equal: $m_1 = m_2 = m_3 = 10$ kg. The third link is a homogeneous parallelepiped of length $\ell = 1.5$ m and width $w = 0.15$ m; its center of mass is located at a distance $l = \ell/2 = 0.75$ m from the joint axis and its inertia moment is $J = (\ell^2/3 + w^2/12)m_3 = 7.51875\ \mathrm{kg\,m^2}$. The goal is to stabilize the system to the equilibrium configuration $(q, \dot{q}) = (0, 0)$ starting at rest ($\dot{q}_0 = 0$) from the initial configuration $q_0 = (-50\ \mathrm{cm}, 75\ \mathrm{cm}, \pi/4)$. A convenient DOF, useful for fine-tuning the transient response, is encompassed by the choice of the controller settings (T, G, and the a_i's), which can be made with the aid of some intuitively deduced "rules of thumb." T controls the length of the periods during which the system operates in open-loop; smaller values of T lead to more frequent updates of the feedback terms. G moderates the control effort exerted on the system due to the oscillatory, time-varying terms; large values of G lead to shorter settling time (to within a given tolerance) but may require larger control efforts. The values of a_i set the position of the poles $\{k_{i1}, k_{i2}\}$, within the unit circle in $\mathbb{C}$, for each of the submatrices A_i in (12). As can be expected, the closer the poles are to the origin, the shorter the settling time is, but also the larger the control effort becomes. In these simulations the settings are $\omega = 1$ rad/s, so $T = 2\pi \approx 6.28$ s; $G = 0.1$ and $k_{i,j} = 0.25$ ($i = 1, 2, 3$, $j = 1, 2$); the gain values $a_1 = a_3 = a_5 \simeq -0.01425$ and $a_2 = a_4 = a_6 \simeq -0.194$ were determined from (13). In order to perform the numerical simulation in the perturbed case, it is assumed that $\widetilde{m}_3 = 1.1m_3$ and $\widetilde{l} = 0.95l$; that is, errors of 10% and -5%, respectively, are present in the knowledge of these two parameters. The latter induce an error of -0.7% in the moment of inertia, so that $\widetilde{J} = 0.993J$. The response of the perturbed system controlled by (9)–(10) appears in Figure 3, which shows the time history of $\log(\|(q(t), \dot{q}(t))\|)$, the configuration variables $q(t)$, and velocities $\dot{q}(t)$, as well as the input forces $\tau(t)$. The differences between the transient responses in the perturbed and nominal cases are barely perceptible, so no simulation for the latter case is included. In order to assess the improved performance of the control law (9)–(10) in the presence of disturbances, let us end this example with a qualitative

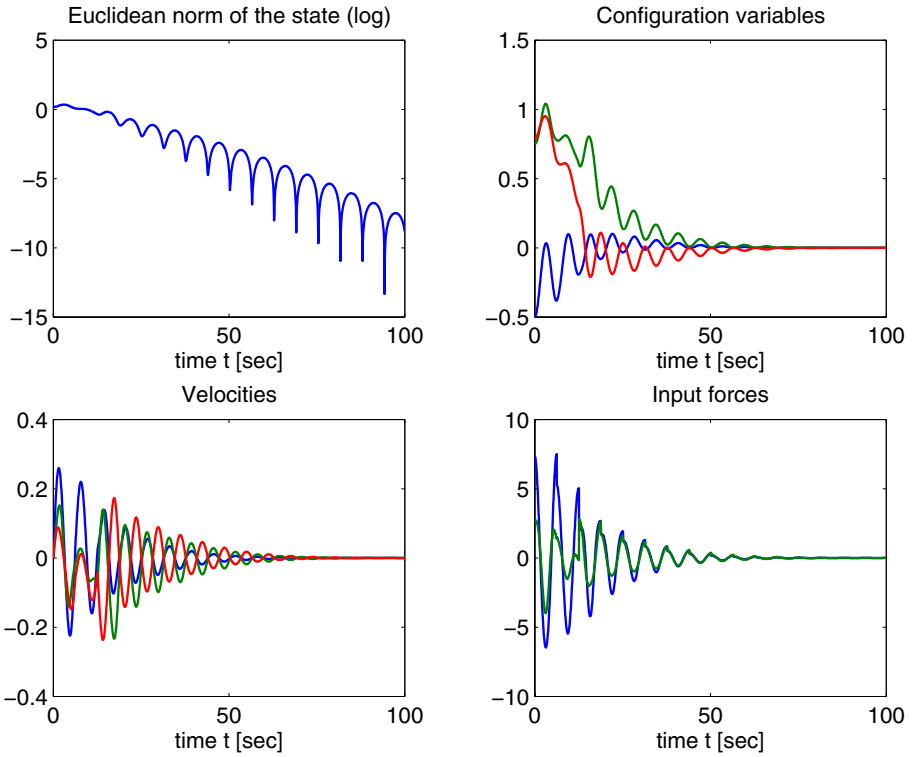


FIG. 3. *Transient response of the* perturbed $PP\overline{R}$ *manipulator using the hybrid control law* (9)–(10).

comparison with another control approach. Recall that in [1], a homogeneous, time-varying feedback law was introduced which ρ-exponentially stabilizes the ECF to the origin. Nevertheless, by virtue of the main result in [13], these control laws are not robust to disturbances in $\mathcal{D}^3$ and, in fact, as illustrated in Figure 4, the disturbances considered in this example make the system's solution tend towards what seems to be a limit cycle (in particular the origin is not Lyapunov-stable).

4.2. Simplified surface vessel. Consider a simplified surface vessel with configuration variables (x, y, θ), as depicted in Figure 5. Research studies concerning this system are reported in several references, including [23], where more details on the modeling assumptions can be found. In particular, it is shown in that reference that the corresponding dynamic model can be written in the form

$$\begin{aligned} \ddot{x} &= u_1, \\ \ddot{\theta} &= u_2, \\ \ddot{y} &= u_1 \tan(\theta) + \frac{c_y}{m}(-\dot{y} + \tan(\theta)\dot{x}). \end{aligned} \tag{22}$$

Clearly this can be viewed as a perturbed ECF system. More precisely, by setting $\varepsilon = c_y/m$ and relabeling the state variables $(x_1, \ldots, x_6) = (x, \dot{x}, \theta, \dot{\theta}, y, \dot{y})$ one can also write system (22) as

$$\dot{x} = b_0(x) + h_0^{\varepsilon}(x) + \sum_{i=1}^{2} u_i(b_i(x) + h_i^{\varepsilon}(x)), \tag{23}$$

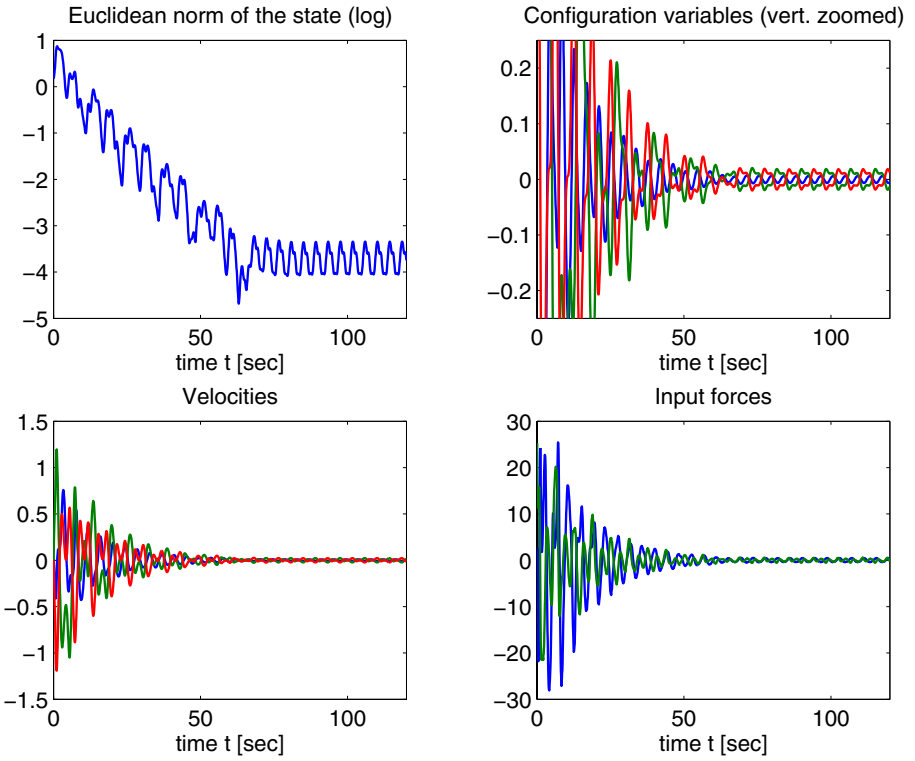


FIG. 4. *Transient response of the* perturbed $PP\overline{R}$ *manipulator using the continuous, homogeneous, time-varying feedback from* [1]. *The time histories of the configuration variables are plotted with a different scale to illustrate their ultimately oscillatory nature.*

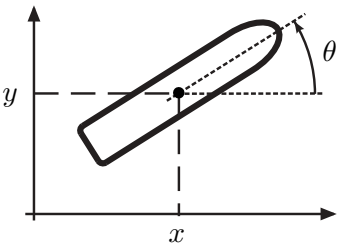


FIG. 5. *Configuration variables for the simplified surface vessel model.*

with b_0, b_1, b_2 given by (2), and the disturbance vector fields defined by

$$h_0^\varepsilon(x) = \varepsilon(-x_6 + x_2\tan(x_3))\frac{\partial}{\partial x_6}, \quad h_1^\varepsilon(x) = (\tan(x_3) - x_3)\frac{\partial}{\partial x_6}, \quad h_2^\varepsilon(x) = 0. \tag{24}$$

Obviously, the family $\boldsymbol{h} = (h_0^\varepsilon, h_1^\varepsilon, h_2^\varepsilon)$ is a disturbance in $\mathcal{D}^3$, but it is not contained in the set $\mathcal{A}$ defined in Proposition 3.1 since $\mathrm{Ord}(h_{0,6}^\varepsilon) = 1$, i.e., $\mathrm{Ord}(h_0^\varepsilon) = 0$. Let us show, however, that $\boldsymbol{h}$ belongs to $\mathcal{A}'$ and hence that it satisfies the assumptions of Conjecture 1. To this end, let $g(x) = \varepsilon(-x_6 + x_2\tan(x_3))$, so that $h_0^\varepsilon(x) = g(x)\partial/\partial x_6$. Note that $\mathrm{Ord}(h_1^\varepsilon) = 2$ and $\mathrm{Ord}(h_2^\varepsilon) = +\infty$; hence we need only certify that all terms

$XY_1 \cdots Y_k\mathrm{id}$, with $X \in \{b_1, b_2\}$, $k \geq 2$ and $Y_1, \ldots, Y_k \in \{b_0, h_0^\varepsilon\}$, satisfy

$$\mathrm{Ord}(XY_1 \cdots Y_k\mathrm{id}) \geq 1. \tag{25}$$

Since $b_1\phi(x) = \partial\phi/\partial x_2 + x_3\partial\phi/\partial x_6$ and $b_2\phi(x) = \partial\phi/\partial x_4$ for any smooth function ϕ, a necessary condition to have $\mathrm{Ord}(b_1\phi) \geq 1$ and $\mathrm{Ord}(b_2\phi) \geq 1$ is that $\mathrm{Ord}(\partial\phi/\partial x_2) \geq 1$ and $\mathrm{Ord}(\partial\phi/\partial x_4) \geq 1$. Naturally, this necessary condition holds whenever $\phi = Y_1 \cdots Y_k\mathrm{id}$ and $\mathrm{Ord}(\phi) \geq 2$. In what follows we shall show that it holds even when the latter is not the case. One has

$$b_0\mathrm{id}_i(x) = \begin{cases} x_{i+1}, & i = 1, 3, 5, \\ 0, & i = 2, 4, 6, \end{cases} \quad \text{and} \quad h_0^\varepsilon\mathrm{id}_i(x) = \begin{cases} 0, & i = 1, \ldots, 5, \\ g(x), & i = 6, \end{cases}$$

from which it follows that $b_0b_0\mathrm{id} = 0$, $h_0^\varepsilon b_0\mathrm{id}_i = 0$ for $i \neq 5$, and $b_0h_0^\varepsilon\mathrm{id}_i = h_0^\varepsilon h_0^\varepsilon\mathrm{id}_i = 0$ for $i \neq 6$. Furthermore,

$$h_0^\varepsilon b_0\mathrm{id}_5(x) = g(x), \quad b_0h_0^\varepsilon\mathrm{id}_6(x) = x_4\frac{\partial g}{\partial x_3}(x) \quad \text{and}$$
$$h_0^\varepsilon h_0^\varepsilon\mathrm{id}_6(x) = g(x)\frac{\partial g}{\partial x_6}(x) = -\varepsilon g(x).$$

By direct calculation one obtains that

$$\frac{\partial g}{\partial x_2}(x) = -\varepsilon\tan(x_3), \quad \frac{\partial g}{\partial x_4}(x) = 0, \tag{26}$$

$$\frac{\partial}{\partial x_2}\left(x_4\frac{\partial g}{\partial x_3}(x)\right) = x_4\frac{\partial^2 g}{\partial x_2\partial x_3}(x), \tag{27}$$
$$\frac{\partial}{\partial x_4}\left(x_4\frac{\partial g}{\partial x_3}(x)\right) = \frac{\partial g}{\partial x_3}(x) = \varepsilon x_2(1 + \tan^2(x_3)).$$

The orders $\mathrm{Ord}(\cdot)$ of all of these functions being ≥ 1, the required condition (25) is satisfied for $k = 2$. Now consider the case $k \geq 3$ and note that, since $\mathrm{Ord}(b_0b_0\mathrm{id}) = +\infty$ and $\mathrm{Ord}(b_0h_0^\varepsilon\mathrm{id}) = 2$, all terms $XY_1 \ldots Y_k\mathrm{id}$ which end with $b_0b_0\mathrm{id}$ or with $b_0h_0^\varepsilon\mathrm{id}$ satisfy (25) for $k \geq 3$. Moreover, $b_0h_0^\varepsilon b_0\mathrm{id}_5 = b_0h_0^\varepsilon\mathrm{id}_6$ and $b_0h_0^\varepsilon h_0^\varepsilon\mathrm{id}_5 = -\varepsilon b_0h_0^\varepsilon\mathrm{id}_6$, so those terms that end with $b_0h_0^\varepsilon h_0^\varepsilon\mathrm{id}$ and $b_0h_0^\varepsilon h_0^\varepsilon\mathrm{id}$ also satisfy (25) for $k \geq 3$. It remains only to consider terms ending with $h_0^\varepsilon h_0^\varepsilon b_0\mathrm{id}$ and $h_0^\varepsilon h_0^\varepsilon h_0^\varepsilon\mathrm{id}$. But $h_0^\varepsilon h_0^\varepsilon b_0\mathrm{id}_5 = h_0^\varepsilon h_0^\varepsilon\mathrm{id}_6$; thus one needs only to analyze terms of the form $X(h_0^\varepsilon)^\ell\mathrm{id}_6$ and $Xb_0(h_0^\varepsilon)^\ell\mathrm{id}_6$. A routine calculation yields, for $\ell \geq 1$,

$$(h_0^\varepsilon)^\ell\mathrm{id}_6(x) = (-\varepsilon)^{\ell-1}g(x) \quad \text{and} \quad b_0(h_0^\varepsilon)^\ell\mathrm{id}_6(x) = (-\varepsilon)^{\ell-1}x_4\frac{\partial g}{\partial x_3}(x). \tag{28}$$

Hence, in view of (26)–(28), those terms also satisfy (25) for every $k \geq 3$. Consequently $\boldsymbol{h} \in \mathcal{A}'$. A numerical simulation of system (23) with the controller (9)–(10) is shown in Figure 6. For this simulation the size of the error is taken to be $\varepsilon = c_y/m = 0.1$, the initial condition is $x = (1, 0, \pi/4, 0, -1, 0)$, and the controller settings are $\omega = 2\pi/T = 1.5$ rad/s, $G = 1$, and $k_{i,j} = 0.1$, $i = 1, 2, 3$, $j = 1, 2$. The gain values $a_1 = a_3 = a_5 \simeq -0.0462$ and $a_2 = a_4 = a_6 \simeq -0.333$ were determined using (13). As depicted in the time-plots, the simulation appears to validate Conjecture 1.

5. Conclusions. A controller scheme, based on well-known hybrid open-loop/feedback techniques, has been introduced for the ECF. This controller exponentially

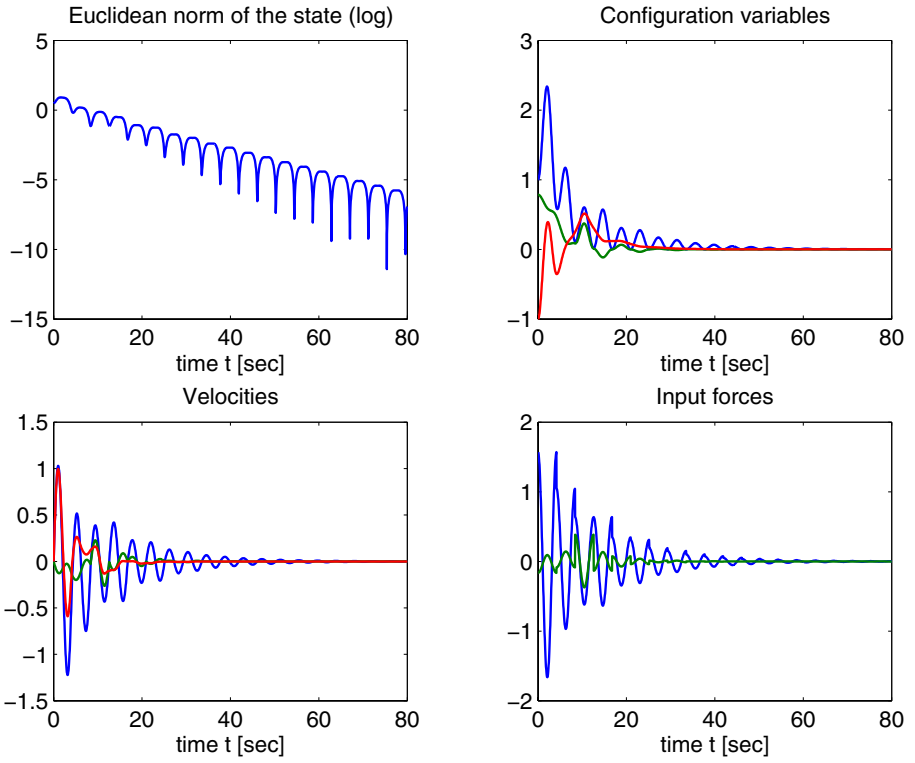


FIG. 6. *Transient response of the system* (23)*—equivalent to the simplified surface vessel model—controlled by* (9)–(10).

stabilizes the origin of a dynamic extension of the ECF, with robustness to a class of additive disturbance vector fields. The class of disturbances includes analytic vector fields added to the control vector fields as well as "high-order" drift perturbations. One positive feature of these results is that, for a class of underactuated systems—whose models need not be feedback-equivalent to the ECF—the problem of local point stabilization with exponential convergence can be effectively tackled by using the same control scheme as for the ECF. The typical performance of the proposed control laws seems qualitatively acceptable, as illustrated by the numerical simulations. On the other hand, these controllers clearly have some limitations regarding their robustness, and instability may be induced by disturbances not contained in the class $\mathcal{A}$ of Proposition 3.1 or by disturbances of a different nature, such as errors in the update time of the control.

A problem that remains open is the extension of the approach in this paper to systems with more inputs and less structure than the ECF. Such an extension would typically involve a design and an analysis stage, the former yielding control laws that stabilize the origin of a dynamically extended, *nominal* system—analogous to (6), but with $h_i^e = 0$, $i = 1, \ldots, m$. The design stage, of an essentially algebraic nature, might be based on techniques related to the design of oscillatory open-loop controls, such as the ones developed in [12]. By contrast, the analysis can be expected to be significantly involved, all the more so as it would be desirable to guarantee robustness to a large class of admissible disturbances.

6. Appendix.

6.1. Notational conventions.

6.1.1. Local order of mappings. Let us recall some definitions and properties about *local order* of mappings, a notion that simplifies the proofs. In this paragraph, n and m represent positive integers, ℓ represents a nonnegative integer, and $\|\cdot\|$ represents Euclidean norm. Given open sets $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$, the symbols $PC(U;V)$, $C^0(U;V)$, $C^\infty(U;V)$, and $C^\omega(U;V)$ denote the sets of *piecewise-continuous*, *continuous*, *smooth*, and *(real-)analytic* mappings from U to V, respectively. Consider a neighborhood U of the origin in $\mathbb{R}^n$. We deal with mappings defined on $U \times \Lambda$, where $\Lambda \subset \mathbb{R}^\ell$, and view the elements of Λ as parameters (e.g., "time" or other parameters). Given a mapping $f : U \times \Lambda \to \mathbb{R}^m$, we write $f(x,\lambda) = \mathrm{o}(\|x\|^k)$ if, for every $\lambda \in \Lambda$,

$$\lim_{x\to 0} \frac{\|f(x,\lambda)\|}{\|x\|^k} = 0. \tag{29}$$

We write $f(x,\lambda) = \mathrm{O}(\|x\|^k)$ if for every $\lambda \in \Lambda$ there is a constant $K > 0$ and a neighborhood $U' \subset U$ of the origin such that, for every $x \in U'\backslash\{0\}$,

$$\frac{\|f(x,\lambda)\|}{\|x\|^k} \le K. \tag{30}$$

Consider a mapping $X = (X_1, \dots, X_n) : U \times \Lambda \to \mathbb{R}^n$ representing a family of *vector fields* $X(\cdot,\lambda) : U \to \mathbb{R}^n$. We write $X(x,\lambda) = \mathrm{o}(\|x\|^k)$ (resp., $X(x,\lambda) = \mathrm{O}(\|x\|^k)$) if $X_i(x,\lambda) = \mathrm{o}(\|x\|^{k+1})$ (resp., $X_i(x,\lambda) = \mathrm{O}(\|x\|^{k+1})$) for $i = 1, \dots, n$. We shall also use the function $\mathrm{Ord} : f \mapsto \mathrm{Ord}(f) \in \mathbb{R} \cup \{+\infty\}$ defined by $\mathrm{Ord}(f) = \sup\{k \in \mathbb{R} : f(x,\lambda) = \mathrm{O}(\|x\|^k)\}$. Every vector field $X(\cdot,\lambda)$ is a *differential operator* acting on $C^\infty(U;\mathbb{R})$; thus, for $\phi \in C^\infty(U;\mathbb{R})$ one has $X\phi(\cdot,\lambda) \in C^\infty(U;\mathbb{R})$, where $X\phi(x,\lambda) = L_X\phi(x,\lambda) = \sum_{i=1}^n \frac{\partial\phi}{\partial x_i}(x,\lambda)$ denotes the Lie derivative of ϕ in the direction of X evaluated at (x,λ). We extend this notation to the case when $\phi \in C^\infty(U;\mathbb{R}^m)$ and use $X\phi(\cdot,\lambda)$ to denote the m-tuple $(X\phi_i)_{i=1,\dots,m}$ of functions $X\phi_i(\cdot,\lambda) \in C^\infty(U;\mathbb{R})$. The following properties are easily established:

LEMMA 6.1. *Assume that, for every $\lambda \in \Lambda$, $f(\cdot,\lambda), g(\cdot,\lambda)$ are C^∞ mappings $U \to \mathbb{R}^m$, and $X(\cdot,\lambda)$, $Y(\cdot,\lambda)$ are C^∞ vector fields $U \to \mathbb{R}^n$. Write μ to denote any of these mappings. Then the following hold:*

(i) $\mathrm{Ord}(f) \ge 0$, $\mathrm{Ord}(X) \ge -1$.
(ii) *If $k \in \mathbb{R}$ and $k \le \mathrm{Ord}(\mu)$, then $\mu(x,\lambda) = \mathrm{O}(\|x\|^k)$.*
(iii) $\mathrm{Ord}(f+g) \ge \min\{\mathrm{Ord}(f), \mathrm{Ord}(g)\}$ *(where $(f+g)(x,\lambda) = f(x,\lambda)+g(x,\lambda)$).*
(iv) $\mathrm{Ord}(fg) = \mathrm{Ord}(f) + \mathrm{Ord}(g)$ *(where $fg(x,\lambda) = f(x,\lambda)g(x,\lambda)$).*
(v) $\mathrm{Ord}(Xf) \ge \mathrm{Ord}(X) + \max\{\mathrm{Ord}(f), 1\}$. *In particular* $\mathrm{Ord}(Xf) \ge 0$.

6.1.2. Iterated differential operators and iterated integrals. Assume that $U \subset \mathbb{R}^n$ is open. Let $\boldsymbol{X} = (X_1, \dots, X_m)$ be a family of real-analytic vector fields $X_i \in C^\omega(U;\mathbb{R}^n)$, and $\phi \in C^\omega(U;\mathbb{R})$ be a real-analytic function. Every element of $\mathcal{I}_{[0,m]} = \bigcup_{k\in\{0,1,2,\dots\}} \{0,1,\dots,m\}^k$ is called a *multi-index*. If $I = (i_1, \dots, i_r) \in \{0,1,\dots,m\}^r$, the multi-index I is said to have *length* r, and this is denoted by $|I| = r$. By convention, $I = \emptyset$ is regarded as a multi-index having zero length.

Let $I = (i_1, \dots, i_r) \in \mathcal{I}_{[0,m]}$ be a multi-index. The *iterated differential operator* $X_I = X_{i_1} \cdots X_{i_r}$ is defined so that the function $X_I\phi \in C^\omega(U;\mathbb{R})$ is given by $X_{i_1} \cdots X_{i_r}\phi$ (each vector field regarded as a first-order differential operator). By convention one sets $X_\emptyset\phi = \phi$. We use $X_I\mathrm{id} : U \to \mathbb{R}^n$ to denote the n-tuple

of functions $(X_I \mathrm{id}_i)_{i=1,\dots,n}$, where $\mathrm{id} : U \to \mathbb{R}^n$ is defined by $\mathrm{id}_i(x) = x_i$ for $x = (x_1, \dots, x_n) \in U$. Given $\alpha \in C^0(U \times \mathbb{R}; \mathbb{R}^m)$ (e.g., a time-varying feedback law), a multi-index $I = (i_1, \dots, i_r)$, and real numbers t_0, t, one defines the *iterated integral* $\int_{t_0}^t \alpha_I : U \to \mathbb{R}$ as follows:

$$\int_{t_0}^t \alpha_I(x) = \int_{t_0}^t \int_{t_0}^{t_r} \cdots \int_{t_0}^{t_2} \alpha_{i_r}(x, t_r) \alpha_{i_{r-1}}(x, t_{r-1}) \cdots \alpha_{i_1}(x, t_1) dt_1 \cdots dt_r.$$

By convention, $\int_{t_0}^t \alpha_\emptyset(x) = 1$ for every $x \in U$.

6.2. Auxiliary lemmas.

6.2.1. Proof of Lemma 3.2. Since the result is local we may assume, without loss of generality, that M is an open subset of $\mathbb{R}^n$ and that $\overline{x} = 0$. We shall appeal to the following two technical lemmas; for improved readability, the proof of Lemma 6.2 is relegated to section 6.2.4, whereas Lemma 6.3 follows from a trivial adaptation of the proof of [10, Theorem 2.6].

LEMMA 6.2. *Let $f_0, \dots, f_m$ be real-analytic vector fields on a real-analytic manifold M, with $\overline{x} \in M$, and let $\phi : M \to \mathbb{R}$ be a real-analytic function. Assume that $f_0(\overline{x}) = 0$. Then there is a constant $C > 0$ with the property that, for every $\eta > 0$, there exists a neighborhood K of $\overline{x}$ such that ϕ and the vector fields $g_0 = (1/\eta) f_0$, $g_i = f_i$ $(i = 1, \dots, m)$ satisfy the estimate*

$$|(g_{i_1} \cdots g_{i_r} \phi)(x)| \leq C^r r!$$

for every $x \in K$ and every multi-index $I = (i_1, \dots, i_r) \in \{0, \dots, m\}^r$ of length $r \geq 1$.

LEMMA 6.3. *Let $f \in C^0(U \times \Lambda \times [t_0, t_1]; \mathbb{R}^n)$, where $U \subset \mathbb{R}^n$ is open and connected, $\Lambda \subset \mathbb{R}^m$ is compact, and $t_0 < t_1$. Assume that $(x_0, \lambda_0) \in U \times \Lambda$ and that* (i) *$f(\cdot, \lambda, t)$ is locally Lipschitz on U, uniformly for $(\lambda, t) \in \Lambda \times [t_0, t_1]$, and* (ii) *$y : [t_0, t_1] \to U$ is a solution to $\dot{y} = f(y, \lambda_0, t)$, with $y(t_0) = x_0$. Then, given $\varepsilon > 0$, there are compact neighborhoods $U' \subset U$ and $\Lambda' \subset \Lambda$ of x_0 and λ_0, respectively, such that for every $x \in U'$ and every function $\varphi \in PC([t_0, t_1]; \Lambda')$, the system $\dot{z} = f(z, \varphi(t), t)$ admits a unique solution $z : [t_0, t_1] \to U$ which satisfies $z(t_0) = x$ and $\|z(t) - y(t)\| \leq \varepsilon$ for all $t \in [t_0, t_1]$.*

Fix $t_0 \in \mathbb{R}$. Let $C > 0$ be the constant whose existence is guaranteed by Lemma 6.2 above, and define $\eta > 0$ such that $CT(m+1)^{\frac{3}{2}} \eta < 1$. Setting $g_0 = \frac{1}{\eta} f_0$ and $g_i = f_i$, $i = 1, \dots, m$, we apply Lemma 6.2 again to deduce that there is a neighborhood K of $0 \in \mathbb{R}^n$ such that $|g_I \phi(x)| \leq C^r r!$ for every $x \in K$ and every multi-index I of length $r \geq 1$. Moreover, by defining $F(x, v, t) = \sum_{i=0}^m g_i(x) v_i$, with $v = (v_0, \dots, v_m)$, we see that F satisfies the assumptions of Lemma 6.3 if one takes $\lambda_0 = (\eta, 0, \dots, 0) \in \mathbb{R}^{m+1}$ and $y : t \mapsto 0 \in \mathbb{R}^n$. Therefore, there exists a constant $\delta' \in (0, \eta)$ such that if $x_0 \in \mathbb{R}^n$, with $\|x_0\| < \delta'$, and if v is a piecewise-continuous function on $[t_0, t_0 + T]$ taking values in $\{u \in \mathbb{R}^m : \|u\| < \delta'\}$, then the solution to $\dot{z} = g_0(z)\eta + \sum_{i=1}^m g_i(z) v_i(t)$ with initial value $z(t_0) = x_0$ satisfies $z(t) \in K$ for $t \in [t_0, t_0 + T]$. But since $\alpha(x, t) \to 0$ as $x \to 0$, uniformly for $t \in \mathbb{R}$, there exists $\delta \in (0, \delta')$ such that $\|\alpha(x, t)\| < \delta'$ whenever $\|x\| < \delta$ and $t \in \mathbb{R}$. It follows that if $\|x_0\| < \delta$, then the solution to system $\dot{x} = f_0(x) + \sum_{i=1}^m \alpha_i(x_0, t) f_i(x)$, $x(t_0) = x_0$, rewritten as $\dot{x} = g_0(x)\eta + \sum_{i=1}^m g_i(x) \alpha_i(x_0, t)$, satisfies $\pi(t, t_0, x_0) \in K$ for $t \in [t_0, t_0 + T]$. Note that, by denoting $v(x, t) = (\eta, \alpha(x, t))$, one has $\|v(x, t)\| < (m+1)^{\frac{1}{2}} \eta$ for $(x, t) \in K \times [t_0, t_0 + T]$. On the other hand, the difference between the

Nth partial sum of the Chen–Fliess expansion $\mathrm{Ser}^N_{\phi,\mathbf{f},\alpha}(t,t_0,x_0)$ and the actual value of ϕ along that solution is (cf. [27, section 4])

$$|\mathrm{Ser}^N_{\phi,\mathbf{f},\alpha}(t,t_0,x_0)-\phi(\pi(t,t_0,x_0))| \quad \le \\ \frac{((m+1)^{\frac{1}{2}}\eta(t-t_0))^{N+1}}{(N+1)!}(m+1)^{N+1}\sup\{|g_I\phi(x)| : x\in K\}.$$

But $\sup\{|f_I\phi(x)| : x\in K\} < C^{N+1}(N+1)!$ so

$$|\mathrm{Ser}^N_{\phi,\mathbf{f},\alpha}(t,t_0,x_0)-\phi(\pi(t,t_0,x_0))| \le (C(m+1)^{\frac{3}{2}}\eta(t-t_0))^{N+1}.$$

Since $t\in[t_0,t_0+T]$, one has $C(m+1)^{\frac{3}{2}}\eta(t-t_0)<1$; hence the series converges uniformly. It is readily checked that the series is absolutely convergent as well. ∎

6.2.2. Proof of Lemma 3.3. (i) Let $U'\subset\mathbb{R}^6$ be a compact set containing 0. From the continuity of α, the T-periodicity of $t\mapsto\alpha(x,t)$, and the definition of ρ, which implies that $\rho(x)=O(\|x\|^{\frac{1}{2}})$, one deduces that $\|\alpha_i(x,t)\|/\|x\|^{\frac{1}{2}}$ $(i=1,2)$ is bounded, say, by $K'>0$, for every $(x,t)\in U'\times\mathbb{R}$. Thus the claim holds for any $K>K'$.

(ii) Set $\alpha_0=1$ and write $\alpha_i(x,t)=U_i(x)+V_i(x)\cos(\omega t)$, $i=0,1,2$, with $U_0=1$, $V_0=0$, and U_1,U_2,V_1,V_2 defined in the obvious way. Note that $\mathrm{Ord}(U_0)=0$, $\mathrm{Ord}(U_1)=\mathrm{Ord}(U_2)=1$, and $\mathrm{Ord}(V_1)=\mathrm{Ord}(V_2)=1/2$. If $I=(i)\in\{1,2\}$, then $\int_0^T\alpha_I(x_0)=U_i(x)T+V_i(x)\int_0^T\cos(\omega\tau)d\tau=U_i(x)T$, so $\int_0^T\alpha_I(x_0)=\mathrm{O}(\|x_0\|)$. If $I=(i,j)\in\{0,1,2\}^2$, then

$$\begin{aligned}\int_0^T\alpha_I(x_0) &= \frac{1}{2}U_i\cdot U_j(x_0)T^2+U_i\cdot V_i(x_0)\int_0^T\int_0^{t_2}\cos(\omega t_1)dt_1dt_2 \\ &\quad + U_j\cdot V_i(x_0)\int_0^T\tau\cos(\omega\tau)d\tau+V_i\cdot V_j(x_0)\int_0^T\cos(\omega t_2)\int_0^{t_2}\cos(\omega t_1)dt_1dt_2 \\ &= \frac{1}{2}U_i\cdot U_j(x_0)T^2,\end{aligned}$$

since the three integrals indicated on the right member of this equation vanish. But then, if $I=(i,j)\neq(0,0)$, one gets $\mathrm{Ord}(\int_0^T\alpha_I)=\mathrm{Ord}(U_i)+\mathrm{Ord}(U_j)\ge 1$, so $\int_0^T\alpha_I(x_0)=\mathrm{O}(\|x_0\|)$.

(iii) One easily shows that if a function $v\in C^0(U\times\mathbb{R};\mathbb{R})$ satisfies $v(x_0,t)=\mathrm{O}(\|x_0\|^\ell)$, then $\int_0^t v(x_0,\tau)d\tau=\mathrm{O}(\|x_0\|^\ell)$ for every $t\in\mathbb{R}$. By writing $\mathrm{Ord}(\alpha_j)=k_j$, with $k_0=0$ and $k_1=k_2=\frac{1}{2}$, one gets

$$\int_0^t\alpha_j(x_0,t_2)\int_0^{t_2}v(x_0,t_1)dt_1dt_2=\mathrm{O}(\|x_0\|^{k_j+\ell}),\quad j=0,1,2. \tag{31}$$

Using these facts and an induction argument, one readily deduces that $\mathrm{Ord}(\int_0^T\alpha_I)\ge\sum_{j=1}^{|I|}k_{i_j}$.

(iv) This is verified directly by induction on the length of I using the fact that, for fixed $T\in\mathbb{R}$, $x_0\mapsto\int_0^T f(x_0,\tau)d\tau$ is continuous whenever $f\in C^0(U\times\mathbb{R};\mathbb{R})$.

(v) This claim follows immediately by inspecting the components of b_0, b_1, and b_2 as defined in (2), and from the definition of the set $\mathcal{A}$.

(vi) It suffices to show (by induction) that for any $x \in U$

$$b_0^k \phi(x) = \underbrace{b_0 \cdots b_0}_{k \text{ times}} \phi(x) = \sum_{\mu \in \{1,2,3\}^k} x_{2\mu} \frac{\partial^k \phi}{\partial x_{2\mu-1}}(x), \tag{32}$$

where we write $x_{2\mu} = x_{2\mu_1} \cdots x_{2\mu_k}$ and $x_{2\mu-1} = x_{2\mu_1-1} \cdots x_{2\mu_k-1}$ for any multi-index $\mu \in \{1,2,3\}^k$. Indeed, (v) follows from (32) since every term $f_\mu(x) = x_{2\mu} \partial^k \phi / \partial x_{2\mu-1}(x)$ in the sum satisfies $\mathrm{Ord}(f_\mu) \geq k$. Using (2), one gets $b_0 \phi(x) = \sum_{i=1}^3 x_{2i} \partial \phi / \partial x_{2i-1}(x)$, i.e., (32) with $k = 1$. If (32) holds for $k = m \geq 1$, then

$$\begin{aligned} b_0(b_0^m \phi)(x) &= \sum_{i=1}^3 x_{2i} \left(\sum_{\mu \in \{1,2,3\}^m} \frac{\partial x_{2\mu}}{\partial x_{2i-1}} \frac{\partial^m \phi}{\partial x_{2\mu-1}}(x) + x_{2\mu} \frac{\partial^{m+1} \phi}{\partial x_{2i-1} \partial x_{2\mu-1}}(x) \right) \\ &= \sum_i \sum_\mu x_{2i} x_{2\mu} \frac{\partial^{m+1} \phi}{\partial x_{2i-1} \partial x_{2\mu-1}}(x) = \sum_{\mu \in \{1,2,3\}^{m+1}} x_{2\mu} \frac{\partial^{m+1} \phi}{\partial x_{2\mu-1}}(x), \end{aligned}$$

since each of the terms $\partial x_{2\mu} / \partial x_{2i-1}$ is zero. Hence (32) holds for all $k \geq 1$. ∎

6.2.3. Proof of Lemma 3.4. (1) Note that, given the finiteness of $\mathcal{I}$, if for every $I \in \mathcal{I}$ the conclusion holds for $(x, \varepsilon) \mapsto a_I(x, \varepsilon) b_I(x)$ and some compact neighborhood $U_I \subset U$ of 0, then the conclusion holds for f by setting $U' = \bigcap_{I \in \mathcal{I}} U_I$. Let us then fix $I \in \mathcal{I}$.

(1)(i) Since a_I is real-analytic, we can write $a_I(x, \varepsilon) b_I(x) = [\frac{\partial a_I}{\partial x}(0, \varepsilon) x + \widetilde{a}_I(x, \varepsilon)] b_I(x)$, where $\varepsilon \mapsto \frac{\partial a_I}{\partial x}(0, \varepsilon)$ is continuous and vanishes at 0, and $(x, \varepsilon) \mapsto \widetilde{a}_I(x, \varepsilon)$ is a continuous mapping satisfying

$$(\forall \varepsilon \in E) \quad \lim_{x \to 0} \frac{\|\widetilde{a}_I(x, \varepsilon)\|}{\|x\|} = 0 \quad \text{and} \quad \widetilde{a}_I(\cdot, 0) = 0.$$

Given any compact neighborhood $U_I \subset U$ of 0, define $q(x, \varepsilon) = \frac{\|\widetilde{a}_I(x, \varepsilon)\|}{\|x\|}$, with $q(0, \cdot) = 0$, so that q is continuous—hence bounded—on $U_I \times E$ and $q(\cdot, 0) = 0$. We claim that $\sup_{x U_I} q(x, \varepsilon) \to 0$ as $\varepsilon \to 0$. Otherwise there would be $\eta > 0$ and a sequence $(\varepsilon_k)_{k \in \mathbb{N}}$, converging to zero, such that $\sup_{x \in U_I} q(x, \varepsilon_k) > 2\eta$ for every $k \in \mathbb{N}$. By the properties of sup, for every $k \in \mathbb{N}$ there would exist $x_k \in U_I$ such that

$$(\forall k \in \mathbb{N}) \quad q(x_k, \varepsilon_k) > \sup_{x \in U_I} q(x, \varepsilon_k) - \eta > \eta. \tag{33}$$

The compactness of U_I would imply the existence of a subsequence $(x_{k_j}, \varepsilon_{k_j})_{j \in \mathbb{N}}$, convergent towards a point $(\overline{x}, 0) \in U_I \times E$. But then, since q is continuous, $q(x_{k_j}, \varepsilon_{k_j})$ should converge to $q(\overline{x}, 0) = 0$, in contradiction to (33). Therefore, in view of the continuity of b_I, the conclusion follows since

$$\frac{\|a_I(x, \varepsilon) b_I(x)\|}{\|x\|} \leq \left(\left\| \frac{\partial a_I}{\partial x}(0, \varepsilon) \right\| + \sup_{x \in U_I} q(x, \varepsilon) \right) \sup_{x \in U_I} \|b_I(x)\|,$$

the right-hand member of which tends to zero as $\varepsilon \to 0$, uniformly for $x \in U_I$.

(1)(ii) The assumption $b_I(x) = \mathrm{O}(\|x\|)$ implies the existence of a constant $K_I > 0$ and a compact neighborhood $U_I \subset U$ of 0 such that $\|b_I(x)\| / \|x\| \leq K_I$ for every $x \in U_I$. Furthermore, by the real-analyticity of a_I, and using $a_I(\cdot, 0) = 0$, one can write $a_I(x, \varepsilon) b_I(x) = [\frac{\partial a_I}{\partial \varepsilon}(x, 0) \varepsilon + \widetilde{a}_I(x, \varepsilon)] b_I(x)$, with $\frac{\partial a_I}{\partial \varepsilon}(\cdot, 0)$ continuous and hence

bounded on U_I, say, $\|\frac{\partial a_I}{\partial \varepsilon}(x,0)\| \le K'$. Moreover, for every $x \in U_I$, the function $q(x,\varepsilon) = \frac{\|\tilde{a}_I(x,\varepsilon)\|}{|\varepsilon|}$ tends to 0 as $\varepsilon \to 0$, so q is also continuous on $U_I \times E$. Proceeding as in the proof of (1)(i) above, one readily shows that $\sup_{x\in U_I} q(x,\varepsilon) \to 0$ as $\varepsilon \to 0$. The conclusion then follows directly from the inequality

$$\begin{aligned}\frac{\|a_I(x,\varepsilon)b_I(x)\|}{\|x\|} &\le \left(\left\|\frac{\partial a_I}{\partial \varepsilon}(x,0)\right\| + q(x,\varepsilon)\right)|\varepsilon|\frac{\|b_I(x)\|}{\|x\|}\\ &\le K_I\left(K' + \sup_{x\in U_I} q(x,\varepsilon)\right)|\varepsilon|.\end{aligned}$$

(2) Let $U' \subset U$ be any compact neighborhood of 0. We claim that if a number $\eta > 0$ exists such that, for every $I \in \mathcal{I}$,

$$(x,\varepsilon) \mapsto \frac{\|a_I(x,\varepsilon)b_I(x)\|}{\|x\|^{1+\eta}} \quad \text{is continuous on } U' \times E, \tag{34}$$

then the conclusion holds for f and U'. Indeed, $\sum_{I\in\mathcal{I}} a_I(x,\varepsilon)b_I(x)$ converges absolutely and uniformly; therefore

$$\frac{\|f(x,\varepsilon)\|}{\|x\|} \le \|x\|^{\eta}\sum_{I\in\mathcal{I}}\frac{\|a_I(x,\varepsilon)b_I(x)\|}{\|x\|^{1+\eta}}, \tag{35}$$

and the series on the right side of (35) converges to a function that is bounded on $U' \times E$. Consequently, the term on the left side of (35) tends to zero as $x \to 0$, uniformly for $\varepsilon \in E$, and this proves the claim. The rest of the proof simply consists of exhibiting such a number η, independent of I, for each case.

(2)(i) The assumption $a_I(x,\varepsilon) = \mathrm{o}(\|x\|)$ and the real-analyticity of a_I imply that all terms in x of degrees < 2 in the Taylor expansion of $x \mapsto a_I(x,\varepsilon)$ at 0 vanish identically. Thus, for every $\varepsilon \in E$, $\|a_I(x,\varepsilon)\|/\|x\|^{1+\eta} \to 0$ as $x \to 0$ whenever $\eta \le 1$. By continuity of b_I, (34) holds with $\eta = 1/2$.

(2)(ii) Since $b_I(x) = \mathrm{O}(\|x\|^{1+c})$, then $\|b_I(x)\|/\|x\|^{1+c/2} \to 0$ as $x \to 0$. Taking $\eta = c/2$ we see that (34) holds.

(2)(iii) The assumptions imply that both $\|a_I(x,\varepsilon)\|/\|x\|^{1-d/3}$ and $\|b_I(x)\|/\|x\|^{2d/3}$ tend to zero as $x \to 0$. Thus, (34) holds with $\eta = d/3$. □

6.2.4. Proof of Lemma 6.2. This is a straightforward adaptation of the proof of [27, Lemma 4.2]. Since the result is local one may assume, without loss of generality, that M is an open subset of $\mathbb{R}^n$ and that $\overline{x} = 0$. By real-analyticity, the mappings ϕ and $f_0, \dots, f_m$ may be extended to complex analytic mappings defined on a polydisc $D(n,\alpha) = \{(z_1,\dots,z_n) \in \mathbb{C}^n : |z_i| < \alpha,\ i = 1,\dots,n\}$ for some $\alpha > 0$. Denote the corresponding extensions by $\tilde{\phi}$ and $\tilde{f}_0, \dots, \tilde{f}_m$, respectively.

By Stirling's formula, there is a constant C'' such that $r^r \le C'' e^r r!$ for all $r \ge 1$. Let $C' = \max\{|\tilde{\phi}(q)| : q \in D(n, \frac{2}{3}\alpha)\}$ and define

$$C = e\max\{1, C'C''\}.$$

Select $\eta > 0$ arbitrarily. Then the vector fields $\tilde{g}_0 = (1/\eta)\tilde{f}_0$, $\tilde{g}_i = \tilde{f}_i$ $(i = 1,\dots,m)$ are analytic extensions of the vector fields $g_0, \dots, g_m$, respectively, to the set $D(n,\alpha)$. Consider the complex control system

$$\dot{z} = \sum_{i=0}^{m} v_i \tilde{g}_i(z), \quad z \in \mathbb{C}^n, \quad v = (v_0,\dots,v_m) \in \mathbb{C}^{m+1}. \tag{36}$$

Clearly, if $v = (\eta, 0, \dots, 0)$, then $(z, v) = (0, v)$ is an equilibrium point; hence the corresponding constant solution $t \mapsto (0, v)$ is defined for $t \in [0, 1]$. Since (36) may be rewritten as a real control system on $\mathbb{R}^{2n}$, one can apply Lemma 6.3 to conclude that there is $\delta \in (0, \frac{2}{3}\alpha)$ with the property that, whenever $z_0 \in D(n, \delta)$, $v_0 \in A = \{w \in \mathbb{C} : |w - \eta| < \delta\}$, and $v_i \in D(1, \delta)$ $(i = 1, \dots, m)$, the system (36) has a unique solution $z : [0, 1] \to \mathbb{C}^n$ satisfying $z(t_0) = z_0$ and $z(t) \in D(n, \frac{2}{3}\alpha)$ for $t \in [0, 1]$.

Let us fix a multi-index $I = (i_1, \dots, i_r) \in \{0, \dots, m\}^r$, $r \geq 1$, and define the set $D^I = \prod_{j=1}^r D_j^I$, where $D_j^I = A$ if $i_j = 0$ and $D_j^I = D(1, \delta)$ otherwise. D^I is thus an open subset of $\mathbb{C}^r$. For any $z \in D^I$, define the input function $v^{I,z} : [0, 1] \to A \times D(m, \delta)$ by setting, for $j = 1, \dots, r$ and $t \in [\frac{j-1}{r}, \frac{j}{r})$, $v^{I,z}(t) = z_j e_{i_j}$ (here $\{e_0, \dots, e_m\}$ denotes the canonical basis of the $\mathbb{C}$-vector space $\mathbb{C}^{m+1}$). The function $v^{I,z}$ thus defined is a piecewise-constant function on $[0, 1]$ taking values in $A \times D(m, \delta)$. Therefore, for any $q \in D(n, \delta)$, the solution $t \mapsto \xi_q^{I,z}(t)$ to system (36), with input $v^{I,z}$ and initial condition $\xi_q^{I,z}(0) = q$, is defined and satisfies $\xi_q^{I,z}(t) \in D(n, \frac{2}{3}\alpha)$ for all $t \in [0, 1]$. Moreover, for $i = 1, \dots, r$, $\xi_q^{I,z}(i/r)$ is analytic in q and z for $(q, z) \in D(n, \delta) \times D^I$ (cf. [27, proof of Lemma 4.2]). Now define a mapping $\psi_I : D(n, \delta) \times D^I \to \mathbb{C}$ by setting $\psi_I(q, z) = \tilde{\phi}(\xi_q^{I,z}(1))$. Then

$$\frac{\partial^r \psi_I}{\partial z_r \cdots \partial z_1}(q, 0) = \left(\frac{1}{r}\right)^r (\tilde{g}_{i_1} \cdots \tilde{g}_{i_r} \tilde{\phi})(q).$$

This is readily shown by extending the following basic argument using induction on the length r of $I = (i_1, \dots, i_r)$. For a vector field $X = (X_1, \dots, X_n)$ defined on some subset B of $\mathbb{C}^n$, denote by $t \mapsto \Phi_q^X(t)$ the local flow of X satisfying $\Phi_q^X(0) = q \in B$. Hence $d\Phi_{q,k}^X/dt = X_k(\Phi_q^X(t))$ for $k = 1, \dots, n$. It is easy to check that $\Phi_q^{zX}(t) = \Phi_q^X(zt)$ when $z \in \mathbb{C}$ and $|z|$ is small enough. Thus, by setting $\psi_{(i)}(q, z) = \tilde{\phi}(\Phi_q^{z\tilde{g}_i}(\tau)) = \tilde{\phi}(\Phi_q^{\tilde{g}_i}(z\tau))$, with $i \in \{0, \dots, m\}$, one gets

$$\begin{aligned}\frac{\partial \psi_{(i)}}{\partial z}(q, z) &= \sum_{k=1}^n \frac{\partial \tilde{\phi}}{\partial r_k}\left(\Phi_q^{\tilde{g}_i}(z\tau)\right) \frac{d\Phi_{q,k}^{\tilde{g}_i}}{dt}(z\tau) \frac{d}{dz}(z\tau) \\ &= \tau \sum_{k=1}^n \frac{\partial \tilde{\phi}}{\partial r_k}\left(\Phi_q^{\tilde{g}_i}(z\tau)\right) \tilde{g}_{i,k}(\Phi_q^{\tilde{g}_i}(z\tau)),\end{aligned}$$

and then, if $\tau = 1$ and $z = 0$ (and since r=1 because $I = (i)$),

$$\frac{\partial \psi_I}{\partial z}(q, 0) = \sum_{k=1}^n \frac{\partial \tilde{\phi}}{\partial r_k}(q) \tilde{g}_{i,k}(q) = \left(\frac{1}{r}\right)^r (\tilde{g}_i \tilde{\phi})(q).$$

Since $q \mapsto \psi_I(q, z)$ is analytic on $D(n, \delta)$ for $z \in D^I$, Cauchy's estimates yield

$$\left|\frac{\partial^r \psi_I}{\partial z_r \cdots \partial z_1}(q, 0)\right| \leq \max\left\{|\tilde{\phi}(q')| : q' \in D\left(n, \frac{2}{3}\alpha\right)\right\} = C',$$

and this implies in turn that $|\tilde{g}_{i_1} \cdots \tilde{g}_{i_r} \tilde{\phi}(q)| \leq C' r^r$ for any $q \in D(n, \frac{2}{3}\alpha)$. By definition of C'', one has $C' r^r \leq C' C'' e^r r!$ Also, $C' C'' \leq \max\{1, C' C''\} = C/e$. Using the fact that $1 \leq C/e$, and hence $C/e \leq (C/e)^r$, one gets $C' C'' e^r r! \leq C^r r!$ for $r \geq 1$. Therefore, by setting $K = D(n, \delta) \cap \mathbb{R}^n$, one concludes that

$$|(g_{i_1} \cdots g_{i_r} \phi)(x)| \leq C^r r!$$

for $x \in K$ and $r \geq 1$. Since the constant C was selected independently of η, r, and I, this finishes the proof. □

REFERENCES

[1] N. Aneke, D. Lizárraga, and H. Nijmeijer, *Homogeneous stabilization of the extended chained form system*, in Proceedings of the IFAC 15th Triennial World Congress, Barcelona, Spain, 2002, paper 498.

[2] M. Bennani and P. Rouchon, *Robust stabilization of flat and chained systems*, in Proceedings of the European Control Conference, Rome, Italy, 1995, pp. 2642–2646.

[3] R. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, Progress in Mathematics 27, R. Brockett, R. Millman, and H. J. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 181–191.

[4] O. Egeland, M. Dalsmo, and O. J. Sørdalen, *Feedback control of a nonholonomic underwater vehicle with constant desired configuration*, Internat. J. Robotics Res., 15 (1996), pp. 24–35.

[5] O. Egeland, E. Berglund, and O. Sørdalen, *Exponential stabilization of a nonholonomic underwater vehicle with constant desired configuration*, in IEEE Conference on Robotics and Automation, 1994, pp. 20–25.

[6] S. Ge, Z. Sun, T. Lee, and M. Spong, *Feedback linearization and stabilization of second-order non-holonomic chained systems*, Internat. J. Control, 74 (2001), pp. 1383–1392.

[7] H. Hermes, *Nilpotent and high-order approximations of vector field systems*, SIAM Review, 33 (1991), pp. 238–264.

[8] J. Imura, K. Kobayashi, and T. Yoshikawa, *Nonholonomic control of* 3 *link planar manipulator with a free joint*, in IEEE Conference on Decision and Control, Kobe, Japan, 1996, pp. 1435–1436.

[9] M. Kawski, *Geometric homogeneity and stabilization*, in Proceedings of the IFAC Nonlinear Control Systems Design Symposium, 1995, pp. 164–169.

[10] H. Khalil, *Nonlinear Systems*, 2nd ed., Prentice–Hall, Upper Saddle River, NJ, 1996.

[11] M.-C. Laiou and A. Astolfi, *Discontinuous control of high-order generalized chained systems*, Systems Control Lett., 37 (1999), pp. 309–322.

[12] W. Liu, *An approximation algorithm for nonholonomic systems*, SIAM J. Control Optim., 35 (1997), pp. 1328–1365.

[13] D. Lizárraga, P. Morin, and C. Samson, *Non-Robustness of Continuous Homogeneous Stabilizers for Affine Control Systems*, in Proceedings of the IEEE Conference on Decision and Control, pp. 855–860.

[14] P. Lucibello and G. Oriolo, *Robust stabilization via iterative state steering with an application to chained-form systems*, Automatica J. IFAC, 37 (2001), pp. 71–79.

[15] N. H. McClamroch, I. Kolmanovsky, S. Cho, and M. Reyhanoglu, *Control problems for planar motion of a rigid body with an unactuated internal degree of freedom*, in Proceedings of the American Control Conference, Philadelphia, 1998, pp. 229–233.

[16] R. M'Closkey and R. M. Murray, *Nonholonomic systems and exponential convergence: Some analysis tools*, in Proceedings of the IEEE Conference on Decision and Control, 1993, pp. 943–948.

[17] P. Morin and C. Samson, *Exponential stabilization of nonlinear driftless systems with robustness to unmodeled dynamics*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 1–35.

[18] P. Morin, C. Samson, J.-B. Pomet, and Z.-P. Jiang, *Time-varying feedback stabilization of the attitude of a rigid spacecraft with two controls*, Systems Control Lett., 25 (1995), pp. 375–385.

[19] R. R. Murray and S. Sastry, *Steering Nonholonomic Systems Using Sinusoids*, in Proceedings of the IEEE Conference on Decision and Control, Honolulu, HI, 1990, pp. 2097–2101.

[20] W. Oelen, H. Berghuis, H. Nijmeijer, and C. Canudas de Wit, *Hybrid stabilizing control on a real mobile robot*, IEEE Robotics & Automation Magazine, 2 (1995), pp. 16–23.

[21] G. Oriolo and Y. Nakamura, *Control of mechanical systems with second-order nonholonomic constraints: Underactuated manipulators*, in Proceedings of the IEEE Conference on Decision and Control, Brighton, England, 1991, pp. 2398–2403.

[22] M. Rathinam and R. M. Murray, *Configuration flatness of Lagrangian systems underactuated by one control*, SIAM J. Control Optim., 36 (1998), pp. 164–179.

[23] M. Reyhanoglu, A. van der Schaft, N. McClamroch, and I. Kolmanovsky, *Nonlinear control of a class of underactuated systems*, in Proceedings of the IEEE Conference on Decision and Control, Kobe, Japan, 1996, pp. 1682–1687.

[24] M. Reyhanoglu, A. van der Schaft, N. H. McClamroch, and I. Kolmanovsky, *Dynamics and Control of a Class of Underactuated Mechanical Systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 1663–1671.

[25] S. Sastry, *Nonlinear Systems. Analysis, Stability and Control*, Interdiscip. Appl. Math. 10, Springer-Verlag, Heidelberg, 1999.

[26] O. Sørdalen, *Conversion of the kinematics of a car with n trailers into a chained form*, in Proceedings of the IEEE Conference on Robotics and Automation, 1993, pp. 382–387.

[27] H. J. Sussmann, *Lie brackets and local controllability: A sufficient condition for scalar-input systems*, SIAM J. Control and Optimization, 21 (1983), pp. 686–713.

[28] A. R. Teel, R. M. Murray, and G. C. Walsh, *Nonholonomic control systems: From steering to stabilization with sinusoids*, Internat. J. Control, 62 (1995), pp. 849–870.

[29] D. Tilbury, O. Sørdalen, L. Bushnell, and S. Sastry, *A multisteering trailer system: Conversion into chained form using dynamic feedback*, IEEE Transactions on Robotics and Automation, 11 (1995), pp. 807–818.

[30] K. Wichlund, O. Sørdalen, and O. Egeland, *Control of vehicles with second-order nonholonomic constraints: Underactuated vechicles*, in Proceedings of the European Control Conference, Rome, Italy, 1995, pp. 3086–3091.

[31] G. Xu and D. Wang, *Exponential stabilization of extended chained forms*, in Proceedings of the American Control Conference, Chicago, IL, 2000, pp. 3259–3263.

[32] T. Yoshikawa, K. Kobayashi, and T. Watanabe, *Design of a desirable trajectory and convergent control for* 3*-D.O.F. manipulator with a nonholonomic constraint*, in Proceedings of the IEEE Conference on Robotics and Automation, San Francisco, CA, 2000, pp. 1805–1810.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 2200–2220

ANALYSIS AND CONTROL OF HOPF BIFURCATIONS*

BOUMEDIENE HAMZI†, WEI KANG‡, AND JEAN-PIERRE BARBOT§

Abstract. In this paper, control systems with two uncontrollable modes on the imaginary axis are studied. The main contributions include the local orientation control of periodic solutions and center manifolds, the quadratic normal form of systems with two imaginary uncontrollable modes, the stabilization of the Hopf bifurcation by state feedback, and the quadratic invariants that characterize the nonlinearity of a system and its Hopf bifurcation.

Key words. nonlinear systems, normal forms, quadratic invariants, Hopf bifurcation, center manifold, control

AMS subject classifications. 93C10, 93C15, 37L10, 37N35

DOI. 10.1137/S0363012900372714

1. Introduction. Nonlinear dynamical systems exhibit complicated performance around bifurcation points. As the parameter of a system is varied, changes may occur in the qualitative structure of its solutions around a point of bifurcation. Using a feedback to stabilize a system with a bifurcation has been studied by many authors (see, for instance, [1], [2], [7], [10], [12], [21], and [23]). Bifurcation phenomenon appears in a large family of engineering systems. The control of bifurcations has attracted increasing attention in recent years, motivated by engineering applications such as the control of surge and rotating stall in engine compressors [20], [21], and the control of voltage instabilities and collapse [25]. It is known that bifurcations in a linearly controllable system can be delayed or stabilized by a linear feedback [1], [8]. However, nonlinear feedback is essential for systems with uncontrollable bifurcation modes. The authors of [3], [1], [2], [5], [9], and [12] studied the bifurcation and/or stability of systems with one and two uncontrollable modes. Using normal forms, the author of [16], [17], [18] developed the analysis and control design algorithm for systems with one uncontrollable mode.

In this paper, control systems with two uncontrollable modes on the imaginary axis are studied. The main contributions include the local orientation control of periodic solutions and center manifolds, the quadratic normal form of systems with the Hopf bifurcation, the stabilization of the Hopf bifurcation by state feedback, and the quadratic invariants that characterize the nonlinearity of a system and its Hopf bifurcation. In this paper, "bifurcation control" means the control of the local orientation of the periodic solution and the stabilization of the Hopf bifurcation. The results in this paper focus on local bifurcations. Global bifurcations are not addressed here.

The Hopf bifurcation studied in this paper occurs in the Moore–Greitzer model of axial flow engine compressors. A simple version of this model is a four-dimensional ordinary differential equation with nonlinear dynamics exhibiting several kinds of

*Received by the editors May 12, 2000; accepted for publication (in revised form) August 18, 2003; published electronically April 7, 2004.
http://www.siam.org/journals/sicon/42-6/37271.html
†Department of Mathematics, University of California, Kerr Hall, One Shields Avenue, Davis, CA 95616 (hamzi@math.ucdavis.edu).
‡Mathematics Department, Naval Postgraduate School, Monterey, CA 93943 (wkang@math.nps.navy.mil).
§ECS, ENSEA, 6 avenue du Ponceau, Cergy 95014, France (barbot@ensea.fr).

bifurcations. When the pressure rise in the system arrives at a critical value, the operating point loses its linear stability, and a Hopf bifurcation takes place. Meanwhile, the critical modes are not linearly controllable. The resulting limit cycle in the system is called a rotating stall, which significantly reduces the efficiency of the compressor [22], [20], [21].

The mathematical analysis in this paper is based on the linear and quadratic normal form of control systems. In section 2, we derive the linear normal form of the control system and its linear center manifold. Then we derive the relationship between the local orientation of center manifold and the state feedback. A necessary and sufficient condition is found in order for the local orientation of center manifold to be achievable by state feedback. In section 3, the quadratic normal form is found for control systems with two imaginary uncontrollable modes. The theorem is proved in the appendix. Then an explicit formula of the quadratic center manifold is derived for the normal form. Sufficient conditions are derived for the tuning of a nonlinear control law to render the Hopf bifurcation supercritical. In section 4, the explicit formula for the coefficients in the normal form is derived based on the Lie operator and the Lie bracket. A set of invariants is found, the value of which does not change under the quadratic change of coordinates and feedback. It is proved that two systems have equivalent quadratic parts if and only if their invariants have the same value. Furthermore, the value of the invariants equals the value of the coefficients in the normal form.

The normal form approach adopted in this paper generalizes Poincaré's normal form method, commonly used in dynamical systems theory, to the area of control systems. It was first introduced in [19], [14]. The results proved in this paper indicate that the linear control determines the local orientation of the periodic solution around the origin, and the quadratic feedback is critical to stabilize the periodic solution. The analysis is based on the center manifold theorem [6] and the Poincaré–Andronov–Hopf theorem [26]. The results in this paper are part of the Ph.D. thesis [13].

2. The orientation of center manifold. Consider the following nonlinear system:

$$\dot{\zeta} = f(\zeta, \mu) + g(\zeta, \mu)\, v. \tag{2.1}$$

The variable $\zeta \in \mathbb{R}^n$ is the state, $v \in \mathbb{R}$ is the input variable, and $\mu \in \mathbb{R}$ is the parameter. The vector fields $f(\zeta, \mu)$ and $g(\zeta, \mu)$ are assumed to be C^k for some sufficiently large k.

Assume $f(0,0) = 0$, $g(0,0) \neq 0$. Suppose that the linearization of the system at the origin is (A, B),

$$A = \frac{\partial f}{\partial \zeta}(0,0), \quad B = g(0,0),$$

with

$$\operatorname{rank}([B\ AB\ A^2B\ \cdots\ A^{n-1}B]) = n - 2. \tag{2.2}$$

Thus, the system is not linearly controllable at the origin. If the uncontrollable modes have nonzero real parts, the stabilizability of the system is determined by the real part of the uncontrollable modes. However, if the real part of the uncontrollable modes is zero, bifurcation occurs even with feedback control. In this paper, we assume that $\pm i\,\omega$ are the uncontrollable modes.

Assumption 1. The linearization of (2.1) has two uncontrollable modes, $\pm i\,\omega$, at the origin.

In this paper, a general nonlinear system is transformed to a simpler system, which is called normal form. Bifurcation analysis based on the Poincaré normal form is a well-known theory in dynamical systems. The control system normal form derived in the present paper is different from those used in the literature of nonlinear dynamical systems without control inputs. Why is it necessary to introduce the control system normal form instead of adopting the Poincaré normal form of vector fields? In fact, even for a linear control system $\dot{x} = Ax + Bu$, the controller normal form is more useful than the diagonal form of A in the feedback design. The normal form of nonlinear control systems generalizes the linear controller form. An affine control system $\dot{x} = f(x) + g(x)u$ has two vector fields $f(x)$ and $g(x)$. Therefore, the normal form of a control system requires the simplification of both f and g simultaneously. The simplification of f does not necessarily result in a simple form for g. Furthermore, the transformation group of control systems consists of changes of coordinates and feedbacks. This is different from the normal form theory of dynamical systems where feedbacks are not considered.

The first step of the analysis is to simplify the linear part of the system (i.e., the determination of the linear normal form). There exist a linear change of coordinates and a feedback independent of μ transforming the system (2.1) into

$$\begin{aligned} \begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} &= A_1 \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \Gamma_1 \mu + O(z,\mu,x,u)^2, \\ \dot{x} \quad &= A_2 x + B_2 u + \Gamma_2 \mu + O(z,\mu,x,u)^2, \end{aligned} \tag{2.3}$$

with $\Gamma_1 = \begin{bmatrix} \gamma_{1\,1} & \gamma_{1\,2} \end{bmatrix}^T$, $\Gamma_2 = \begin{bmatrix} \gamma_{2\,1} & \gamma_{2\,2} & \cdots & \gamma_{2\,n-2} \end{bmatrix}^T$, and $z \in \mathbb{R}^2$, $x \in \mathbb{R}^{n-2}$:

$$A_1 = \begin{bmatrix} 0 & -\omega \\ \omega & 0 \end{bmatrix},\ A_2 = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}_{(n-2)\times(n-2)},\ B_2 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}_{(n-2)\times 1}. \tag{2.4}$$

The subsystem (A_2, B_2) is the Brunovsky form of controllable linear systems.

The system can be simplified further. The μ-terms can be canceled by the change of coordinates: $\bar{z}_1 = z_1 + \frac{\gamma_{1\,2}}{\omega}\mu$, $\bar{z}_2 = z_2 - \frac{\gamma_{1\,1}}{\omega}\mu$, $\bar{x}_1 = x_1$, $\bar{x}_i = x_i + \gamma_{2\,i-1}\mu$ for $i = 2, \ldots, n$ and a feedback $\bar{u} = u + \gamma_{2\,n}\mu$. The linear normal form of (2.1) is summarized in the following lemma

LEMMA 2.1. *There exists a linear change of coordinates and feedback which transforms* (2.1) *into*

$$\begin{aligned} \dot{z} &= A_1 z + f_1^{[2]}(z,\mu,x) + g_1^{[1]}(z,\mu,x)u + O(z,\mu,x,u)^3, \\ \dot{x} &= A_2 x + B_2 u + f_2^{[2]}(z,\mu,x) + g_2^{[1]}(z,\mu,x)\,u + O(z,\mu,x,u)^3. \end{aligned} \tag{2.5}$$

2.1. The linear center manifold. Consider

$$\begin{aligned} \dot{z} &= A_1 z + O(z,\mu,x,u)^2, \\ \dot{x} &= A_2 x + B_2 u + O(z,\mu,x,u)^2, \end{aligned} \tag{2.6}$$

and the feedback

$$u = F_1 z_1 + F_2 z_2 + F_3 \mu + \sum_{i=1}^{n-2} a_i x_i + O(z,\mu,x)^2. \tag{2.7}$$

To stabilize the system around the bifurcation point, the controllable part has to be stable. So, we assume the following.

Assumption 2. The matrix $A_2 + B_2 \begin{bmatrix} a_1 & \cdots & a_{n-2} \end{bmatrix}$ is Hurwitz.

Suppose that the center manifold of the closed-loop system (2.6)–(2.7) is

$$x = \Pi(z,\mu).$$

Suppose that the linear part of Π is

$$\Pi^{[1]}(z,\mu) = \Pi^{[1]} \begin{bmatrix} z \\ \mu \end{bmatrix}, \qquad \Pi^{[1]} = \begin{bmatrix} \Pi_{11}^{[1]} & \Pi_{12}^{[1]} & \Pi_{13}^{[1]} \\ \vdots & \vdots & \vdots \\ \Pi_{n-2,1}^{[1]} & \Pi_{n-2,2}^{[1]} & \Pi_{n-2,3}^{[1]} \end{bmatrix} \in \mathbb{R}^{(n-2)\times 3}, \tag{2.8}$$

where $\Pi^{[1]}(z,\mu)$ is a function and $\Pi^{[1]}$ represents the matrix of the linear function. The center manifold equation is

$$\frac{\partial \Pi(z,\mu)}{\partial (z,\mu)} \begin{bmatrix} \dot z \\ \dot\mu \end{bmatrix} = A_2 \Pi + B_2 \left(\sum_{i=1}^{n-2} a_i \Pi_i(z,\mu) + F_1 z_1 + F_2 z_2 + F_3 \mu \right) + O(z,\mu,x)^2.$$

The linear part of the equation is equivalent to

$$\begin{aligned} &\Pi_{i+1}^{[1]}(z,\mu) = \frac{\partial \Pi_i^{[1]}}{\partial (z,\mu)} \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} z \\ \mu \end{bmatrix}, \quad i \le n-3, \\ &\sum_{i=1}^{n-2} a_i \Pi_i^{[1]}(z,\mu) + F_1 z_1 + F_2 z_2 + F_3 \mu = \frac{\partial \Pi_{n-2}^{[1]}}{\partial (z,\mu)} \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} z \\ \mu \end{bmatrix}, \end{aligned}$$

where $\Pi_i^{[1]}$ represents the ith row of $\Pi^{[1]}$. So,

$$\begin{aligned} &\Pi_{i+1}^{[1]} = \Pi_i^{[1]} \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix}, \quad i \le n-3, \\ &\sum_{i=1}^{n-2} a_i \, \Pi_i^{[1]} + \begin{bmatrix} F_1 & F_2 & F_3 \end{bmatrix} = \Pi_{n-2}^{[1]} \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

Hence,

$$\begin{aligned} &\Pi_{i+1}^{[1]} = \Pi_1^{[1]} \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix}^i, \quad i \le n-3, \\ &\sum_{i=1}^{n-2} a_i \, \Pi_1^{[1]} \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix}^{i-1} + \begin{bmatrix} F_1 & F_2 & F_3 \end{bmatrix} = \Pi_1^{[1]} \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix}^{n-2}. \end{aligned}$$

Let $P(\lambda)$ be the characteristic polynomial of $A_2 + B_2 \begin{bmatrix} a_1 & \cdots & a_{n-2} \end{bmatrix}$; then

$$P(\lambda) = \lambda^{n-2} - \sum_{i=1}^{n-2} a_i \lambda^{i-1}. \tag{2.9}$$

So

$$\begin{cases} \Pi_1^{[1]} P\left(\begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} F_1 & F_2 & F_3 \end{bmatrix}, \\ \Pi_{i+1}^{[1]} = \Pi_1^{[1]} \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix}^i. \end{cases} \tag{2.10}$$

If $a_1 \neq 0$, from the expression of $P(\lambda)$ we obtain

$$\begin{cases} \Pi_1^{[1]} = \begin{bmatrix} F_1 & F_2 & F_3 \end{bmatrix} \begin{bmatrix} P(A_1)^{-1} & 0 \\ 0 & -\frac{1}{a_1} \end{bmatrix}, \\ \Pi_{i+1}^{[1]} = \Pi_1^{[1]} \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix}^i. \end{cases} \tag{2.11}$$

This formula holds only if $P(A_1)$ is invertible, and $a_1 \neq 0$. This is always true. Because of Assumption 2, the eigenvalues of $A_2 + B_2[a_1 \ a_2 \ \cdots \ a_{n-2}]$ are not on the imaginary axis. Therefore, the roots of the characteristic polynomial $P(\lambda)$ are not on the imaginary axis. Therefore, $P(\pm i\omega) \neq 0$. On the other hand, the eigenvalues of $P(A_2)$ are $P(\pm i\omega)$, which are nonzero. Therefore, the matrix $P(A_2)$ is invertible. Furthermore, the value $(-1)^{n-1}a_1$ equals the product of all eigenvalues of $A_2 + B_2[a_1 \ a_2 \ \cdots \ a_{n-2}]$. If all the eigenvalues are on the left half plane, it is easy to check that $a_1 < 0$.

THEOREM 2.2. *Suppose* $\begin{bmatrix} a_1, \ldots, a_{n-2} \end{bmatrix}$ *stabilizes the x-subsystem. Given any vector* $\Pi_1^{[1]} \in \mathbb{R}^{1\times 3}$*, there exist unique vectors* $\Pi_i^{[1]} \in \mathbb{R}^{1\times 3}$*,* $i > 1$*, and* $\begin{bmatrix} F_1, F_2, F_3 \end{bmatrix}$ *given by* (2.11)*, such that*

$$x = \Pi^{[1]} \begin{bmatrix} z \\ \mu \end{bmatrix} \tag{2.12}$$

is the linear part of the center manifold of the closed-loop system.

2.2. Orientation. In this subsection, a center manifold $x = \Pi(z, \mu)$ is treated as a submanifold in the space of $(z, \mu, x) \in \mathbb{R}^{n+1}$. A row vector in $\mathbb{R}^{n+1}$ is orthogonal to the manifold at the origin if the vector is orthogonal to the tangent space of the manifold at the origin. The orientation of a center manifold at the origin is a set of vectors. The vectors are orthogonal to the manifold, linearly independent, and they generate a complement subspace of the manifold. In other words, the orientation of the center manifold at the origin is a basis of the orthogonal complement subspace of the tangent space of the center manifold.

THEOREM 2.3. *Given any* $(n-2) \times (n+1)$ *matrix of the form*

$$[\mathcal{M}_{(n-2)\times 3} \ \mathcal{N}_{(n-2)\times(n-2)}],$$

its row vectors define the center manifold orientation at the origin for (2.6)–(2.7) *if and only if* $\mathcal{N}^{-1}$ *exists and* $\Pi^{[1]} = -\mathcal{N}^{-1}\mathcal{M}$ *satisfies* (2.11)*.*

Proof. Suppose that $[\mathcal{M}_{(n-2)\times 3} \ \mathcal{N}_{(n-2)\times(n-2)}]$ defines the orientation of a center manifold. Then it is orthogonal to the tangent space of the center manifold.It is

known that the tangent space of the center manifold is given by its linear part,

$$x - \Pi^{[1]} \begin{bmatrix} z \\ \mu \end{bmatrix} = 0,$$

where $\Pi^{[1]}$ satisfies (2.11). In the (z, μ, x) space, a set of orthogonal vectors of the tangent space is the row vectors of $[-\Pi^{[1]}\ I]$. Therefore, both $[-\Pi^{[1]}\ I]$ and $[\mathcal{M}_{(n-2)\times 3}\ \mathcal{N}_{(n-2)\times(n-2)}]$ generate the same space, which is orthogonal to the tangent space of the center manifold. Therefore, the row vectors of $[-\Pi^{[1]}\ I]$ are linear combinations of the row vectors in $[\mathcal{M}_{(n-2)\times 3}\ \mathcal{N}_{(n-2)\times(n-2)}]$, i.e.,

$$[-\Pi^{[1]}\ I] = \mathcal{N}^{-1}[\mathcal{M}_{(n-2)\times 3}\ \mathcal{N}_{(n-2)\times(n-2)}].$$

So, $\Pi^{[1]} = -\mathcal{N}^{-1}\mathcal{M}$, and it satisfies (2.11).

On the other hand, suppose $-\mathcal{N}^{-1}\mathcal{M}$ satisfies (2.11). By Lemma 2.2, the linear space

$$\mathcal{N}^{-1}\mathcal{M} \begin{bmatrix} z \\ \mu \end{bmatrix} + x = 0$$

represents the linear part of the center manifold. It is the tangent space of the center manifold. Therefore, $[\mathcal{N}^{-1}\mathcal{M}\ \ I]$, the row vectors in the coefficient matrix of this equation, forms a basis of the orthogonal space. It is easy to check that the row vectors of $[\mathcal{M}\ \mathcal{N}]$ and $[\mathcal{N}^{-1}\mathcal{M}\ \ I]$ generate the same vector space. Therefore, $[\mathcal{M}\ \ \mathcal{N}]$ defines the orientation of the center manifold. □

3. Quadratic normal form and Hopf bifurcation. This section has two parts. The quadratic normal form is introduced in the first part. Then the relationship between the quadratic feedback and the stability of the Hopf bifurcation is derived. As a corollary, a design algorithm is found. The feedback designed using this method stabilizes the periodic solution in the Hopf bifurcation.

3.1. Quadratic normal form and center manifold. The following quadratic transformations are employed to simplify the quadratic part of a system into its normal form while leaving the linear part invariant:

$$\begin{bmatrix} z_1 \\ z_2 \\ x \end{bmatrix} = \begin{bmatrix} \bar{z}_1 \\ \bar{z}_2 \\ \bar{x} \end{bmatrix} + \phi^{[2]}(\bar{z}, \mu, \bar{x}), \tag{3.1}$$

$$\bar{u} = u + \alpha^{[2]}(\bar{z}, \mu, \bar{x}) + \beta^{[1]}(\bar{z}, \mu, \bar{x})u. \tag{3.2}$$

The normal form is given in the following theorem.

THEOREM 3.1. *Consider system* (2.5). *There exist a quadratic change of coordinates* (3.1) *and feedback* (3.2) *that transform the system into a unique system of the form*

$$\begin{aligned} \dot{z} &= A_1 z + f_1^{[2,0]}(z, \mu) + f_1^{[1,1]}(z, \mu, x_1) + f_1^{[0,2]}(x) + O(z, \mu, x)^3, \\ \dot{x} &= A_2 x + B_2 u + f_2^{[0,2]}(x) + O(z, \mu, x, u)^3, \end{aligned} \tag{3.3}$$

where

$$
\begin{aligned}
f_1^{[2,0]}(z,\mu) &= \sum_{i=1}^{2}\sum_{j=1}^{2} \beta_i^j e_1^i\, z_j\, z_3, \quad \beta_1^1 = \beta_2^2,\ \beta_1^2 = -\beta_2^1,\\
f_1^{[1,1]}(z,\mu,x_1) &= \sum_{i=1}^{2}\sum_{j=1}^{3} \gamma_i^j e_1^i\, z_j\, x_1,\\
f_1^{[0,2]}(x) &= \sum_{i=1}^{2}\sum_{j=1}^{n-2} \delta_i^j e_1^i\, x_j^2,\\
f_2^{[0,2]}(x) &= \sum_{i=1}^{n-2}\sum_{j=i+2}^{n-2} \rho_i^j e_2^i\, x_j^2.
\end{aligned}
\tag{3.4}
$$

For reasons of simplicity, we use (z,μ,x) instead of $(\bar{z},\mu,\bar{x})$ for the variables in the normal form. In the notation, $z_3 = \mu$. We will continue to use both z_3 and μ in the rest of this paper. The notation e_1^i is the ith unit vector in z space. The vector e_2^i is the ith unit vector in x space. The superscript $[2,0]$ implies that $f_i^{[2,0]}(z,\mu)$ is quadratic in (z,μ), and the variable x does not appear in $f_i^{[2,0]}(z,\mu)$. The superscript $[1,1]$ implies that $f_i^{[1,1]}(z,\mu,x)$ consists of quadratic terms in which both the degree of (z,μ) and the degree of x equal one, i.e., the cross terms of x and (z,μ). Similarly, $f_i^{[0,2]}(x)$ consists of quadratic terms of x. A typical quadratic normal form of dimension five is given in the following example:

$$
\begin{aligned}
\dot{z}_1 &= -\omega z_2 + z_1\mu + z_2\mu + z_1x_1 + z_2x_1 + \mu x_1 + x_1^2 + x_2^2 + x_3^2,\\
\dot{z}_2 &= \omega z_1 - z_1\mu + z_2\mu + z_1x_1 + z_2x_1 + \mu x_1 + x_1^2 + x_2^2 + x_3^2,\\
\dot{x}_1 &= x_2 + x_3^2,\\
\dot{x}_2 &= x_3,\\
\dot{x}_3 &= u.
\end{aligned}
$$

In this example,

$$
f_1^{[2,0]} = \begin{bmatrix} z_1\mu + z_2\mu \\ -z_1\mu + z_2\mu \end{bmatrix}, \quad f_1^{[1,1]} = \begin{bmatrix} z_1x_1 + z_2x_1 + \mu x_1 \\ z_1x_1 + z_2x_1 + \mu x_1 \end{bmatrix},
$$

$$
f_1^{[0,2]} = \begin{bmatrix} x_1^2 + x_2^2 + x_3^2 \\ x_1^2 + x_2^2 + x_3^2 \end{bmatrix}, \quad f_2^{[0,2]} = \begin{bmatrix} x_3^2 \\ 0 \\ 0 \end{bmatrix}.
$$

The normal form is used to develop an algorithm of feedback design to stabilize the Hopf bifurcation. The proof of the quadratic normal form is not used in the feedback design. So, the proof of Theorem 3.1 is given in the appendix.

Now let us determine the quadratic part of the center manifold. In the following, $ad_X(Y)$ represents the Lie bracket of two matrices X and Y, i.e.,

$$
ad_X(Y) = XY - YX.
$$

THEOREM 3.2. *Suppose* $\begin{bmatrix} a_1, \dots, a_{n-2} \end{bmatrix}$ *stabilizes the x-subsystem, and suppose $F_i = 0$ for $i = 1,2,3$. The linear part of the center manifold has been determined as in Theorem* 2.2. *Given any matrix Q_1 and a quadratic function $\Pi_i^{[2]}$, $i = 1,\dots,n-2$, defined by*

$$
\Pi_i^{[2]}(z,\mu) = \begin{bmatrix} z & \mu \end{bmatrix} (-1)^{i-1} ad^{i-1}_{\left[\begin{smallmatrix} A_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]}(Q_1) \begin{bmatrix} z \\ \mu \end{bmatrix},
\tag{3.5}
$$

there exists a unique quadratic feedback

$$u(z,x,\mu)=\sum_{i=1}^{n-2}a_ix_i+\begin{bmatrix} z & \mu \end{bmatrix}Q_{fb}\begin{bmatrix} z \\ \mu \end{bmatrix}+O(z,\mu,x)^3, \tag{3.6}$$

with

$$Q_{fb}=P\Big(-ad_{\left[\begin{smallmatrix} A_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]}\Big)(Q_1), \tag{3.7}$$

such that (3.5) *is the quadratic part of the center manifold.* $P(\lambda)$ *is the characteristic polynomial of* A_2+B_2K *for* $K=\begin{bmatrix} a_1 & \cdots & a_{n-2} \end{bmatrix}$.

Remark. Theorem 3.2 implies that for any given matrix Q_1, there always exists Q_{fb} given by (3.7) so that the feedback (3.6) yields a center manifold satisfying (3.5). In the next section it will be proved that the stability of the Hopf bifurcation is determined by Q_1 and the invariants. Theorem 3.2 guarantees that if a Q_1 stabilizes a Hopf bifurcation, then Q_1 is always achievable by a suitable quadratic feedback. Thus, the problem of finding the stabilizing feedback is converted to the problem of maneuvering the quadratic part of the center manifold.

Remark. The spectrum of the operator $ad_{\left[\begin{smallmatrix} A_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]}$ consists of $\{\pm i\omega,\pm 2i\omega,0\}$, all on the imaginary axis. The spectrum of $P(-ad_{\left[\begin{smallmatrix} A_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]})$ is

$$\{P(\pm i\omega),P(\pm 2i\omega),P(0)\}.$$

Since the roots of $P(\lambda)$ are all in the left half plane, the spectrum of $P(-ad_{\left[\begin{smallmatrix} A_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]})$ does not contain zero. So, $P(-ad_{\left[\begin{smallmatrix} A_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]})$ is an invertible linear operator.

Proof. Consider

$$\begin{aligned}
\dot z &= A_1z+f_1^{[2,0]}(z,\mu)+f_1^{[1,1]}(z,\mu,x)+f_1^{[0,2]}(x)+O(z,\mu,x)^3,\\
\dot x &= A_2x+B_2\left(\sum_{i=1}^{n-2}a_ix_i+\begin{bmatrix} z & \mu \end{bmatrix}Q_{fb}\begin{bmatrix} z \\ \mu \end{bmatrix}\right)+f_2^{[0,2]}(x)+O(z,\mu,x)^3.
\end{aligned}$$

Since $F_1=F_2=F_3=0$, the linear part of the center manifold vanishes (Theorem 2.2). The quadratic part of the center manifold has the form

$$\begin{aligned}
x_i &= \Pi_i(z,\mu),\\
\Pi_i^{[2]} &= \begin{bmatrix} z & \mu \end{bmatrix}Q_i\begin{bmatrix} z \\ \mu \end{bmatrix},
\end{aligned}$$

where Q_i, $i=1,\dots,n-2$, are real 3×3 symmetric matrices. From the center manifold equation

$$\frac{\partial\Pi}{\partial(z\ \mu)}\begin{bmatrix} \dot z \\ \dot\mu \end{bmatrix}=A_2\Pi+Bu+O(z,\mu,x)^2$$

we have

$$\begin{aligned}
&\begin{bmatrix} z & \mu \end{bmatrix}\left(Q_i\begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix}-\begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix}Q_i\right)\begin{bmatrix} z \\ \mu \end{bmatrix}\\
&\quad=\begin{bmatrix} z & \mu \end{bmatrix}Q_{i+1}\begin{bmatrix} z \\ \mu \end{bmatrix}\quad\text{for } 1\le i\le n-3,
\end{aligned}$$

$$\begin{bmatrix} z & \mu \end{bmatrix}\left(Q_{n-2}\begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix} Q_{n-2}\right)\begin{bmatrix} z \\ \mu \end{bmatrix} = \sum_{i=1}^{n-2} a_i \begin{bmatrix} z & \mu \end{bmatrix} Q_i \begin{bmatrix} z \\ \mu \end{bmatrix} + \begin{bmatrix} z & \mu \end{bmatrix} Q_{fb} \begin{bmatrix} z \\ \mu \end{bmatrix}.$$

In this equation, we used the fact that

$$\begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix}^T = -\begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix}.$$

This equation is equivalent to

$$\begin{aligned} Q_{i+1} &= -ad_{\left[\begin{smallmatrix} A_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]}(Q_i), \\ Q_{fb} &= -ad_{\left[\begin{smallmatrix} A_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]}(Q_{n-2}) - \sum_{i=1}^{n-2} a_i Q_i, \end{aligned}$$

hence,

$$\begin{aligned} Q_{i+1} &= (-1)^i ad^i_{\left[\begin{smallmatrix} A_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]}(Q_1), \\ Q_{fb} &= (-1)^{n-2} ad^{n-2}_{\left[\begin{smallmatrix} A_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]}(Q_1) - \sum_{i=1}^{n-2} a_i (-1)^{i-1} ad^{i-1}_{\left[\begin{smallmatrix} A_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]}(Q_1). \end{aligned} \tag{3.8}$$

So

$$\begin{aligned} Q_{i+1} &= (-1)^i ad^i_{\left[\begin{smallmatrix} A_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]}(Q_1), \\ Q_{fb} &= P\Big(-ad_{\left[\begin{smallmatrix} A_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]}\Big)(Q_1). \end{aligned}$$ □

In Theorem 3.2, we assume that the linear feedback is independent of (z, μ). However, it does not mean that we lose any generality. The next theorem shows that if a closed-loop system has nonzero F_i terms, it can be transformed into a system in which the linear part of the controllable system is not explicitly a function of (z, μ). Then, the formulae in Theorem 3.2 are applicable to the new system.

PROPOSITION 3.3. *Given a system*

$$\begin{aligned} \dot{z} &= A_1 z + O(z, x)^2, \\ \dot{x} &= A_2 x + B_2\left(F_1 z_1 + F_2 z_2 + F_3 z_3 + \sum_{i=1}^{n-2} a_i x_i\right) + O(z, x)^2, \end{aligned} \tag{3.9}$$

there exists a linear change of coordinates to transform the system into the form

$$\begin{aligned} \dot{\tilde{z}} &= A_1 \tilde{z} + O(\tilde{z}, \tilde{x})^2, \\ \dot{\tilde{x}} &= A_2 \tilde{x} + B_2\left(\sum_{i=1}^{n-2} a_i \tilde{x}_i\right) + O(\tilde{z}, \tilde{x})^2. \end{aligned} \tag{3.10}$$

Proof. Consider the following change of coordinates:

$$\begin{aligned}
\tilde{x}_1 &= x_1 + \begin{bmatrix} \varphi_1 & \varphi_2 \end{bmatrix} z + \varphi_3 \mu,\\
\tilde{x}_2 &= x_2 + \begin{bmatrix} \varphi_1 & \varphi_2 \end{bmatrix} A_1 z,\\
&\vdots\\
\tilde{x}_{n-2} &= x_{n-2} + \begin{bmatrix} \varphi_1 & \varphi_2 \end{bmatrix} A_1^{n-3} z.
\end{aligned}$$

In the new coordinates, the system has the form

$$\begin{aligned}
\dot{\tilde{x}}_1 &= \tilde{x}_2,\\
\dot{\tilde{x}}_2 &= \tilde{x}_3,\\
&\vdots\\
\dot{\tilde{x}}_{n-2} &= \sum_{i=1}^{n-2} a_i \tilde{x}_i + (F_3 - a_1\varphi_3)\mu\\
&\quad + \left\{ \begin{bmatrix} F_1 & F_2 \end{bmatrix} + \begin{bmatrix} \varphi_1 & \varphi_2 \end{bmatrix} A_1^{n-2} - \sum_{i=1}^{n-2} a_i \begin{bmatrix} \varphi_1 & \varphi_2 \end{bmatrix} A_1^{i-1} \right\} z\\
&= \sum_{i=1}^{n-2} a_i \tilde{x}_i + (F_3 - a_1\varphi_3)\mu + \left(\begin{bmatrix} F_1 & F_2 \end{bmatrix} + \begin{bmatrix} \phi_1 & \phi_2 \end{bmatrix} P(A_1) \right) z. \qquad (3.11)
\end{aligned}$$

Define

$$\varphi_3 = \frac{F_3}{a_1}, \qquad \begin{bmatrix} \varphi_1 & \varphi_2 \end{bmatrix} = - \begin{bmatrix} F_1 & F_2 \end{bmatrix} P(A_1)^{-1}.$$

Then we obtain a dynamics possessing the form (3.10). ∎

Remark. When $F_i \neq 0$, the quadratic part of the center manifold is given by

$$Q_{i+1} = -ad_{\left[\begin{smallmatrix} A_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]}(Q_i) + \sum_{j=i+2}^{n-2} \rho_i^j {\Pi_j^{[1]}}^T \Pi_j^{[1]} - \sum_{j=2}^{2} \Pi_{i,j}^{[1]} P_j, \qquad (3.12)$$

$$Q_{fb} = -ad_{\left[\begin{smallmatrix} A_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]}(Q_{n-2}) - \sum_{i=1}^{n-2} a_i Q_i + \sum_{j=i+2}^{n-2} \rho_i^j {\Pi_j^{[1]}}^T \Pi_j^{[1]}, \qquad (3.13)$$

with $P_i \in \mathbb{R}^{3\times 3}$ given by

$$P_i = \begin{bmatrix} 0 & 0 & \frac{\beta_i^1}{2} \\ 0 & 0 & \frac{\beta_i^2}{2} \\ \frac{\beta_i^1}{2} & \frac{\beta_i^2}{2} & 0 \end{bmatrix} + \begin{bmatrix} \gamma_i^1 \\ \gamma_i^2 \\ \gamma_i^3 \end{bmatrix} \Pi_1^{[1]} + \sum_{j=1}^{n-2} \delta_i^j {\Pi_j^{[1]}}^T \Pi_j^{[1]}.$$

3.2. Control of the Hopf bifurcation. The center manifold has dimension two. To determine its stability we use the Poincaré–Andronov–Hopf theorem. Let us recall the following result.

THEOREM 3.4 (see [11]). *Consider the system*

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} 0 & -\omega \\ \omega & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} \Psi(z_1, z_2, \mu) \\ \tilde{\Psi}(z_1, z_2, \mu) \end{bmatrix}. \qquad (3.14)$$

If

$$\Psi_{\mu z_1} + \tilde{\Psi}_{\mu z_2} \neq 0,$$
$$\breve{a} \neq 0,$$

where $\breve{a}$ *is a constant defined below, a curve of periodic solutions bifurcates from the origin into* $\mu < 0$ *if* $\breve{a}(\Psi_{\mu z_1} + \tilde{\Psi}_{\mu z_2}) > 0$ *or* $\mu > 0$ *if* $\breve{a}(\Psi_{\mu z_1} + \tilde{\Psi}_{\mu z_2}) < 0$.

The periodic solution is stable if $\breve{a} < 0$. *The periodic solution is unstable if* $\breve{a} > 0$. *The origin is stable for* $\mu > 0$ *(resp.,* $\mu < 0$*) and unstable for* $\mu < 0$ *(resp.,* $\mu > 0$*) if* $\Psi_{\mu z_1} + \tilde{\Psi}_{\mu z_2} < 0$ *(resp.,* $\Psi_{\mu z_1} + \tilde{\Psi}_{\mu z_2} > 0$*).*

The coefficient $\breve{a}$ is a constant involving partial derivatives evaluated at the bifurcation point, i.e., $(z_1, z_2, \mu) = (0, 0, 0)$. It is given by

$$\begin{aligned}\breve{a} = {} & \frac{1}{16}(\Psi_{z_1 z_1 z_1} + \tilde{\Psi}_{z_1 z_1 z_2} + \Psi_{z_1 z_2 z_2} + \tilde{\Psi}_{z_2 z_2 z_2}) \\ & + \frac{1}{16\omega}(\Psi_{z_1 z_2}(\Psi_{z_1 z_1} + \Psi_{z_2 z_2}) - \tilde{\Psi}_{z_1 z_2}(\tilde{\Psi}_{z_1 z_1} + \tilde{\Psi}_{z_2 z_2}) \\ & \quad - \Psi_{z_1 z_1}\tilde{\Psi}_{z_1 z_1} + \Psi_{z_2 z_2}\tilde{\Psi}_{z_2 z_2}).\end{aligned} \tag{3.15}$$

Remark. When $\Psi_{\mu z_1} + \tilde{\Psi}_{\mu z_2} > 0$ (resp., $\Psi_{\mu z_1} + \tilde{\Psi}_{\mu z_2} < 0$), if $\breve{a} < 0$, then a stable periodic orbit of amplitude approximately

$$R = \left(\frac{(\Psi_{\mu z_1} + \tilde{\Psi}_{\mu z_2})\mu}{|\breve{a}|}\right)^{\frac{1}{2}} \tag{3.16}$$

bifurcates from the origin into $\mu > 0$ (resp., $\mu < 0$) as μ passes through zero. The origin itself is a stable focus if $\mu < 0$ (resp., $\mu > 0$).

If $\breve{a} > 0$, then an unstable periodic orbit of amplitude

$$R = \left(\frac{(\Psi_{\mu z_1} + \tilde{\Psi}_{\mu z_2})\mu}{\breve{a}}\right)^{\frac{1}{2}} \tag{3.17}$$

bifurcates into $\mu < 0$ (resp., $\mu > 0$), where the origin is a stable focus, and there are no periodic orbits in a small neighborhood of the origin of $\mu > 0$ (resp., $\mu < 0$).

Remark. If $\breve{a} < 0$, a stable periodic solution attracts local trajectories when the origin is an unstable equilibrium point. Hence, the local trajectories of the system stay around the origin even if the origin is not a stable point. In this case, the bifurcation is called *supercritical Hopf bifurcation.*

Using the expressions of the center manifold (3.5) and the normal form (3.3), we can determine the dynamics of (2.1) on the center manifold. A straightforward application of Theorem 3.4 permits us to compute the feedback coefficients to stabilize the Hopf bifurcation. Substituting (3.5) into the $\dot{z}$ dynamics of the normal form (3.3), the critical coefficients in the reduced system can be found. In this case,

$$\begin{aligned}\breve{a} = {} & \frac{1}{16}\Bigg(2(Q_{11}(3\gamma_1^1 + \gamma_2^2) + 2Q_{12}(\gamma_1^2 + \gamma_2^1) + Q_{22}(3\gamma_2^2 + \gamma_1^1)) \\ & + \frac{\partial^3 f_{11}^{[3]}}{\partial z_1^3} + \frac{\partial^3 f_{12}^{[3]}}{\partial z_1^2 \partial z_2} + \frac{\partial^3 f_{11}^{[3]}}{\partial z_1 \partial z_2^2} + \frac{\partial^3 f_{12}^{[3]}}{\partial z_2^3}\Bigg), \\ \Psi_{\mu z_1} + \tilde{\Psi}_{\mu z_2} = {} & 2\,\beta_1^1,\end{aligned} \tag{3.18}$$

where

$$Q_1 = \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{12} & Q_{22} & Q_{23} \\ Q_{13} & Q_{23} & Q_{33} \end{bmatrix}$$

is the matrix of the quadratic center manifold, and

$$f_1^{[3]} = \begin{bmatrix} f_{11}^{[3]} \\ f_{12}^{[3]} \end{bmatrix}$$

represents the cubic part of the $\dot{z}$ equation in (3.3). The next theorem of bifurcation control is a straightforward corollary of Theorem 3.4 and (3.18).

THEOREM 3.5. *Given a system in the normal form* (3.3)*, suppose* $\beta_1^1 \neq 0$*. If one of the following conditions is satisfied,*

(1) $3\gamma_1^1 + \gamma_2^2 \neq 0$,
(2) $\gamma_1^2 + \gamma_2^1 \neq 0$,
(3) $3\gamma_2^2 + \gamma_1^1 \neq 0$,

then there always exists a nonlinear feedback (3.6) *that renders the Hopf bifurcation supercritical. The feedback coefficient* Q_{fb} *is determined by* (3.7)*, in which* Q_1 *is any symmetric matrix satisfying* $\breve{a} < 0$.

Remark. In some cases where $\beta_1^1 = 0$, it may be possible to use the feedback (F_1, F_2, F_3) in (2.7) to modify the value of $\Psi_{\mu z_1} + \tilde{\Psi}_{\mu z_2}$. Indeed, using this feedback we find that

$$\Psi_{\mu z_1} + \tilde{\Psi}_{\mu z_2} = 2\beta_1^1 + (\gamma_1^1 + \gamma_2^2)\Pi_{13}^{[1]} + \sum_{j=1}^{n-2} \left(\delta_1^j \Pi_{j1}^{[1]} + \delta_2^j \Pi_{j2}^{[1]}\right) \Pi_{j3}^{[1]},$$

where $\Pi_{ji}^{[1]}$ for $i = 1, 2, 3$ and $j = 1, \ldots, n-2$ are given by (2.8) and (2.11).

4. Invariants. Given a system in the form of (2.5), how do we find its normal form? In this section, a set of numbers associated with (2.5) is found. The numbers are invariant under any quadratic change of coordinates and feedback. They are called the quadratic invariants. It is also proved that these invariants equal the coefficients in the normal form. Two systems are equivalent under a quadratic change of coordinates and feedback if and only if the invariants of the systems are equal. Therefore, the set of quadratic invariants completely characterizes the quadratic part of a nonlinear control system. For a given system (2.5), the values of the invariants are the coefficients in the normal form.

Consider a system in the form of (2.5). Denote by C_x, C_z the following row vector and matrix:

$$C_x = \begin{bmatrix} 0 & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}_{1 \times n}, \quad C_z = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \end{bmatrix}_{2 \times n}. \tag{4.1}$$

Given two vector fields $X(\xi)$ and $Y(\xi)$ defined in $\mathbb{R}^n$, the operator ad_X is defined by

$$ad_X(Y) = [X, Y] = \frac{\partial Y}{\partial \xi} X - \frac{\partial X}{\partial \xi} Y.$$

The Lie operator L_X is defined by

$$L_X(\kappa(\xi)) = \frac{\partial \kappa(\xi)}{\partial \xi} X$$

for C^1 functions defined in $\mathbb{R}^n$.

DEFINITION 4.1. *Given a system* (2.5)*, the quadratic invariants are defined by*

$$\rho_t^{n-r-1} = \frac{1}{2} C_x A^{t-1} \left[ad_f^r(g), ad_f^{r-1}(g) \right] \Big|_{z=0,x=0,\mu=0}, \qquad \begin{array}{l} 1 \le r \le n-4, \\ 1 \le t \le n-r-3, \end{array}$$

$$\begin{bmatrix} \delta_1^{n-r-1} \\ \delta_2^{n-r-1} \end{bmatrix} = \frac{1}{2} C_z \left[ad_f^r(g), ad_f^{r-1}(g) \right] \Big|_{z=0,x=0,\mu=0}, \quad 1 \le r \le n-2,$$

$$\begin{bmatrix} \gamma_1^j \\ \gamma_2^j \end{bmatrix} = (-1)^n C_z \frac{\partial}{\partial z_j} ad_f^{n-2}(g) \big|_{x=0,\ z=0,\ \mu=0}, \quad j = 1, 2, 3,$$

$$\begin{bmatrix} \beta_1^1 \\ \beta_1^2 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \dfrac{\partial^2 f_{11}}{\partial z_3 \partial z_1} + \dfrac{\partial^2 f_{12}}{\partial z_3 \partial z_2} \\ \dfrac{\partial^2 f_{11}}{\partial z_3 \partial z_2} - \dfrac{\partial^2 f_{12}}{\partial z_3 \partial z_1} \end{bmatrix} \Bigg|_{x=0,\ z=0,\ \mu=0},$$

(4.2)

where f and g are the right hand side of (2.5)*. The vector*

$$f_1 = \begin{bmatrix} f_{11} \\ f_{12} \end{bmatrix}$$

represents the vector field of $\dot{z}$ in (2.5).

Notation such as ρ_i^j, δ_i^j, γ_i^j, and β_i^j are used for both the invariants and the coefficients in the normal form (3.3). In the following, it is proved that they are actually equal to each other.

THEOREM 4.2. *Consider a system in the form of* (2.5)*.*

(i) *The quadratic transformation* (3.1)–(3.2) *does not change the value of quadratic invariants.*

(ii) *For a system in normal form* (3.3)*, its quadratic invariants* (4.2) *are equal to the coefficients of the quadratic terms in the normal form.*

(iii) *Given two systems in the form of* (2.5) *with the same linearization (same ω), the quadratic part of one system can be transformed into that of another system by a suitable transformation* (3.1)–(3.2) *if and only if they have the same quadratic invariants.*

Proof. (i) Suppose the system (2.5) is transformed into the following system by a quadratic change of coordinates (3.1)–(3.2):

$$\begin{bmatrix} \dot{\bar{z}}_1 \\ \dot{\bar{z}}_2 \end{bmatrix} = A_1 \begin{bmatrix} \bar{z}_1 \\ \bar{z}_2 \end{bmatrix} + \begin{bmatrix} \bar{f}_{11}^{[2]}(\bar{z}, \mu, \bar{x}) \\ \bar{f}_{12}^{[2]}(\bar{z}, \mu, \bar{x}) \end{bmatrix} + \begin{bmatrix} \bar{g}_{11}^{[1]}(\bar{z}, \mu, \bar{x}) \\ \bar{g}_{12}^{[1]}(\bar{z}, \mu, \bar{x}) \end{bmatrix} \bar{u} + O(\bar{z}, \mu, \bar{x}, \bar{u})^3,$$

$$\dot{\bar{x}} = A_2 \bar{x} + B_2 \bar{u} + \bar{f}_2^{[2]}(\bar{z}, \mu, \bar{x}) + \bar{g}_2^{[1]}(\bar{z}, \mu, \bar{x})\, u + O(\bar{z}, \mu, \bar{x}, \bar{u})^3. \tag{4.3}$$

Denote the invariants of (2.5) and (4.3) by ρ_i^j, δ_i^j, γ_i^j, β_1^j, and $\bar{\rho}_i^j$, $\bar{\delta}_i^j$, $\bar{\gamma}_i^j$, $\bar{\beta}_1^j$, respectively. Notice that if we treat $f(z, \mu, x)$ and $g(z, \mu, x)$ as vector fields in $\mathbb{R}^n$, then f and $\bar{f}$ represent the same vector field. Similarly, g and $\bar{g}$ represent the same vector field. Since Lie bracket and Lie operators are independent of the choice of coordinate systems, sometimes we use f and g to represent these two vector fields without mentioning the coordinate system (z, x or $\bar{z}$, $\bar{x}$). On a manifold, the operators $\frac{\partial}{\partial z_j}$ and $\frac{\partial}{\partial \bar{z}_j}$ are equivalent to vector fields. They depend on the selection of coordinate

systems. The invariants can be expressed in the following way using Lie brackets and Lie operators:

$$\rho_t^{n-r-1} = \tfrac{1}{2} L_{[ad_f^r(g), ad_f^{r-1}(g)]} L_f^{t-1}(x_1)|_{z=0,x=0,\mu=0}, \qquad \begin{matrix} 1 \le r \le n-4, \\ 1 \le t \le n-r-3, \end{matrix}$$

$$\begin{bmatrix} \delta_1^{n-r-1} \\ \delta_2^{n-r-1} \end{bmatrix} = \tfrac{1}{2} L_{[ad_f^r(g), ad_f^{r-1}(g)]}(z)|_{z=0,x=0,\mu=0}, \quad 1 \le r \le n-2, \tag{4.4}$$

$$\begin{bmatrix} \gamma_1^j \\ \gamma_2^j \end{bmatrix} = (-1)^n L_{[\frac{\partial}{\partial z_j}, \mathrm{ad}_f^{n-2}(g)]}(z)\big|_{x=0,\ z=0,\ \mu=0}, \quad j = 1,2,3.$$

Under the new coordinates, we have

$$x_1 = \bar{x}_1 + O(\bar{z},\mu,\bar{x})^2, \quad z = \bar{z} + O(\bar{z},\mu,\bar{x})^2, \quad \frac{\partial}{\partial z_j} = \frac{\partial}{\partial \bar{z}_j} + O(\bar{z},\mu,\bar{x})^2. \tag{4.5}$$

From (4.4) and (4.5),

$$\begin{aligned} \rho_t^{n-r-1} = {} & \frac{1}{2} L_{[ad_f^r(g), ad_f^{r-1}(g)]} L_f^{t-1}(\bar{x}_1)|_{z=0,x=0,\mu=0} \\ & + \frac{1}{2} L_{[ad_f^r(g), ad_f^{r-1}(g)]} L_f^{t-1}(O(\bar{z},\mu,\bar{x})^2)|_{z=0,x=0,\mu=0}. \end{aligned}$$

In this relation, the second term on the right side is zero. The first term on the right side is $\bar{\rho}_t^{n-r-1}$. This proves $\rho_t^{n-r-1} = \bar{\rho}_t^{n-r-1}$. Similarly, we can prove that $\delta_i^j = \bar{\delta}_i^j$ for all feasible i and j.

Now let us consider γ_i^j. By (4.4) and (4.5), we have

$$\begin{aligned} \begin{bmatrix} \gamma_1^j \\ \gamma_2^j \end{bmatrix} = (-1)^n \Bigg(& L_{[\frac{\partial}{\partial \bar{z}_j}, \mathrm{ad}_f^{n-2}(g)]}(\bar{z})\Big|_{\bar{x}=0,\ \bar{z}=0,\ \mu=0} + L_{[\frac{\partial}{\partial \bar{z}_j}, \mathrm{ad}_f^{n-2}(g)]}(O(\bar{z},\mu,\bar{x})^2)\Big|_{\bar{x}=0,\ \bar{z}=0,\ \mu=0} \\ & + L_{[O(\bar{z},\mu,\bar{x}), \mathrm{ad}_f^{n-2}(g)]}(\bar{z} + O(\bar{z},\mu,\bar{x})^2)\Big|_{\bar{x}=0,\ \bar{z}=0,\ \mu=0} \Bigg). \end{aligned} \tag{4.6}$$

By (4.4), we know that

$$L_{[\frac{\partial}{\partial \bar{z}_j}, \mathrm{ad}_f^{n-2}(g)]}(\bar{z})\Big|_{\bar{x}=0,\bar{z}=0,\mu=0} = \begin{bmatrix} \bar{\gamma}_1^j \\ \bar{\gamma}_2^j \end{bmatrix}. \tag{4.7}$$

It is easy to check that

$$L_{[\frac{\partial}{\partial \bar{z}_j}, \mathrm{ad}_f^{n-2}(g)]}(O(\bar{z},\mu,\bar{x})^2)\Big|_{\bar{x}=0,\bar{z}=0,\mu=0} = 0 \tag{4.8}$$

and

$$ad_f^r(g) = (-1)^r \begin{bmatrix} 0 \\ A_2^r B_2 \end{bmatrix} + O(\bar{z},\mu,\bar{x}).$$

Therefore,

$$ad_f^{n-2}(g) = O(\bar{z},\mu,\bar{x}).$$

So

$$L_{[O(\bar{z},\mu,\bar{x}),\mathrm{ad}_f^{n-2}(g)]}(\bar{z}+O(\bar{z},\mu,\bar{x})^2)\Big|_{\bar{x}=0,\bar{z}=0,\mu=0}=0. \tag{4.9}$$

Equations (4.7), (4.8), and (4.9) imply $\gamma_i^j=\bar{\gamma}_i^j$.

Now, let us consider β_1^1 and β_1^2. The homological equation (A.2) in the appendix for $f_1^{[2,0]}(\bar{z},\mu)$ is

$$f_1^{[2,0]}(z,\mu)-\bar{f}_1^{[2,0]}(z,\mu)=\frac{\partial\phi_1^{[2,0]}}{\partial z}A_1z-A_1\phi_1^{[2,0]},$$

where the change of coordinates $\phi_1^{[2,0]}$ is

$$\phi_1^{[2,0]}=\begin{bmatrix} a_1^1z_1\mu+a_1^2z_2\mu \\ a_2^1z_1\mu+a_2^2z_2\mu \end{bmatrix}.$$

Straightforward computation shows that

$$f_1^{[2,0]}(z,\mu)-\bar{f}_1^{[2,0]}(z,\mu)=\begin{bmatrix} \omega(a_2^1+a_1^2)z_1\mu+\omega(a_2^2-a_1^1)z_2\mu \\ \omega(a_2^2-a_1^1)z_1\mu-\omega(a_1^2+a_2^1)z_2\mu \end{bmatrix}.$$

From the definition (4.2),

$$\begin{bmatrix} \beta_1^1-\bar{\beta}_1^1 \\ \beta_1^2-\bar{\beta}_1^2 \end{bmatrix}=\begin{bmatrix} \dfrac{\partial^2 f_{11}-\bar{f}_{11}}{\partial z_3\partial z_1}+\dfrac{\partial^2 f_{12}-\bar{f}_{12}}{\partial z_3\partial z_2} \\ \dfrac{\partial^2 \bar{f}_{11}-\bar{f}_{11}}{\partial z_3\partial z_2}-\dfrac{\partial^2 f_{12}-\bar{f}_{12}}{\partial z_3\partial z_1} \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

This proves that β_i^j is invariant.

(ii) The proof of the second part of Theorem 4.2 is based on calculation. Suppose that the vectors f and g are in the normal form (3.3). By mathematical induction we can prove that for $1\le r\le n-4$, we have

$$\mathrm{ad}_f^r(g)=(-1)^r\left(\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}\Bigg\}\begin{matrix} n-r \\ \end{matrix}+\begin{bmatrix} 2\delta_1^{n-r-1}x_{n-r-1} \\ 2\delta_2^{n-r-1}x_{n-r-1} \\ 2\rho_1^{n-r-1}x_{n-r-1} \\ \vdots \\ 2\rho_{n-r-3}^{n-r-1}x_{n-r-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}\Bigg\}n-r-1\right)$$
$$+\,h_r(x_{n-r},\dots,x_{n-2},z,\mu)+O(x,z,\mu)^2.$$

Similarly, we can derive that

$$\mathrm{ad}_f^{n-3}(g)=(-1)^{n-3}\begin{bmatrix} 2\delta_1^2x_2 \\ 2\delta_2^2x_2 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}+h_{n-3}(x_3,\dots,x_{n-2},\mu,z)+O(z,\mu,x)^2,$$

$$\mathrm{ad}_f^{n-2}(g) = (-1)^{n-2} \begin{bmatrix} \gamma_1^1 z_1 + \gamma_1^2 z_2 + \gamma_1^3 \mu + 2\delta_1^1 x_1 \\ \gamma_2^1 z_1 + \gamma_2^2 z_2 + \gamma_2^3 \mu + 2\delta_2^1 x_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + h_{n-2}(x_2, \ldots, x_{n-2}, \mu, z) + O(z, \mu, x)^2.$$

Therefore,

$$\begin{aligned} &\left[ad_f^r(g), ad_f^{r-1}(g)\right] \\ &\quad = \begin{bmatrix} 2\delta_1^{n-r-1} & 2\delta_2^{n-r-1} & 2\rho_1^{n-r-1} & \cdots & 2\rho_{n-r-3}^{n-r-1} & 0 & \cdots & 0 \end{bmatrix}^T \\ &\qquad + O(z, \mu, x), \quad r < n-3, \left[ad_f^r(g), ad_f^{r-1}(g)\right] \\ &\quad = \begin{bmatrix} 2\delta_1^{n-r-1} & 2\delta_2^{n-r-1} & 0 & \cdots & 0 \end{bmatrix}^T + O(z, \mu, x), \quad r = n-3, n-2. \end{aligned}$$

This implies

$$\begin{aligned} \rho_t^{n-r-1} &= \frac{1}{2} C_x A^{t-1} \left[ad_f^r(g), ad_f^{r-1}(g)\right] \Big|_{z=0, x=0, \mu=0}, \quad 1 \le r \le n-4, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad 1 \le t \le n-r-3, \\ \begin{bmatrix} \delta_1^{n-r-1} \\ \delta_2^{n-r-1} \end{bmatrix} &= \frac{1}{2} C_z \left[ad_f^r(g), ad_f^{r-1}(g)\right] \Big|_{z=0, x=0, \mu=0}, \quad 1 \le r \le n-2, \\ \begin{bmatrix} \gamma_1^j \\ \gamma_2^j \end{bmatrix} &= (-1)^n C_z \left[\frac{\partial}{\partial z_j}, \mathrm{ad}_f^{n-2}(g)\right] \Big|_{x=0,\ z=0,\ \mu=0}, \quad j = 1, 2, 3. \end{aligned}$$

The formula for the invariants β_1^1 and β_1^2 of the normal form is a straightforward corollary of the definition of invariants.

(iii) By (i) and (ii), the value of the invariants of a system uniquely determines the coefficients of its normal form. On the other hand, Theorem 3.1 implies that one system can be transformed into another system by quadratic transformation if and only if they have the same normal form. Therefore, a system can be transformed into another system by quadratic transformation if and only if they have the same values of invariants. □

5. Conclusion. In this paper, linear and quadratic normal forms of nonlinear systems with a pair of imaginary uncontrollable modes are derived. Based on the normal form, formulae of feedbacks are found to control the bifurcation of the system. The Hopf bifurcation cannot be removed from the closed-loop system because the imaginary eigenvalues are uncontrollable. However, it is proved that both the orientation and the stability of the periodic solution can be controlled by state feedback. It is proved in this paper that a linear feedback determines the local orientation of the periodic solution around the bifurcation point, and the quadratic feedback controls the stability of the periodic solution. The explicit relations between the feedback and the performance of the periodic solution, such as the local orientation and stability, are derived. A set of formulae is derived which computes the coefficients in the normal form of systems with two imaginary uncontrollable modes.

Appendix. Proof of normal form theorem. Given a system (2.5), the quadratic functions $f_1^{[2]}(z,\mu,x)$ and $f_2^{[2]}(z,\mu,x)$ have a decomposition denoted by

$$f_i^{[2]}(z,\mu,x) = f_i^{[2,0]}(z,\mu) + f_i^{[1,1]}(z,\mu,x) + f_i^{[0,2]}(x), \qquad i=1,2.$$

The superscript $[2,0]$ implies that $f_i^{[2,0]}(z,\mu)$ is quadratic in (z,μ), and the variable x does not appear in $f_i^{[2,0]}(z,\mu)$. The superscript $[1,1]$ implies that $f_i^{[1,1]}(z,\mu,x)$ consists of quadratic terms in which both the degree of (z,μ) and the degree of x equal one, i.e., the cross terms of x and (z,μ). Similarly, $f_i^{[0,2]}(x)$ consists of quadratic terms of x. For example, z_1x_1 and μx_2 are terms of $f_i^{[1,1]}$, and x_2x_4 is a term in $f_i^{[0,2]}(x)$. The subscript $i=1$ implies that the vector has 2 components, representing the uncontrollable dynamics, i.e., the right side of $\dot z$ equation in (2.5). The subscript $i=2$ implies that the vector has $n-2$ components. It represents the vector in the $\dot x$ equation of (2.5). Similar notation applies to other vector fields and functions, such as the change of coordinates $\phi^{[2]}(z,\mu,x)$ and the feedback $\alpha^{[2]}(z,\mu,x)$ and $\beta^{[1]}(z,\mu,x)$.

Let us consider the system

$$\begin{bmatrix} \dot z \\ \dot x \end{bmatrix} = A \begin{bmatrix} z \\ x \end{bmatrix} + Bu + f^{[2]}(z,\mu,x) + g^{[1]}(z,\mu,x)u.$$

In [14], it was proved that this system can be transformed into

$$\begin{bmatrix} \dot{\bar z} \\ \dot{\bar x} \end{bmatrix} = A \begin{bmatrix} \bar z \\ \bar x \end{bmatrix} + Bv + \bar f^{[2]}(\bar z,\mu,\bar x) + \bar g^{[1]}(\bar z,\mu,\bar x)v$$

by the change of coordinates and feedback

$$\begin{aligned} \begin{bmatrix} z \\ x \end{bmatrix} &= \begin{bmatrix} \bar z \\ \bar x \end{bmatrix} + \phi^{[2]}(\bar z,\mu,\bar x) + O(\bar z,\mu,\bar x)^2, \\ v &= u + \alpha^{[2]}(\bar z,\mu,\bar x) + \beta^{[1]}(\bar z,\mu,\bar x)u \end{aligned} \tag{A.1}$$

if and only if the following equation holds,

$$\begin{aligned} &\left[A\begin{bmatrix} z \\ x \end{bmatrix}, \phi^{[2]}(z,\mu,x)\right] + B\alpha^{[2]}(z,\mu,x) = f^{[2]}(z,\mu,x) - \bar f^{[2]}(z,\mu,x), \\ &\left[B, \phi^{[2]}(z,\mu,x)\right] + B\beta^{[1]}(z,\mu,x) = g^{[1]}(z,\mu,x) - \bar g^{[1]}(z,\mu,x), \end{aligned} \tag{A.2}$$

where the Lie bracket between two vectors $X_1(\xi)$ and $X_2(\xi)$ is defined by $[X_1,X_2] = \frac{\partial X_1}{\partial \xi}X_2 - \frac{\partial X_2}{\partial \xi}X_1$. Following the terminology of Poincaré's normal form theory, (A.2) is called the homological equation. For a system in the form of (2.5), the homological equation (A.2) is equivalent to equation

$$\begin{aligned} &\frac{\partial \phi_i^{[j,k]}}{\partial z}A_1 z + \frac{\partial \phi_i^{[j,k]}}{\partial x}A_2 x - A_i\phi_i^{[j,k]} + B_i\alpha^{[j,k]} = f_i^{[j,k]} - \bar f_i^{[j,k]}, \\ &\frac{\partial \phi_i^{[2]}}{\partial x_{n-2}} + B_i\beta^{[1]} = g_i^{[1]} - \bar g_i^{[1]}, \qquad i=1,2,3,\ \ j=0,1,2,\ \ j+k=2. \end{aligned} \tag{A.3}$$

In this equation,

$$\phi(z,\mu,x) = \begin{bmatrix} \phi_1(z,\mu,x) \\ \phi_2(z,\mu,x) \end{bmatrix} = \begin{bmatrix} \phi_1^{[2,0]}(z,\mu) + \phi_1^{[1,1]}(z,\mu,x) + \phi_1^{[0,2]}(x) \\ \phi_2^{[2,0]}(z,\mu) + \phi_2^{[1,1]}(z,\mu,x) + \phi_2^{[0,2]}(x) \end{bmatrix}.$$

Notice that B_1 is actually a zero vector. The following remark is a corollary of the homological equation (A.3). In [15], the remark is called the *separation principle*.

Remark. The homological equation (A.3) implies that $f_i^{[j,k]} - \bar{f}_i^{[j,k]}$ is determined by the corresponding $\phi_i^{[j,k]}$. For example, the transformation $\phi_1^{[2,0]}$ changes the term $f_1^{[2,0]}$ but leaves the other $f_i^{[j,k]}$ invariant. The homological equation also implies that $(\phi_2, \alpha^{[2]}, \beta^{[1]})$ does not change anything in f_1 and g_1, and ϕ_1 does not change anything in f_2 and g_2.

To prove Theorem 3.1, we will prove that for any $f^{[2]}$ and $g^{[1]}$ in (2.5), there always exists $(\phi(z,\mu,x), \alpha^{[2]}(z,\mu,x), \beta^{[1]}(z,\mu,x))$ so that the homological equation (A.3) is satisfied for a unique pair of vectors $\bar{f}^{[2]}$ and $\bar{g}^{[1]}$ in the normal form (3.3)–(3.4). To study the solvability of (A.3), we consider the set of all vectors $f^{[2]}(z,\mu,x)$ and $g^{[1]}(z,\mu,x)$ as linear spaces, denoted by $W(f^{[2]})$ and $W(g^{[1]})$, respectively. Similarly, the linear spaces consisting of $\phi^{[2]}$, $\alpha^{[2]}$, and $\beta^{[1]}$ are denoted by $V(\phi^{[2]})$, $V(\alpha^{[2]})$, and $V(\beta^{[1]})$. These linear spaces have subspace decomposition corresponding to the decomposition of the vector fields. For example, $W(f_1^{[2,0]})$ is a subspace of $W(f^{[2]})$ consisting of all the vectors $f_1^{[2,0]}(z,\mu)$. The linear spaces of the vectors in normal form are denoted by $W(\bar{f}^{[2]})$ and $W(\bar{g}^{[1]})$, where $\bar{f}^{[2]}$ and $\bar{g}^{[1]}$ satisfy the formulae in (3.3) and (3.4).

The left side of (A.3) can be considered as a linear mapping from $V(\phi^{[2]}) \times V(\alpha^{[2]}) \times V(\beta^{[1]})$ to $W(f^{[2]}) \times W(g^{[2]})$, denoted by Π. So,

$$\Pi\left(\begin{bmatrix}\phi_1^{[2]}\\ \phi_2^{[2]}\end{bmatrix}, \alpha^{[2]}, \beta^{[1]}\right) = \left(\begin{bmatrix} \dfrac{\partial \phi_1^{[2]}}{\partial z} A_1 z + \dfrac{\partial \phi_1^{[2]}}{\partial x} A_2 x - A_1 \phi_1^{[2]} \\ \dfrac{\partial \phi_2^{[2]}}{\partial z} A_1 z + \dfrac{\partial \phi_2^{[2]}}{\partial x} A_2 x - A_2 \phi_2^{[2]} + B_2 \alpha^{[2]} \end{bmatrix} \begin{bmatrix} \dfrac{\partial \phi_1^{[2]}}{\partial x_{n-2}} \\ \dfrac{\partial \phi_2^{[2]}}{\partial x_{n-2}} + B_2 \beta^{[1]} \end{bmatrix}\right). \tag{A.4}$$

Given $(f^{[2]}, g^{[1]})$ from (2.5), there exist a transformation (A.1) and a unique normal form $(\bar{f}^{[2]}, \bar{g}^{[1]})$ satisfying the homological equation (A.3) if and only if

$$\begin{aligned} &\mathrm{Im}(\Pi) \cap \left(W(\bar{f}^{[2]}) \times W(\bar{g}^{[1]})\right) = \{0\}, \\ &\mathrm{Im}(\Pi) + \left(W(\bar{f}^{[2]}) \times W(\bar{g}^{[1]})\right) = W(f^{[2]}) \times W(g^{[1]}). \end{aligned}$$

It is equivalent to the conditions

$$\begin{aligned} &\mathrm{Im}(\Pi) \cap \left(W(\bar{f}^{[2]}) \times W(\bar{g}^{[1]})\right) = \{0\}, \\ &\dim(\mathrm{Im}(\Pi)) + \dim\left(W(\bar{f}^{[2]}) \times W(\bar{g}^{[1]})\right) = \dim\left(W(f^{[2]}) \times W(g^{[1]})\right). \end{aligned} \tag{A.5}$$

From (A.4) and the separation principle,

$$\begin{aligned} &\Pi\left(\begin{bmatrix}\phi_1^{[2]}\\ 0\end{bmatrix}, 0, 0\right) \in W(f_1^{[2]}) \times W(g_1^{[1]}), \\ &\Pi\left(\begin{bmatrix}0\\ \phi_2^{[2]}\end{bmatrix}, \alpha^{[2]}, \beta^{[1]}\right) \in W(f_2^{[2]}) \times W(g_2^{[1]}). \end{aligned}$$

Therefore, the condition (A.5) can be proved separately for the controllable part involving $V(\phi_2^{[2]})$, $\alpha^{[2]}$, $\beta^{[1]}$, and $W(f_2^{[2]}) \times W(g_2^{[1]})$, and the uncontrollable part involving

$V(\phi_1^{[2]})$ and $W(f_1^{[2]}) \times W(g_1^{[1]})$. The proof for the controllable part can be found in [15]. In the following, we focus on the proof of (A.5) for the restriction of Π on $V(\phi_1^{[2]})$. In this case, $W(\bar{g}_1^{[1]}) = \{0\}$ and $W(\bar{f}_1^{[0,2]}) \times W(\bar{g}_1^{[1]}) = W(\bar{f}_1^{[0,2]})$. Equation (A.5) is reduced to

$$\begin{aligned} &W(\bar{f}_1^{[2]}) \cap \Pi\left(V(\phi_1^{[2]})\right) = \{0\}, \\ &\dim\left(\Pi(V(\phi_1^{[2]}))\right) + \dim\left(W(\bar{f}_1^{[2]})\right) = \dim\left(W(f_1^{[2]})\right). \end{aligned} \tag{A.6}$$

Once again, by the separation principle, (A.6) can be proved separately for $\phi_1^{[2,0]}$, $\phi_1^{[1,1]}$, and $\phi_1^{[0,2]}$.

Given any $\phi_1^{[2]} \neq 0$ in the subspace $V(\phi_1^{[1,1]}) + V(\phi_1^{[0,2]})$, let j be the largest integer so that

$$\frac{\partial \phi_1^{[2]}}{\partial x_j} \neq 0. \tag{A.7}$$

Now, we prove that $\ker(\Pi)$ in $V(\phi_1^{[1,1]}) + V(\phi_1^{[0,2]})$ is $\{0\}$. Suppose this is not true, and suppose $\Pi(\phi_1^{[2]}) = 0$. Then (A.4) implies

$$\begin{aligned} &\frac{\partial \phi_1^{[2]}}{\partial z} A_1 z + \frac{\partial \phi_1^{[2]}}{\partial x} A_2 x - A_1 \phi_1^{[2]} = 0, \\ &\frac{\partial \phi_1^{[2]}}{\partial x_{n-2}} = 0. \end{aligned} \tag{A.8}$$

The second equation of (A.8) and the condition (A.7) imply that $1 \leq j < n-2$. Because

$$A_2 x = \begin{bmatrix} x_2 & x_3 \cdots x_{n-2} & 0 \end{bmatrix}^T, \tag{A.9}$$

the first equation of (A.8) implies that

$$\frac{\partial}{\partial x_{j+1}} \left(\frac{\partial \phi_1^{[2]}}{\partial x} A_2 x \right) \neq 0. \tag{A.10}$$

Therefore,

$$\frac{\partial \phi_1^{[2]}}{\partial z} A_1 z + \frac{\partial \phi_1^{[2]}}{\partial x} A_2 x - A_1 \phi_1^{[2]} \neq 0.$$

This contradicts (A.8). Therefore, $\ker(\Pi) = \{0\}$ in $V(\phi_1^{[1,1]}) + V(\phi_1^{[0,2]})$. So,

$$\dim\left(\Pi(V(\phi_1^{[1,1]}) + V(\phi_1^{[0,2]}))\right) = \dim\left(V(\phi_1^{[1,1]}) + V(\phi_1^{[0,2]})\right). \tag{A.11}$$

If Π is restricted to $V(\phi_1^{[2,0]})$, the mapping (A.4) is reduced to

$$\Pi\left(\phi_1^{[2,0]}(z,\mu)\right) = [A_1 z, \ \ \phi_1^{[2,0]}] \in W\left(f_1^{[2,0]}\right).$$

This is the homological equation of Poincaré normal form in the classical theory of dynamical systems (see [4]). It is a known fact that

$$\dim\left(\Pi(V(\phi_1^{[2,0]}))\right) = \dim\left(V(\phi_1^{[2,0]})\right) - 2. \tag{A.12}$$

Notice that the dimension of the normal form for the uncontrollable part, i.e., the dimension of $W(\bar{f}_1^{[2]})$ in (3.3), is $2n+4$. The dimension of $W(g_1^{[1]})$ is $2(n+1)$. In summary, (3.3), (A.11), and (A.12) imply

$$\begin{aligned}
&\dim\left(\Pi(V(\phi_1^{[2]})\right) + \dim\left(W(\bar{f}_1^{[2]})\right) \\
&\quad = \left(\dim(V(\phi_1^{[2]})) - 2\right) + 2n + 4 \\
&\quad = \dim\left(W(f_1^{[2]})\right) + 2n + 2 \\
&\quad = \dim\left(W(f_1^{[2]})\right) + \dim\left(W(g_1^{[1]})\right).
\end{aligned}$$

This equality proves the second equation in (A.6).

We complete the proof by proving the first equation of (A.6). By the separation principle, it can be proved separately for the three components in

$$V\left(\phi_1^{[2]}\right) = V\left(\phi_1^{[2,0]}\right) + V\left(\phi_1^{[1,1]}\right) + V\left(\phi_1^{[0,2]}\right).$$

Equation (A.6) holds if and only if the normal form is not in the image space of Π. From the classical normal form theory of dynamical systems, it is known that the subspace $W(\bar{f}_1^{[2,0]})$ is not in $\Pi(V(\phi_1^{[2,0]}))$ given any $\Pi(\phi_1^{[1,1]})$ in $\Pi(V(\phi_1^{[1,1]}))$. From (A.10), x_{j+1} appears in $\Pi(\phi_1^{[1,1]})$. However, no $x_2, x_3, \ldots, x_{n-2}$ appears in the normal form $W(\bar{f}_1^{[1,1]})$. Therefore,

$$W\left(\bar{f}_1^{[1,1]}\right) \cap \Pi\left(V(\phi_1^{[1,1]})\right) = \{0\}.$$

Given any $\phi_1^{[0,2]}$ in $V(\phi_1^{[0,2]})$, suppose that j is the largest integer satisfying (A.7). Let $x_i x_j$ be any term in $\phi_1^{[0,2]}$ with nonzero coefficient. If $j < n-2$, then

$$\frac{\partial x_i x_i}{\partial x} A_2 x = x_{i+1} x_j + x_i x_{j+1}.$$

Since $i \le j$, we know that x_{j+1}^2 does not appear in $\Pi(\phi_1^{[0,2]})$. Since $f_1^{[0,2]}$ consists of x_i^2 terms only, we know that $\Pi(V(\phi_1^{[0,2]}))$ is not in normal form. If $j = n-2$, then

$$\Pi(\phi_1^{[0,2]}) = \left(\begin{bmatrix} \dfrac{\partial \phi_1^{[0,2]}}{\partial x} A_2 x - A_1 \phi_1^{[0,2]} \\ 0 \end{bmatrix}, \begin{bmatrix} \dfrac{\partial \phi_1^{[0,2]}}{\partial x_{n-1}} \\ 0 \end{bmatrix} \right).$$

Since

$$\frac{\partial \phi_1^{[0,2]}}{\partial x_{n-1}} \neq 0$$

and

$$W(\bar{g}_1^{[1]}) = 0,$$

we know that $\Pi(\phi_1^{[0,2]})$ is not in normal form. Therefore,

$$W(\bar{f}_1^{[0,2]}) \cap \Pi\left(V(\phi_1^{[0,2]})\right) = \{0\}.$$

REFERENCES

[1] E. H. Abed and J.-H. Fu, *Local feedback stabilization and bifurcation control, part* I. *Hopf bifurcation*, Systems Control Lett., 7 (1986), pp. 11–17.

[2] E. H. Abed and J.-H. Fu, *Local Feedback stabilization and bifurcation control, part* II. *Stationary bifurcation*, Systems Control Lett., 8 (1987), pp. 467–473.

[3] D. Aeyels, *Stabilization of a class of nonlinear systems by a smooth feedback control*, Systems Control Lett., 5 (1985), pp. 289–294.

[4] V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, 2nd ed., Springer-Verlag, New York, 1988.

[5] S. Behtash and S. Sastry, *Stabilization of nonlinear systems with uncontrollable linearization*, IEEE Trans. Automat. Control, 33 (1988), pp. 585–590.

[6] J. Carr, *Application of Centre Manifold Theory*, Springer-Verlag, New York, 1981.

[7] G. Chen and J. L. Moiola, *An overview of bifurcation, chaos and nonlinear dynamics in control systems*, J. Franklin Inst., 331B (1994), pp. 819–858.

[8] G. Chen and X. Dong, *From Chaos to Order*, World Scientific, River Edge, NJ, 1998.

[9] F. Colonius and W. Kliemann, *Controllability and stabilization of one-dimensional systems near bifurcation points*, Systems Control Lett., 24 (1995), pp. 87–95.

[10] R. Genesio, A. Tesi, H. O. Wang, and E. H. Abed, *Control of period doubling bifurcations using harmonic balance*, in Proceedings of the 32nd IEEE Conference on Decision and Control, San Antonio, TX, 1993, pp. 492–497.

[11] P. Glendinning, *Stability, Instability and Chaos: An Introduction to the Theory of Nonlinear Differential Equations*, Cambridge University Press, Cambridge, MA, 1994.

[12] G. Gu, X. Chen, A. G. Sparks, and S. S. Banda, *Bifurcation stabilization with local output feedback*, SIAM J. Control Optim., 37 (1999), pp. 934–956.

[13] B. Hamzi, *Analyse et commande des systèmes non linéaires non commandables en première approximation dans le cadre de la théorie des bifurcations*, Ph.D. Thesis, University of Paris XI-Orsay, France, 2001.

[14] W. Kang and A. J. Krener, *Extended quadratic controller normal form and dynamic state feedback linearization of nonlinear systems*, SIAM J. Control Optim., 30 (1992), pp. 1319–1337.

[15] W. Kang, *Quadratic normal forms of nonlinear control systems with uncontrollable linearization*, in Proceedings of the 34th IEEE Conference on Decision and Control, 1995, pp. 608–612.

[16] W. Kang, *Bifurcation and normal form of nonlinear control systems.* I, SIAM J. Control Optim., 36 (1998), pp. 193–212.

[17] W. Kang, *Bifurcation and normal form of nonlinear control systems.* II, SIAM J. Control Optim., 36 (1998), pp. 213–232.

[18] W. Kang, *Bifurcation control via state feedback for systems with a single uncontrollable mode*, SIAM J. Control Optim., 38 (2000), pp. 1428–1452.

[19] A. J. Krener, *Approximate linearization by state feedback and coordinate change*, Systems Control Lett., 5 (1984), pp. 181–185.

[20] A. J. Krener, *The Feedbacks Which Soften the Primary Bifurcation of MG* 3 *Approximate Linearization by State Feedback and Coordinate Change*, PRET Working Paper D95-9-11 (1995), 181–185.

[21] D.-C. Liaw and E. H. Abed, *Stability analysis and control of rotating stall*, in Proceedings of the 2nd IFAC Symposium on Nonlinear Control Systems, Bordeaux, France, 1992, pp. 57–62.

[22] F. E. McCaughan, *Bifurcation analysis of axial flow compressor stability*, SIAM J. Appl. Math., 50 (1990), pp. 1232–1253.

[23] H. O. Wang and E. H. Abed, *Bifurcation control of a chaotic system*, Automatica, 31 (1995), pp. 1213–1226.

[24] Y. Wang and R. M. Murray, *Feedback stabilization of steady-state feedback and Hopf bifurcations*, in Proceedings of the 37th IEEE Conference on Decision and Control, 1998, pp. 2431–2437.

[25] M. A. Pinsky and I. Shumyatsky, *Feedback stabilization of bifurcation phenomena and its application to the control of voltage instabilities and collapse*, in Proceedings of the 3rd IFAC Symposium on Nonlinear Control Systems, Lake Tahoe, CA, 1995, pp. 58–63.

[26] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Texts Appl. Math. 2, Springer-Verlag, New York, 1990.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 2221–2238

GLOBAL ASYMPTOTIC CONTROLLABILITY IMPLIES INPUT-TO-STATE STABILIZATION*

MICHAEL MALISOFF†, LUDOVIC RIFFORD‡, AND EDUARDO SONTAG§

Abstract. The main problem addressed in this paper is the design of feedbacks for globally asymptotically controllable (GAC) control affine systems that render the closed-loop systems input-to-state stable (ISS) with respect to actuator errors. Extensions for fully nonlinear GAC systems with actuator errors are also discussed. Our controllers have the property that they tolerate small observation noise as well.

Key words. asymptotic controllability, Lyapunov functions, input-to-state stability, nonsmooth analysis

AMS subject classifications. 93B52, 93D15, 93D20

DOI. 10.1137/S0363012903422333

1. Introduction.

The theory of input-to-state stability (ISS) forms the basis for much of the current research in mathematical control theory (see [15, 22, 23]). The ISS property was introduced in [19]. In the past decade, there has been a great deal of research done on the problem of finding ISS stabilizing control laws (see [7, 8, 9, 12]). This paper is concerned with the ISS of control systems of the form

$$\dot{x} = f(x) + G(x)u, \tag{1.1}$$

where f and G are locally Lipschitz vector fields on $\mathbb{R}^n$, $f(0) = 0$, and the control u is valued in $\mathbb{R}^m$ (but see also section 5 for extensions for fully nonlinear systems). We assume throughout that (1.1) is globally asymptotically controllable (GAC), and we construct a feedback $K : \mathbb{R}^n \to \mathbb{R}^m$ for which

$$\dot{x} = f(x) + G(x)K(x) + G(x)u \tag{1.2}$$

is ISS. As pointed out in [3, 24], a continuous stabilizing feedback K fails to exist in general. This fact forces us to consider discontinuous feedbacks K, so our solutions will be taken in the more general sense of sampling and Euler solutions for dynamics that are discontinuous in the state. By an Euler solution, we mean a uniform limit of sampling solutions, taken as the frequency of sampling becomes infinite (see section 2 for precise definitions). This will extend [19, 20], which show how to make C^o-stabilizable systems ISS to actuator errors. In particular, our results apply to the nonholonomic integrator (see [3, 10] and section 4 below) and other applications where

*Received by the editors January 31, 2003; accepted for publication (in revised form) August 18, 2003; published electronically April 7, 2004.
http://www.siam.org/journals/sicon/42-6/42233.html

†Corresponding author. Department of Mathematics, 304 Lockett Hall, Louisiana State University and A & M College, Baton Rouge, LA 70803-4918 (malisoff@lsu.edu). This author was supported in part by Louisiana Board of Regents Support Fund Contract LEQSF(2002-04)-ENH-TR-26 as part of the project "Interdisciplinary Education, Outreach, and Research in Control Theory at LSU."

‡Institut Girard Desargues, Université Lyon 1, Bâtiment Braconnier, 21 Avenue Claude Bernard, 69622 Villeurbanne Cedex, France (rifford@desargues.univ-lyon1.fr).

§Department of Mathematics, Rutgers-New Brunswick, Hill Center-Busch Campus, 110 Frelinghuysen Road, Piscataway, NJ 08854-8019 (sontag@control.rutgers.edu). This author was supported by USAF Grant F49620-01-1-0063 and by NSF Grant CCR-0206789.

Brockett's condition is not satisfied, and which therefore cannot be stabilized by continuous feedbacks (see [21, 22, 25]).

Our results also *strengthen* [6], which constructed feedbacks for GAC systems that render the closed-loop systems globally asymptotically stable (GAS). Our main tool will be the recent constructions of semiconcave control Lyapunov functions (CLFs) for GAC systems from [16, 17]. Our results also apply in the more general situation where measurement noise may occur. In particular, our feedback K will have the additional feature that the *perturbed* system

$$\dot{x} = f(x) + G(x)K(x+e) + G(x)u \tag{1.3}$$

is also ISS when the observation error $e : [0,\infty) \to \mathbb{R}^n$ in the controller is *sufficiently small*. In this context, the precise value of $e(t)$ is unknown to the controller, but information about upper bounds on the magnitude of $e(t)$ can be used to design the feedback. We will prove the following:

THEOREM 1.1. *If* (1.1) *is GAC, then there exists a feedback* K *for which* (1.3) *is ISS for Euler solutions.*

The preceding theorem characterizes the uniform limits of sampling solutions of (1.3) (see section 2 for the precise definitions of Euler and sampling solutions). From a computational standpoint, it is also desirable to know how frequently to sample in order to achieve ISS for sampling solutions. This information is provided in the following semidiscrete version of Theorem 1.1 for sampling solutions:

THEOREM 1.2. *If* (1.1) *is GAC, then there exists a feedback* K *for which* (1.3) *is ISS for sampling solutions.*

This paper is organized as follows. In section 2, we review the relevant background on CLFs, ISS, nonsmooth analysis, and discontinuous feedbacks. In section 3, we prove our main results. This is followed in section 4 by a comparison of our feedback construction with the known feedback constructions for C^o-stabilizable systems, and an application of our results to the nonholonomic integrator. We close in section 5 with an extension for fully nonlinear systems.

2. Definitions and main lemmas. Let $\mathcal{K}_\infty$ denote the set of all continuous functions $\rho : [0,\infty) \to [0,\infty)$ for which (i) $\rho(0) = 0$ and (ii) ρ is strictly increasing and unbounded. Note for future reference that $\mathcal{K}_\infty$ is closed under inverse and composition (i.e., if $\rho_1, \rho_2 \in \mathcal{K}_\infty$, then $\rho_1^{-1}, \rho_1 \circ \rho_2 \in \mathcal{K}_\infty$). We let $\mathcal{KL}$ denote the set of all continuous functions $\beta : [0,\infty) \times [0,\infty) \to [0,\infty)$ for which (1) $\beta(\cdot, t) \in \mathcal{K}_\infty$ for each $t \geq 0$, (2) $\beta(s,\cdot)$ is nonincreasing for each $s \geq 0$, and (3) $\beta(s,t) \to 0$ as $t \to +\infty$ for each $s \geq 0$.

For each $k \in \mathbb{N}$ and $r > 0$, we define

$$\mathcal{M}^k = \{\text{measurable } u : [0,\infty) \to \mathbb{R}^k : |u|_\infty < \infty\}$$

and $\mathcal{M}^k_r := \{u \in \mathcal{M}^k : |u|_\infty \leq r\}$, where $|\cdot|_\infty$ is the essential supremum. We let $\|u(s)\|_I$ denote the essential supremum of a function u restricted to an interval I. Let $|\cdot|$ denote the Euclidean norm, in the appropriate dimension, and

$$r\mathcal{B}_k := \{x \in \mathbb{R}^k : |x| < r\}$$

for each $k \in \mathbb{N}$ and $r > 0$. The closure of $r\mathcal{B}_k$ is denoted by $r\bar{\mathcal{B}}_k$, and $\mathrm{bd}(S)$ denotes the boundary of any subset S in Euclidean space. We also set

$$\mathcal{O} := \{e : [0,\infty) \to \mathbb{R}^n\}, \ \ \sup(e) = \sup\{|e(t)| : t \geq 0\}$$

for all $e \in \mathcal{O}$, and $\mathcal{O}_\eta := \{e \in \mathcal{O} : \sup(e) \leq \eta\}$ for each $\eta > 0$. For any compact set $\mathcal{F} \subseteq \mathbb{R}^n$ and $\varepsilon > 0$, we define the compact set

$$\mathcal{F}^\varepsilon := \{x \in \mathbb{R}^n : \min\{|x - p| : p \in \mathcal{F}\} \leq \varepsilon\},$$

i.e., the "ε-enlargement of $\mathcal{F}$." Given a continuous function

$$h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n : (x, u) \mapsto h(x, u)$$

that is locally Lipschitz in x uniformly on compact subsets of $\mathbb{R}^n \times \mathbb{R}^m$, we let $\phi_h(\cdot, x_o, u)$ denote the trajectory of $\dot{x} = h(x, u)$ starting at $x_o \in \mathbb{R}^n$ for each choice of $u \in \mathcal{M}^m$. In this case, $\phi_h(\cdot, x_o, u)$ is defined on some maximal interval $[0, t)$, with $t > 0$ depending on u and x_o. Let C^k denote the set of all continuous functions $\varphi : \mathbb{R}^n \to \mathbb{R}$ that have at least k continuous derivatives (for $k = 0, 1$). We use the following controllability notion, which was introduced in [18] and later reformulated in terms of $\mathcal{KL}$ functions in [22]:

DEFINITION 2.1. *We call the system $\dot{x} = h(x, u)$* globally asymptotically controllable (GAC) *provided there are a nondecreasing function $\sigma : [0, \infty) \to [0, \infty)$ and a function $\beta \in \mathcal{KL}$ satisfying the following: for each $x_o \in \mathbb{R}^n$, there exists $u \in \mathcal{M}^m$ such that*

(a) $|\phi_h(t, x_o, u)| \leq \beta(|x_o|, t)$ *for all $t \geq 0$; and*
(b) $|u(t)| \leq \sigma(|x_o|)$ *for a.e. $t \geq 0$.*

In this case, we call σ the GAC modulus *of $\dot{x} = h(x, u)$.*

In our main results, the controllers will be taken to be discontinuous feedbacks, so the dynamics will be discontinuous in the state variable. Therefore, we will form our trajectories through sampling and through uniform limits of sampling trajectories, as follows. We say that $\pi = \{t_o, t_1, t_2, \ldots\} \subset [0, \infty)$ is a *partition* of $[0, \infty)$ provided $t_o = 0$, $t_i < t_{i+1}$ for all $i \geq 0$, and $t_i \to \infty$ as $i \to +\infty$. The set of all partitions of $[0, \infty)$ is denoted by Par. Let

$$F : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^n : (x, p, u) \mapsto F(x, p, u)$$

be a continuous function that is locally Lipschitz in x uniformly on compact subsets of $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$. A *feedback* for F is defined to be any locally bounded function $K : \mathbb{R}^n \to \mathbb{R}^m$ for which $K(0) = 0$. In particular, we allow discontinuous feedbacks. The arguments x, p, and u in F are used to represent the state, feedback value, and actuator error, respectively.

Given a feedback $K : \mathbb{R}^n \to \mathbb{R}^m$, $\pi = \{t_o, t_1, t_2, \ldots\} \in \text{Par}$, $x_o \in \mathbb{R}^n$, $e \in \mathcal{O}$, and $u \in \mathcal{M}^m$, the *sampling solution* for the initial value problem (IVP)

$$\dot{x}(t) = F(x(t), K(x(t) + e(t)), u(t)), \tag{2.1}$$

$$x(0) = x_o \tag{2.2}$$

is the continuous function defined by recursively solving

$$\dot{x}(t) \;=\; F(x(t), K(x(t_i) + e(t_i)), u(t)) \tag{2.3}$$

from the initial time t_i up to time $s_i = t_i \vee \sup\{s \in [t_i, t_{i+1}] : x(\cdot) \text{ is defined on } [t_i, s)\}$, where $x(0) = x_o$. In this case, the sampling solution of (2.1)–(2.2) is defined on the right-open interval from time zero up to time $\bar{t} = \inf\{s_i : s_i < t_{i+1}\}$. This sampling solution will be denoted by $t \mapsto x_\pi(t; x_o, u, e)$ to exhibit its dependence on $\pi \in \text{Par}$,

$x_o \in \mathbb{R}^n$, $u \in \mathcal{M}^m$, and $e \in \mathcal{O}$, or simply by x_π, when the dependence is clear from the context. Note that if $s_i = t_{i+1}$ for all i, then $\bar{t} = +\infty$ (as the infimum of the empty set), so in that case, the sampling solution $t \mapsto x_\pi(t; x_o, u, e)$ is defined on $[0, \infty)$.

We also define the *upper diameter* and the *lower diameter* of a given partition $\pi = \{t_o, t_1, t_2, \ldots\}$ by

$$\overline{\mathbf{d}}(\pi) = \sup_{i \geq 0}(t_{i+1} - t_i), \quad \underline{\mathbf{d}}(\pi) = \inf_{i \geq 0}(t_{i+1} - t_i),$$

respectively. We let $\mathrm{Par}(\delta) := \left\{\pi \in \mathrm{Par} : \overline{\mathbf{d}}(\pi) < \delta\right\}$ for each $\delta > 0$. We will say that a function $y : [0, \infty) \to \mathbb{R}^n$ is an *Euler solution (robust to small observation errors)* of

$$\dot{x}(t) = F(x(t), K(x(t)), u(t)), \quad x(0) = x_o \tag{2.4}$$

for $u \in \mathcal{M}^m$ provided there are sequences $\pi_r \in \mathrm{Par}$ and $e_r \in \mathcal{O}$ such that

(a) $\overline{\mathbf{d}}(\pi_r) \to 0$;
(b) $\sup(e_r)/\underline{\mathbf{d}}(\pi_r) \to 0$; and
(c) $t \mapsto x_{\pi_r}(t; x_o, u, e_r)$ converges uniformly to y as $r \to +\infty$.

Note that the approximating trajectories in the preceding definition all use the same input u (but see Remark 2.4 for a more general notion of Euler solutions, which also involves sequences of inputs).

This paper will design feedbacks that make closed-loop GAC systems ISS with respect to actuator errors. More precisely, we will use the following definition:

Definition 2.2. *We say that* (2.1) *is* ISS for sampling solutions *provided there are* $\beta \in \mathcal{KL}$ *and* $\gamma \in \mathcal{K}_\infty$ *satisfying: For each* $\varepsilon, M, N > 0$ *with* $0 < \varepsilon < M$*, there exist positive* $\delta = \delta(\varepsilon, M, N)$ *and* $\kappa = \kappa(\varepsilon, M, N)$ *such that for each* $\pi \in \mathrm{Par}(\delta)$*,* $x_o \in M\bar{\mathcal{B}}_n$*,* $u \in \mathcal{M}^m_N$*, and* $e \in \mathcal{O}$ *for which* $\sup(e) \leq \kappa \underline{\mathbf{d}}(\pi)$*,*

$$|x_\pi(t; x_o, u, e)| \leq \max\{\beta(M, t) + \gamma(N), \varepsilon\} \tag{2.5}$$

for all $t \geq 0$.

Roughly speaking, condition (2.5) says that the system is ISS, modulo small overflows, if the sampling is done "quickly enough," as determined by the condition $\pi \in \mathrm{Par}(\delta)$, but "not too quickly," as determined by the additional requirement that $\underline{\mathbf{d}}(\pi) \geq (1/\kappa)\sup(e)$. In the special case where the observation error $e \equiv 0$, the condition on $\underline{\mathbf{d}}(\pi)$ in Definition 2.2 is no longer needed; our results are new even for this particular case.

Notice that the bounds on e are in the supremum, not the essential supremum. It is easy to check that Definition 2.2 remains unchanged if we replace the right-hand side in (2.5) by $\beta(M, t) + \gamma(N) + \varepsilon$. We also use the following analogue of Definition 2.2 for Euler solutions:

Definition 2.3. *We say that the system* (2.1) *is* ISS for Euler solutions *provided there exist* $\beta \in \mathcal{KL}$ *and* $\gamma \in \mathcal{K}_\infty$ *satisfying: If* $u \in \mathcal{M}^m$ *and* $x_o \in \mathbb{R}^n$*, and if* $t \mapsto x(t)$ *is an Euler solution of* (2.4)*, then*

$$|x(t)| \leq \beta(|x_o|, t) + \gamma(|u|_\infty) \tag{2.6}$$

for all $t \geq 0$.

Remark 2.4. In the definition of Euler solutions we gave above, all of the approximating trajectories $t \mapsto x_{\pi_r}(t; x_o, u, e_r)$ use the same input $u \in \mathcal{M}^m$. A different way to define Euler solutions, which gives rise to a more general class of limiting solutions, is as follows: A function $y : [0, \infty) \to \mathbb{R}^n$ is a *generalized Euler solution* of (2.4) for $u \in \mathcal{M}^m$ provided there are sequences $\pi_r \in \mathrm{Par}$, $e_r \in \mathcal{O}$, and $u_r \in \mathcal{M}^m$ such that

(a) $\overline{\mathbf{d}}(\pi_r) \to 0$;
(b) $\sup(e_r)/\underline{\mathbf{d}}(\pi_r) \to 0$;
(c) $|u_r|_\infty \leq |u|_\infty$ for all r; and
(d) $t \mapsto x_{\pi_r}(t; x_o, u_r, e_r)$ converges uniformly to y as $r \to +\infty$.

We can then define ISS for generalized Euler solutions exactly as in Definition 2.3, by merely replacing "Euler solution" with "generalized Euler solution" throughout the definition. Our proof of Theorem 1.1 will actually show the following slightly more general result: If (1.1) is GAC, then there exists a feedback K for which (1.3) is ISS for generalized Euler solutions.

Our main tools in this paper will be nonsmooth analysis and nonsmooth Lyapunov functions. The following definitions will be used. Let Ω be an arbitrary open subset of $\mathbb{R}^n$. Recall the following definition:

DEFINITION 2.5. *Let $g : \Omega \to \mathbb{R}$ be a continuous function on Ω; it is said to be* semiconcave *on Ω provided for each point $x_o \in \Omega$, there exist $\rho, C > 0$ such that*

$$g(x) + g(y) - 2g\left(\frac{x+y}{2}\right) \leq C||x-y||^2$$

for all $x, y \in x_o + \rho\mathcal{B}_n$.

The *proximal superdifferential* (respectively, *proximal subdifferential*) of a function $V : \Omega \to \mathbb{R}$ at $x \in \Omega$, which is denoted by $\partial^P V(x)$ (resp., $\partial_P V(x)$), is defined to be the set of all $\zeta \in \mathbb{R}^n$ for which there exist $\sigma, \eta > 0$ such that

$$V(y) - V(x) - \sigma|y-x|^2 \leq \langle \zeta, y-x\rangle \quad (\text{resp., } V(y) - V(x) + \sigma|y-x|^2 \geq \langle \zeta, y-x\rangle)$$

for all $y \in x + \eta\mathcal{B}_n$. The *limiting subdifferential* of a continuous function $V : \Omega \to \mathbb{R}$ at $x \in \Omega$ is

$$\partial_L V(x) := \{q \in \mathbb{R}^n : \exists x_n \to x \quad \text{and} \quad q_n \in \partial_P V(x_n) \ \text{ s.t. } \ q_n \to q\}.$$

In what follows, we assume $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n : (x,u) \mapsto h(x,u)$ is continuous, that it is locally Lipschitz in x uniformly on compact subsets of $\mathbb{R}^n \times \mathbb{R}^m$, and that $h(0,0) = 0$. The following definition was introduced in [18] and reformulated in proximal terms in [22].

DEFINITION 2.6. *A* control-Lyapunov function (CLF) *for*

$$\dot{x} = h(x,u) \tag{2.7}$$

is a continuous, positive definite, proper function $V : \mathbb{R}^n \to \mathbb{R}$ for which there exist a continuous, positive definite function $W : \mathbb{R}^n \to \mathbb{R}$, and a nondecreasing function $\alpha : [0,\infty) \to [0,\infty)$, satisfying

$$\forall \zeta \in \partial_P V(x), \quad \inf_{|u| \leq \alpha(|x|)} \langle \zeta, h(x,u)\rangle \leq -W(x)$$

for all $x \in \mathbb{R}^n$. In this case, we call (V,W) a Lyapunov pair *for* (2.7).

Recall the following lemmas (see [17]):

LEMMA 2.7. *If* (2.7) *is GAC, then there exist a CLF V for* (2.7) *that is semiconcave on $\mathbb{R}^n \setminus \{0\}$ and a nondecreasing function $\alpha : [0,\infty) \to [0,\infty)$ that satisfy*

$$\forall \zeta \in \partial_L V(x), \quad \min_{|u| \leq \alpha(|x|)} \langle \zeta, h(x,u)\rangle \leq -V(x) \tag{2.8}$$

for all $x \in \mathbb{R}^n$.

LEMMA 2.8. *Let $V : \Omega \to \mathbb{R}$ be semiconcave. Then V is locally Lipschitz, and $\emptyset \neq \partial_L V(x) \subseteq \partial^P V(x)$ for all $x \in \Omega$. Moreover, for each compact set $Q \subset \Omega$, there exist constants $\sigma, \mu > 0$ such that $V(y) - V(x) - \sigma|y - x|^2 \leq \langle \zeta, y - x \rangle$ for all $y \in x + \mu\mathcal{B}_n$, all $x \in Q$, and all $\zeta \in \partial^P V(x)$.*

Notice that Lemma 2.8 allows the constants in the definition of $\partial^P V(x)$ to be chosen uniformly on compact sets.

Remark 2.9. In [17], the controls u take all their values in a given compact metric space U. The precise version of the CLF existence theorem in [17] is the same as our Lemma 2.7, except that the infimum in the decay condition (2.8) is replaced by the infimum over all $u \in U$. The version of Lemma 2.7 we gave above follows from a slight modification of the arguments of [16, 17], using the GAC modulus in the GAC definition (see Definition 2.1). The existence theory [16] for semiconcave CLFs is a strengthening of the proof that continuous CLFs exist for any GAC system (see [18]).

3. Proofs of theorems. Let V be a CLF satisfying the requirements of Lemma 2.7 for the dynamics

$$h(x, u) = f(x) + G(x)u. \tag{3.1}$$

Define the functions $\underline{\alpha}, \overline{\alpha} \in \mathcal{K}_\infty$ by

$$\underline{\alpha}(s) = \min\{|x| : V(x) \geq s\} \text{ and } \overline{\alpha}(s) = \max\{|x| : V(x) \leq s\}. \tag{3.2}$$

One can easily check that

$$\forall x \in \mathbb{R}^n, \quad \underline{\alpha}(V(x)) \leq |x| \text{ and } \overline{\alpha}(V(x)) \geq |x|. \tag{3.3}$$

Moreover, by reducing $\underline{\alpha}$, we may assume that $\underline{\alpha}(s) \leq s$ for all $s \geq 0$, while still satisfying (3.3).

Let $x \mapsto \zeta(x)$ be any selection of $\partial_L V(x)$ on $\mathbb{R}^n$ and let $\zeta(0) \in \mathbb{R}^n$ be arbitrary. For each $x \in \mathbb{R}^n$, we can choose $u = u_x \in \alpha(|x|)\mathcal{B}_m$ that satisfies the inequality in (2.8) for the dynamics (3.1) and $\zeta = \zeta(x)$. Define the feedback $K_1 : \mathbb{R}^n \to \mathbb{R}^m$ by $K_1(x) = u_x$ for all $x \neq 0$ and $K_1(0) = 0$. We use the functions

$$\begin{array}{l} a(x) = \langle \zeta(x), f(x) + G(x)K_1(x) \rangle, \;\; b_j(x) = \langle \zeta(x), g_j(x) \rangle \quad \forall j, \\ K_2(x) = -V(x)(\operatorname{sgn}\{b_1(x)\}, \operatorname{sgn}\{b_2(x)\}, \ldots, \operatorname{sgn}\{b_m(x)\})^T, \end{array} \tag{3.4}$$

where g_j is the jth column of G for $j = 1, 2, \ldots, m$, and

$$\operatorname{sgn}\{s\} \;\; = \;\; \begin{cases} 1, & s > 0, \\ -1, & s < 0, \\ 0, & s = 0. \end{cases}$$

We remark that our results remain true, with minor changes in the proofs, if the factor $-V(x)$ in the definition of K_2 is replaced by $-W(x)$ for an arbitrary positive definite proper continuous function $W : \mathbb{R}^n \to \mathbb{R}$. In particular, $K := K_1 + K_2$ is a feedback for the dynamics

$$F(x, p, u) = f(x) + G(x)(p + u).$$

Moreover,

$$a(x) \leq -V(x) < 0 \;\; \forall x \in \mathbb{R}^n \setminus \{0\}. \tag{3.5}$$

We next show that

$$\dot{x}(t) = F(x(t), K(x(t) + e(t)), u(t)) \tag{3.6}$$

is ISS for sampling solutions.

To this end, choose $\varepsilon, M, N > 0$ for which $0 < \varepsilon < M$. It clearly suffices to verify the ISS property (2.5) for $\varepsilon < 1$, since that would imply the property for *all* overflows $\varepsilon > 0$. Choose

$$u \in \mathcal{M}_N^m, \;\; e \in \mathcal{O}_{\varepsilon/16}, \;\; x_o \in M\bar{\mathcal{B}}_n. \tag{3.7}$$

In what follows, x_π denotes the sampling solution for (3.6) for the choices (3.7) and $\pi \in \mathrm{Par}$, and $\tilde{x}_\pi$ is the (possibly discontinuous) function that is inductively defined by solving the IVP

$$\dot{x}(t) = f(x(t)) + G(x(t))[K(\tilde{x}_i) + u(t)], \;\; x(t_i) = \tilde{x}_i$$

on $[t_i, t_{i+1})$, where $\tilde{x}_i := x_i + e(t_i)$, $x_i := x_\pi(t_i)$, and $\pi = \{t_o, t_1, t_2, \ldots\}$. We later restrict the choice of π so that x_π and $\tilde{x}_\pi$ are defined on $[0, \infty)$. We will use the compact set

$$Q = \left\{ \left[\overline{\alpha} \circ \underline{\alpha}^{-1}(N + M) + 1 \right] \bar{\mathcal{B}}_n \right\} \setminus \varepsilon \mathcal{B}_n.$$

Notice that $Q, Q^{\varepsilon/2} \subseteq \mathbb{R}^n \setminus \{0\}$, and that $x_o \in Q^\varepsilon$. Using Lemma 2.8 and the semiconcavity of V on $\mathbb{R}^n \setminus \{0\}$, we can find $\sigma, \mu > 0$ such that

$$V(y) - V(x) \leq \langle \zeta(x), y - x \rangle + \sigma |y - x|^2 \tag{3.8}$$

for all $y \in x + \mu \mathcal{B}_n$ and $x \in Q^{\varepsilon/2}$. Let $\mathcal{L}_\varepsilon > 1$ be a Lipschitz constant for V on $Q^{\varepsilon/2}$, the existence of which is also guaranteed by Lemma 2.8. It follows from the definition of a CLF that

$$\begin{aligned} \lambda_- &:= \min \left\{ V(p) : p \in Q^{\varepsilon/2} \right\}, \\ \lambda_+ &:= \max \left\{ V(p) : p \in Q^{\varepsilon} \right\} \end{aligned} \tag{3.9}$$

are finite positive numbers. Therefore, we can choose $\tilde{\varepsilon} \in (0, \varepsilon)$ for which

$$\overline{\alpha}\left(p + \frac{\mathcal{L}_\varepsilon}{4} \tilde{\varepsilon} \right) \leq \overline{\alpha}(p) + \frac{\varepsilon}{8} \qquad \forall p \in \left[0, \underline{\alpha}^{-1}(N) + \lambda_+ \right]. \tag{3.10}$$

We can also find

$$\delta = \delta(\varepsilon, M, N) \in \left(0, \frac{\tilde{\varepsilon}}{16 + \lambda_+ + 16\lambda_+} \right) \tag{3.11}$$

such that if

$$\pi \in \mathrm{Par}(\delta), \;\; e \in \mathcal{O}_{\tilde{\varepsilon}/16}, \;\; x_i \in Q^\varepsilon, \tag{3.12}$$

and if $t \in [t_i, t_{i+1})$ is such that $x_\pi(s)$ and $\tilde{x}_\pi(s)$ remain in $Q^{2\varepsilon}$ for all $s \in [t_i, t]$, then

$$\max\{|x_\pi(t) - x_i|, |\tilde{x}_\pi(t) - \tilde{x}_i|\} \leq \min \left\{ \mu, \frac{\tilde{\varepsilon}}{16(1 + \mathcal{L}_\varepsilon)}, \sqrt{\frac{\lambda_-}{8\sigma}(t - t_i)} \right\}. \tag{3.13}$$

This follows from the local boundedness of K, f, and G. It follows from (3.13) that $\tilde{x}_\pi(t) \in Q^{\varepsilon/4}$ (resp., $x_\pi(t) \in Q^{\varepsilon/4}$) for all $t \in [t_i, t_{i+1})$ and all i such that $\tilde{x}_i \in Q$ (resp., $x_i \in Q$), since the trajectories cannot move the initial value more than $\frac{\varepsilon}{16}$ and there are no blow up times for the trajectories. In particular, (3.13) will show that x_π and $\tilde{x}_\pi$ are defined on $[0, \infty)$, since the argument we are about to give shows that $x_i \in Q^\varepsilon$ for all i. By reducing δ as necessary, we can assume

$$\begin{array}{l} \|\zeta(\tilde{x}_i) \cdot (F(\tilde{x}_i, K(\tilde{x}_i), u(s)) - f(\tilde{x}_\pi(s)) \\ -G(\tilde{x}_\pi(s))[u(s) + K(\tilde{x}_i)]) \|_{[t_i, t_{i+1})} \leq \frac{\lambda_-}{8} \end{array} \tag{3.14}$$

for all i such that $\tilde{x}_i \in Q^{\varepsilon/2}$. This follows from the Lipschitzness of f and G on Q^ε. Having chosen δ to satisfy the preceding requirements, pick any $\pi \in \mathrm{Par}(\delta)$. It follows from (3.8) and (3.13) that

$$\begin{array}{rcl} V(\tilde{x}_\pi(t)) - V(\tilde{x}_i) & \leq & \langle \zeta(\tilde{x}_i), \tilde{x}_\pi(t) - \tilde{x}_i \rangle + \sigma |\tilde{x}_\pi(t) - \tilde{x}_i|^2 \\ & \leq & \langle \zeta(\tilde{x}_i), \tilde{x}_\pi(t) - \tilde{x}_i \rangle + \frac{\lambda_-}{8}(t - t_i) \end{array} \tag{3.15}$$

for all $t \in [t_i, t_{i+1})$ and all i such that $\tilde{x}_i \in Q^{\varepsilon/4}$. Moreover, if $\tilde{x}_i \in Q^{\varepsilon/4}$ and $t \in [t_i, t_{i+1})$, and if

$$V(\tilde{x}_i) \geq N, \tag{3.16}$$

then

$$\begin{array}{rcl} \langle \zeta(\tilde{x}_i), \tilde{x}_\pi(t) - \tilde{x}_i \rangle & \leq & \left\langle \zeta(\tilde{x}_i), \displaystyle\int_{t_i}^{t} F(\tilde{x}_i, K(\tilde{x}_i), u(s)) \mathrm{d}s \right\rangle + \frac{\lambda_-}{8}(t - t_i) \quad \text{(by (3.14))} \\ & = & (t - t_i) \langle \zeta(\tilde{x}_i), f(\tilde{x}_i) + G(\tilde{x}_i) K(\tilde{x}_i) \rangle \\ & & + \displaystyle\int_{t_i}^{t} \langle \zeta(\tilde{x}_i), G(\tilde{x}_i) u(s) \rangle \mathrm{d}s + \frac{\lambda_-}{8}(t - t_i) \\ & \leq & (t - t_i) a(\tilde{x}_i) - (t - t_i) V(\tilde{x}_i) \displaystyle\sum_{j=1}^{m} |b_j(\tilde{x}_i)| \\ & & + N(t - t_i) \displaystyle\sum_{j=1}^{m} |b_j(\tilde{x}_i)| + \frac{\lambda_-}{8}(t - t_i) \\ & \leq & (t - t_i) a(\tilde{x}_i) + \frac{\lambda_-}{8}(t - t_i) \quad \text{(by (3.16))} \\ & \leq & -(t - t_i) V(\tilde{x}_i) + \frac{\lambda_-}{8}(t - t_i) \quad \text{(by (3.5))}. \end{array} \tag{3.17}$$

Let

$$S = \{x \in \mathbb{R}^n : V(x) \leq \underline{\alpha}^{-1}(N)\}.$$

Then $S \subset Q^\varepsilon$. Indeed, $x \in S$ implies

$$\underline{\alpha} \circ \overline{\alpha}^{-1}(|x|) \leq \underline{\alpha} \circ \overline{\alpha}^{-1} \circ \overline{\alpha} \circ V(x) \leq N,$$

and therefore $|x| \leq \overline{\alpha} \circ \underline{\alpha}^{-1}(N)$. By further reducing ε, we can assume $(2\varepsilon)\mathcal{B}_n \subset S$. If $\tilde{x}_i \in Q^{\varepsilon/4}$ but $\tilde{x}_i \notin S$, then $V(\tilde{x}_i) \geq \underline{\alpha}^{-1}(N) \geq N$, so (3.9) and (3.15) give

$$\begin{array}{rcl} V(\tilde{x}_\pi(t)) - V(\tilde{x}_i) & \leq & -(t - t_i) \dfrac{V(\tilde{x}_i)}{2} + (t - t_i) \dfrac{\lambda_-}{4} \\ & \leq & -(t - t_i) \dfrac{V(\tilde{x}_i)}{4} \qquad \forall t \in [t_i, t_{i+1}). \end{array} \tag{3.18}$$

Let $\mathcal{L}_f$ and $\mathcal{L}_G$ be Lipschitz constants for f and G restricted to Q^ε, respectively. Define the constants

$$\begin{aligned} &R = N + \sup\left\{|K(x)| : x \in Q^{\varepsilon/2}\right\}, \\ &L = \mathcal{L}_f + R\mathcal{L}_G, \quad \kappa = \kappa(\varepsilon, M, N) := \frac{\min\{\lambda_-, \varepsilon\}}{16\mathcal{L}_\varepsilon(e^{L\delta}+1)}. \end{aligned} \tag{3.19}$$

We will presently show that

$$\sup_{t_i \le t < t_{i+1}} |x_\pi(t) - \tilde{x}_\pi(t)| \le |e(t_i)|e^{L\delta} \qquad \forall i \text{ s.t. } x_i \in Q^{\varepsilon/4}. \tag{3.20}$$

Using (3.20), we will now find $\beta \in \mathcal{KL}$ and $\gamma \in \mathcal{K}$ to satisfy the ISS estimate

$$|x_\pi(t)| \le \beta(|x_o|, t) + \gamma(N) + \varepsilon \ \ \forall t \ge 0 \tag{3.21}$$

which will prove Theorem 1.2.

To this end, assume $x_i \in Q$, but that $x_i \notin S^{\tilde{\varepsilon}/16}$. Then (3.13) implies $x_\pi(t)$ and $\tilde{x}_\pi(t)$ both remain in $Q^{\varepsilon/4}$ on $[t_i, t_{i+1})$. Moreover, $\tilde{x}_i \in Q^{\varepsilon/4} \setminus S$, by the choice of e in (3.12). Therefore, if $t \in [t_i, t_{i+1})$, and if

$$\sup(e) \le \kappa \underline{\mathbf{d}}(\pi), \tag{3.22}$$

then the choice of κ gives

$$\begin{aligned} V(x_{i+1}) - V(x_i) &= V(x_{i+1}) - V(\tilde{x}_\pi(t_{i+1}^-)) + V(\tilde{x}_\pi(t_{i+1}^-)) - V(\tilde{x}_i) \\ &\quad + V(\tilde{x}_i) - V(x_i) \\ &\le \mathcal{L}_\varepsilon |x_{i+1} - \tilde{x}_\pi(t_{i+1}^-)| - \frac{t_{i+1}-t_i}{4} V(\tilde{x}_i) \\ &\quad + \mathcal{L}_\varepsilon |e(t_i)| \ \text{(by (3.18))} \\ &\le \mathcal{L}_\varepsilon |e(t_i)| e^{L\delta} - \frac{t_{i+1}-t_i}{4} V(\tilde{x}_i) + \mathcal{L}_\varepsilon |e(t_i)| \ \text{(by (3.20))} \\ &\le \frac{\lambda_-}{16}(t_{i+1}-t_i) - \frac{t_{i+1}-t_i}{4} V(\tilde{x}_i) \ \text{(by (3.22))} \\ &\le -\frac{t_{i+1}-t_i}{8} V(\tilde{x}_i) \ \text{(by (3.9))} \\ &\le -\frac{t_{i+1}-t_i}{8} V(x_i) + \frac{t_{i+1}-t_i}{8} |e(t_i)| \mathcal{L}_\varepsilon \\ &\le -\frac{t_{i+1}-t_i}{8} V(x_i) + \frac{(t_{i+1}-t_i)^2}{16} \lambda_- \\ &\le -\frac{t_{i+1}-t_i}{16} V(x_i), \end{aligned} \tag{3.23}$$

where we use

$$t_{i+1} - t_i \ \le \ \overline{\mathbf{d}}(\pi) \ \le \ \delta < 1$$

to get the last inequality. Set

$$J(t) = \frac{16}{16+t}$$

for all $t \ge 0$. One can easily check that Q^ε contains the set

$$S_V := \{p : V(p) \le \max\{V(q) : |q| \le M + N\}\}.$$

In fact, $p \in S_V$ implies

$$\begin{aligned} |p| &\leq \overline{\alpha}\,(\max\{V(q) : |q| \leq M+N\}) \\ &= \max\{\overline{\alpha} \circ \underline{\alpha}^{-1} \circ \underline{\alpha}(V(q)) : |q| \leq M+N\} \\ &\leq \overline{\alpha} \circ \underline{\alpha}^{-1}(M+N). \end{aligned}$$

In particular, $x_o \in S_V$. It follows from (3.23) that if none of $x_o, x_1, \ldots, x_j$ lies in $S^{\tilde{\varepsilon}/16}$, then

$$\begin{aligned} V(x_1) - V(x_0) &\leq -\frac{t_1}{16} V(x_j), \\ V(x_2) - V(x_1) &\leq -\frac{t_2 - t_1}{16} V(x_j), \\ &\vdots \\ V(x_j) - V(x_{j-1}) &\leq -\frac{t_j - t_{j-1}}{16} V(x_j). \end{aligned}$$

Summing the preceding inequalities would then give

$$V(x_j) - V(x_o) \;\leq\; -\frac{t_j}{16} V(x_j), \quad \text{so} \quad V(x_j) \;\leq\; J(t_j) V(x_o)\,.$$

Hence,

$$V(x_i) \leq J(t_i) V(x_o) \quad \text{for} \quad i = 0, 1, \ldots, j.$$

By the choice of δ in (3.11), it would then follow from (3.13) that

$$V(x_\pi(t)) \leq J(t) V(x_o) + \frac{\tilde{\varepsilon}}{8}$$

up to the least time t at which $x_\pi(t) \in S^{\tilde{\varepsilon}/16}$. Hence, for such t, the choice of $\tilde{\varepsilon}$ (see (3.10)) gives

$$\begin{aligned} |x_\pi(t)| &\leq \overline{\alpha}\left(J(t)V(x_o) + \tfrac{\tilde{\varepsilon}}{8}\right), \\ &\leq \overline{\alpha}\,(J(t)V(x_o)) + \tfrac{\varepsilon}{8}. \end{aligned}$$

On the other hand, (3.23) also shows that if $x_\pi(t) \in S^{\tilde{\varepsilon}/8}$ for some t, then

$$|x_\pi(s)| \leq \overline{\alpha} \circ \underline{\alpha}^{-1}(N) + \varepsilon \;\; \forall s \geq t. \tag{3.24}$$

Indeed, let s_1 be the first sample time above such a time t. Assume $x_\pi(t) \notin \varepsilon \mathcal{B}_n$. By (3.13), $x_\pi(s_1) \in S^{\tilde{\varepsilon}/4}$ and $x_\pi(s_1) \notin \frac{\varepsilon}{2}\mathcal{B}_n$. Therefore, there exists $p \in S$ for which

$$\begin{aligned} V(x_\pi(s_1)) &= V(x_\pi(s_1)) - V(p) + V(p) \\ &\leq \mathcal{L}_\varepsilon \frac{\tilde{\varepsilon}}{4} + \underline{\alpha}^{-1}(N). \end{aligned}$$

In fact, we can pick $p = x_\pi(s_1)$ if $x_\pi(s_1) \in S$ and $p \in \partial S$ otherwise, so $p \notin \frac{\varepsilon}{2}\mathcal{B}_n$. It follows from (3.13) and (3.23) that for the next sample time s_i, we either have $x_\pi(s_i) \in S^{\tilde{\varepsilon}/8}$, or else we have

$$V(x_\pi(s_i)) \leq \mathcal{L}_\varepsilon \frac{\tilde{\varepsilon}}{4} + \underline{\alpha}^{-1}(N).$$

In the first case,

$$|x_\pi(s_i)| \le \overline{\alpha} \circ \underline{\alpha}^{-1}(N) + \frac{\varepsilon}{8},$$

while in the second case,

$$|x_\pi(s_i)| \le \overline{\alpha}\left(\frac{\tilde{\varepsilon}\mathcal{L}_\varepsilon}{4} + \underline{\alpha}^{-1}(N)\right) \le \overline{\alpha} \circ \underline{\alpha}^{-1}(N) + \frac{\varepsilon}{8},$$

by the choice of $\tilde{\varepsilon}$. If $x_\pi(s_i) \notin S^{\tilde{\varepsilon}/16}$, then $V(x_\pi(s_{i+1})) \le V(x_\pi(s_i))$ (by (3.23)), so the preceding argument also gives

$$|x_\pi(s_{i+1})| \le \overline{\alpha} \circ \underline{\alpha}^{-1}(N) + \frac{\varepsilon}{8}.$$

By repeating this argument for subsequent sample times, the assertion (3.24) then follows from (3.13). Defining $\beta \in \mathcal{KL}$ and $\gamma \in \mathcal{K}_\infty$ by

$$\beta(s,t) = \overline{\alpha}\left(\underline{\alpha}^{-1}(s)J(t)\right), \quad \gamma(s) = \overline{\alpha} \circ \underline{\alpha}^{-1}(s), \tag{3.25}$$

it follows that (3.21) holds for all $x_o \in M\bar{\mathcal{B}}_n$, $u \in \mathcal{M}_N^m$, $\pi \in \mathrm{Par}(\delta)$, and $e \in \mathcal{O}$ for which $\sup(e) \le \kappa \underline{\mathbf{d}}(\pi)$. Therefore, Theorem 1.2 will follow once we check (3.20), which is a consequence of Gronwall's inequality.

To this end, notice that if $x_i \in Q^{\varepsilon/4}$, then

$$|x_\pi(t) - \tilde{x}_\pi(t)| \le |x_i - \tilde{x}_i| + \int_{t_i}^t (\mathcal{L}_f|x_\pi(s) - \tilde{x}_\pi(s)| + R\mathcal{L}_G|x_\pi(s) - \tilde{x}_\pi(s)|)\,\mathrm{d}s$$

for all $t \in [t_i, t_{i+1})$, where we are using the constants in (3.19). It follows from Gronwall's inequality that

$$|x_\pi(t) - \tilde{x}_\pi(t)| \le |x_i - \tilde{x}_i| e^{L|t_i - t_{i+1}|} \le |x_i - \tilde{x}_i| e^{L\overline{\mathbf{d}}(\pi)} \le |e(t_i)| e^{L\delta}$$

for all $t \in [t_i, t_{i+1})$, which is (3.20). This proves Theorem 1.2.

We turn next to Theorem 1.1. We need to show the ISS property (2.6) for all Euler solutions $x(t)$ of (2.4). We will actually prove the slightly stronger version of the theorem for generalized Euler solutions, as asserted in Remark 2.4. To this end, choose $u \in \mathcal{M}^m$, $x_o \in \mathbb{R}^n$, and $\varepsilon > 0$. Using our previous conclusion that (1.3) is ISS for sampling solutions, we can let

$$\delta_\varepsilon = \delta(\varepsilon, |x_o|, |u|_\infty) \quad \text{and} \quad \kappa_\varepsilon = \kappa(\varepsilon, |x_o|, |u|_\infty)$$

be the constants from Definition 2.2. Let $x(t)$ be a generalized Euler solution of (2.4), and let π_r, u_r, and e_r satisfy the requirements of the generalized Euler solution definition. It follows from the definition that there is an $\bar{r} \in \mathbb{N}$ such that

$$\overline{\mathbf{d}}(\pi_r) \le \delta_\varepsilon, \quad \sup(e_r) \le \kappa_\varepsilon \underline{\mathbf{d}}(\pi_r)$$

for all $r \ge \bar{r}$. It then follows from (3.21) that

$$|x_{\pi_r}(t; x_o, u_r, e_r)| \le \beta(|x_o|, t) + \gamma(|u|_\infty) + \varepsilon \tag{3.26}$$

for all $t \ge 0$ and $r \ge \bar{r}$, where β and γ are in (3.25). The ISS condition (2.6) now follows by passing to the limit in (3.26) as $r \to \infty$, since $\varepsilon > 0$ was arbitrary. This concludes the proof of Theorem 1.1.

4. Stabilization of the nonholonomic integrator. In this section, we illustrate how the feedback constructed in section 3 can be used to stabilize Brockett's nonholonomic integrator control system (see [3, 10, 22]). We will also use the nonholonomic integrator to compare our feedback construction to the feedbacks from [19, 20]. The nonholonomic integrator was introduced in [3], as an example of a system that cannot be stabilized using continuous feedback. It is well known that if the state space of a system contains obstacles (e.g., if the state space is $\mathbb{R}^2 \setminus (-1,1)^2$, and therefore has a topological obstacle around the origin), then it is impossible to stabilize the system using continuous feedback. In fact, this is a special case of a theorem of Milnor, which asserts that the domain of attraction of an asymptotically stable vector field must be diffeomorphic to Euclidean space, and therefore cannot be the complement $\mathbb{R}^2 \setminus (-1,1)^2$ (see [21]).

Brockett's example illustrates how, even if we assume that the state evolves in Euclidean space, similar obstructions to stabilization may occur. These obstructions are not due to the topology of the state space, but instead arise from "virtual obstacles" that are implicit in the form of the control system (see [22]). Such obstacles occur when it is impossible to move *instantly* in some directions, even though it is possible to move *eventually* in every direction ("nonholonomy"). This gives rise to Brockett's criterion (see [3]), which is a necessary condition for the existence of a continuous stabilizer, in terms of the vector fields that define the system (see [21, 22, 25]). The nonholonomic integrator does not satisfy Brockett's criterion, and therefore cannot be stabilized by continuous feedbacks.

The physical model for Brockett's example is as follows. Consider a three-wheeled shopping cart whose front wheel acts as a castor. The state variable is $(x_1, x_2, \theta)^T$, where $(x_1, x_2)^T$ is the midpoint of the rear axle of the cart, and θ is the cart's orientation. The front wheel is free to rotate, but there is a "nonslipping" constraint that $(\dot{x}_1, \dot{x}_2)^T$ must always be parallel to $(\cos(\theta), \sin(\theta))^T$. This gives the equations

$$\begin{aligned} \dot{x}_1 &= v_1 \cos(\theta), \\ \dot{x}_2 &= v_1 \sin(\theta), \\ \dot{\theta} &= v_2, \end{aligned} \tag{4.1}$$

where v_1 is a "drive" command and v_2 is a steering command. Using the feedback transformation

$$\begin{aligned} &z_1 := \theta,\ z_2 := x_1 \cos(\theta) + x_2 \sin(\theta),\ z_3 := x_1 \sin(\theta) - x_2 \cos(\theta), \\ &u_1 := v_2,\ u_2 := v_1 - v_2 z_3 \end{aligned}$$

followed by a second transformation brings the equations (4.1) into the form

$$\begin{aligned} \dot{x}_1 &= u_1, \\ \dot{x}_2 &= u_2, \\ \dot{x}_3 &= x_1 u_2 - x_2 u_1, \end{aligned} \tag{4.2}$$

which is called the nonholonomic integrator control system.

One can show (see [11]) that (4.2) is a GAC system. However, since Brockett's condition is not satisfied for (4.2), the system has no continuous stabilizer. While there does not exist a C^1 CLF for the system (4.2) (see [11]), it is now well known that every GAC system admits a continuous CLF (see [18]). In fact, it was shown in [10] that the nonholonomic integrator dynamics (4.2) has the nonsmooth CLF

$$V(x) = \max\left\{ \sqrt{x_1^2 + x_2^2}, |x_3| - \sqrt{x_1^2 + x_2^2} \right\}, \tag{4.3}$$

which is semiconcave outside the cone $x_3^2 = 4(x_1^2+x_2^2)$ (see [17] for a detailed discussion of some special properties of this CLF). For the special case of the dynamics (4.2) and CLF (4.3), the feedback $K = K_1 + K_2$ we constructed in section 3 is as follows.

To simplify notation, we use the radius $r(x) := \sqrt{x_1^2 + x_2^2}$. We also use the sets

$$\begin{array}{rcl} S_o & = & \{x \in \mathbb{R}^3 : \ x_3 \neq 0, \ r(x) = 0\}, \\ S_+ & = & \{x \in \mathbb{R}^3 : \ x_3^2 \geq 4r^2(x) > 0\}, \\ S_- & = & \{x \in \mathbb{R}^3 : \ x_3^2 < 4r^2(x)\}, \end{array}$$

which form a partition of $\mathbb{R}^3 \setminus \{0\}$. Notice that $V(x) = r(x)$ on S_-, and also that $V(x) = |x_3| - r(x)$ on $\mathbb{R}^3 \setminus S_-$. To find our selection $\zeta(x) \in \partial_L V(x)$, we first choose $\zeta(0) = 0$, and $\zeta(x) = (0, -1, \mathrm{sgn}\{x_3\})^T$ for all $x \in S_o$. Using the notation of (3.4), this gives

$$b(x) = \begin{cases} (\, -x_2 \mathrm{sgn}\{x_3\} - x_1/r(x), \ x_1 \mathrm{sgn}\{x_3\} - x_2/r(x)\,)^T, & x \in S_+, \\ (\, x_1/r(x), \ x_2/r(x)\,)^T, & x \in S_-, \\ (0, -1)^T, & x \in S_o, \end{cases} \tag{4.4}$$

and $b(0) = 0$. Notice that $1 \leq |b(x)|^2 \leq r^2(x) + 1$ for all $x \neq 0$. We also have

$$K_1(x) = \begin{cases} \mu_1(x)\,(\, -x_2 \mathrm{sgn}\{x_3\} - x_1/r(x), \ x_1 \mathrm{sgn}\{x_3\} - x_2/r(x)\,)^T, & x \in S_+, \\ -\,(x_1, x_2)^T, & x \in S_-, \\ (0, |x_3|)^T, & x \in S_o, \end{cases}$$

with $K_1(0) = 0$, where we have set

$$\mu_1(x) := \frac{r(x) - |x_3|}{r^2(x) + 1}.$$

In this case, we have taken

$$K_1(x) = -b(x)V(x)/|b(x)|^2$$

for $x \neq 0$, where $b(x)$ is defined in (4.4), and K_1 is continuous at the origin. On the other hand, our feedback K_2 from (3.4) becomes

$$K_2(x) = -\begin{cases} (\, \mu_2(x_1, -x_2, x), \ \mu_2(x_2, x_1, x)\,)^T, & x \in S_+, \\ r(x)\,(\, \mathrm{sgn}\{x_1\}, \ \mathrm{sgn}\{x_2\}\,)^T, & x \in S_-, \\ |x_3|\,(0, \ -1)^T, & x \in S_0, \end{cases}$$

with $K_2(0) = 0$, where we have set

$$\mu_2(a, b, x) \ := \ (|x_3| - r(x)) \,\mathrm{sgn}\{\, b\, r(x) \,\mathrm{sgn}\{x_3\} - a\,\}.$$

Since V is semiconcave on $\Omega := \mathbb{R}^3 \setminus \mathrm{bd}(S_-)$, the argument from section 3 applies to sampling solutions that satisfy the additional requirement that $\tilde{x}_\pi(s) \in \Omega$ for all $s \geq 0$. It follows from the proof of Theorem 1.2 that the nonholonomic integrator system (4.2) can be stabilized for both actuator errors and small observation errors (for this restricted set of sampling solutions), using the combined feedback $K = K_1 + K_2$.

Remark 4.1. In this example, we chose to work with the CLF (4.3) because it has been explicitly proven in [10] to be a CLF for the control system (4.2). The

example illustrates how to extend our results to more general CLFs that may not be semiconcave on $\mathbb{R}^3 \setminus \{0\}$. For such cases, the ISS estimates hold for those sampling solutions that remain in the domain of semiconcavity of the CLF. On the other hand, we let the reader prove that the nonholonomic integrator system also has the CLF

$$\tilde{V}(x) = \left(\sqrt{x_1^2 + x_2^2} - |x_3|\right)^2 + x_3^2,$$

which is semiconcave on $\mathbb{R}^3 \setminus \{0\}$ (as the sum of the smooth function $x_1^2 + x_2^2 + 2x_3^2$ and a semiconcave function). Therefore, if we use $\tilde{V}$ to form our feedbacks, instead of the CLF (4.3), then our theorems apply directly, without any state restrictions on the sampling solutions.

Remark 4.2. The results in [19] designed feedbacks that make C^o-stabilizable systems ISS with respect to actuator errors. For the case of C^o-stabilizable systems, a smooth (i.e., C^∞) Lyapunov function is known to exist (see [1]). In [19], the system was rendered ISS using the feedback

$$\hat{K}(x) := -L_G V(x) = -\nabla V(x) G(x), \tag{4.5}$$

where V is a smooth CLF for the dynamics (1.1). In that case, (4.5) is continuous at the origin. However, in the more general situation where the system is merely GAC, there may not exist a smooth Lyapunov function, so V must be taken to be nonsmooth. In this case, the use of the nonsmooth analogue

$$\tilde{K}(x) := -\zeta(x) G(x) \tag{4.6}$$

of (4.5) (where $\zeta(x) \in \partial_L V(x)$ for all $x \neq 0$) could give rise to a feedback that would not be continuous at the origin. For example, if we use the nonholonomic integrator (4.2) and the CLF (4.3), then $\tilde{K}$ takes the values

$$\tilde{K}\left((\varepsilon, \varepsilon, 0)^T\right) = -\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^T, \quad \tilde{K}\left((\varepsilon, \varepsilon, 3\sqrt{2}\varepsilon)^T\right) = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^T + \varepsilon(1, -1)^T,$$

so $\tilde{K}$ is discontinuous at the origin. On the other hand, our choice of K_2 is automatically continuous at the origin.

Remark 4.3. Under the additional hypothesis that (1.1) satisfies the small control property (see [21]), the system can be stabilized by a feedback that is continuous at the origin (see [17]). More precisely, suppose there exists a semiconcave CLF V satisfying the following: For each $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that $0 < |x| \leq \delta$ implies

$$\exists u_x \in \varepsilon \mathcal{B}_m \quad \text{s.t. } \forall \zeta \in \partial_P V(x), \ \ \langle \zeta, f(x) + G(x) u_x \rangle \leq -V(x).$$

Then the system can be rendered GAS by a feedback that is continuous at the origin (see [17]). For the case of the nonholonomic integrator (4.2), the system is GAS under the feedback K_1, which is continuous at the origin, so our total feedback $K = K_1 + K_2$ is continuous at the origin as well.

5. ISS for fully nonlinear GAC systems. We conclude with an extension of our results for fully nonlinear GAC systems

$$\dot{x} = f(x, u), \tag{5.1}$$

where we assume for simplicity that the observation error e in the controller is zero. We assume throughout this section that

$$f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n : (x, u) \mapsto f(x, u)$$

is continuous and locally Lipschitz in x uniformly on compact subsets of $\mathbb{R}^n \times \mathbb{R}^m$ and $f(0, 0) = 0$. It is natural to ask whether these hypotheses are sufficient for the existence of a continuous feedback $K(x)$ for which

$$\dot{x} = f(x, K(x) + u) \tag{5.2}$$

is ISS for Euler solutions. However, one can easily construct examples for which such feedbacks cannot exist. Here is an example from [20] where this situation occurs. Consider the GAC system $\dot{x} = -x + u^2 x^2$ on $\mathbb{R}$. If $K(x)$ is any continuous feedback for which

$$\dot{x} = -x + (K(x) + u)^2 x^2 \tag{5.3}$$

is ISS, then $|K(x)| < x^{-1/2}$ for all $x > 0$. It follows that the solution of

$$\dot{x} = -x + (K(x) + 1)^2 x^2$$

starting at $x(0) = 4$ is unbounded. Therefore, there does not exist a continuous feedback K for which (5.3) is ISS. On the other hand, one *can* find a (possibly discontinuous) feedback that makes (5.1) ISS. We use the following weaker sense of ISS for fully nonlinear systems that was introduced in [20].

DEFINITION 5.1. *We say that* (5.1) *is* input-to-state stabilizable in the weak sense *provided there exist a feedback K, and an $m \times m$ matrix G of continuously differentiable functions which is invertible at each point, such that*

$$\dot{x} = F(x, K(x), u)$$

is ISS for sampling and Euler solutions, where $F(x, p, u) = f(x, p + G(x)u)$.

We will prove the following.

PROPOSITION 5.2. *If* (5.1) *is GAC, then* (5.1) *is also input-to-state stabilizable in the weak sense.*

Proof. We modify the proof from section 3. We define V, ζ, $\underline{\alpha}$, $\overline{\alpha}$, and K_1 as in the proof of Theorem 1.2, except we use the fully nonlinear dynamics $h = f$ from (5.1). Next we follow the proof of the main result in [20], with the following modifications. Define the (possibly discontinuous) function $\mathcal{D}$ by

$$\mathcal{D}(s, r) = \sup \left\{ \langle \zeta(x), f(x, K_1(x) + p) \rangle + \frac{V(x)}{2} : |x| = s, |p| = r \right\}. \tag{5.4}$$

For any interval I of the form $[i, i+1]$, or of the form $[\frac{1}{i+1}, \frac{1}{i}]$, for $i \in \mathbb{N}$, one can find $r = r(i) > 0$ such that $s \in I$ implies $\mathcal{D}(s, b) < 0$ for all $b \in [0, r]$. This follows from the positive definiteness of V, the local Lipschitzness of f, and the local boundedness of $\partial_P V$ on compact subsets of $\mathbb{R}^n \setminus \{0\}$.

The argument of [20] therefore gives $\alpha_4 \in \mathcal{K}_\infty$ and a smooth, everywhere invertible matrix-valued function $G : \mathbb{R}^n \to \mathbb{R}^{m \times m}$ satisfying the following: If

$$|x| > \alpha_4(|u(s)|_\infty), \tag{5.5}$$

then for a.e. $t \geq 0$,

$$\langle \zeta(x), f(x, K_1(x) + G(x)u(t))\rangle + \frac{V(x)}{2} \ \leq \ \mathcal{D}(|x|, |G(x)u(s)|_\infty) < 0.$$

(See Remark 5.3 for a characterization of the set of matrices G for which ISS can be expected, in terms of $\mathcal{D}$.) We can evidently assume that $\alpha_4(s) \geq s$ for all $s \geq 0$ (e.g., by replacing $\alpha_4(s)$ by $\max\{\alpha_4(s), s\}$, which makes the condition (5.5) more restrictive). Fix M, N, $\varepsilon \in (0, M)$, $u \in \mathcal{M}_N^m$, and $x(t) = x_\pi(t)$ as before, with $e = 0$. Define the compact sets

$$S := \left\{x \in \mathbb{R}^n : V(x) \leq \underline{\alpha}^{-1} \circ \alpha_4(N)\right\}, \quad Q = \left\{(\overline{\alpha} \circ \underline{\alpha}^{-1}(M + \alpha_4(N)) + 1)\bar{\mathcal{B}}_n\right\} \setminus \varepsilon\mathcal{B}_n.$$

Notice that $S \subseteq Q^\varepsilon$. We choose $\tilde{\varepsilon}$ as before, and we choose $\delta = \delta(\varepsilon, M, N)$, satisfying (3.11), such that if $\overline{\mathbf{d}}(\pi) < \delta$, then

$$|x_\pi(t) - x_i| \ \leq \ \min\left\{\mu, \frac{\tilde{\varepsilon}}{16(1 + \mathcal{L}_\varepsilon)}, \sqrt{\frac{\lambda_-}{8\sigma}(t - t_i)}\right\} \tag{5.6}$$

for all indices i such that $x_i \in Q^\varepsilon$ and all $t \in [t_i, t_{i+1}]$, where σ and μ are as defined before, and $\lambda_- = \min\{V(x) : x \in Q^{\varepsilon/4}\}$. Reducing δ as necessary, we can assume

$$\|\zeta(x_i) \cdot [f(x_i, K_1(x_i) + G(x_i)u(s)) - f(x(s), K_1(x_i) + G(x(s))u(s))]\|_{[t_i, t_{i+1}]} \leq \frac{\lambda_-}{8}$$

for all indices i satisfying $x_i \in Q^{\varepsilon/2}$. Reasoning as in the earlier proof gives

$$V(x_\pi(t)) - V(x_i) \leq -(t - t_i)\frac{V(x_i)}{16} \quad \forall t \in [t_i, t_{i+1}]$$

for all i such that $x_i \in Q^{\varepsilon/4} \setminus S$. The remainder of the proof is as before, except with $\overline{\alpha} \circ \underline{\alpha}^{-1}(N)$ replaced by $\overline{\alpha} \circ \underline{\alpha}^{-1}(\alpha_4(N))$, and with $\overline{\alpha} \circ \underline{\alpha}^{-1}(s)$ replaced by $\overline{\alpha} \circ \underline{\alpha}^{-1}(\alpha_4(s))$ in the definition of γ. This proves Proposition 5.2. □

Remark 5.3. The statement of Proposition 5.2 is an existence result in terms of the invertible matrix G. However, we can strengthen the proposition by using the function $\mathcal{D}$ in (5.4) to characterize the class of G for which ISS can be expected, as follows. Following [20], we first choose strictly decreasing sequences $\{r_i\}$ and $\{r_i'\}$ of positive numbers such that $0 < r_{i+1} < r_i' < r_i$ for all $i \in \mathbb{N}$, and such that

$$\mathcal{D}(s, r) < 0 \quad \forall (s, r) \in ([i, i + 1] \times [0, r_i]) \cup ([1/(i + 1), 1/i] \times [0, r_i'])$$

for all $i \in \mathbb{N}$. The existence of these sequences follows from the argument we gave in the proof of the proposition. Define $\rho : [0, \infty) \to [0, \infty)$ by setting

($\rho 1$) $\rho(s) = r_k$ for all $s \in [k, k + 1)$ and $k \in \mathbb{N}$;
($\rho 2$) $\rho(s) = r_k'$ for all $s \in [1/(k + 1), 1/k)$ and $k \in \mathbb{N}$; and
($\rho 3$) $\rho(0) = 0$.

We then choose any smooth function $g : [0, \infty) \to (0, \infty)$ satisfying

($g1$) $g(s) = 1$ for all $s \in [0, 1]$;
($g2$) $g(s) \leq \rho(s)/s$ for all $s \geq 2$; and
($g3$) $g(s) \leq 1$ for all $s \geq 0$.

The existence of such a function g follows from exactly the same argument used in [20]. It then also follows from the argument of [20] that we can satisfy the conditions of the proposition by choosing $G(\xi) = g(|\xi|)I$.

Proposition 5.2 allows us to characterize GAC for fully nonlinear systems in terms of feedback equivalence, as follows. First recall that two systems $\dot{x} = f(x,u)$ and $\dot{x} = h(x,u)$, evolving on $\mathbb{R}^n \times \mathbb{R}^m$, are called *feedback equivalent* provided there exist a locally bounded function $K : \mathbb{R}^n \to \mathbb{R}^m$ and an everywhere invertible function $G : \mathbb{R}^n \to \mathbb{R}^{m\times m}$ for which

$$h(x,u) = f(x, K(x) + G(x)u)$$

for all $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$. In this case, we also say $\dot{x} = f(x,u)$ is feedback equivalent to (2.1) with $e = 0$ and $F(x,p,u) = f(x,p+G(x)u)$. The following elegant statement follows directly from Proposition 5.2.

COROLLARY 5.4. *The fully nonlinear system* (5.1) *is GAC if and only if it is feedback equivalent to a system which is ISS for sampling and Euler solutions.*

Remark 5.5. Although, as shown by the counterexample (5.3), it is in general impossible to obtain input-to-state stabilization (in the nonweak sense) for systems that are not affine in controls, it is still the case that for some restricted classes of systems this objective can be attained, under appropriate neutral-stability assumptions on the dynamics. One such class is that of systems in which the input appears inside a saturation nonlinearity, such as $\dot{x} = f(x,u) = f_0(x) + g(x)\sigma(u)$. The papers [14] and [5] (see [26] for an application of these results to the recursive design of stabilizers for a large class of systems with saturation) as well as [4] and [13] dealt with such questions, for systems that are linear in the absence of the saturation (the f_0 and g vector fields are linear and constant, respectively), while [2] obtained similar results for more general nonlinear systems.

REFERENCES

[1] Z. ARTSTEIN, *Stabilization with relaxed controls*, Nonlinear Anal., 7 (1983), pp. 1163–1173.

[2] X. BAO AND Z. LIN, *On L_p input to state stabilizability of affine in control, nonlinear systems subject to actuator saturation*, J. Franklin Inst., 337 (2000), pp. 691–712.

[3] R. BROCKETT, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. Brockett, R. Millman, and H. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 181–191.

[4] Y. CHITOUR, *On the L^p stabilization of the double integrator subject to input saturation*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 291–331.

[5] Y. CHITOUR, W. LIU, AND E. SONTAG, *On the continuity and incremental-gain properties of certain saturated linear feedback loops*, Internat. J. Robust Nonlinear Control, 5 (1995), pp. 413–440.

[6] F. CLARKE, YU. S. LEDYAEV, E. SONTAG, AND A.I. SUBBOTIN, *Asymptotic controllability implies feedback stabilization*, IEEE Trans. Automat. Control, 42 (1997), pp. 1394–1407.

[7] R. FREEMAN AND P. KOKOTOVIĆ, *Inverse optimality in robust stabilization*, SIAM J. Control Optim., 34 (1996), pp. 1365–1391.

[8] M. KRSTIĆ, I. KANELLAKOPOULOS, AND P. KOKOTOVIĆ, *Nonlinear and Adaptive Control Design*, Wiley, New York, 1995.

[9] M. KRSTIĆ AND Z. LI, *Inverse optimal design of input-to-state stabilizing nonlinear controllers*, IEEE Trans. Automat. Control, 43 (1998), pp. 336–350.

[10] Y. LEDYAEV AND L. RIFFORD, *Robust Stabilization of the Nonholonomic Integrator*, unpublished.

[11] Y. LEDYAEV AND E. SONTAG, *A Lyapunov characterization of robust stabilization*, Nonlinear Anal., 37 (1999), pp. 813–840.

[12] D. LIBERZON, E. SONTAG, AND Y. WANG, *Universal construction of feedback laws achieving ISS and integral-ISS disturbance attenuation*, Systems Control Lett., 46 (2002), pp. 111–127.

[13] Z. LIN, *H_∞-almost disturbance decoupling with internal stability for linear systems subject to input saturation*, IEEE Trans. Automat. Control, 42 (1997), pp. 992–995.

[14] W. LIU, Y. CHITOUR, AND E. SONTAG, *On finite-gain stabilizability of linear systems subject to input saturation*, SIAM J. Control Optim., 34 (1996), pp. 1190–1219.

[15] L. Praly and Y. Wang, *Stabilization in spite of matched unmodelled dynamics and an equivalent definition of input-to-state stability*, Math. Control Signals Systems, 9 (1996), pp. 1–33.
[16] L. Rifford, *Existence of Lipschitz and semiconcave control-Lyapunov functions*, SIAM J. Control Optim., 39 (2000), pp. 1043–1064.
[17] L. Rifford, *Semiconcave control-Lyapunov functions and stabilizing feedbacks*, SIAM J. Control Optim., 41 (2002), pp. 659–681.
[18] E. Sontag, *A Lyapunov-like characterization of asymptotic controllability*, SIAM J. Control Optim., 21 (1983), pp. 462–471.
[19] E. Sontag, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.
[20] E. Sontag, *Further facts about input to state stabilization*, IEEE Trans. Automat. Control, 35 (1990), pp. 473–476.
[21] E. Sontag, *Mathematical Control Theory, Deterministic Finite Dimensional Systems*, 2nd ed., Texts Appl. Math. 6, Springer-Verlag, New York, 1998.
[22] E. Sontag, *Stability and stabilization: Discontinuities and the effect of disturbances*, in Nonlinear Analysis, Differential Equations, and Control, F. Clarke and R. J. Stern, eds., Kluwer, Dordrecht, The Netherlands, 1999, pp. 551–598.
[23] E. Sontag, *The ISS philosophy as a unifying framework for stability-like behavior*, in Nonlinear Control in the Year 2000, Vol. 2, Lecture Notes in Control and Inform. Sci. 259, A. Isidori, F. Lamnabhi-Lagarrigue, and W. Respondek, eds., Springer-Verlag, London, 2000, pp. 443–468.
[24] E. Sontag and H. Sussmann, *Remarks on continuous feedback*, in Proceedings of the 19th Conference on Decision and Control, Albuquerque, NM, IEEE Control Systems Society, 1980, pp. 916–921.
[25] R. Stern, *Brockett's stabilization condition under state constraints*, Systems Control Lett., 47 (2002), pp. 335–341.
[26] H. Sussmann, E. D. Sontag, and Y. Yang, *A general result on the stabilization of linear systems using bounded controls*, IEEE Trans. Automat. Control, 39 (1994), pp. 2411–2425.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 2239–2263

SECOND ORDER SUFFICIENT CONDITIONS FOR TIME-OPTIMAL BANG-BANG CONTROL*

HELMUT MAURER† AND NIKOLAI P. OSMOLOVSKII‡

Abstract. We study second order sufficient optimality conditions (SSC) for optimal control problems with control appearing linearly. Specifically, time-optimal bang-bang controls will be investigated. In [N. P. Osmolovskii, *Sov. Phys. Dokl.*, 33 (1988), pp. 883–885; *Theory of Higher Order Conditions in Optimal Control*, Doctor of Sci. thesis, Moscow, 1988 (in Russian); *Russian J. Math. Phys.*, 2 (1995), pp. 487–516; *Russian J. Math. Phys.*, 5 (1997), pp. 373–388; *Proceedings of the Conference "Calculus of Variations and Optimal Control*," Chapman & Hall/CRC, Boca Raton, FL, 2000, pp. 198–216; A. A. Milyutin and N. P. Osmolovskii, *Calculus of Variations and Optimal Control*, Transl. Math. Monogr. 180, AMS, Providence, RI, 1998], SSC have been developed in terms of the positive definiteness of a quadratic form on a critical cone or subspace. No systematical numerical methods for verifying SSC are to be found in these papers. In the present paper, we study explicit representations of the critical subspace. This leads to an easily implementable test for SSC in the case of a bang-bang control with one or two switching points. In general, we show that the quadratic form can be simplified by a transformation that uses a solution to a linear matrix differential equation. Particular conditions even allow us to convert the quadratic form to perfect squares. Three numerical examples demonstrate the numerical viability of the proposed tests for SSC.

Key words. optimal bang-bang control, second order sufficient conditions, Q-transformation to perfect squares, numerical verification, applications

AMS subject classifications. 49K15, 49K30, 65L10, 94C99

DOI. 10.1137/S0363012902402578

1. Introduction.

Second order sufficient optimality conditions (SSC) for optimal control problems subject to mixed control-state constraints have been studied by various authors; cf. Dunn [8, 9]; Malanowski [22]; Maurer and Pickenhain [30]; Maurer and Oberle [29]; Milyutin and Osmolovskii [31]; Osmolovskii [35, 36, 37, 38, 39, 40]; and Zeidan [48]. SSC amount to testing the positive definiteness of a certain quadratic form on the so-called critical cone or subspace. Provided that the strict Legendre–Clebsch condition holds, a well-known numerical recipe allows the conversion of the quadratic form to a perfect square. Namely, it suffices to check that an associated Riccati matrix differential equation has a bounded solution along the extremal trajectory. This test has been performed in a number of practical examples and has played a crucial role in sensitivity analysis of parametric control problems; cf., e.g., Augustin, Malanowski, and Maurer [2, 21, 22, 23, 24, 25, 27, 28]. Recently, the Riccati approach has been also extended to discontinuous controls (broken extremals) by Osmolovskii and Lempio [42].

The above mentioned tests for SSC are not applicable to optimal control problems with control appearing *linearly*. Bang-bang controls do belong to this class of problems. Though first and higher order *necessary* optimality conditions for bang-bang controls have been studied, e.g., in Bressan [3], Schättler [44], and Sussmann

*Received by the editors February 13, 2002; accepted for publication (in revised form) October 15, 2003; published electronically May 17, 2004.

http://www.siam.org/journals/sicon/42-6/40257.html

†Westfälische Wilhelms-Universität Münster, Institut für Numerische Mathematik, Einsteinstrasse 62, D–48149 Münster, Germany (maurer@math.uni-muenster.de). This author was supported by the Deutsche Forschungsgemeinschaft under grant MA 691/8–3.

‡Department of Applied Mathematics, Moscow State Civil Engineering University (MISI), Jaroslavskoe sh. 26, 129337 Moscow, Russia (nikolai@osmolovskii.msk.ru). This author was supported by the Russian Foundation for Fundamental Research under grant 00-15-96109.

[45, 46, 47], there is no systematic study of *sufficient* optimality conditions and their numerical verification. A general set of second order necessary and sufficient conditions for an extremal with a discontinuous control (cf. Osmolovskii [37]) can be derived from the theory of higher order conditions in Levitin, Milyutin, and Osmolovskii [20]. The main results for bang-bang controls which follow from these general conditions are given in Milyutin and Osmolovskii [31]. Some proofs missing in that book will appear in Osmolovskii [40]. Only recently, other authors have derived SSC for general bang-bang control problems with *fixed* final time (cf. Agrachev, Stefani, and Zezza [1]; Ledzewicz and Schättler [19]; and Noble and Schättler [33]).

In this paper, we shall consider the special class of *time-optimal* bang-bang controls with given initial and terminal state. To our knowledge, the paper of Sarychev [43] seems to be the only study on SSC for this class of problems. However, it is not clear how one might apply the SSC in this article to practical examples. Thus our aim is to derive SSC in a form that is also suitable for practical verification. The two main tools to achieve this goal will be (1) a detailed study of the critical subspace and (2) an adaptation of the above mentioned Riccati approach to bang-bang controls. The organization of the paper then is as follows. In section 2, Pontryagin's minimum principle and the bang-bang property are discussed. The accessory problem, respectively, the quadratic form and the critical subspace are introduced in section 3. SSC are given in a general form that is evaluated particularly for bang-bang controls with one or two switching points. Section 4 presents the Q-transformation whereby the quadratic form is simplified with the help of the solution Q of a linear differential equation. Positive definiteness conditions are given under which the quadratic form can be transformed into perfect squares. In section 5, we shall discuss three numerical examples that illustrate several numerical procedures for verifying positive definiteness of the corresponding quadratic forms.

2. Time-optimal bang-bang control problems.

2.1. Statement of the problem, strong minimum. We consider time-optimal control problems with control appearing linearly. Let $x(t) \in \mathbb{R}^{d(x)}$ denote the state variable and $u(t) \in \mathbb{R}^{d(u)}$ the control variable in the time interval $t \in \Delta = [0, T]$ with a nonfixed final time $T > 0$. For simplicity, the initial and terminal states are fixed in the following control problem:

$$\text{Minimize the final time} \quad T \tag{2.1}$$

subject to the constraints on the interval $\Delta = [0, T]$,

$$dx/dt = \dot{x} = f(t, x, u) = a(t, x) + B(t, x)u, \tag{2.2}$$

$$x(0) = x_0, \quad x(T) = x_1, \tag{2.3}$$

$$u(t) \in U, \quad (t, x(t)) \in \mathcal{Q}\,. \tag{2.4}$$

Here, x_0, x_1 are given points in $\mathbb{R}^{d(x)}$, $\mathcal{Q} \subset \mathbb{R}^{1+d(x)}$ is an open set, and $U \subset \mathbb{R}^{d(u)}$ is a convex polyhedron. The functions a, B are twice continuously differentiable on $\mathcal{Q}$ with B being a $d(x) \times d(u)$ matrix function. A trajectory or control process

$$\mathcal{T} = \{\, (x(t), u(t)) \mid t \in [0, T] \,\}$$

is said to be *admissible* if $x(\cdot)$ is absolutely continuous, $u(\cdot)$ is measurable and essentially bounded, and the pair of functions $(x(t), u(t))$ satisfies the constraints (2.2)–(2.4)

on the interval $\Delta = [0, T]$. The component $x(t)$ will be called the *state trajectory.*

DEFINITION 2.1. *An admissible trajectory* $\mathcal{T}^0 = \{(x^0(t), u^0(t)) \mid t \in [0, T^0]\}$ *is said to be strongly (resp., strictly strongly) locally time-optimal if there exists* $\varepsilon > 0$ *such that* $T \geq T^0$ *(resp.,* $T > T^0$*) holds for all admissible* $\mathcal{T} = \{(x(t), u(t)) \mid t \in [0, T]\}$ *(resp., different from* $\mathcal{T}^0$*) with* $|T - T^0| < \varepsilon$ *and* $\max_{[0,T^0]\cap[0,T]} |x(t) - x^0(t)| < \varepsilon$.

2.2. Minimum principle. Let

$$\mathcal{T} = \{\, (x(t), u(t)) \mid t \in [0, T] \,\}$$

be a fixed admissible trajectory such that the control $u(\cdot)$ is a piecewise constant function on the interval $\Delta = [0, T]$ with *finitely many* points of discontinuity. In order to simplify notation we shall not use such symbols and indices as zero, hat, or asterisk to distinguish this trajectory from others. Denote by

$$\theta = \{t_1, \ldots, t_s\}, \quad 0 < t_1 < \cdots < t_s < T\,,$$

the finite set of all discontinuity points (jump points) of the control $u(t)$. Then $\dot{x}(t)$ is a piecewise continuous function whose discontinuity points belong to the set θ and, thus, $x(t)$ is a piecewise smooth function on Δ. Henceforth, we shall use the notation

$$[u]^k = u^{k+} - u^{k-}$$

to denote the jump of the function $u(t)$ at the point $t_k \in \theta$, where

$$u^{k-} = u(t_k - 0), \quad u^{k+} = u(t_k + 0)$$

are, respectively, the left-hand and the right-hand values of the control $u(t)$ at t_k. Similarly, we denote by $[\dot{x}]^k$ the jump of the function $\dot{x}(t)$ at the same point.

Let us formulate the first order necessary conditions of optimality for the trajectory $\mathcal{T}$, the *Pontryagin minimum principle.* To this end we introduce the *Pontryagin function* or Hamiltonian function

$$H(t, x, u, \psi) = \psi f(t, x, u) = \psi a(t, x) + \psi B(t, x) u, \tag{2.5}$$

where ψ is a row-vector of dimension $d(x)$, while x, u, f are column-vectors. In what follows, partial derivatives of the Pontryagin function and all other functions will be denoted by subscripts referring to the respective variables.

The factor of the control u in the Pontryagin function is called the *switching function*

$$\sigma(t, x, \psi) = \psi B(t, x).$$

Consider the pair of functions

$$\psi_0(\cdot) : \Delta \to \mathbb{R}^1, \quad \psi(\cdot) : \Delta \to \mathbb{R}^{d(x)},$$

which are continuous on Δ and continuously differentiable on each interval of the set $\Delta \setminus \theta$. Denote by M_0 the set of normed pairs of functions $(\psi_0(\cdot), \psi(\cdot))$ satisfying the conditions

$$\psi_0(T) \geq 0, \quad |\psi(0)| = 1, \tag{2.6}$$

$$\dot{\psi}(t) = -H_x(t, x(t), u(t), \psi(t)) \quad \forall\, t \in \Delta \setminus \theta, \tag{2.7}$$

$$\dot{\psi}_0(t) = -H_t(t, x(t), u(t), \psi(t)) \quad \forall\, t \in \Delta \setminus \theta, \tag{2.8}$$

$$\min_{u \in U} H(t, x(t), u, \psi(t)) = H(t, x(t), u(t), \psi(t)) \quad \forall\, t \in \Delta \setminus \theta, \tag{2.9}$$

$$H(t, x(t), u(t), \psi(t)) + \psi_0(t) = 0 \quad \forall\, t \in \Delta \setminus \theta\,. \tag{2.10}$$

Then the condition $M_0 \neq \emptyset$ is equivalent to the Pontryagin minimum principle. We assume that this condition is satisfied for the trajectory $\mathcal{T}$. We say in this case that $\mathcal{T}$ is an *extremal trajectory* for the problem. M_0 is a finite-dimensional *compact* set since in (2.6) the initial values $\psi(0)$ are assumed to belong to the unit ball of $\mathbb{R}^{d(x)}$. The case that there exists a multiplier $(\psi_0, \psi) \in M_0$ with $\psi_0(T) > 0$ will be called the *nondegenerate* or *normal* case.

Henceforth, it will be convenient to use the simple abbreviation (t) for all arguments $(t, x(t), u(t), \psi(t))$, e.g., $H(t) = H(t, x(t), u(t), \psi(t))$, $\sigma(t) = \sigma(t, x(t), \psi(t))$. The continuity of the pair of functions $(\psi_0(t), \psi(t))$ at the points $t_k \in \theta$ constitutes the Weierstrass–Erdmann necessary conditions for nonsmooth extremals. We formulate one more important condition of this type. Namely, for $(\psi_0, \psi) \in M_0$ and $t_k \in \theta$ consider the function

$$(\Delta_k H)(t) = H(t, x(t), u^{k+}, \psi(t)) - H(t, x(t), u^{k-}, \psi(t)) = \sigma(t, x(t), \psi(t))[u]^k.$$

This function has a derivative

$$D^k(H) := -\frac{d}{dt}(\Delta_k H)(t_k) = -\dot\sigma(t_k^{\pm})[u]^k,$$

where the values on the right-hand side are the same for the derivative $\dot\sigma(t_k^+)$ from the right and the derivative $\dot\sigma(t_k^-)$ from the left. In the case of a *scalar* control u, the total derivative $\sigma_t + \sigma_x \dot x + \sigma_\psi \dot\psi$ does not contain the control variable explicitly [17, 18] and, hence, the derivative of the switching function $\dot\sigma(t)$ is *continuous* at t_k. Then the minimum condition (2.9) immediately implies the following property.

PROPOSITION 2.2. *For each $(\psi_0, \psi) \in M_0$ the following conditions hold:*

$$D^k(H) = -\dot\sigma(t_k^{\pm})[u]^k \geq 0 \quad \textit{for } k = 1, \dots, s. \tag{2.11}$$

2.3. Bang-bang control. The classical definition of a bang-bang control is that of a control which assumes values in the vertex set of the admissible polyhedron U in (2.4). We need a slightly more restrictive definition of a bang-bang control to obtain the sufficient conditions in Theorem 3.3. Let

$$\operatorname{Arg\,min}_{v \in U} \sigma(t)v$$

be the set of points $v \in U$ where the minimum of the linear function $\sigma(t)v$ is attained. For a given extremal trajectory $\mathcal{T} = \{\, (x(t), u(t)) \mid t \in \Delta \,\}$ with piecewise constant control $u(t)$ we shall say that $u(t)$ is a *bang-bang control* if there exists $(\psi_0, \psi) \in M_0$ such that

$$\operatorname{Arg\,min}_{v \in U} \sigma(t)v = [u(t-0), u(t+0)]\,, \tag{2.12}$$

where $[u(t-0), u(t+0)] = \{\alpha u(t-0) + (1-\alpha)u(t+0) \mid 0 \leq \alpha \leq 1\}$ denotes the line segment in $\mathbb{R}^{d(u)}$. Notice that $[u(t-0), u(t+0)]$ is a singleton $\{u(t)\}$ at each continuity point of the control $u(t)$ with $u(t)$ being a vertex of the polyhedron U. Only at the points $t_k \in \theta$ does the line segment $[u^{k-}, u^{k+}]$ coincide with an edge of the polyhedron.

If the control is *scalar*, $d(u) = 1$ and $U = [u_{min}, u_{max}]$, then the bang-bang property is equivalent to

$$\sigma(t, x(t), \psi(t)) \neq 0 \quad \forall\, t \in \Delta \setminus \theta,$$

which implies the following control law:

$$u(t) = \left\{ \begin{array}{ll} u_{min} & \text{if} \quad \sigma(t) > 0 \\ u_{max} & \text{if} \quad \sigma(t) < 0 \end{array} \right\} \quad \forall\, t \in \Delta \setminus \theta. \tag{2.13}$$

For vector-valued control inputs, condition (2.12) imposes further restrictions. For example, if U is the unit cube in $\mathbb{R}^{d(u)}$, condition (2.12) precludes simultaneous switching of the control components. However, this property holds for most examples; cf., e.g., the time-optimal control of a robot manipulator with $d(u) = 2$ in Chernousko, Akulenko, and Bolotnik [6]. Moreover, condition (2.12) will be indispensable in the sensitivity analysis of optimal bang-bang controls, a topic that we are currently investigating.

3. Critical subspace, quadratic form, and sufficient optimality conditions for bang-bang controls. In order to formulate quadratic sufficient optimality conditions for a given extremal $\mathcal{T}$ with bang-bang control $u(\cdot)$ we shall introduce the space $\mathcal{Z}(\theta)$, the critical subspace $\mathcal{K} \subset \mathcal{Z}(\theta)$, and the quadratic form Ω defined in $\mathcal{Z}(\theta)$.

3.1. Critical subspace. Denote by $P_\theta C^1(\Delta, R^n)$ the space of piecewise continuous functions

$$\bar{x}(\cdot) : \Delta \to \mathbb{R}^n$$

that are continuously differentiable on each interval of the set $\Delta \setminus \theta$. For each $\bar{x} \in P_\theta C^1(\Delta, R^n)$ and for $t_k \in \theta$ we use the abbreviation

$$[\bar{x}]^k = \bar{x}^{k+} - \bar{x}^{k-}, \quad \text{where} \quad \bar{x}^{k-} = \bar{x}(t_k - 0), \quad \bar{x}^{k+} = \bar{x}(t_k + 0).$$

Putting

$$\bar{z} = (\bar{T}, \xi, \bar{x}) \quad \text{with} \quad \bar{T} \in R^1, \quad \xi = (\xi_1, \dots, \xi_s) \in R^s, \quad \bar{x} \in P_\theta C^1(\Delta, R^n),$$

we have

$$\bar{z} \in \mathcal{Z}(\theta) := \mathbb{R}^1 \times \mathbb{R}^s \times P_\theta C^1(\Delta, \mathbb{R}^n).$$

Denote by $\mathcal{K}$ the set of all $\bar{z} \in \mathcal{Z}(\theta)$ satisfying the following conditions:

$$\dot{\bar{x}}(t) = f_x(t, x(t), u(t))\bar{x}(t), \qquad [\bar{x}]^k = [\dot{x}]^k \xi_k, \quad k = 1, \dots, s, \tag{3.1}$$

$$\bar{x}(0) = 0, \quad \bar{x}(T) + \dot{x}(T)\bar{T} = 0\,. \tag{3.2}$$

Then $\mathcal{K}$ is a subspace of the space $Z(\theta)$ which we call the *critical subspace*. Each element $\bar{z} \in \mathcal{K}$ is uniquely defined by the number $\bar{T}$ and the vector ξ. Consequently, the subspace $\mathcal{K}$ is *finite-dimensional*.

An explicit representation of the variations $\bar{x}(t)$ in (3.1) is obtained as follows. For each $k = 1, \dots, s$, define the vector functions $y^k(t)$ as the solutions to the system

$$\dot{y} = f_x(t)y, \quad y(t_k) = [\dot{x}]^k, \quad t \in [t_k, T]. \tag{3.3}$$

For $t < t_k$ we put $y^k(t) = 0$ which yields the jump $[y^k]^k = [\dot{x}]^k$. It follows from the superposition principle for linear ODEs that

$$\bar{x}(t) = \sum_{k=1}^{s} y^k(t)\xi_k \tag{3.4}$$

from which we obtain the representation

$$\bar{x}(T) + \dot{x}(T)\bar{T} = \sum_{k=1}^{s} y^k(T)\xi_k \; + \dot{x}(T)\bar{T}\,. \tag{3.5}$$

Furthermore, denote by $x(t; t_1, \ldots, t_s)$ the solution of the state equation (2.2) using the optimal bang-bang control with switching points $t_1, \ldots, t_s$. It easily follows from elementary properties of ODEs that the partial derivatives of state trajectories w.r.t. the switching points are given by

$$\frac{\partial x}{\partial t_k}(t; t_1, \ldots, t_s) = -y^k(t) \quad \text{for } \; t \geq t_k, \;\; k = 1, \ldots, s. \tag{3.6}$$

This relation holds for all $t \in [0, T] \setminus \{t_k\}$, because for $t < t_k$ we have $\frac{\partial x}{\partial t_k}(t) = 0$ and $y^k(t) = 0$. Hence, (3.4) yields

$$\bar{x}(t) = -\sum_{k=1}^{s} \frac{\partial x}{\partial t_k}(t)\xi_k \; . \tag{3.7}$$

In the nondegenerate case $\psi_0(T) > 0$, the critical subspace simplifies as follows.

PROPOSITION 3.1. *If there exists $(\psi_0, \psi) \in M_0$ such that $\psi_0(T) > 0$, then $\bar{T} = 0$ holds for each $\bar{z} = (\bar{T}, \xi, \bar{x}) \in \mathcal{K}$.*

Proof. For arbitrary $(\psi_0, \psi) \in M_0$ and $\bar{z} = (\bar{T}, \xi, \bar{x}) \in \mathcal{K}$ we have

$$\frac{d}{dt}(\psi\bar{x}) = \dot{\psi}\bar{x} + \psi\dot{\bar{x}} = -\psi f_x(t)\bar{x} + \psi f_x(t)\bar{x} = 0,$$

and also

$$[\psi\bar{x}]^k = \psi(t_k)[\bar{x}]^k = \psi(t_k)[\dot{x}]^k\xi_k = [\psi\dot{x}]^k\xi_k = -[\psi_0]^k\xi_k = 0.$$

Consequently, $\psi(t)\bar{x}(t)$ is a constant function on $[0, T]$ which yields in view of (3.2)

$$0 = (\psi\bar{x})(0) = (\psi\bar{x})(T) = -\psi(T)\dot{x}(T)\bar{T} = \psi_0(T)\bar{T}.$$

Hence the inequality $\psi_0(T) > 0$ implies that $\bar{T} = 0$.

In section 3.2, we shall conclude from Theorem 3.3 that the property $\mathcal{K} = \{0\}$ essentially represents a first order sufficient condition. Since $\bar{x}(T) + \dot{x}(T)\bar{T} = 0$ by (3.2), the representations (3.4), (3.5), and Proposition 3.1 induce the following conditions for $\mathcal{K} = \{0\}$.

PROPOSITION 3.2. *Assume that one of the following conditions is satisfied:*

(a) *the $s+1$ vectors $y^k(T) = -\frac{\partial x}{\partial t_k}(T)$, $k = 1, \ldots, s$, $\dot{x}(T)$, are linearly independent,*
(b) *there exists $(\psi_0, \psi) \in M_0$ with $\psi_0(T) > 0$, and the s vectors $y^k(T) = -\frac{\partial x}{\partial t_k}(T)$, $k = 1, \ldots, s$, are linearly independent,*
(c) *there exists $(\psi_0, \psi) \in M_0$ with $\psi_0(T) > 0$, and the bang-bang control has one switching point, i.e., $s = 1$.*

Then the critical subspace is $\mathcal{K} = \{0\}$.

Now we discuss the case of two switching points, i.e., $s = 2$, to prepare the numerical example in section 5.2. Let us assume that $\psi_0(T) > 0$ and $[\dot{x}]^1 \neq 0$, $[\dot{x}]^2 \neq 0$.

By virtue of Proposition 3.1, we have $\bar{T} = 0$ and hence $\bar{x}(T) = 0$ for each element $\bar{z} \in \mathcal{K}$. Then the relations (3.2) and (3.4) yield

$$0 = \bar{x}(T) = y^1(T)\xi_1 + y^2(T)\xi_2\,. \tag{3.8}$$

The conditions $[\dot{x}]^1 \neq 0$ and $[\dot{x}]^2 \neq 0$ imply that $y^1(T) \neq 0$ and $y^2(T) \neq 0$, respectively. Futhermore, assume that $\mathcal{K} \neq \{0\}$. Then (3.8) shows that the nonzero vectors $y^1(T)$ and $y^2(T)$ are collinear, i.e.,

$$y^2(T) = \alpha y^1(T) \tag{3.9}$$

with some factor $\alpha \neq 0$. As a consequence, the relation $y^2(t) = \alpha y^1(t)$ is valid for all $t \in (t_2, T]$ since the functions $y^1(t)$ and $y^2(t)$ are continuous solutions to the system $\dot{y} = f_x(t)y$ in $(t_2, T]$. In particular, we have $y^2(t_2 + 0) = \alpha y^1(t_2)$ and thus

$$[\dot{x}]^2 = \alpha y^1(t_2) \tag{3.10}$$

which is equivalent to (3.9). In addition, the equalities (3.8) and (3.9) imply that

$$\xi_2 = -\frac{1}{\alpha}\xi_1. \tag{3.11}$$

We shall use these formulas in the next subsection.

3.2. Quadratic form. In the sequel, second order partial derivatives will be denoted by double subscripts, e.g., $H_{xx} = D_x^2 H$. For $(\psi_0, \psi) \in M_0$ and $\bar{z} \in \mathcal{K}$ we define the functional

$$\begin{aligned}\Omega(\psi_0, \psi, \bar{z}) = {} & \sum_{k=1}^{s}(D^k(H)\xi_k^2 + 2[H_x]^k \bar{x}^k_{\mathrm{av}}\xi_k) + \int_0^T \langle H_{xx}(t)\bar{x}(t), \bar{x}(t)\rangle\, dt \\ & -(\dot{\psi}_0(T) - \dot{\psi}(T)\dot{x}(T))\bar{T}^2\,,\end{aligned} \tag{3.12}$$

where

$$\bar{x}^k_{\mathrm{av}} := \frac{1}{2}(\bar{x}^{k-} + \bar{x}^{k+}).$$

Note that the functional $\Omega(\psi_0, \psi, \bar{z})$ is linear in ψ_0 and ψ and quadratic in $\bar{z}$.

Now we introduce SSC for a bang-bang control which have been obtained by Osmolovskii; see [31, Part 2, chapter 3, section 12.4]. Some proofs missing in this book will appear in Osmolovskii [40].

THEOREM 3.3. *Let the following Condition $\mathcal{B}$ be fulfilled for the trajectory $\mathcal{T}$:*

(a) *$u(t)$ is a bang-bang control such that (2.12) holds;*

(b) *there exists $(\psi_0, \psi) \in M_0$ such that $D^k(H) > 0$ for $k = 1, \ldots, s$;*

(c) $\max_{(\psi_0,\psi)\in M_0} \Omega(\psi_0, \psi, \bar{z}) > 0 \quad \forall\, \bar{z} \in \mathcal{K} \setminus \{0\}$.

Then $\mathcal{T}$ is a strict strong minimum.

Remarks.

1. In this theorem, the sufficient Condition $\mathcal{B}$ is a natural strengthening of the corresponding necessary quadratic condition in the same problem; see [31, Part 2].
2. Condition (c) is automatically fulfilled if $\mathcal{K} = \{0\}$ holds (cf. Proposition 3.2), which gives a first order sufficient condition for a strong minimum.

3. If there exists $(\psi_0, \psi) \in M_0$ such that

$$\Omega(\psi_0, \psi, \bar{z}) > 0 \quad \forall \bar{z} \in \mathcal{K} \setminus \{0\},$$

then condition (c) is obviously fulfilled.

For boxes $U = \{u = (u_1, \dots, u_{d(u)}) \in \mathbb{R}^{d(u)} \,|\, u_i^{\min} \le u_i \le u_i^{\max},\ i = 1, \dots, d(u)\}$, condition (b) is equivalent to the property $\dot{\sigma}_i(t_k) \neq 0$ if t_k is a switching point of the ith control component $u_i(t)$. Note again that condition (2.12) precludes the simultaneous switching of two or more control components. A further remark concerns the case that the set M_0 of Pontryagin multipliers is not a singleton. This case has been illustrated in Osmolovskii [38, pp. 377–380] by the following time-optimal control problem for a linear system:

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = x_3, \quad \dot{x}_3 = x_4, \quad \dot{x}_4 = u, \quad |u| \le 1, \quad x(0) = a, \quad x(T) = b,$$

where $x = (x_1, x_2, x_3, x_4)$. It was shown in this paper that for some a and b there exists an extremal in this problem with two switching points of the control such that, under an appropriate normalization, the set M_0 is a segment. For this extremal, the maximum of the quadratic forms Ω over M_0 is positive on each nonzero element of the critical subspace and hence the sufficient conditions of Theorem 3.3 are satisfied.

3.3. Nondegenerate case. Let us assume the *nondegenerate* or *normal* case that there exists $(\psi_0, \psi) \in M_0$ such that the cost function multiplier $\psi_0(T)$ is positive. By virtue of Proposition 3.1 we have in this case that $\bar{T} = 0$ for all $\bar{z} \in \mathcal{K}$. Thus the critical subspace $\mathcal{K}$ is defined by the conditions

$$\dot{\bar{x}} = f_x(t)\bar{x}, \quad [\bar{x}]^k = [\dot{x}]^k \xi_k \quad (k = 1, \dots, s), \quad \bar{x}(0) = 0, \quad \bar{x}(T) = 0. \tag{3.13}$$

In particular, these conditions imply $\bar{x}(t) \equiv 0$ on $[0, t_1)$ and $(t_s, T]$. Hence, we have $\bar{x}^{1-} = \bar{x}^{s+} = 0$ for all $\bar{z} \in \mathcal{K}$. Then the quadratic form (3.12) is equal to

$$\Omega(\psi, \bar{z}) = \sum_{k=1}^{s} (D^k(H)\xi_k^2 + 2[H_x]^k \bar{x}^k_{\mathrm{av}} \xi_k) + \int_{t_1}^{t_s} \langle H_{xx}(t)\bar{x}(t), \bar{x}(t)\rangle \, dt. \tag{3.14}$$

Just this case of a time-optimal (autonomous) control problem was studied by Sarychev [43]. He used a special transformation of the problem and obtained sufficient optimality condition for the transformed problem. It is not easy but possible to reformulate his results in terms of the original problem. The comparison of both types of conditions reveals that Sarychev used the same critical subspace, but his quadratic form is a lower bound for Ω. Namely, in his quadratic form the positive term $D^k(H)\xi_k^2$ has the factor $\frac{1}{4}$ instead of the factor 1 for the same term in Ω. Therefore, the sufficient Condition $\mathcal{B}$ is always fulfilled whenever Sarychev's condition is fulfilled. However, Osmolovskii has constructed an example of a control problem where the optimal solution satisfies Condition $\mathcal{B}$, but does not satisfy Sarychev's condition. Finally, Sarychev proved that his condition is sufficient for an L_1-minimum w.r.t. the control (which is a "Pontryagin minimum" [31] in this problem). In fact it could be proved that his condition is sufficient for a strong minimum.

3.4. Cases of one or two switching points of the control. From Theorem 3.3 and Proposition 3.2(c) we immediately deduce sufficient conditions for a bang-bang control with one switching point. The result will be used for the example in section 5.1 and is also applicable to the time-optimal control of an image converter discussed in Kim et al. [15].

THEOREM 3.4. *Let the following conditions be fulfilled for the trajectory* $\mathcal{T}$*:*

(a) *$u(t)$ is a bang-bang control with one switching point;*

(b) *there exists $(\psi_0, \psi) \in M_0$ such that $\psi_0(T) > 0$ and $D^1(H) > 0$.*

Then $\mathcal{T}$ is a strict strong minimum.

Now we turn our attention to the case of two switching points where $s = 2$. Assume the nondegenerate case $\psi_0(T) > 0$ and suppose that $[\dot{x}]^1 \neq 0$, $[\dot{x}]^2 \neq 0$ and $y^2(T) = \alpha y^1(T)$ as in (3.9). Otherwise, $\mathcal{K} = \{0\}$ holds and, hence, the first order sufficient condition for a strong minimum is satisfied. For any element $\bar{z} \in \mathcal{K}$ we have $\bar{T} = 0$, $\bar{x}^{1-} = 0$, $\bar{x}^{2+} = 0$. Consequently,

$$\bar{x}^1_{\rm av} = \frac{1}{2}[\bar{x}]^1 = \frac{1}{2}[\dot{x}]^1\xi_1, \quad \bar{x}^2_{\rm av} = \frac{1}{2}\bar{x}^{2-} = \frac{1}{2}y^1(t_2)\xi_1 = \frac{1}{2\alpha}[\dot{x}]^2\xi_1$$

in view of $\bar{x}(t) = y^1(t)\xi_1 + y^2(t)\xi_2$, $y^2(t_2 - 0) = 0$ and (3.10). Using these relations in the quadratic form (3.14) together with (3.11) and the conditions $y^2(t) = 0$ for all $t < t_2$, $[H_x]^k = -[\dot{\psi}]^k$, $k = 1, 2$, we compute the quadratic form for an element of the critical subspace as

$$\begin{aligned}
\Omega &= D^1(H)\xi_1^2 + D^2(H)\xi_2^2 - 2[\dot{\psi}]^1\bar{x}^1_{\rm av}\xi_1 - 2[\dot{\psi}]^2\bar{x}^2_{\rm av}\xi_2 + \int_{t_1}^{t_2}\langle H_{xx}\bar{x}, \bar{x}\rangle\, dt \\
&= D^1(H)\xi_1^2 + \tfrac{1}{\alpha^2}D^2(H)\xi_1^2 - [\dot{\psi}]^1[\dot{x}]^1\xi_1^2 + \tfrac{1}{\alpha^2}[\dot{\psi}]^2[\dot{x}]^2\xi_1^2 + \left(\int_{t_1}^{t_2}\langle H_{xx}y^1, y^1\rangle\, dt\right)\xi_1^2 \\
&= \rho\xi_1^2,
\end{aligned}$$

where

$$\rho := \left(D^1(H) - [\dot{\psi}]^1[\dot{x}]^1\right) + \frac{1}{\alpha^2}\left(D^2(H) + [\dot{\psi}]^2[\dot{x}]^2\right) + \int_{t_1}^{t_2}\langle H_{xx}y^1, y^1\rangle\, dt. \tag{3.15}$$

Thus, we obtain the following proposition.

PROPOSITION 3.5. *Assume that $\psi_0(T) > 0$, $s = 2$, $[\dot{x}]^1 \neq 0$, $[\dot{x}]^2 \neq 0$, and $y^2(T) = \alpha y^1(T)$ (which is equivalent to* (3.10)*) with some factor α. Then the condition of the positive definiteness of Ω on $\mathcal{K}$ is equivalent to the inequality $\rho > 0$, where ρ is defined by* (3.15).

4. Sufficient conditions for positive definiteness of the quadratic form Ω on the critical subspace $\mathcal{K}$. In this section we consider the nondegenerate case in section 3.3 and assume

(i) $u(t)$ is a bang-bang control with $s > 1$ switching points;

(ii) there exists $(\psi_0, \psi) \in M_0$ such that $\psi_0(T) > 0$ and $D^k(H) > 0$, $k = 1, \ldots, s$.

Under these assumptions the critical subspace $\mathcal{K}$ is defined by (3.13). Let $(\psi_0, \psi) \in M_0$ be a fixed element (possibly, different from that in assumption (ii)) and denote by $\Omega = \Omega(\psi_0, \psi, \cdot)$ the quadratic form for this element. Recall that Ω is given by (3.14). According to Theorem 3.3 *the positive definiteness of the quadratic form* (3.14) *on the subspace $\mathcal{K}$ in* (3.13) *is a sufficient condition for a strict strong minimum of the trajectory.* Now our aim is to find conditions that guarantee this property of positive definiteness. In what follows we shall use some ideas and results presented in Osmolovskii and Lempio [42], who have extended the Riccati approach in [4, 30, 22, 48] to broken extremals.

4.1. Q-transformation of Ω on $\mathcal{K}$. Let $Q(t)$ be a symmetric matrix on $[t_1, t_s]$ with piecewise continuous entries which are absolutely continuous on each interval of the set $[t_1, t_s] \setminus \theta$. Therefore, Q may have a jump at each point $t_k \in \theta$ including t_1, t_s, and thus the symmetric matrices Q^{1-} and Q^{s+} are also defined.

For $\bar{z} \in \mathcal{K}$ we obviously have

$$\int_{t_1}^{t_s} \frac{d}{dt}\langle Q\bar{x}, \bar{x}\rangle \, dt = \langle Q\bar{x}, \bar{x}\rangle \Big|_{t_1-0}^{t_s+0} - \sum_{k=1}^{s} \left[\langle Q\bar{x}, \bar{x}\rangle\right]^k,$$

where $\left[\langle Q\bar{x}, \bar{x}\rangle\right]^k$ is the jump of the function $\langle Q\bar{x}, \bar{x}\rangle$ at the point $t_k \in \theta$. Using the conditions

$$\dot{\bar{x}} = f_x(t)\bar{x}, \quad \bar{x}^{1-} = \bar{x}^{s+} = 0,$$

we obtain

$$\sum_{k=1}^{s} \left[\langle Q\bar{x}, \bar{x}\rangle\right]^k + \int_{t_1}^{t_s} \langle (\dot{Q} + f_x^* Q + Q f_x)\bar{x}, \bar{x}\rangle \, dt = 0, \tag{4.1}$$

where the asterisk denotes transposition. Adding this zero-form to Ω we get

$$\Omega = \sum_{k=1}^{s} \Big(D^k(H)\xi_k^2 - 2[\dot{\psi}]^k \bar{x}_{\mathrm{av}}^k \xi_k + \left[\langle Q\bar{x}, \bar{x}\rangle\right]^k \Big) + \int_{t_1}^{t_s} \langle (H_{xx} + \dot{Q} + f_x^* Q + Q f_x)\bar{x}, \bar{x}\rangle \, dt. \tag{4.2}$$

We shall call this formula the *Q-transformation of* Ω *on* $\mathcal{K}$.

In order to eliminate the integral term in Ω we assume that $Q(t)$ satisfies the following linear matrix differential equation:

$$\dot{Q} + f_x^* Q + Q f_x + H_{xx} = 0 \quad \text{on } [t_1, t_s] \setminus \theta. \tag{4.3}$$

It is interesting to note that the same equation is obtained from the modified Riccati equation in [30, equation (47)] when all control variables are on the boundary of the control constraints. Using (4.3) the quadratic form (4.2) reduces to

$$\Omega = \sum_{k=1}^{s} \omega_k, \quad \omega_k := D^k(H)\xi_k^2 - 2[\dot{\psi}]^k \bar{x}_{\mathrm{av}}^k \xi_k + \left[\langle Q\bar{x}, \bar{x}\rangle\right]^k. \tag{4.4}$$

Thus, we have proved the following statement.

PROPOSITION 4.1. *Let $Q(t)$ satisfy the linear differential equation* (4.3) *on $[t_1, t_s]\setminus\theta$. Then for each $\bar{z} \in \mathcal{K}$ the representation* (4.4) *holds.*

Now our goal is to derive conditions such that $\omega_k > 0$ holds on $\mathcal{K} \setminus \{0\}$ for $k = 1, \ldots, s$. We shall transform ω_k as in [42]. First we shall express it via the vector $(\xi_k, \bar{x}^{k-})$ and then via $(\xi_k, \bar{x}^{k+})$. To express ω_k as a quadratic form of $(\xi_k, \bar{x}^{k-})$, we use the formula

$$\bar{x}^{k+} = \bar{x}^{k-} + [\dot{x}]^k \xi_k, \tag{4.5}$$

which implies

$$\langle Q^{k+}\bar{x}^{k+}, \bar{x}^{k+}\rangle = \langle Q^{k+}\bar{x}^{k-}, \bar{x}^{k-}\rangle + 2\langle Q^{k+}[\dot{x}]^k, \bar{x}^{k-}\rangle \xi_k + \langle Q^{k+}[\dot{x}]^k, [\dot{x}]^k\rangle \xi_k^2.$$

Consequently,

$$\left[\langle Q\bar{x}, \bar{x}\rangle\right]^k = \langle [Q]^k \bar{x}^{k-}, \bar{x}^{k-}\rangle + 2\langle Q^{k+}[\dot{x}]^k, \bar{x}^{k-}\rangle \xi_k + \langle Q^{k+}[\dot{x}]^k, [\dot{x}]^k\rangle \xi_k^2.$$

Using this relation together with

$$\bar{x}^k_{\rm av} = \bar{x}^{k-} + \frac{1}{2}[\dot{x}]^k \xi_k$$

in the definition (4.4) of ω_k, we obtain

$$\begin{aligned} \omega_k &= \{D^k(H) + \left(([\dot{x}]^k)^* Q^{k+} - [\dot{\psi}]^k\right)[\dot{x}]^k\}\,\xi_k^2 \\ &\quad +2\left(([\dot{x}]^k)^* Q^{k+} - [\dot{\psi}]^k\right)\bar{x}^{k-}\xi_k + (\bar{x}^{k-})^*[Q]^k\bar{x}^{k-}. \end{aligned} \tag{4.6}$$

Here $[\dot{x}]^k$ and $\bar{x}^{k-}$ are column-vectors while $([\dot{x}]^k)^*$, $(\bar{x}^{k-})^*$, and $[\dot{\psi}]^k$ are row-vectors. Putting

$$q_{k+} = ([\dot{x}]^k)^* Q^{k+} - [\dot{\psi}]^k \tag{4.7}$$

we get

$$\omega_k = \left(D^k(H) + (q_{k+})[\dot{x}]^k\right)\xi_k^2 + 2(q_{k+})\bar{x}^{k-}\xi_k + (\bar{x}^{k-})^*[Q]^k\bar{x}^{k-}. \tag{4.8}$$

We immediately see from this representation that one way to enforce $\omega_k > 0$ is to impose the following conditions:

$$D^k(H) > 0, \quad q_{k+} = ([\dot{x}]^k)^* Q^{k+} - [\dot{\psi}]^k = 0, \quad [Q]^k \geq 0. \tag{4.9}$$

In practice, however, it might be difficult to check these conditions since it is necessary to satisfy the $d(x)$ equality constraints $q_{k+} = ([\dot{x}]^k)^* Q^{k+} - [\dot{\psi}]^k = 0$ together with the inequality constraints $[Q]^k \geq 0$. It is more convenient to express ω_k as a quadratic form in the variables $(\xi_k, \bar{x}^{k-})$ with the matrix

$$M_{k+} = \begin{pmatrix} D^k(H) + (q_{k+})[\dot{x}]^k & q_{k+} \\ (q_{k+})^* & [Q]^k \end{pmatrix}, \tag{4.10}$$

where q_{k+} is a row-vector and $(q_{k+})^*$ is a column-vector.

Similarly, using the relation

$$\bar{x}^{k-} = \bar{x}^{k+} - [\dot{x}]^k \xi_k,$$

we obtain

$$[\langle Q\bar{x}, \bar{x}\rangle]^k = \langle [Q]^k \bar{x}^{k+}, \bar{x}^{k+}\rangle + 2\langle Q^{k-}[\dot{x}]^k, \bar{x}^{k+}\rangle \xi^k - \langle Q^{k-}[\dot{x}]^k, [\dot{x}]^k\rangle \xi_k^2.$$

This formula together with the relation

$$\bar{x}^k_{\rm av} = \bar{x}^{k+} - \frac{1}{2}[\dot{x}]^k \xi_k$$

leads to the representation

$$\begin{aligned} \omega_k &= \{D^k(H) - \left(([\dot{x}]^k)^* Q^{k-} - [\dot{\psi}]^k\right)[\dot{x}]^k\}\,\xi_k^2 \\ &\quad +2\left(([\dot{x}]^k)^* Q^{k-} - [\dot{\psi}]^k\right)\bar{x}^{k+}\xi_k + (\bar{x}^{k+})^*[Q]^k\bar{x}^{k+}. \end{aligned} \tag{4.11}$$

Defining

$$q_{k-} = ([\dot{x}]^k)^* Q^{k-} - [\dot{\psi}]^k, \tag{4.12}$$

we get

$$\omega_k = \left(D^k(H) - (q_{k-})[\dot{x}]^k\right)\xi_k^2 + 2(q_{k-})\bar{x}^{k+}\xi_k + (\bar{x}^{k+})^*[Q]^k\bar{x}^{k+}. \tag{4.13}$$

Again, we see that $\omega_k > 0$ holds if we require the conditions

$$D^k(H) > 0, \quad q_{k-} = ([\dot{x}]^k)^* Q^{k-} - [\dot{\psi}]^k = 0, \quad [Q]^k \geq 0. \tag{4.14}$$

To find a more general condition for $\omega_k > 0$, we consider (4.13) as a quadratic form in the variables $(\xi_k, \bar{x}^{k+})$ with the matrix

$$M_{k-} = \begin{pmatrix} D^k(H) - (q_{k-})[\dot{x}]^k & q_{k-} \\ (q_{k-})^* & [Q]^k \end{pmatrix}. \tag{4.15}$$

Since the right-hand sides of equalities (4.8) and (4.13) are connected by the relation (4.5), the following statement obviously holds.

PROPOSITION 4.2. *For each $k = 1, \dots, s$, the positive (semi)definiteness of the matrix M_{k-} is equivalent to the positive (semi)definiteness of the matrix M_{k+}.*

Now we can prove the following theorem.

THEOREM 4.3. *Let $Q(t)$ be a solution of the linear differential equation* (4.3) *on $[t_1, t_s] \setminus \theta$ which satisfies the following conditions:*

(a) *the matrix M_{k+} is positive semidefinite for each $k = 2, \dots, s$;*

(b) $b_{k+} := D^k(H) + (q_{k+})[\dot{x}]^k > 0$ *for each $k = 1, \dots, s-1$.*

Then Ω is positive on $\mathcal{K} \setminus \{0\}$.

Proof. Take an arbitrary element $\bar{z} = (\xi, \bar{x}) \in \mathcal{K}$. Let us show that $\Omega \geq 0$ for this element. Condition (a) implies that $\omega_k \geq 0$ for $k = 2, \dots, s$. Condition (b) for $k = 1$ together with condition $\bar{x}^{1-} = 0$ implies that $\omega_1 \geq 0$. Consequently, $\Omega \geq 0$.

Assume that $\Omega = 0$. Then $\omega_k = 0$ for $k = 1, \dots, s$. The conditions $\omega_1 = 0$, $\bar{x}^{1-} = 0$, and $b_{1+} > 0$ by formula (4.8) (with $k = 1$) yield $\xi_1 = 0$. Then $[\bar{x}]^1 = 0$ and hence $\bar{x}^{1+} = 0$. The last equality together with equation $\dot{\bar{x}} = f_x(t)\bar{x}$ shows that $\bar{x}(t) = 0$ in (t_1, t_2) and hence $\bar{x}^{2-} = 0$. Similarly, the conditions $\omega_2 = 0$, $\bar{x}^{2-} = 0$ and $b_{2+} > 0$ by formula (4.8) (with $k = 2$) imply that $\xi_2 = 0$ and $\bar{x}(t) = 0$ in (t_2, t_3). Therefore, $\bar{x}^{3-} = 0$, etc. Continuing this process we get $\bar{x} \equiv 0$ and $\xi_k = 0$ for $k = 1, \dots, s-1$. Now using formula (4.4) for $\omega_s = 0$, as well as the conditions $D^s(H) > 0$ and $\bar{x} \equiv 0$, we obtain that $\xi_s = 0$. Consequently, we have $\bar{z} = 0$ which means that Ω is positive on $\mathcal{K} \setminus \{0\}$.

Similarly, using representation (4.13) for ω_k we can prove the following statement.

THEOREM 4.4. *Let $Q(t)$ be a solution of the linear differential equation* (4.3) *on $[t_1, t_s] \setminus \theta$ which satisfies the following conditions:*

(a) *the matrix M_{k-} is positive semidefinite for each $k = 1, \dots, s-1$,*

(b) $b_{k-} := D^k(H) - (q_{k-})[\dot{x}]^k > 0$ *for each $k = 2, \dots, s$.*

Then Ω is positive on $\mathcal{K} \setminus \{0\}$.

4.2. Q-transformation of Ω to perfect squares. We shall formulate special jump conditions for the matrix Q at each point $t_k \in \theta$. This will make it possible to transform Ω to perfect squares and thus to prove its positive definiteness on $\mathcal{K}$.

PROPOSITION 4.5 (see [42]). *Suppose that*

$$b_{k-} := D^k(H) - (q_{k-})[\dot{x}]^k > 0 \tag{4.16}$$

and that Q satisfies the jump condition at t_k

$$b_{k-}[Q]^k = (q_{k-})^*(q_{k-}), \tag{4.17}$$

where $(q_{k-})^*$ is a column-vector while q_{k-} is a row-vector. Then ω_k can be written as the perfect square

$$\omega_k = (b_{k-})^{-1}\left((b_{k-})\xi_k + (q_{k-})(\bar{x}^{k+})\right)^2 = (b_{k-})^{-1}\left(D^k(H)\xi_k + (q_{k-})(\bar{x}^{k-})\right)^2. \tag{4.18}$$

Proof. Using (4.13) and (4.17), we obtain

$$\begin{aligned}\omega_k &= (b_{k-})\xi_k^2 + 2(q_{k-})\bar{x}^{k+}\xi_k + (\bar{x}^{k+})^*[Q]^k\bar{x}^{k+}\\ &= (b_{k-})^{-1}\left((b_{k-})^2\xi_k^2 + 2(q_{k-})\bar{x}^{k+}(b_{k-})\xi_k + \left((q_{k-})\bar{x}^{k+}\right)^2\right)\\ &= (b_{k-})^{-1}\left((b_{k-})\xi_k + (q_{k-})(\bar{x}^{k+})\right)^2.\end{aligned}$$

Since

$$(b_{k-})\xi_k + (q_{k-})\bar{x}^{k+} = \left(D^k(H) - (q_{k-})[\dot{x}]^k\right)\xi_k + (q_{k-})\bar{x}^{k+}$$

$$= D^k(H)\xi_k - (q_{k-})[\bar{x}]^k + (q_{k-})\bar{x}^{k+} = D^k(H)\xi_k + (q_{k-})\bar{x}^{k-},$$

we see that equality (4.18) holds.

THEOREM 4.6. *Let $Q(t)$ satisfy the linear differential equation* (4.3) *on $[t_1, t_s]\setminus\theta$. Let condition* (4.16) *hold for each $k = 1, \dots, s$ and condition* (4.17) *hold for each $k = 1, \dots, s-1$. Then Ω is positive on $\mathcal{K}\setminus\{0\}$.*

Proof. By Proposition 4.5 and formulae (4.13), (4.15) the matrix M_{k-} is positive semidefinite for each $k = 1, \dots, s-1$, and hence both conditions (a) and (b) of Theorem 4.4 are fulfilled. Then by this theorem, Ω is positive on $\mathcal{K}\setminus\{0\}$.

Similar assertions hold for the jump conditions that use right-hand values of Q at each point $t_k \in \theta$.

PROPOSITION 4.7 (see [42]). *Suppose that*

$$b_{k+} := D^k(H) + (q_{k+})[\dot{x}]^k > 0 \tag{4.19}$$

and that Q satisfies the jump condition at point t_k

$$b_{k+}[Q]^k = (q_{k+})^*(q_{k+}). \tag{4.20}$$

Then

$$\omega_k = (b_{k+})^{-1}\left((b_{k+})\xi_k + (q_{k+})(\bar{x}^{k-})\right)^2 = (b_{k+})^{-1}\left(D^k(H)\xi_k + (q_{k+})(\bar{x}^{k+})\right)^2. \tag{4.21}$$

THEOREM 4.8. *Let $Q(t)$ satisfy the linear differential equation* (4.3) *on $[t_1, t_s]\setminus\theta$. Let condition* (4.19) *hold for each $k = 1, \dots, s$ and condition* (4.20) *hold for each $k = 2, \dots, s$. Then Ω is positive on $\mathcal{K}\setminus\{0\}$.*

4.3. Case of two switching points of the control. Let $s = 2$, i.e., $\theta = \{t_1, t_2\}$, and let $Q(t)$ be a symmetric matrix with absolutely continuous entries on $[t_1, t_2]$. Put

$$Q^k = Q(t_k), \quad q_k = ([\dot{x}]^k)^*Q^k - [\dot{\psi}]^k, \quad k = 1, 2.$$

THEOREM 4.9. *Let $Q(t)$ satisfy the linear differential equation* (4.3) *on (t_1, t_2) such that the following inequalities hold at t_1, t_2:*

$$D^1(H) + q_1[\dot{x}]^1 > 0, \quad D^2(H) - q_2[\dot{x}]^2 > 0. \tag{4.22}$$

Then Ω is positive on $\mathcal{K} \setminus \{0\}$.

Proof. In the case considered we have

$$Q^{1+} = Q^1, \quad q_{1+} = q_1, \quad Q^{2-} = Q^2, \quad q_{2-} = q_2$$

and

$$b_{1+} := D^1(H) + q_1[\dot{x}]^1 > 0, \quad b_{2-} := D^2(H) - q_2[\dot{x}]^2 > 0. \tag{4.23}$$

Define the jumps $[Q]^1$ and $[Q]^2$ by the conditions

$$b_{1+}[Q]^1 = (q_{1+})^*(q_{1+}), \quad b_{2-}[Q]^2 = (q_{2-})^*(q_{2-}). \tag{4.24}$$

Then $[Q]^1$ and $[Q]^2$ are symmetric matrices. Put

$$Q^{1-} = Q^{1+} - [Q]^1, \quad Q^{2+} = Q^{2-} + [Q]^2.$$

Then Q^{1-} and Q^{2+} are also symmetric matrices. Thus, we obtain a symmetric matrix $Q(t)$ satisfying (4.3) on (t_1, t_2), the inequalities (4.23), and the jump conditions (4.24). By Propositions 4.7 and 4.5, the terms ω_1 and ω_2 are nonnegative. In view of (4.4) we see that $\Omega = \omega_1 + \omega_2$ is nonnegative on $\mathcal{K}$. Suppose that $\Omega = 0$ for some $\bar{z} = (\xi, \bar{x}) \in \mathcal{K}$. Then $\omega_k = 0$ for $k = 1, 2$ and thus Propositions 4.7 and 4.5 give

$$b_{1+}\xi_1 + (q_{1+})\bar{x}^{1-} = 0, \quad b_{2-}\xi_2 + (q_{2-})\bar{x}^{2+} = 0.$$

But $\bar{x}^{1-} = 0$ and $\bar{x}^{2+} = 0$. Consequently, $\xi_1 = \xi_2 = 0$ and then conditions $\bar{x}^{1-} = 0$ and $[\bar{x}]^1 = 0$ imply that $\bar{x}^{1+} = 0$. The last equality together with equation $\dot{\bar{x}} = f_x(t)\bar{x}$ implies that $\bar{x}(t) = 0$ on (t_1, t_2). Thus $\bar{x} \equiv 0$ and then $\bar{z} = 0$. We have proved that Ω is positive on $\mathcal{K} \setminus \{0\}$.

4.4. Control system with a constant matrix B. In the case that $B(t, x) = B$ is a constant matrix, the adjoint equation has the form

$$\dot{\psi} = -\psi a_x,$$

which implies that

$$[\dot{\psi}]^k = 0, \quad k = 1, \dots, s.$$

Therefore,

$$\begin{aligned} q_{k-} &= ([\dot{x}]^k)^* Q^{k-}, & q_{k+} &= ([\dot{x}]^k)^* Q^{k+}, \\ (q_{k-})^* q_{k-} &= Q^{k-}[\dot{x}]^k([\dot{x}]^k)^* Q^{k-}, & (q_{k+})^* q_{k+} &= Q^{k+}[\dot{x}]^k([\dot{x}]^k)^* Q^{k+}, \\ b_{k-} &= D^k(H) - ([\dot{x}]^k)^* Q^{k-}[\dot{x}]^k, & b_{k+} &= D^k(H) + ([\dot{x}]^k)^* Q^{k+}[\dot{x}]^k, \end{aligned}$$

where

$$D^k(H) = \dot{\psi}(t_k) B[u]^k, \quad k = 1, \dots, s.$$

In case of two switching points with $s = 2$, the conditions (4.22) take the form

$$D^1(H) + \langle Q^1[\dot{x}]^1, [\dot{x}]^1) \rangle > 0\,, \quad D^2(H) - \langle Q^2[\dot{x}]^2, [\dot{x}]^2) \rangle > 0. \tag{4.25}$$

Now assume, in addition, that u is one-dimensional and that with $n = d(x)$

$$B = \beta e_n := \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \beta \end{pmatrix}, \quad \beta > 0, \qquad U = [-c, c], \quad c > 0\,.$$

In this case we get

$$[\dot{x}]^k = B[u]^k = \beta e_n [u]^k, \quad k = 1, \dots, s,$$

and thus

$$\langle Q^k [\dot{x}]^k, [\dot{x}]^k) \rangle = \beta^2 \langle Q^k e_n, e_n \rangle |[u]^k|^2 = 4\beta^2 c^2\, Q_{nn}(t_k),$$

where Q_{nn} is the element of matrix

$$Q = \begin{pmatrix} Q_{11} & \dots & Q_{1n} \\ \vdots & \vdots & \vdots \\ Q_{n1} & \dots & Q_{nn} \end{pmatrix}.$$

Moreover, in the last case we obviously have

$$D^k(H) = 2\beta c\, |\dot{\psi}_n(t_k)|, \quad k = 1, \dots, s. \tag{4.26}$$

For $s = 2$ conditions (4.25) then yield the estimates

$$Q_{nn}(t_1) > -\frac{|\dot{\psi}_n(t_1)|}{2\beta c}\,, \qquad Q_{nn}(t_2) < \frac{|\dot{\psi}_n(t_2)|}{2\beta c}\,. \tag{4.27}$$

5. Numerical examples. In this section, we shall discuss three time-optimal control problems with fixed initial and final states $x(0) = x_0$ and $x(T) = x_1$. To solve these problems numerically, we need to reduce them to control problems with *fixed* final time. The procedure to achieve this goal is well known [11, 29] and consists of introducing a new time variable $\tau \in [0, 1]$ according to the transformation

$$t = \tau \cdot T\,, \quad \tau \in [0, 1]\,. \tag{5.1}$$

In what follows, we shall identify the function $y(\tau)$ with the function $y(\tau \cdot T)$ for all $y \in \{x, u, \psi\}$. This time transformation leads to the *augmented* state variable

$$\tilde{x} := \begin{pmatrix} x \\ T \end{pmatrix} \in \mathbb{R}^{d(x)+1}$$

for which we obtain the ODE and boundary conditions

$$\begin{aligned} &dx/d\tau = T \cdot f(\tau \cdot T, x(\tau), u(\tau)), \quad dT/d\tau = 0, \quad \tau \in [0, 1]\,, \\ &x(0) = x_0, \quad x(1) = x_1. \end{aligned} \tag{5.2}$$

In the same way, the adjoint equation (2.7) is rewritten as

$$d\psi/d\tau = -T \cdot H_x(\tau \cdot T, x(\tau), u(\tau), \psi(\tau)). \tag{5.3}$$

All examples in this section will treat *autonomous* problems for which we will be able to compute *nondegenerate* solutions with $\psi_0(T) > 0$ in (2.6). Then we may scale the equations such that $\psi_0(T) = 1$ holds. Furthermore, in the autonomous case it follows from (2.8) that $\psi_0(t) \equiv \psi_0(T) = 1$. Hence, (2.10) yields the following condition expressed in the new time variable τ:

$$\psi(\tau)\, f(x(\tau), u(\tau)) + 1 \equiv 0 \quad \forall\, \tau \in [0,1]\,. \tag{5.4}$$

Moreover, u can be expressed via x and ψ from the minimum principle (2.9),

$$\min_{u \in U} \psi(\tau)\, f(x(\tau), u) + 1 = 0 \quad \forall\, \tau \in [0,1]\,. \tag{5.5}$$

In the following examples, we shall use shooting methods (cf. Bulirsch [5] and Oberle and Grimm [34]) for solving the boundary value problem (5.2)–(5.5). Shooting methods are known to provide highly accurate solutions for which we shall carry out the second order test.

5.1. Time-optimal control of a Van der Pol oscillator. The following time-optimal control of a Van der Pol oscillator has been treated by several authors; cf., e.g., Kaya and Noakes [13, 14]. The state variables are the voltage $x_1(t) = U(t)$ at time $t \in [\,0, T\,]$ and $x_2(t) := \dot{x}_1(t)$. The control $u(t)$ is the voltage at the generator; cf. the tunneldiode oscillator in [29, Figure 5.1 in section 5].

The control problem is to minimize the endtime T subject to the constraints

$$\dot{x}_1(t) = x_2(t)\,,\ \dot{x}_2(t) = -x_1(t) + x_2(t)(1 - x_1^2(t)) + u(t)\,, \tag{5.6}$$

$$x_1(0) = -0.4,\ x_2(0) = 0.6\,, \quad x_1(T) = 0.6,\ x_2(T) = 0.4\,, \tag{5.7}$$

$$|\,u(t)\,| \le 1 \quad \text{for } \ t \in [0, T\,]\,. \tag{5.8}$$

The Pontryagin function or Hamiltonian (2.5) becomes

$$H(x, u, \psi) = \psi_1 x_2 + \psi_2\,(-x_1 + x_2(1 - x_1^2) + u)\,. \tag{5.9}$$

The time transformation (5.1) yields the transformed state and adjoint equations (5.2), (5.3) in the time interval $\tau \in [0,1]$; for simplicity, the time argument τ will be omitted:

$$\begin{array}{rclcrcl} dx_1/d\tau &=& T \cdot x_2\,, & & dx_2/d\tau &=& T \cdot \big(-x_1 + x_2(1 - x_1^2) + u\big)\,, \\ d\psi_1/d\tau &=& T \cdot \psi_2(1 + 2x_1\, x_2)\,, & & d\psi_2/d\tau &=& -T \cdot (\psi_1 + \psi_2(1 - x_1^2))\,, \\ dT/d\tau &=& 0\,. & & & & \end{array} \tag{5.10}$$

The boundary conditions (5.7) and the condition (5.4) yield

$$\begin{aligned} &x_1(0) = -0.4,\ x_2(0) = 0.6, \quad x_1(1) = 0.6,\ x_2(1) = 0.4, \\ &0.4\psi_1(1) + \psi_2(1)(-0.344 + u(1)) + 1 = 0\,. \end{aligned} \tag{5.11}$$

The switching function $\sigma(x, \psi) = \psi_2$ determines the optimal control according to the control law (2.13),

$$u(\tau) = \left\{ \begin{array}{rcl} 1 & \text{if} & \psi_2(\tau) < 0 \\ -1 & \text{if} & \psi_2(\tau) > 0 \end{array} \right\}\,. \tag{5.12}$$

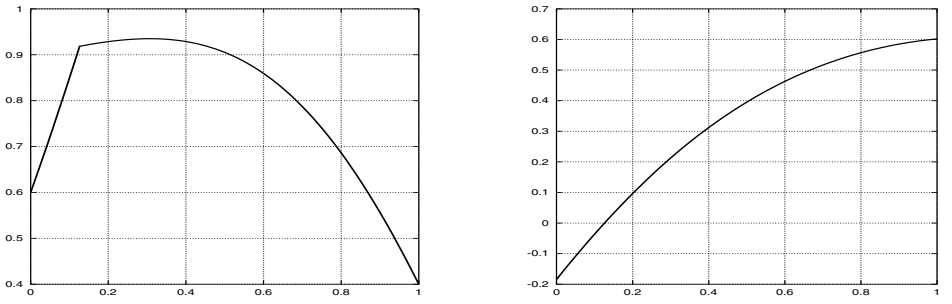


FIG. 5.1. *Van der Pol oscillator: state $x_2(\tau)$ and switching function $\sigma(\tau) = \psi_2(\tau)$, $\tau \in [0,1]$.*

It can easily be seen that the *singular case*, where $\psi_2(\tau) \equiv 0$ holds in a time interval $[\tau_1, \tau_2]$, does not occur. In fact, $\psi_2(\tau) \equiv 0$ would imply $\psi_1(\tau) \equiv 0$ and thus $H[\tau] \equiv 0$ which would contradict the condition (5.4) in the autonomous case. Computations show that the optimal bang-bang control has the following structure with two bang-bang arcs and only one switching point τ_1:

$$u(\tau) = \left\{ \begin{array}{rll} 1 & \text{for} & 0 \le \tau \le \tau_1 \\ -1 & \text{for} & \tau_1 \le \tau \le 1 \end{array} \right\}. \tag{5.13}$$

Hence, we have to impose the switching condition

$$\sigma[\tau_1] = \psi_2(\tau_1) = 0 \tag{5.14}$$

to determine the switching point τ_1.

The task now is to solve the boundary value problem with the following components: the state and adjoint equations (5.10) using the optimal control structure (5.13), the boundary conditions (5.11) and the switching condition (5.14). Employing the code BNDSCO in [34] we obtain the state variables and adjoint variables displayed in Figure 5.1. The optimal final time, the switching point, and some selected values for the adjoint variables are

$$\begin{array}{lll} T = 1.25407473, & \tau_1 = 0.12624458, & t_1 = \tau_1 \cdot T = 0.1583201376, \\ \psi_1(0) = -1.08160561, & \psi_2(0) = -0.18436798, & \psi_1(\tau_1) = -1.08863205, \\ \psi_1(1) = -0.47781383, & \psi_2(1) = 0.60184112\,. & \end{array} \tag{5.15}$$

Since the bang-bang control has only one switching point, we are in the position to apply Theorem 3.4. To check the assumptions of this theorem it remains to verify the condition $D^1(H) = |\dot\sigma(t_1)[u]^1| > 0$. Indeed, in view of the adjoint equation (5.10) and the switching condition $\psi_2(\tau_1) = 0$ we find for the original time variable $t_1 = \tau_1 \cdot T$,

$$D^1(H) = |\dot\sigma(t_1)[u]^1| = 2|\psi_1(t_1)| = 2 \cdot 1.08863205 > 0\,.$$

Then Theorem 3.4 asserts that the computed solution is a strict strong minimum.

Let us briefly discuss the optimal solution for the following boundary values (cf. Kaya and Noakes [14]) different from those in (5.7),

$$x_1(0) = x_2(0) = 1, \quad x_1(T) = x_2(T) = 0\,. \tag{5.16}$$

The optimal bang-bang control has two bang-bang arcs with one switching point τ_1. However, the control structure is reversed as compared to the one in (5.13):

$$u(\tau) = \left\{ \begin{array}{rll} -1 & \text{for} & 0 \le \tau \le \tau_1 \\ 1 & \text{for} & \tau_1 \le \tau \le 1 \end{array} \right\}. \tag{5.17}$$

We get the following numerical results,

$$\begin{array}{lll} T = 3.09520234, & \tau_1 = 0.23358852, & t_1 = \tau_1 \cdot T = 0.72300373, \\ \psi_1(0) = 0.94728449, & \psi_2(0) = 0.97364224, & \psi_1(\tau_1) = 1.70467637, \\ \psi_1(1) = 0.19669125, & \psi_2(1) = -1\,, & \end{array}$$

for which we obtain

$$D^1(H) = |\dot\sigma(t_1)[u]^1| = 2|\psi_1(t_1)| = 2 \cdot 1.70467637 > 0\,.$$

Theorem 3.4 shows again that the computed solution is a strict strong minimum.

5.2. Time-optimal control of the Rayleigh problem. The Rayleigh problem is concerned with the same electric circuit as treated in the previous section. However, the state variables are different since now the state variable $x_1(t) = I(t)$ denotes the electric current; cf. the dynamical model in [12, 27, 28, 29].

The control problem is to minimize the endtime T subject to

$$\dot x_1(t) = x_2(t)\,, \quad \dot x_2(t) = -x_1(t) + x_2(t)(1.4 - 0.14x_2(t)^2) + 4u(t)\,, \tag{5.18}$$

$$x_1(0) = x_2(0) = -5\,, \quad x_1(T) = x_2(T) = 0\,, \tag{5.19}$$

$$|\,u(t)\,| \le 1 \quad \text{for } t \in [0, T\,]\,. \tag{5.20}$$

The Pontryagin function (2.5) for this problem is

$$H(x, u, \psi) = \psi_1 x_2 + \psi_2\,(-x_1 + x_2(1.4 - 0.14x_2^2) + 4u). \tag{5.21}$$

The time transformation (5.1) and the transformed state and adjoint equations (5.2), (5.3) in the time interval $\tau \in [0, 1]$ lead to the following equations; again, the time argument τ will be omitted:

$$\begin{array}{rclrcl} dx_1/d\tau &=& T \cdot x_2\,, & dx_2/d\tau &=& T \cdot (-x_1 + x_2(1.4 - 0.14x_2^2) + 4u)\,, \\ d\psi_1/d\tau &=& T \cdot \psi_2\,, & d\psi_2/d\tau &=& -T \cdot (\psi_1 + \psi_2(1.4 - 0.42x_2^2))\,, \\ dT/d\tau &=& 0\,. & & & \end{array} \tag{5.22}$$

The boundary conditions (5.19) and the condition (5.4) yield, in view of (5.21),

$$x_1(0) = x_2(0) = -5, \quad x_1(1) = x_2(1) = 0, \quad 4\,\psi_2(1)\,u(1) + 1 = 0. \tag{5.23}$$

The switching function $\sigma(x, \psi) = 4\psi_2$ determines the optimal control via the minimum condition (2.13):

$$u(\tau) = \left\{ \begin{array}{rll} 1 & \text{if} & \psi_2(\tau) < 0 \\ -1 & \text{if} & \psi_2(\tau) > 0 \end{array} \right\}. \tag{5.24}$$

Again, the *singular case* with $\psi_2(\tau) \equiv 0$ holding in a time interval $[\tau_1, \tau_2]$ can be eliminated. Hence, the optimal control is bang-bang. In view of the special terminal

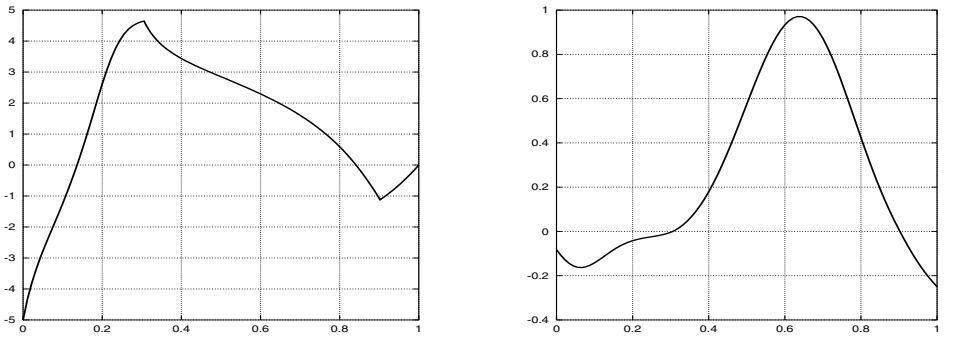


FIG. 5.2. *Rayleigh problem: state $x_2(\tau)$ and switching function $\psi_2(\tau) = \sigma(\tau)/4$, $\tau \in [0,1]$.*

conditions for the state, a simple reasoning reveals that the optimal control cannot be composed of only two bang-bang arcs. Computations show that the optimal control comprises the following three bang-bang arcs:

$$u(\tau) = \left\{ \begin{array}{rll} 1 & \text{for} & 0 \le \tau \le \tau_1 \\ -1 & \text{for} & \tau_1 \le \tau \le \tau_2 \\ 1 & \text{for} & \tau_2 \le \tau \le 1 \end{array} \right\}. \tag{5.25}$$

This control structure yields the two switching conditions

$$\psi_2(\tau_1) = 0\,, \quad \psi_2(\tau_2) = 0\,. \tag{5.26}$$

Thus we have to solve the multipoint boundary value problem consisting of the state and adjoint equations (5.22) with the optimal control structure (5.25), the boundary conditions (5.23), and the switching conditions (5.26).

The code BNDSCO in [34] yields the final time, the switching points, and some selected values for the adjoint variables as follows:

$$\begin{array}{rclrcl} T &=& 3.66817339\,, & & & \\ \tau_1 &=& 0.30546718\,, & \tau_2 &=& 0.90236928\,, \\ t_1 &=& \tau_1 \cdot T = 1.12050658\,, & t_2 &=& \tau_2 \cdot T = 3.31004698\,, \\ \psi_1(0) &=& -0.12234128\,, & \psi_2(0) &=& -0.08265161\,, \\ \psi_1(\tau_1) &=& -0.21521225\,, & \psi_1(\tau_2) &=& 0.89199176\,, \\ \psi_1(1) &=& 0.84276186\,, & \psi_2(1) &=& -0.25\,. \end{array} \tag{5.27}$$

Figure 5.2 displays the state variable $x_2(\tau)$ and the switching function $\psi_2(\tau)$ which match precisely the control laws (5.24) and (5.25).

We are going to show now in two different ways that the computed control provides a strict strong minimum. First, we compute the quantities $D^k(H) = -\dot\sigma(t_k)[u]^k$, $k = 1,2$, where $-\dot\sigma(t_k) = -4\dot\psi_2(t_k) = 4\psi_1(t_k)$ holds in view of the adjoint equation in (5.22) evaluated in the original time variable $t \in [0,T]$. Inserting the values from (5.27) we find

$$D^1(H) = 8 \cdot 0.21521225 = 1.7269800 > 0\,,\ D^2(H) = 8 \cdot 0.89199176 = 7.1359341 > 0\,.$$

The variational system $\dot y = f_x(t)y$ with $y = (y_1, y_2)$ in (3.3) reads explicitly

$$\dot y_1 = y_2, \quad \dot y_2 = -y_1 + (1.4 - 0.42x_2^2)y_2\,.$$

The initial values for the variations $y^1(t)$, $y^2(t)$ w.r.t. the switching points t_1, t_2 are

$$y^1(t_1) = [\dot{x}]^1 = \begin{pmatrix} 0 \\ -8 \end{pmatrix}, \quad y^2(t_2) = [\dot{x}]^2 = \begin{pmatrix} 0 \\ 8 \end{pmatrix}.$$

At the second switching point t_2 we find $y^1(t_2) = (0, 2.517130)$. In view of the initial value $y^2(t_2) = (0, 8)$, this already implies that the vectors $y^1(T)$ and $y^2(T)$ are *linearly dependent*. Explicitly, we get $y^1(T) = (1.084614, 3.656286)$, $y^2(T) = (3.447153, 11.620490)$ which gives $y^2(T) = \alpha y^1(T)$ with $\alpha = 3.17823$ in relation (3.9). Thus, condition (b) in Proposition 3.2 asserting the zero critical subspace is not satisfied here. Here, the critical subspace is a one-dimensional subspace and the test for optimality proceeds via Proposition 3.5 by verifying that the number ρ in (3.15) is positive. Using the above variational vectors we compute

$$\int_{t_1}^{t_2} \langle H_{xx}(t) y^1(t), y^1(t) \rangle \, dt = -0.84 \int_{t_1}^{t_2} x_2(t) \psi_2(t) (y_2^1(t))^2 dt \; = \; -0.97063758.$$

Finally, observing the relations $[\dot{\psi}]^1 = [\dot{\psi}]^2 = 0$ and inserting the computed values of $D^1(H)$, $D^2(H)$ and α we obtain

$$\rho = 1.726980 + 0.706448 - 0.970638 = 1.462790 > 0.$$

Hence, we have shown that the solution described by (5.27) is a strict strong minimum.

An alternative proof of optimality proceeds via Theorem 4.9. Consider the symmetric 2×2 matrix

$$Q(t) = \begin{pmatrix} Q_{11}(t) & Q_{12}(t) \\ Q_{12}(t) & Q_{22}(t) \end{pmatrix}.$$

The linear equation (4.3), $\dot{Q} = -Q f_x - f_x^* Q - H_{xx}$, in the original time variable $t \in [t_1, t_2]$ leads to the following three ODEs:

$$\begin{aligned} \dot{Q}_{11} &= 2\, Q_{12}, \\ \dot{Q}_{12} &= -Q_{11} - Q_{12}(1.4 - 0.42 x_2^2) + Q_{22}, \\ \dot{Q}_{22} &= -2\,(Q_{12} + Q_{22}(1.4 - 0.42 x_2^2)) + 0.84 \psi_2 x_2. \end{aligned} \tag{5.28}$$

We have to find a solution $Q(t)$ that satisfies the estimates (4.22), respectively, (4.27) at the switching points t_1 and t_2,

$$Q_{22}(t_1) > -\frac{|\psi_1(t_1)|}{8} = -0.026901531\,, \quad Q_{22}(t_2) < \frac{|\psi_1(t_2)|}{8} = 0.11149897\,. \tag{5.29}$$

These conditions hold if we choose, e.g., the following initial values at the switching point t_1,

$$Q_{11}(t_1) = 0, \quad Q_{12}(t_1) = 0, \quad Q_{22}(t_1) = -0.02,$$

which produce the value $Q_{22}(t_2) = -0.048826568$ at the second switching point. Then Theorem 4.9 assures us that the computed solution (5.27) provides a strict strong minimum.

5.3. Time-optimal control of a nuclear reactor. Hassan, Ghonaimy, and Abdel Malek [10] have presented a model for the time-optimal control of a nuclear reactor. A detailed solution has been given in Maurer [26]. Now our aim is to verify second order conditions for this specific solution. The model comprises the state variables x_1, neutron density; x_2, delayed neutron concentration; and x_3, reactivity. The control problem is to minimize the final time T subject to

$$\begin{array}{lll} \dot{x}_1(t) = k_1(x_3(t)-1)x_1(t) + k_2 x_2(t)\,, & x_1(0) = n_0, & x_1(T) = n_f, \\ \dot{x}_2(t) = k_1 x_1(t) - k_2 x_2(t)\,, & x_2(0) = n_0 k_1/k_2, & x_2(T) = n_f k_1/k_2, \\ \dot{x}_3(t) = u(t)\,, & x_3(0) = 0, & x_3(T) = 0, \\ |\,u(t)\,| \le 0.2 & \text{for } t \in [0, T\,]\,. & \end{array} \tag{5.30}$$

The constants are $k_1 = 5.0$, $k_2 = 0.1$, $n_0 = 0.04$, $n_f = 0.06$. The Pontryagin function or Hamiltonian (2.5) becomes

$$H(x, u, \psi) = \psi_1(k_1(x_3 - 1)x_1 + k_2 x_2) + \psi_2(k_1 x_1 - k_2 x_2) + \psi_3 u\,. \tag{5.31}$$

The time transformation (5.1) and the scaled equations (5.2)–(5.4) yield the following state and adjoint equations and boundary conditions:

$$\begin{array}{rcll} dx_1/d\tau &=& T\cdot(k_1(x_3-1)x_1 + k_2 x_2)\,, & x_1(0) = 0.04, \quad x_1(1) = 0.06, \\ dx_2/d\tau &=& T\cdot(k_1 x_1 - k_2 x_2)\,, & x_2(0) = 2, \quad x_2(1) = 3, \\ dx_3/d\tau &=& T\cdot u(\tau)\,, & x_3(0) = 0, \quad x_3(1) = 0, \\ d\psi_1/d\tau &=& -T\cdot(\psi_1 k_1(x_3-1) + \psi_2 k_1)\,, & \\ d\psi_2/d\tau &=& T\cdot k_2(\psi_2 - \psi_1)\,, & \\ d\psi_3/d\tau &=& -T\psi_1 k_1 x_1\,, & \psi_3(0) = -5\,, \quad \psi_3(1) = -5\,. \end{array} \tag{5.32}$$

The switching function $\sigma(x, \psi) = \psi_3(t)$ determines the optimal control via $u(t) = -0.2\,sign(\psi_3(t))$. The optimal control computed in [26] is composed of three bang-bang arcs,

$$u(\tau) = \left\{ \begin{array}{rcl} 0.2 & \text{for} & 0 \le \tau \le \tau_1 \\ -0.2 & \text{for} & \tau_1 \le \tau \le \tau_2 \\ 0.2 & \text{for} & \tau_2 \le \tau \le 1 \end{array} \right\}, \tag{5.33}$$

which imply the two further switching conditions

$$\psi_3(\tau_1) = 0, \quad \psi_3(\tau_2) = 0\,. \tag{5.34}$$

The earlier computations in [26] are confirmed by the code BNDSCO in [34] which yields the following solution of the boundary value problem (5.32)–(5.34):

$$\begin{array}{rcl@{\qquad}rcl} T &=& 7.04780685, & & & \\ \tau_1 &=& 0.47987830\,, & t_1 &=& \tau_1 \cdot T = 3.38208957\,, \\ \tau_2 &=& 0.97987830\,, & t_2 &=& \tau_2 \cdot T = 6.90599299\,, \\ \psi_1(0) &=& -2.97015515\,, & \psi_2(0) &=& -2.84546900\,, \\ \psi_1(\tau_1) &=& -5.22557130\,, & \psi_2(\tau_1) &=& -2.22864972\,, \\ x_1(\tau_1) &=& 0.11014294\,, & x_1(\tau_2) &=& 0.06078025\,, \\ \psi_1(\tau_2) &=& 78.6539693\,, & \psi_2(\tau_2) &=& -3.53032114\,, \\ \psi_1(1) &=& 165.786058\,, & \psi_2(1) &=& -5.25230261\,. \end{array} \tag{5.35}$$

The state variable $x_3(\tau)$ and the switching function $\sigma(\tau) = \psi_3(\tau)$ are displayed in Figure 5.3.

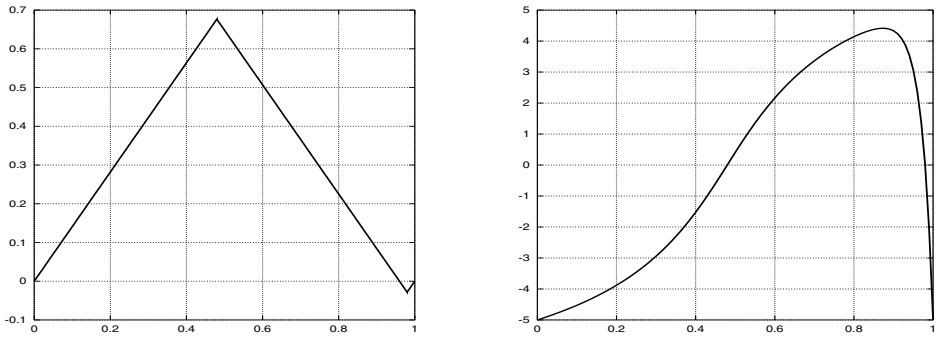


FIG. 5.3. *Nuclear reactor: state $x_3(\tau)$ and switching function $\sigma(\tau) = \psi_3(\tau)$.*

As in the foregoing example, we can show in two different ways that the computed control provides a strict strong minimum. The quantities

$$D^k(H) = -\dot\sigma(t_k)[u]^k = 0.4\,|\dot\psi_3(t_k)| = 0.4\,|\psi_1(t_k)k_1x_1(t_k)|, \quad k = 1, 2,$$

are computed on the basis of solution data in (5.35) as

$$D^1(H) = 1.15111957 > 0, \quad D^2(H) = 9.56121580 > 0\,.$$

Evaluating the variational system (3.3), $\dot y = f_x(t)y$ with $y = (y_1, y_2, y_3)$, we get

$$\dot y_1 = k_1(x_3 - 1)y_1 + k_2y_2 + k_1x_1y_3, \quad \dot y_2 = k_1y_1 - k_2y_2, \quad \dot y_3 = 0.$$

The initial values for the variations $y^1(t)$, $y^2(t)$ w.r.t. t_1, t_2 are

$$y^1(t_1) = [\dot x]^1 = \begin{pmatrix} 0 \\ 0 \\ -0.4 \end{pmatrix}, \quad y^2(t_2) = [\dot x]^2 = \begin{pmatrix} 0 \\ 0 \\ 0.4 \end{pmatrix}.$$

This leads to the following variational vectors at the terminal time T:

$$y^1(T) = \begin{pmatrix} -0.04508835 \\ -1.0424039 \\ -0.4 \end{pmatrix}, \quad y^2(T) = \begin{pmatrix} 0.012216498 \\ 0.0048217532 \\ 0.4 \end{pmatrix}, \quad \dot x(T) = \begin{pmatrix} 0 \\ 0 \\ 0.2 \end{pmatrix},$$

which obviously are *linearly independent*. Thus, either condition (a) or (b) in Proposition 3.2 implies that the critical cone is $\mathcal{K} = \{0\}$. Hence, Theorem 3.3 asserts that the solution candidate characterized by (5.35) provides indeed a strict strong minimum.

Alternatively, it is instructive to use also the test of optimality in Theorem 4.9. Since $d(x) = 3$ we consider the symmetric 3×3 matrix $Q(t) = (Q_{ik})_{1\le i,k\le 3}$. By evaluating the linear equation (4.3) one immediately recognizes that the equations for Q_{11}, Q_{12}, Q_{22} are homogeneous in these variables and can thus be satisfied by $Q_{11}(t) = Q_{12}(t) = Q_{22}(t) \equiv 0$. The remaining three equations then simplify to

$$\begin{aligned} \dot Q_{13} &= -Q_{13}k_1(x_3 - 1) - Q_{23}k_1 - \psi_1k_1, \\ \dot Q_{23} &= -Q_{13}k_2 + Q_{23}k_2, \\ \dot Q_{33} &= -2Q_{13}k_1x_1. \end{aligned} \tag{5.36}$$

Our task is to find a solution to these ODEs which satisfies the estimates (4.22) or (4.27) at the switching points t_1 and t_2. Since

$$\frac{|\dot{\psi}_3|}{2\beta c} = \frac{k_1|\psi_1 x_1|}{0.4} = 12.5|\psi_1 x_1|,$$

conditions (4.27) require that the following estimates be satisfied:

$$\begin{array}{rcrcr} Q_{33}(t_1) & > & -12.5\,|\psi_1(t_1)x_1(t_1)| & = & -7.1944973, \\ Q_{33}(t_2) & < & 12.5\,|\psi_1(t_2)x_1(t_2)| & = & 59.788260. \end{array} \tag{5.37}$$

The strategy for finding appropriate initial values at the point t_1 is the following: we fix the initial values

$$Q_{13}(t_1) = 0, \quad Q_{33}(t_1) = 0,$$

and determine $Q_{23}(t_1)$ in such a way that the inequality $Q_{33}(t_2) < 59.788260$ holds. We found that the initial value $Q_{23}(t_1) = 4.23$ produced the value $Q_{33}(t_2) = -96.953435$. Hence, the inequalities (5.37) hold and Theorem 4.9 asserts that the computed solution is a strict strong minimum.

6. Conclusion. We have considered time-optimal bang-bang control problems with finitely many switching points. SSC for such problems amount to the requirement that a certain quadratic form be positive on a finite-dimensional critical subspace. An explicit representation of the critical subspace has been derived in terms of the variations of the state trajectories w.r.t. the switching points. For bang-bang controls with one or two switching points, this approach results in a rather straightforward test of SSC. To treat the general case, we have shown that the so-called Q-transformation allows us to convert the quadratic form to another quadratic form which might be better suited for practical verification. The resulting numerical test then consists in determining a solution of a linear matrix differential equation which satisfies additional jump conditions at the switching points. The viability of the presented tests has been demonstrated by three numerical examples. Further examples with applications of bang-bang control to the design of lasers may be found in the dissertation of Kim [16].

Though the techniques have been developed in this paper only for time-optimal bang-bang controls with fixed terminal conditions, the basic ideas apply as well to arbitrary bang-bang control problems with general cost functionals and boundary conditions. Results for this general approach will be presented in a future paper that will also highlight a more detailed analysis of the boundary conditions.

During the revision of this paper we became aware of the work of Agrachev, Stefani, and Zezza [1], where a different approach to SSC for bang-bang controls is presented for problems with *fixed* final time. Agrachev and his coauthors reduce the bang-bang control problem to a finite-dimensional optimization problem w.r.t. the switching times and show that it suffices to test SSC for this optimization problem. Currently, we are implementing this approach and are in the process of comparing it with the numerical methods given in the present paper. Recently, we have been able to show that the SSC given in Theorem 3.3 are equivalent to the SSC in Agrachev, Stefani, and Zezza [1] in the case when the set M_0 of Lagrange multipliers is a singleton which is not assumed in Theorem 3.3. The SSC developed in this paper and in [1] will pave the way to a theoretical and computational sensitivity analysis for bang-bang control problems which is similar in spirit to that developed in [2, 21, 22, 23, 24, 25, 27, 28].

Acknowledgments. We are grateful to the referees for helpful remarks and suggestions to improve the paper. We are also indebted to Dirk Augustin for providing us the bang-bang control for the Rayleigh problem in section 5.2.

REFERENCES

[1] A. A. Agrachev, G. Stefani, and P. L. Zezza, *Strong optimality for a bang-bang trajectory*, SIAM J. Control Optim., 41 (2002), pp. 991–1014.

[2] D. Augustin and H. Maurer, *Computational sensitivity analysis for state constrained optimal control problems*, Ann. Oper. Res., 101 (2001), pp. 75–99.

[3] A. Bressan, *A high order test for optimality of bang-bang controls*, SIAM J. Control Optim., 23 (1985), pp. 38–48.

[4] A. E. Bryson and Y. C. Ho, *Applied Optimal Control*, Revised printing, Hemisphere Publishing, Washington, D.C., 1975.

[5] R. Bulirsch, *Die Mehrzielmethode zur numerischen Lösung von nichtlinearen Randwertproblemen und Aufgaben der optimalen Steuerung*, Report of the Carl-Cranz Gesellschaft, Oberpfaffenhofen, Germany, 1971.

[6] F. L. Chernousko, L. D. Akulenko, and N. N. Bolotnik, *Time-optimal control for robotic manipulators*, Optimal Control Appl. Methods, 10 (1989), pp. 293–311.

[7] A. L. Dontchev, W. W. Hager, A. B. Poore, and B. Yang, *Optimality, stability, and convergence in nonlinear control*, Applied Math. Optim., 31 (1995), pp. 297–326.

[8] J. C. Dunn, *Second-order optimality conditions in sets of L^∞ functions with range in a polyhedron*, SIAM J. Control Optim., 33 (1995), pp. 1603–1635.

[9] J. C. Dunn, *On L^2 sufficient conditions and the gradient projection method for optimal control problems*, SIAM J. Control Optim., 34 (1996), pp. 1270–1290.

[10] M. A. Hassan, M. A. R. Ghonaimy, and N. R. Abdel Malek, *Computational solution of the nuclear reactor minimum time start-up problem with state constraints*, in 2nd IFAC Symposium on Multivariable Technical Control Systems, Düsseldorf, Germany, Oct. 11–13, North–Holland, Amsterdam, 1971.

[11] M. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

[12] D. H. Jacobson and D. Q. Mayne, *Differential Dynamic Programming*, American Elsevier, New York, 1970.

[13] Y. C. Kaya and J. L. Noakes, *Computations and time-optimal controls*, Optimal Control Appl. Methods, 17 (1996), pp. 171–185.

[14] Y. C. Kaya and J. L. Noakes, *Computational method for time-optimal switching control*, J. Optim. Theory Appl., 117 (2003), pp. 69–92.

[15] J.-H. R. Kim, H. Maurer, Yu. A. Astrov, M. Bode, and H.-G. Purwins, *High-speed switch-on of a semiconductor gas discharge image converter using optimal control methods*, J. Comput. Phys., 170 (2001), pp. 395–414.

[16] J.-H. R. Kim, *Optimierungsmethoden und Sensitivitätsanalyse für optimale bang-bang Steuerungen mit Anwendungen in der Nichtlinearen Optik*, Dissertation, Institut für Numerische Mathematik, Universität Münster, Münster, Germany, 2002.

[17] H. W. Knobloch, *Higher Order Necessary Conditions in Optimal Control Theory*, Lecture Notes in Comput. Sci. 34, Springer-Verlag, Berlin, 1981.

[18] A. J. Krener, *The high order maximum principle and its application to singular externals*, SIAM J. Control Optim., 15 (1977), pp. 256–293.

[19] U. Ledzewicz and H. Schättler, *Optimal bang-bang controls for a two-compartment model in cancer chemotherapy*, J. Optim. Theory Appl., 114 (2002), pp. 609–637.

[20] E. S. Levitin, A. A. Milyutin, and N. P. Osmolovskii, *Conditions of high order for a local minimum in problems with constraints*, Uspeki Mat. Nauk, 33 (1978), pp. 85–148 (in Russian); Russian Math. Surveys, 33 (1978), pp. 97–168 (in English).

[21] K. Malanowski, *Stability and sensitivity of solutions to nonlinear optimal control problems*, Appl. Math. Optim., 32 (1995), pp. 111–141.

[22] K. Malanowski, *Stability and sensitivity analysis for optimal control problems with control-state constraints*, Dissertationes Math. (Rozprawy Mat.), 394, (2001), pp. 1–51.

[23] K. Malanowski and H. Maurer, *Sensitivity analysis for parametric control problems with control-state constraints*, Comput. Optim. Appl., 5 (1996), pp. 253–283.

[24] K. Malanowski and H. Maurer, *Sensitivity analysis for state constrained optimal control problems*, Discrete Contin. Dynam. Systems, 4 (1998), pp. 241–272.

[25] K. Malanowski and H. Maurer, *Sensitivity analysis for optimal control problems subject to higher order state constraints. Optimization with data perturbations*, II, Ann. Oper. Res., 101 (2001), pp. 43–73.

[26] H. Maurer, *On optimal control problems with bounded state variables and control appearing linearly*, SIAM J. Control Optim., 15 (1977), pp. 345–362.

[27] H. Maurer and D. Augustin, *Second order sufficient conditions and sensitivity analysis for the controlled Rayleigh problem*, in Parametric Optimization and Related Topics, IV, J. Guddat et al., eds., Lang, Frankfurt, 1997, pp. 245–259.

[28] H. Maurer and D. Augustin, *Sensitivity analysis and real-time control of parametric optimal control problems using boundary value methods*, in On-line Optimization of Large Scale Systems, M. Grötschel et al., eds., Springer-Verlag, Berlin, 2001, pp. 17–55.

[29] H. Maurer and H. J. Oberle, *Second order sufficient conditions for optimal control problems with free final time: The Riccati approach*, SIAM J. Control Optim., 41 (2002), pp. 380–403.

[30] H. Maurer and S. Pickenhain, *Second-order sufficient conditions for optimal control problems with mixed control-state constraints*, J. Optim. Theory Appl., 86 (1995), pp. 649–667.

[31] A. A. Milyutin and N. P. Osmolovskii, *Calculus of Variations and Optimal Control*, Transl. Math. Monogr. 180, AMS, Providence, RI, 1998.

[32] L. W. Neustadt, *Optimization: A Theory of Necessary Conditions*, Princeton University Press, Princeton, NJ, 1976.

[33] J. Noble and H. Schättler, *Sufficient conditions for relative minima of broken extremals in optimal control theory*, J. Math. Anal. Appl., 269 (2002), pp. 98–128.

[34] H. J. Oberle and W. Grimm, *BNDSCO — A Program for the Numerical Solution of Optimal Control Problems*, Internal report 515–89/22, Institute for Flight Systems Dynamics, DLR, Oberpfaffenhofen, Germany, 1989.

[35] N. P. Osmolovskii, *High-order necessary and sufficient conditions for Pontryagin and bounded-strong minima in the optimal control problems*, Dokl. Akad. Nauk SSSR, Ser. Cybernetics and Control Theory, 303 (1988), pp. 1052–1056 (in Russian); Sov. Phys. Dokl., 33 (1988), pp. 883–885 (in English).

[36] N. P. Osmolovskii. *Theory of Higher Order Conditions in Optimal Control*, Doctor of Sci. thesis, Moscow, Russia, 1988 (in Russian).

[37] N. P. Osmolovskii, *Quadratic conditions for nonsingular extremals in optimal control (a theoretical treatment)*, Russian J. Math. Phys., 2 (1995), pp. 487–516.

[38] N. P. Osmolovskii, *Quadratic conditions for nonsingular extremals in optimal control (examples)*, Russian J. Math. Phys., 5 (1998), pp. 373–388.

[39] N. P. Osmolovskii, *Second-order conditions for broken extremals*, in Calculus of Variations and Optimal Control, Chapman & Hall/CRC, Boca Raton, FL, 2000, pp. 198–216.

[40] N. P. Osmolovskii, *Quadratic extremality conditions for broken extremals in the general problem of calculus of variations*, J. Math. Sci., submitted.

[41] N. P. Osmolovskii and F. Lempio, *Jacobi conditions and the Riccati equation for a broken extremal*, J. Math. Sci., 100 (2000), pp. 2572–2592.

[42] N. P. Osmolovskii and F. Lempio, *Transformation of quadratic forms to perfect squares for broken extremals*, Set-Valued Anal., 10 (2002), pp. 209-232.

[43] A. Sarychev, *First- and second-order sufficient optimality conditions for bang-bang controls*, SIAM J. Control Optim., 35 (1997), pp. 315–340.

[44] H. Schättler, *On the local structure of time-optimal bang-bang trajectories in* $\mathbb{R}^3$, SIAM J. Control Optim., 26 (1988), pp. 186–204.

[45] H. J. Sussmann, *A bang-bang theorem with bounds on the number of switchings*, SIAM J. Control Optim., 17 (1979), pp. 629–651.

[46] H. J. Sussmann, *The structure of time-optimal trajectories for single-input systems in the plane: The* C^∞ *nonsingular case*, SIAM J. Control Optim., 25 (1987), pp. 433–465.

[47] H. J. Sussmann, *The structure of time-optimal trajectories for single-input systems in the plane: The general real analytic case*, SIAM J. Control Optim., 25 (1987), pp. 868–904.

[48] V. Zeidan, *The Riccati equation for optimal control problems with mixed state-control constraints: Necessity and sufficiency*, SIAM J. Control Optim., 32 (1994), pp. 1297–1321.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 2264–2285

ON A GENERAL STRUCTURE OF THE STABILIZING CONTROLLERS BASED ON STABLE RANGE*

A. QUADRAT†

Abstract. In this paper, we prove that some stabilizing controllers of a plant, which admits a left/right-coprime factorization, have a special form where their stable and unstable parts are separated. The dimension of the unstable part depends on the algebraic concept of stable range of the ring A of SISO stable plants. Moreover, we prove that, if the stable range of A is equal to 1, then every plant—defined by a transfer matrix with entries in the quotient field of A and admitting a left/right-coprime factorization—can be stabilized by a stable controller (strong stabilization). In particular, using a result of Treil proving that the stable range of $H_\infty(\mathbb{D})$ is equal to 1, we show that every stabilizable plant—defined by a transfer matrix with entries in the quotient field of $H_\infty(\mathbb{D})$ or $H_\infty(\mathbb{C}_+)$—is strongly stabilizable and, equivalently, every couple of stabilizable plants can be simultaneously stabilized by a controller (simultaneous stabilization). Finally, using the fact that the topological stable range of $H_\infty(\mathbb{D})$ is equal to 2, a result due to Suárez, we show that every unstabilizable SISO plant—defined by a transfer function with entries in the quotient field of $H_\infty(\mathbb{D})$—is as close as we want to a stabilizable plant in the product topology.

Key words. stabilizing controllers, strong stabilization, simultaneous stabilization, stable range, k-stability, topological stable range, unit 1-stable range, n-fold rings, H_∞, K-theory

AMS subject classifications. 93D15, 93D25, 93C20, 19B10, 30D55

DOI. 10.1137/S0363012902408277

1. Introduction.

The fractional representation approach to analysis and synthesis problems was developed in the eighties in order to express in a unique mathematical framework several questions on stabilization problems. In that framework, we can study internal stabilization (existence of an internally stabilizing controller), parametrization of all stabilizing controllers, strong stabilization (possibility of stabilizing a plant by means of a stable controller), simultaneous stabilization (possibility of stabilizing a set of plants by means of a single controller), metrics of robustness (gap or graph topologies), H_∞ or H_2-optimal controllers, etc. See [2, 6, 42] for more details.

Recently, the reformulation of the fractional representation approach to analysis and synthesis problems within an *algebraic analysis approach* has allowed us to obtain new necessary and sufficient conditions for internal stabilizability and for the existence of (weakly) left/right/doubly coprime factorizations in the general setting [25, 26, 24]. Moreover, all the rings of SISO stable plants (used in this framework) over which one of the previous properties is satisfied were completely characterized [25, 26, 24]. In [27, 28], a new parametrization of all stabilizing controllers of a stabilizable plant was developed. It generalizes the Youla–Kučera parametrization [42] for stabilizable plants which do not necessarily admit doubly coprime factorizations. All these results show that a natural mathematical framework for the study of stabilization problems is the so-called *K-theory* [22, 32]. See [29] for more details.

*Received by the editors May 23, 2002; accepted for publication (in revised form) August 27, 2003; published electronically May 17, 2004. An initial French version of this paper first appeared in the *Proceedings of the Conférence Internationale Francophone d' Automatique*, Nantes, France, 2002.

http://www.siam.org/journals/sicon/42-6/40827.html

†INRIA Sophia Antipolis, CAFE, 2004 route des lucioles, BP 93, 06902 Sophia Antipolis cedex, France (Alban.Quadrat@sophia.inria.fr).

The purpose of this paper is to show that the concept of *stable range* developed in K-theory also plays an important role in the study of the strong and simultaneous stabilization problems [42]. Using the fractional representation approach to synthesis problems [6, 42], we show that, if the transfer matrix P, with entries in the quotient field of an integral domain A of SISO stable plants (e.g., $A = RH_\infty$, $H_\infty(\mathbb{C}_+)$ or W_+), admits a left-coprime factorization $P = D^{-1}N$, then there exist some stabilizing controllers of P having separated stable and unstable parts. In particular, we show that the dimension of the unstable part is related to the concept of *k-stability* of the matrix $R = (D : -N)$ with entries in A [17, 41]. Moreover, using some relations between the k-stability of a matrix with entries in A and the concept of *stable range* $\mathrm{sr}(A)$ *of* A [1, 7, 41], we prove that there exist some stabilizing controllers of P which are such that their unstable parts are defined by $\mathrm{sr}(A) - 1$ unstable rows. Therefore, if the stable range $\mathrm{sr}(A)$ of A is 1, then every transfer matrix which admits a left-coprime factorization is strongly stabilizable; i.e., it is internally stabilized by a stable controller. In particular, using the fact that the stable range of $H_\infty(\mathbb{D})$ is equal to 1 (see [38]), we prove that every stabilizable plant, defined by means of a transfer matrix with entries in the quotient field of $H_\infty(\mathbb{D})$ or $H_\infty(\mathbb{C}_+)$, is strongly stabilizable (strong stabilization). Let us notice that this result answers one of the questions asked in [9]. Moreover, using a result of Vidyasagar [42], we prove that every couple of plants, defined by transfer matrices with entries in $H_\infty(\mathbb{D})$ or $H_\infty(\mathbb{C}_+)$, is simultaneously stabilized by a controller (simultaneous stabilization). Finally, introducing the concept of *topological stable range*, we show that every unstabilizable SISO plant, defined by a transfer function $p = n/d$, with $0 \neq d,\ n \in H_\infty(\mathbb{D})$, is as close as we want to a stabilizable plant in the product topology.

Plan of the paper. In section 2, we give the definition of the stable range of a ring A and present some examples which will be used in the rest of the paper. In section 3, we introduce the concept of k-stability of a matrix with entries in a ring A. We recall the fractional representation approach to analysis and synthesis problems in section 4. In section 5, we give the first main result of this paper concerning the form of certain stabilizing controllers (Theorem 5.1) and examples in order to illustrate this result. Exploiting the relations between k-stability of a matrix with entries in a ring A and the stable range of A, we give the second main result of the paper (Corollary 6.4) and its corollaries (Corollaries 6.5 and 6.6). In the last section, we introduce the definitions of topological stable range, unit 1-stable range, and n-fold ring, and give some applications of these concepts to some stabilization problems.

Notation. A will denote a commutative ring with a unit [33], $A^{q\times p}$ the set of $q \times p$ matrices with entries in A, I_p the identity matrix of $A^{p\times p}$, and

$$\mathrm{GL}_p(A) = \{R \in A^{p\times p} \mid \exists\, S \in A^{p\times p} :\ RS = SR = I_p\}$$

the group of invertible elements of $A^{p\times p}$. If $R \in A^{q\times p}$, then $R^T \in A^{p\times q}$ is the transposed matrix. If A is an integral domain (i.e., $a\,b = 0,\ a \neq 0 \Rightarrow b = 0$), then we shall denote the *field of fractions* of A by $K = Q(A) = \{n/d \mid d \neq 0,\ n \in A\}$. Finally, p and q will always denote two positive integers satisfying $p \geq q$ ($p - q$ will denote the number of input variables for the transfer matrices) and $\triangleq$ will mean "by definition."

2. Stable range of a commutative ring.

2.1. Definition. Let us give some definitions that will be constantly used in this paper.

Definition 2.1 (see [1, 4, 7, 41]). *We have the following definitions and notation:*

- *A vector* $a = (a_1 : \cdots : a_n) \in A^{1\times n}$ *is said to be* unimodular *if there exists a vector* $b = (b_1 : \cdots : b_n) \in A^{1\times n}$ *such that* $a\,b^T = \sum_{i=1}^n a_i\,b_i = 1$.
- *We denote the set of all the unimodular vectors of* $A^{1\times n}$ *by* $\mathrm{U}_n(A)$.

Let us notice that $\mathrm{U}_1(A)$ is the set of the units $\mathrm{U}(A) = \{a \in A \mid a^{-1} \in A\}$ of A.

Example 2.1. Let us take $A = H_\infty(\mathbb{C}_+)$, where $H_\infty(\mathbb{C}_+)$ is the algebra of $\mathbb{C}$-valued holomorphic functions on the open right half plane $\mathbb{C}_+ = \{s \in \mathbb{C} \mid \mathrm{Re}\, s > 0\}$ which are bounded w.r.t. the norm $\| f \|_\infty = \sup_{s\in\mathbb{C}_+} |f(s)|$. See [5] for more details. The vector $a = (\frac{s-1}{s+1} : \frac{e^{-s}}{s+1}) \in A^{1\times 2}$ is unimodular because we have

$$\left(\tfrac{s-1}{s+1}\right)\left(1 + 2\left(\tfrac{1-e^{-(s-1)}}{s-1}\right)\right) + \left(\tfrac{e^{-s}}{s+1}\right) 2\,e = 1, \qquad 1 + 2\left(\tfrac{1-e^{-(s-1)}}{s-1}\right),\ 2\,e \in A.$$

Definition 2.2 (see [1, 4, 7, 41]). *A vector* $a = (a_1 : \cdots : a_n) \in \mathrm{U}_n(A)$ *is called* stable *(or* reductible*) if there exists an* $(n-1)$*-tuple* $b = (b_1 : \cdots : b_{n-1}) \in A^{1\times(n-1)}$ *such that*

$$(a_1 + a_n\,b_1 : \cdots : a_{n-1} + a_n\,b_{n-1}) \in \mathrm{U}_{n-1}(A);$$

i.e., there exists $(c_1 : \cdots : c_{n-1}) \in A^{1\times(n-1)}$ *such that* $\sum_{i=1}^{n-1}(a_i + a_n\,b_i)\,c_i = 1$.

Example 2.2. We have the following examples:

- Let us consider $A = H_\infty(\mathbb{C}_+)$ and $a = (1 - e^{-2s} : 1 + e^{-2s}) \in A^{1\times 2}$. We have

  (2.1)
  $$\tfrac{1}{2}\,(1 - e^{-2s}) + \tfrac{1}{2}\,(1 + e^{-2s}) = 1 \Rightarrow (1 - e^{-2s}) + (1 + e^{-2s}) = 2 \in \mathrm{U}_1(A),$$

  and thus, a is a stable vector of $\mathrm{U}_2(A)$.
- Let $A = RH_\infty = \mathbb{R}(s) \cap H_\infty(\mathbb{C}_+)$ be the $\mathbb{R}$-algebra of proper and stable real rational functions [42]. The vector

  $$a = \left(\tfrac{(s-1)(s-2)}{(s+1)^2} : \tfrac{s}{(s+1)^2}\right) \in A^{1\times 2}$$

  is stable because we have

  (2.2)
  $$\tfrac{(s-1)(s-2)}{(s+1)^2} + \tfrac{6\,s}{(s+1)^2} = \tfrac{(s+2)}{(s+1)} \in \mathrm{U}_1(A).$$

Remark 2.1. If a vector $(a_1 : a_2) \in \mathrm{U}_2(A)$ is stable, then, in general, this is not the case for $(a_2 : a_1) \in \mathrm{U}_2(A)$. For instance, if $A = \mathbb{R}[s]$, then $(s^2 + 1 : s) \in A^{1\times 2}$ is a stable vector because we have $(s^2 + 1) + s\,(-s) = 1 \in \mathrm{U}_1(A)$, whereas the vector $(s : s^2 + 1) \in U_2(A)$ is not stable because there does not exist $b \in A$ such that $r \triangleq s + (s^2 + 1)\,b(s) \in A$ is invertible, i.e., is a nonzero real constant (the degree of the polynomial r is at least 1).

Definition 2.3 (see [31, 34, 38, 41]). *We call the* stable range $\mathrm{sr}(A)$ *of* A *the smallest* $n \in \mathbb{N} \cup \{+\infty\}$ *such that every vector of* $\mathrm{U}_{n+1}(A)$ *is stable.*

Let us notice that the stable range $\mathrm{sr}(A)$ is also called the *stable rank* of A.

Remark 2.2. Let us notice that if $\mathrm{sr}(A) = n$, then, for $m \geq n$, every element of $U_m(A)$ is stable [11]. Indeed, if $(a_1 : \cdots : a_{n+2}) \in U_{n+2}(A)$, then there exist $b_1, \ldots, b_{n+2} \in A$ such that $\sum_{i=1}^{n+2} a_i\,b_i = 1$. Hence, the vector

$$(a_1 : \cdots : a_n : a_{n+1}\,b_{n+1} + a_{n+2}\,b_{n+2}) \in A^{1\times(n+1)}$$

is unimodular. Using the fact that $\mathrm{sr}(A) = n$, there exist $c_1, \ldots, c_n \in A$ such that the vector

$$(a_1 + c_1\,(a_{n+1}\,b_{n+1} + a_{n+2}\,b_{n+2}) : \cdots : a_n + c_n\,(a_{n+1}\,b_{n+1} + a_{n+2}\,b_{n+2})) \in A^{1\times n}$$

is unimodular; i.e., there exist $d_1, \ldots, d_n \in A$ such that

$$\sum_{i=1}^{n} (a_i + c_i\,(a_{n+1}\,b_{n+1} + a_{n+2}\,b_{n+2}))\,d_i = 1$$

$$\Rightarrow \sum_{i=1}^{n} (a_i + c_i\,a_{n+2}\,b_{n+2})\,d_i + a_{n+1}\left(\sum_{i=1}^{n} b_{n+1}\,c_i\,d_i\right) = 1,$$

which shows that $(a_1 + (c_1\,b_{n+2})\,a_{n+2} : \cdots : a_n + (c_n\,b_{n+2})\,a_{n+2} :\ a_{n+1})$ is unimodular, and thus the vector $(a_1 : \cdots : a_{n+2}) \in U_{n+2}(A)$ is a stable vector. The result directly follows by induction on n.

Example 2.3. We have the following interpretations of $\mathrm{sr}(A) = 2$ and $\mathrm{sr}(A) = 1$:

- A ring A has a stable range $\mathrm{sr}(A) = 2$ iff, $\forall\, n \geq 3$, every element of $\mathrm{U}_n(A)$ is stable and there exists a vector $(a_1 :\ a_2) \in \mathrm{U}_2(A)$ such that, for every $b \in A$, $a_1 + a_2\,b \notin \mathrm{U}_1(A)$, i.e., $a_1 + a_2\,b$ is not invertible.
- A ring A has a stable range $\mathrm{sr}(A) = 1$ iff, for every $(a_1 :\ a_2) \in \mathrm{U}_2(A)$, there exists $b \in A$ such that $a_1 + a_2\,b \in \mathrm{U}_1(A)$, i.e., $a_1 + a_2\,b$ is invertible.

2.2. Examples.

THEOREM 2.4 (see [38]). *If $\mathbb{D}$ denotes the open unit disc and $H_\infty(\mathbb{D})$ the ring of $\mathbb{C}$-valued holomorphic functions on $\mathbb{D}$ which are bounded w.r.t. the norm $\| f \|_\infty = \sup_{z\in\mathbb{D}} |f(z)|$, then we have*

$$\mathrm{sr}(H_\infty(\mathbb{D})) = 1.$$

COROLLARY 2.5. *With the notation of Example* 2.1*, we have*

$$\mathrm{sr}(H_\infty(\mathbb{C}_+)) = 1.$$

Proof. Let us consider a unimodular matrix $a = (a_1 :\ a_2) \in U_2(H_\infty(\mathbb{C}_+))$. Let us denote by $(b_1 :\ b_2)^T \in H_\infty(\mathbb{C}_+)^{2\times 1}$ a right-inverse of a; i.e., we have

$$a_1(s)\,b_1(s) + a_2(s)\,b_2(s) = 1. \tag{2.3}$$

The fractional linear transformation $s = \psi(z) = (1+z)/(1-z)$ bijectively maps the open unit disc $\mathbb{D}$ on the open right half plane $\mathbb{C}_+$ and $z = \psi^{-1}(s) = (s-1)/(s+1)$. Moreover, from Lemma A.6.15 of [5], we have $f \in H_\infty(\mathbb{C}_+) \Leftrightarrow f \circ \psi \in H_\infty(\mathbb{D})$. Thus, from (2.3), we deduce

$$(a_1 \circ \psi)(z)\,(b_1 \circ \psi)(z) + (a_2 \circ \psi)(z)\,(b_2 \circ \psi)(z) = 1 \circ \psi = 1, \tag{2.4}$$

i.e., $(a_1 \circ \psi :\ a_2 \circ \psi) \in \mathrm{U}_2(\mathrm{H}_\infty(\mathbb{D}))$. By Theorem 2.4, we know that $\mathrm{sr}(H_\infty(\mathbb{D})) = 1$, and thus there exist $c,\, d \in H_\infty(\mathbb{D})$ such that

$$((a_1 \circ \psi)(z) + (a_2 \circ \psi)(z)\,c(z))\,d(z) = 1 \Leftrightarrow (a_1(s) + a_2(s)\,c(\psi^{-1}(s)))\,d(\psi^{-1}(s)) = 1;$$

i.e., $a = (a_1 :\ a_2)$ is 1-stable, and thus $\mathrm{sr}(H_\infty(\mathbb{C}_+)) = 1$. ∎

THEOREM 2.6 (see [1]). *If A is a principal ideal domain, namely, an integral domain such that every ideal of A can be generated by a single element of A, then* $\mathrm{sr}(A) \leq 2$.

COROLLARY 2.7. *Let RH_∞ be the ring of proper and stable real rational functions. Then, we have*

$$\mathrm{sr}(RH_\infty) = 2.$$

Proof. It is well known that RH_∞ is a principal ideal domain [42]. Therefore, by Theorem 2.6, we obtain that $\mathrm{sr}(RH_\infty) \leq 2$. Finally, let $(d:\ n) \in \mathrm{U}_2(RH_\infty)$ with $d \neq 0$ and let us define the transfer function $P = n/d \in \mathbb{R}(s) = Q(RH_\infty)$. Let us notice that $P = n/d$ is a coprime factorization of P because $(d:\ n) \in \mathrm{U}_2(RH_\infty)$. Now, it is also well known that there exists $c \in RH_\infty$ such that $d + c\,n$ is a unit of RH_∞ iff P has the *parity interlacing property* [2, 42], namely, P has an even number of real poles between every pair of real zeros in $\{\mathrm{Re}\, s \geq 0\} \cup \{\infty\}$. Hence, there exist vectors $(d:\ n) \in \mathrm{U}_2(RH_\infty)$ which are not stable in the sense of Definition 2.2 (e.g., $((s-1)/(s+1):\ s/(s+1)^2) \in \mathrm{U}_2(RH_\infty)$ is not stable because the transfer function $P = s/((s+1)\,(s-1))$ does not have the parity interlacing property—see Example 4 of section 3.2 of [42]). Therefore, we have $\mathrm{sr}(RH_\infty) = 2$. □

Let us give more examples of stable ranges of integral domains.

THEOREM 2.8. *We have the following results:*

- [12, 41] $\mathrm{sr}(\mathbb{R}[x_1, \ldots, x_n]) = n + 1$.
- [19] *The ring of entire functions*

$$E(k) = \left\{ f(s) = \sum_{n=0}^{+\infty} a_n\, s^n \;\middle|\; s \in \mathbb{C},\ a_n \in k,\ \lim_{n \to +\infty} |a_n|^{1/n} = 0 \right\}$$

  satisfies $\mathrm{sr}(E(k)) = 1$ *if* $k = \mathbb{C}$ *and* 2 *if* $k = \mathbb{R}$.
- [20] *The disc algebra* $A(\mathbb{D})$*, i.e., the ring of functions which are holomorphic in the open unit disc* $\mathbb{D}$ *and continuous on the unit circle* $\mathbb{T}$*, satisfies* $\mathrm{sr}(A(\mathbb{D})) = 1$.
- [34] *If we denote by* W_+ *the Wiener algebra defined by*

$$W_+ = \left\{ \sum_{n=0}^{+\infty} a_n\, z^n \;\middle|\; \sum_{n=0}^{+\infty} |\, a_n \,| < +\infty \right\},$$

  then we have $\mathrm{sr}(W_+) = 1$.

Let us recall that the polynomial ring $\mathbb{R}[x_1, \ldots, x_n]$ is used in the study of multidimensional systems, W_+ represents the sets of l_∞-stable (bounded input bounded output stability) shift-invariant causal digital filters [42], and the disc algebra $A(\mathbb{D})$ is used for interpolation problems and discrete-time control systems [42]. Finally, $E(\mathbb{R})$ is used in the study of a certain class of time-delay systems $\mathcal{E} = E(\mathbb{R}) \cap \mathbb{R}(s)[e^{-s}]$ [21].

3. *k*-stability for matrices. Let us extend the definition of k-stability for matrices with entries in A.

DEFINITION 3.1 (see [11, 17, 41]). *A matrix* $R \in A^{q \times p}$ *is* unimodular *if there exists a matrix* $S \in A^{p \times q}$ *such that* $R\,S = I_q$*, i.e.,* R *has a right-inverse* S.

Remark 3.1. First, let us notice that the previous concept of a unimodular matrix is standard in commutative algebra, whereas, in control theory, a unimodular matrix usually denotes a square matrix $R \in A^{p \times p}$ such that there exists $S \in A^{p \times p}$ satisfying $R\,S = S\,R = I_p$. The reader should be careful not to confuse these two different definitions (only Definition 3.1 will be used in the course of the paper).

Second, if $R \in A^{q \times p}$ is a unimodular matrix, then it is clear that R has *full row rank*, namely its rows are A-linearly independent. Moreover, the A-submodule $A^{1 \times q}\, R$ of $A^{1 \times p}$ generated by the A-linear combinations of the rows of R is isomorphic to $A^{1 \times q}$, and thus we have $1 \leq q \leq p$.

If $R_i \in A^{q \times 1}$ is a column vector, then we shall denote by $\mathrm{col}(R_1, \ldots, R_p)$ the $q \times p$ matrix R whose first column is R_1, whose second one is $R_2, \ldots$, and whose last column is R_p.

LEMMA 3.2. *$R = \mathrm{col}(R_1 : \cdots : R_p) \in A^{q\times p}$ is unimodular iff the A-module*

$$R\,A^p \triangleq \sum_{i=1}^{p} R_i\,A = \left\{ \sum_{i=1}^{p} R_i\,a_i \in A^q \mid a_i \in A \right\}$$

is equal to A^q.

Proof. $\Rightarrow$ Let R be unimodular. Then there exists $S \in A^{p\times q}$ such that $R\,S = I_q$. Therefore, for every $\lambda \in A^q$, the vector $\mu = S\,\lambda \in A^p$ is such that $\lambda = R\,\mu$, and thus $\lambda = \sum_{i=1}^{p} R_i\,\mu_i \in R\,A^p$, where $\mu = (\mu_1 : \cdots : \mu_p)^T$. Hence, we have $R\,A^p = A^q$.

$\Leftarrow$ Let us suppose that $R\,A^p = A^q$. Then, for every $\lambda \in A^q$, there exists $(a_i)_{1\leq i\leq p}$, with $a_i \in A$, such that $\lambda = \sum_{i=1}^{p} R_i\,a_i$. In particular, for $j = 1, \ldots, q$, let us consider the vector e_j of A^q defined by 1 in the jth component and 0 elsewhere. Then, for $j = 1, \ldots, q$, there exists $S_j \in A^p$ such that $e_j = R\,S_j$, and thus, if we define $S = \mathrm{col}(S_1 : \cdots : S_q) \in A^{p\times q}$, then we have $R\,S = I_q$; i.e., R is unimodular. □

Let us introduce the concept of k-stability for unimodular matrices.

DEFINITION 3.3 (see [17, 41]). *A unimodular matrix $R = \mathrm{col}(R_1, \ldots, R_p) \in A^{q\times p}$ is called* k-stable $(1 \leq k \leq p - q)$ *if there exists a $(p-k)$-tuple $(c_i)_{1\leq i\leq p-k}$ belonging to the A-module*

$$R_{p-k+1}\,A + \cdots + R_p\,A \triangleq \left\{ \sum_{i=1}^{k} R_{p-k+i}\,b_i \mid b_i \in A \right\} \tag{3.1}$$

such that the matrix

$$\mathrm{col}(R_1 + c_1 : R_2 + c_2 : \cdots : R_{p-k} + c_{p-k}) \in A^{q\times(p-k)}$$

is unimodular.

Remark 3.2. Let us notice that a vector $a \in \mathrm{U}_n(A)$ is 1-stable iff a is stable in the sense of Definition 2.2.

LEMMA 3.4. *A unimodular matrix $R \in A^{q\times p}$ is k-stable iff there exists a matrix $T_k \in A^{k\times(p-k)}$ such that the matrix*

$$R_k = \mathrm{col}(R_1 : \cdots : R_{p-k}) + \mathrm{col}(R_{p-k+1} : \cdots : R_p)\,T_k \in A^{q\times(p-k)} \tag{3.2}$$

is unimodular.

Proof. $\Rightarrow$ Let R be a k-stable matrix; then there exists a $(p-k)$-tuple $(c_i)_{1\leq i\leq p-k}$ of elements of the A-module (3.1) such that $\mathrm{col}(R_1 + c_1 : \cdots : R_{p-k} + c_k) \in A^{q\times(p-k)}$ is a unimodular matrix. By definition of the c_i, there exists $b_{ij} \in A$ such that

$$c_i = \sum_{j=1}^{k} R_{p-k+j}\,b_{i(p-k+j)}.$$

Therefore, we have

$$\mathrm{col}(R_1 + c_1 : \cdots : R_{p-k} + c_k) = \mathrm{col}(R_1 : \cdots : R_{p-k}) + \mathrm{col}(R_{p-k+1} : \cdots : R_p)\,T_k,$$

where $T_k \in A^{k\times(p-k)}$ is defined by

$$T_k = \begin{pmatrix} b_{1(p-k+1)} & b_{2(p-k+1)} & \cdots & b_{(p-k)(p-k+1)} \\ \vdots & \vdots & & \vdots \\ b_{1p} & b_{2p} & \cdots & b_{(p-k)p} \end{pmatrix}.$$

$\Leftarrow$ All the columns c_i of the matrix $\mathrm{col}(R_{p-k+1} : \cdots : R_p)\, T_k$ belong to the A-module (3.1). Thus, R_k has the form $\mathrm{col}(R_1 + c_1 : \cdots : R_{p-k} + c_k)$; i.e., R is k-stable. □

Example 3.1. Let us consider $A = RH_\infty$ and the following matrix:

$$R = \begin{pmatrix} \frac{s-1}{s+1} & 0 & -\frac{1}{s+1} \\ \frac{1}{s+1} & -\frac{s}{s+1} & 0 \end{pmatrix} \in A^{2\times 3}.$$

The matrix

$$R_1 = \begin{pmatrix} \frac{s+2}{s+1} & \frac{1}{s+1} \\ \frac{1}{s+1} & -\frac{s}{s+1} \end{pmatrix} = \begin{pmatrix} \frac{s-1}{s+1} & 0 \\ \frac{1}{s+1} & -\frac{s}{s+1} \end{pmatrix} + \begin{pmatrix} -\frac{1}{s+1} \\ 0 \end{pmatrix} (-3 : -1) \tag{3.3}$$

is invertible ($\det R_1 = -1$), and thus R is 1-stable.

PROPOSITION 3.5. *If R is k-stable, then R is $(k-1)$-stable.*

Proof. Using the fact that R is k-stable, then there exist

$$c_1, \ldots, c_{p-k} \in R_{p-k+1}\, A + \cdots + R_p\, A$$

such that $R_k = \mathrm{col}(R_1 + c_1 : \cdots : R_{p-k} + c_{p-k})$ is unimodular. Let us decompose c_i as $c_i = d_i + e_i$, where $d_i \in R_{p-k+1}\, A$ and $e_i \in R_{p-k+2}\, A + \cdots + R_p\, A$, and let us define $R_{k+1} = \mathrm{col}(R_1 + e_1 : \cdots : R_{p-k} + e_{p-k} : R_{p-k+1})$. Then we claim that R_{k+1} is unimodular, and thus R is $(k-1)$-stable. Indeed, we have

$$\sum_{i=1}^{p-k} (R_i + c_i)\, A \subseteq \sum_{i=1}^{p-k} (R_i + e_i)\, A + R_{p-k+1}\, A \subseteq A^q.$$

Then, applying Lemma 3.2 to R_k, we obtain that $\sum_{i=1}^{p-k} (R_i + c_i)\, A = A^q$, and thus $\sum_{i=1}^{p-k} (R_i + e_i)\, A + R_{p-k+1}\, A = A^q$, which proves that R_{k+1} is unimodular by Lemma 3.2. □

4. Internal stabilization. Let A be an integral domain and let its *field of fractions* be

$$K = Q(A) = \{n/d \mid n \in A,\, 0 \neq d \in A\}.$$

In *the fractional representation approach to analysis and synthesis problems* [5, 6, 42], we consider a class of plants which are defined by means of transfer matrices whose entries belong to the quotient field $K = Q(A)$ of an integral domain of stable SISO plants (see [25, 26, 24, 27] for more details).

Example 4.1. We have the following examples of algebras of SISO stable plants:

- For finite-dimensional systems, we usually consider the integral domain of proper and stable real rational functions $A = RH_\infty = \mathbb{R}(s) \cap H_\infty(\mathbb{C}_+)$ and $K = \mathbb{R}(s)$ [42]. Then, A corresponds to the set of proper and stable real rational transfer functions, whereas an element of $K \backslash A$ represents either an unstable or an improper transfer function. For instance,

  $$P = s/((s-1)\,(s-2)) \in \mathbb{R}(s)$$

  belongs to $K = Q(A)$ because we have $P = n/d$, where $n = s/(s+1)^2 \in A$ by $d = ((s-1)\,(s-2))/(s+1)^2 \in A$.

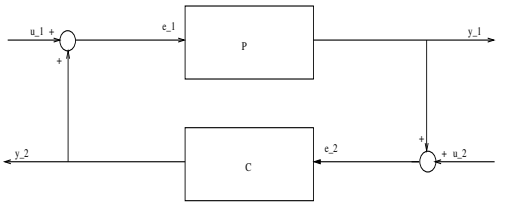


FIG. 4.1. *Closed-loop.*

- For infinite-dimensional systems, we can consider $A = H_\infty(\mathbb{C}_+)$ [36], which gives a class of unstable plants defined by transfer matrices with entries in the quotient field $K = Q(H_\infty(\mathbb{C}_+))$. For instance, the transfer function

$$P = (1 + e^{-2\,s})/(1 - e^{-2\,s})$$

of a wave equation (see, e.g., Exercise 4.24 of [5]) satisfies $P = n/d$, where $n = 1 + e^{-2\,s} \in A$ and $d = 1 - e^{-2\,s} \in A$, and thus we have $P \in K$.

Let us consider a plant defined by the transfer matrix $P \in K^{q\times(p-q)}$, a controller defined by $C \in K^{(p-q)\times q}$, and the closed-loop given by Figure 4.1. We have the following equations:

$$\begin{pmatrix} I_{p-q} & -C \\ -P & I_q \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}.$$

DEFINITION 4.1 (see [5, 6, 42]). *A plant defined by the transfer matrix* $P \in K^{q\times(p-q)}$ *is* internally stabilizable *if there exists a controller* $C \in K^{(p-q)\times q}$ *such that all the entries of the matrix*

$$\begin{pmatrix} I_{p-q} & -C \\ -P & I_q \end{pmatrix}^{-1} = \begin{pmatrix} (I_{p-q} - C\,P)^{-1} & (I_{p-q} - C\,P)^{-1}\,C \\ P\,(I_{p-q} - C\,P)^{-1} & I_q + P\,(I_{p-q} - C\,P)^{-1}\,C \end{pmatrix} \tag{4.1}$$

$$= \begin{pmatrix} I_{p-q} + C\,(I_q - P\,C)^{-1}\,P & C\,(I_q - P\,C)^{-1} \\ (I_q - P\,C)^{-1}\,P & (I_q - P\,C)^{-1} \end{pmatrix} \tag{4.2}$$

belong to A. *Such a controller,* $C \in K^{(p-q)\times q}$, *is called a* stabilizing controller *of* P.

Example 4.2. The controller $C = -(s-1)/(s+1)$ is not a stabilizing controller of the plant $P = s/(s-1)$ because we have

$$\begin{cases} e_1 = \frac{(s+1)}{(2\,s+1)}\,u_1 + \frac{(-s+1)}{(2\,s+1)}\,u_2, \\ e_2 = \frac{s\,(s+1)}{(2\,s+1)(s-1)}\,u_1 + \frac{(s+1)}{(2\,s+1)}\,u_2, \end{cases}$$

and the transfer function between e_2 and u_1 has the unstable pole 1; i.e., it does not belong to RH_∞.

DEFINITION 4.2. *We have the following definitions* [5, 6, 42]:

- *A transfer matrix* $P \in K^{q\times(p-q)}$ *admits a* left-coprime factorization *if there exist* $R = (D : -N) \in A^{q\times p}$ *and* $S = (X^T : Y^T)^T \in A^{p\times q}$ *such that*

$$\begin{cases} P = D^{-1}\,N, \\ R\,S = D\,X - N\,Y = I_q. \end{cases}$$

- *A transfer matrix* $P \in K^{q\times(p-q)}$ *admits a* right-coprime factorization *if there exist* $\tilde{R} = (\tilde{N}^T : \tilde{D}^T)^T \in A^{p\times(p-q)}$ *and* $\tilde{S} = (-\tilde{Y} : \tilde{X}) \in A^{(p-q)\times p}$ *such that*

$$\left\{ \begin{array}{l} P = \tilde{N}\,\tilde{D}^{-1}, \\ \tilde{S}\,\tilde{R} = -\tilde{Y}\,\tilde{N} - \tilde{X}\,\tilde{D} = I_{p-q}. \end{array} \right.$$

- *A transfer matrix* $P \in K^{q\times(p-q)}$ *admits a* doubly coprime factorization *if* P *admits both a left and right-coprime factorization.*

PROPOSITION 4.3 (see [42, Theorem 25, p. 105]). *Every transfer matrix* $P \in K^{q\times(p-q)}$ *which admits a left-coprime factorization* $P = D^{-1}\,N$, $D\,X - N\,Y = I_q$, $\det X \neq 0$, *is internally stabilized by the controller* $C = Y\,X^{-1}$.

If $P = D_1^{-1}\,N_1 = D_2^{-1}\,N_2$ are two left-coprime factorizations of P and $R_i = (D_i : -N_i)$, for $i = 1, 2$, then there exists a matrix $U \in \mathrm{GL}_q(A)$ such that $R_2 = U\,R_1$. Hence, we deduce that R_q is k-stable iff R_2 is k-stable. A similar result also holds for right-coprime factorizations.

DEFINITION 4.4. *We have the following definitions* [2, 42]:

- *A plant* $P \in K^{q\times(p-q)}$ *is* strongly stabilizable *if there exists a stable controller* $C \in A^{(p-q)\times q}$ *which internally stabilizes* P.
- *Two plants* P_1, $P_2 \in K^{q\times(p-q)}$ *are* simultaneously stabilizable *if there exists a controller* $C \in K^{(p-q)\times q}$ *which internally stabilizes* P_1 *and* P_2.

The next proposition is a reformulation of Lemma 7 of section 5.3 of [42] (we thank an anonymous associate editor for pointing out this reference to us).

PROPOSITION 4.5. *A transfer matrix* $P \in K^{q\times(p-q)}$ *is strongly stabilizable iff* P *admits a doubly coprime factorization* $P = D^{-1}\,N = \tilde{N}\,\tilde{D}^{-1}$ *such that the matrices* $(D : -N) \in A^{q\times p}$ *and* $(\tilde{D}^T : \tilde{N}^T) \in A^{(p-q)\times p}$ *are, respectively,* $(p-q)$ *and* q*-stable.*

In particular, $P \in K(A)$ *is strongly stabilizable iff there exists a coprime factorization* $P = n/d$ *such that the vector* $(d : n) \in U_2(A)$ *is* 1*-stable.*

Proof. Let us suppose that there exists a stable controller $C \in A^{(p-q)\times q}$ which internally stabilizes P. Then, all the entries of the matrix (4.1) belong to A and, in particular, $P\,(I_{p-q} - C\,P)^{-1} = (I_q - P\,C)^{-1}\,P = V \in A^{q\times(p-q)}$.

Then, from the fact that

$$I_{p-q} + C\,V = I_{p-q} + C\,(I_q - P\,C)^{-1}\,P = (I_{p-q} - C\,P)^{-1},$$

we deduce that $I_{p-q} + C\,V$ is an invertible matrix, and thus we have

$$P\,(I_{p-q} - C\,P)^{-1} = V \Leftrightarrow P = V\,(I_{p-q} + C\,V)^{-1}.$$

Then, P admits the right-coprime factorization $P = V\,(I_{p-q} + C\,V)^{-1}$ because

$$(-C : I_{p-q}) \left(\begin{array}{c} V \\ I_{p-q} + C\,V \end{array} \right) = I_{p-q}.$$

The matrix $((I_{p-q} + C\,V)^T : V^T)$ is q-stable because $I_{p-q} + V^T\,C^T - V^T\,C^T = I_{p-q}$.

Moreover, from the fact that

$$I_q + V\,C = I_q + P\,(I_{p-q} - C\,P)^{-1}\,C = (I_q - P\,C)^{-1},$$

we deduce that $I_q + V\,C$ is an invertible matrix, and thus we have

$$(I_q - P\,C)^{-1}\,P = V \Leftrightarrow P = (I_q + V\,C)^{-1}\,V.$$

Then, P admits the left-coprime factorization $P = (I_q + V\,C)^{-1}\,V$, and the matrix $(I_q + V\,C : -V)$ satisfies $I_q + V\,C - V\,C = I_q$; i.e., $(I_q + V\,C : -V)$ is $(p-q)$-stable.

Conversely, if P admits a left-coprime factorization $P = D^{-1}\,N$ such that the matrix $R = (D : -N) \in A^{q\times p}$ is $(p-q)$-stable, then there exists $T_1 \in A^{(p-q)\times q}$ such that $U \triangleq D - N\,T_1 \in \mathrm{GL}_q(A)$. In particular, we have $D\,U^{-1} - N\,(T_1\,U^{-1}) = I_q$, where $U^{-1} \in A^{q\times q}$. Thus, by Proposition 4.3, $C = (T_1\,U^{-1})\,(U^{-1})^{-1} = T_1$ is a stable controller which internally stabilizes P, and thus P is strongly stabilizable. □

5. A general structure of the stabilizing controllers. In the next theorem, we show that there exists a stabilizing controller C of P such that the dimension of its unstable part depends on the k-stability of the matrix $R = (D : -N) \in A^{q\times p}$, where $P = D^{-1}\,N$ is a left-coprime factorization of P. Moreover, the unstable part of C is isolated into a single transfer matrix $V\,U^{-1} \in K^{r\times(p-q)}$, where $r = p - q - k$.

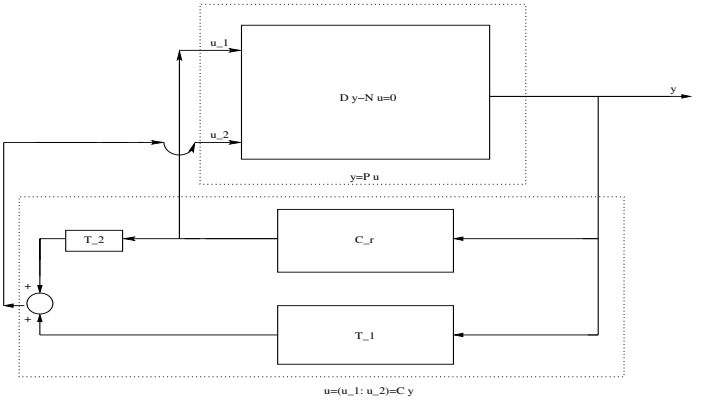


FIG. 5.1. *Closed-loop $y = P\,u$ and $u = C\,y$.*

THEOREM 5.1. *Let A be an integral domain of SISO stable plants, $K = Q(A)$, and let $P \in K^{q\times(p-q)}$ be a transfer matrix admitting a left-coprime factorization $P = D^{-1}\,N$ with $R = (D : -N) \in A^{q\times p}$. If R is k-stable and $r \triangleq p - q - k \geq 0$, then there exist two stable matrices*

$$\left\{ \begin{array}{l} T_1 \in A^{k\times q}, \\ T_2 \in A^{k\times r} \end{array} \right. \tag{5.1}$$

such that the matrix $R_k = (D - \Lambda\,T_1 : -(N_r + \Lambda\,T_2)) \in A^{q\times(p-k)}$ admits a right-inverse with entries in A, with the notation

$$R = (D : -N) = (\underset{\leftrightarrow\ q}{D} \quad : \underset{\leftrightarrow\ r}{-N_r} \quad : \underset{\leftrightarrow\ k}{-\Lambda}) \quad \in A^{q\times p}. \tag{5.2}$$

Let us define by $S_k = (U^T : V^T)^T \in A^{(p-k)\times q}$, $U \in A^{q\times q}$, $V \in A^{r\times q}$ any right-inverse of R_k such that $\det U \neq 0$. Then, the controller $C \in K^{(p-q)\times q}$ defined by

$$C = \left(\begin{array}{c} V\,U^{-1} \\ T_1 + T_2\,(V\,U^{-1}) \end{array} \right), \quad \begin{array}{l} \updownarrow\ r = p - q - k \\ \updownarrow\ k \end{array} \tag{5.3}$$

internally stabilizes P (see Figure 5.1). Moreover, if $\det(D - \Lambda\,T_1) \neq 0$, then the controller $C_r = V\,U^{-1} \in K^{r\times q}$ internally stabilizes the plant

$$P_r = (D - \Lambda\,T_1)^{-1}\,(N_r + \Lambda\,T_2) \in K^{q\times r} \tag{5.4}$$

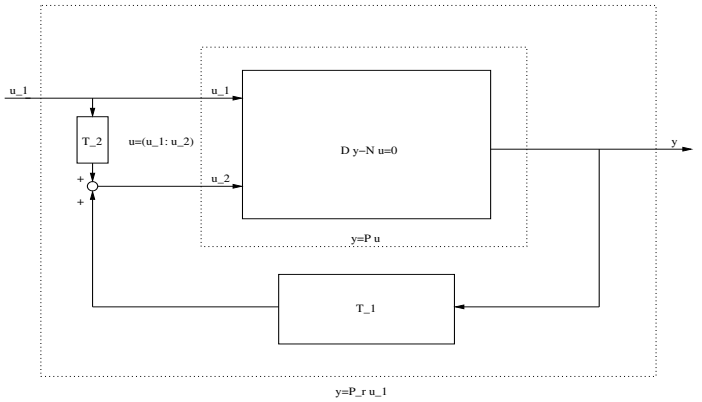


FIG. 5.2. *Plant* $y = P_r\, u_1$.

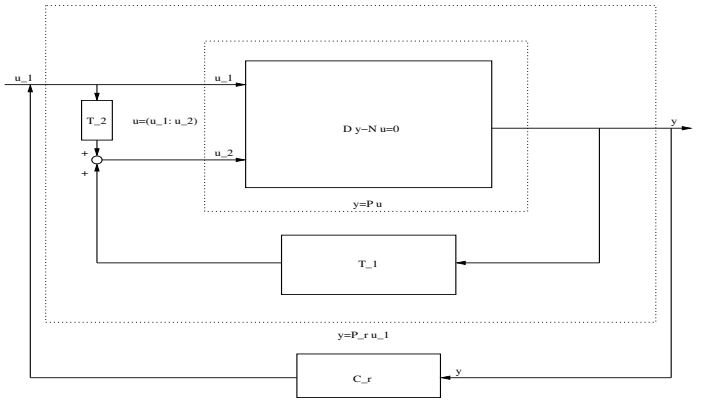


FIG. 5.3. *Closed-loop* $y = P_r\, u_1$ *and* $u_1 = C_r\, y$.

(see Figures 5.2 *and* 5.3*). The unstable part of the controller* (5.3) *corresponds to* $C_r = V\, U^{-1}$*, and its dimension is equal to* $r \times q$*.*

Similar results also hold for a transfer matrix P *admitting a right-coprime factorization* $P = \tilde{N}\, \tilde{D}^{-1}$ $(\tilde{R} = (\tilde{N}^T :\ \tilde{D}^T)^T \in A^{p\times(p-q)})$.

Proof. P admits a left-coprime factorization $P = D^{-1}\, N$, and thus the matrix $R = (D : -N) \in A^{q\times p}$ has a right-inverse $S = (X^T : Y^T)^T \in A^{p\times q}$; i.e., R is unimodular in the sense of Definition 3.1. Also, by hypothesis, R is k-stable, and thus, by Lemma 3.4, there exists $T_k \in A^{k\times(p-k)}$ such that the matrix $R_k \in A^{q\times(p-k)}$ defined by (3.2) is unimodular. Let us denote by $S_k \in A^{(p-k)\times q}$ a right-inverse of R_k; i.e., we have

$$R_k\, S_k = I_q. \tag{5.5}$$

Using expressions (3.2) and (5.5), we obtain that

$$\mathrm{col}(R_1 : \cdots : R_p) \begin{pmatrix} S_k \\ T_k\, S_k \end{pmatrix} = I_q \Leftrightarrow (D :\ -N) \begin{pmatrix} S_k \\ T_k\, S_k \end{pmatrix} = I_q.$$

If we write $S_k = (U_k^T : V_k^T)^T$, with $U_k \in A^{q\times q}$ and $V_k \in A^{r\times q}$, then we have

$$D\,U_k - N \begin{pmatrix} V_k \\ T_k\,S_k \end{pmatrix} = I_q.$$

If $\det U \neq 0$, then by Proposition 4.3, the controller C defined by

$$\begin{aligned} C &= \begin{pmatrix} V_k \\ T_k\,S_k \end{pmatrix} U_k^{-1} = \begin{pmatrix} V_k\,U_k^{-1} \\ T_k \begin{pmatrix} U_k \\ V_k \end{pmatrix} U_k^{-1} \end{pmatrix} \\ &= \begin{pmatrix} V_k\,U_k^{-1} \\ T_k \begin{pmatrix} I_q \\ V_k\,U_k^{-1} \end{pmatrix} \end{pmatrix} = \begin{pmatrix} V_k\,U_k^{-1} \\ T_{k1} + T_{k2}\,(V_k\,U_k^{-1}) \end{pmatrix} \end{aligned}$$

internally stabilizes $P = D^{-1}\,N$, where $T_k = (T_{k1} : T_{k2}) \in A^{k\times(q+r)}$ and the dimensions of T_{k1} and T_{k2} are defined by (5.1). With the notation of (5.2), we have

$$\begin{aligned} R_k &= \operatorname{col}(R_1 : \cdots : R_{p-k}) - \Lambda\,(T_{k1} : T_{k2}) \\ &= (\operatorname{col}(R_1 : \cdots : R_q) - \Lambda\,T_{k1} : \operatorname{col}(R_{q+1} : \cdots : R_{p-k}) - \Lambda\,T_{k2}) \\ &= (D - \Lambda\,T_{k1} : -(N_r + \Lambda\,T_{k2})). \end{aligned}$$

Using the fact that $R_k\,S_k = I_q$, by Proposition 4.3, we obtain that $C_r = V_k\,U_k^{-1}$ is a stabilizing controller of the plant $P_r = (D - \Lambda\,T_{k1})^{-1}\,(N_r + \Lambda\,T_{k2})$. □

Example 5.1. Let us consider $A = H_\infty(\mathbb{C}_+)$ and the following transfer matrix:

$$P = \begin{pmatrix} \frac{e^{-s}}{s+1} & \frac{s-1}{s+1} \\ 0 & \frac{1}{s-1} \end{pmatrix} \in K^{2\times 2},$$

where $K = Q(A)$. In [25, 26], it is shown that P admits the left-coprime factorization $P = D^{-1}\,N$, where $R = (D : -N) \in A^{2\times 4}$ is defined by

$$R = \begin{pmatrix} 1 & 0 & -\frac{e^{-s}}{s+1} & -\frac{s-1}{s+1} \\ 0 & \frac{s-1}{s+1} & 0 & -\frac{1}{s+1} \end{pmatrix}.$$

The matrix R_1, defined by

$$\begin{aligned} R_1 &= \begin{pmatrix} 1 & 0 & -\frac{e^{-s}}{s+1} \\ 0 & \frac{s-1}{s+1} & 0 \end{pmatrix} + \begin{pmatrix} -\frac{s-1}{s+1} \\ -\frac{1}{s+1} \end{pmatrix} \underbrace{\begin{pmatrix} 0 & -2 & 0 \end{pmatrix}}_{T_1} \\ &= \begin{pmatrix} 1 & 2\,\frac{(s-1)}{(s+1)} & -\frac{e^{-s}}{s+1} \\ 0 & 1 & 0 \end{pmatrix}, \end{aligned} \tag{5.6}$$

is unimodular because we have

$$\begin{pmatrix} 1 & 2\,\frac{(s-1)}{(s+1)} & -\frac{e^{-s}}{s+1} \\ 0 & 1 & 0 \end{pmatrix} \underbrace{\begin{pmatrix} 1 - \frac{e^{-s}}{s+1} & -2\,\frac{(s-1)}{(s+1)} \\ 0 & 1 \\ -1 & 0 \end{pmatrix}}_{S_1} = I_2. \tag{5.7}$$

Thus, the matrix R is 1-stable, and we can apply Theorem 5.1 to P with $p=4$, $q=2$, $k=1$, and $r=1$. We know that $(S_1^T : (T_1\, S_1)^T)^T$ is a left inverse of R; i.e., we have

$$\begin{pmatrix} 1 & 0 & -\frac{e^{-s}}{s+1} & -\frac{s-1}{s+1} \\ 0 & \frac{s-1}{s+1} & 0 & -\frac{1}{s+1} \end{pmatrix} \begin{pmatrix} 1-\frac{e^{-s}}{s+1} & -2\,\frac{(s-1)}{(s+1)} \\ 0 & 1 \\ -1 & 0 \\ 0 & -2 \end{pmatrix} = I_3. \tag{5.8}$$

If we define

$$U_1 = \begin{pmatrix} 1-\frac{e^{-s}}{s+1} & -2\,\frac{(s-1)}{(s+1)} \\ 0 & 1 \end{pmatrix}, \quad V_1 = (-1 :\ 0), \quad T_{11} = (0 :\ -2) \in A^{1\times 2}, \quad T_{12} = 0 \in A,$$

then a stabilizing controller C of P has the form

$$C = \begin{pmatrix} V_1\, U_1^{-1} \\ T_{11} + T_{12}\,(V_1\, U_1^{-1}) \end{pmatrix} = \begin{pmatrix} -\left(1-\frac{e^{-s}}{s+1}\right)^{-1} & -2\,\frac{(s-1)}{(s+1)}\left(1-\frac{e^{-s}}{s+1}\right)^{-1} \\ 0 & -2 \end{pmatrix}. \tag{5.9}$$

Let us notice that $\inf_{s\in\mathbb{C}_+} |1-\frac{e^{-s}}{s+1}| = 0$ (take the sequence $(s_n = 1/n)_{n\in\mathbb{N}}$), and thus, by the Corona theorem [16], we have $(1-\frac{e^{-s}}{s+1})^{-1} \notin A$. Therefore, the first row of the controller C is unstable, whereas its second row is stable. Now, we may wonder if P is strongly stabilizable. Let us notice that the matrix

$$R_2 = \begin{pmatrix} 1 & 2\,\frac{(s-1)}{(s+1)} \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} -\frac{e^{-s}}{s+1} \\ 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 2\,\frac{(s-1)}{(s+1)} \\ 0 & 1 \end{pmatrix}$$

is unimodular because we have

$$\begin{pmatrix} 1 & 2\,\frac{(s-1)}{(s+1)} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -2\,\frac{(s-1)}{(s+1)} \\ 0 & 1 \end{pmatrix} = I_2. \tag{5.10}$$

Then, the matrix R_1 is 1-stable, and thus R is 2-stable:

$$R_2 = \begin{pmatrix} 1 & 0 \\ 0 & \frac{s-1}{s+1} \end{pmatrix} + \begin{pmatrix} -\frac{e^{-s}}{s+1} & -\frac{s-1}{s+1} \\ 0 & -\frac{1}{s+1} \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 2\,\frac{(s-1)}{(s+1)} \\ 0 & 1 \end{pmatrix} \in \mathrm{U}_2(A). \tag{5.11}$$

By Theorem 5.1, we obtain that P is strongly stabilizable ($p=4$, $q=2$, $k=2$, $r=0$). From (5.11), we obtain

$$\begin{pmatrix} 1 & 0 & -\frac{e^{-s}}{s+1} & -\frac{s-1}{s+1} \\ 0 & \frac{s-1}{s+1} & 0 & -\frac{1}{s+1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & -2 \end{pmatrix} \begin{pmatrix} 1 & -2\,\frac{(s-1)}{(s+1)} \\ 0 & 1 \end{pmatrix} = I_2,$$

which shows that

$$S_2 = U_2 = \begin{pmatrix} 1 & -2\,\frac{(s-1)}{(s+1)} \\ 0 & 1 \end{pmatrix}, \quad T_2 = T_{21} = \begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix} \in A^{2\times 2},$$

and thus a stable stabilizing controller C' of P is defined by

$$C' = T_2 = \begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix} \in A^{2\times 2}.$$

To finish, let us show how, using parametrization of all stabilizing controllers of the plant $P_1 = (D - \Lambda_1 T_{11})^{-1}(N_1 + \Lambda_1 T_{12})$, where

$$\Lambda_1 = \begin{pmatrix} \frac{s-1}{s+1} \\ \frac{1}{s+1} \end{pmatrix}, \quad N_1 = \begin{pmatrix} \frac{e^{-s}}{s+1} \\ 0 \end{pmatrix},$$

it was already possible to find C'. First, let us notice that we have

$$R_1 = (D - \Lambda_1 T_{11} : \ -(N_1 + \Lambda_1 T_{12})) \in A^{2\times 3}.$$

Now, from (5.7), we know that $S_1 = (U_1^T : V_1^T)^T$ is a right-inverse of R_1. Computing a doubly coprime factorization of P_1, we obtain the following parametrization of all right inverses of R_1 (see [25, 26] for more details):

$$S_1 = \begin{pmatrix} U_1(k_1, k_2) \\ V_1(k_1, k_2) \end{pmatrix} = \begin{pmatrix} 1 + (k_1 - 1)\frac{e^{-s}}{s+1} & -2\frac{(s-1)}{(s+1)} + \frac{e^{-s}}{s+1} k_2 \\ 0 & 1 \\ k_1 - 1 & k_2 \end{pmatrix} \quad \forall\, k_1,\, k_2 \in A.$$

Therefore, some stabilizing controllers of P are of the form

$$C = \begin{pmatrix} V_1 U_1^{-1} \\ T_{11} + T_{12}(V_1 U_1^{-1}) \end{pmatrix} = \begin{pmatrix} a(k_1 - 1) & a(2(k_1 - 1)\frac{(s-1)}{(s+1)} + k_2) \\ 0 & -2 \end{pmatrix}, \tag{5.12}$$

where $a = (1 + (k_1 - 1)\frac{e^{-s}}{s+1})^{-1}$. Then, taking $k_1 = 1$ and $k_2 = 0$, we recover the stable controller C' of P.

The first difficulty in computing the controllers of the form (5.3) is to be able to determine explicitly the k-stability of a given matrix whose entries belong to a ring A. In section 6, we shall see that it is possible to give a lower bound for it by studying the stable range of the ring A. The second main difficulty is to compute T_k such that R_k, defined by (3.2), satisfies (5.5). In the following corollary of Theorem 5.1, we study the particular case where $T_k = 0$.

COROLLARY 5.2. *Let $P = D^{-1}N \in K^{q\times(p-q)}$ be a transfer matrix. If there exists an integer k satisfying $0 \leq k \leq p - q$ such that $P_r = D^{-1}N_r$ admits a left-coprime factorization, $DX - N_r Y = I_q$, with $\det X \neq 0$ and*

$$R = (D : -N) = (\underset{\leftrightarrow\, q}{D} \ : \underset{\leftrightarrow\, r}{-N_r} \ : \underset{\leftrightarrow\, k}{-\Lambda}) \in A^{q\times p},$$

then the controller

$$C = \begin{pmatrix} Y X^{-1} \\ 0 \end{pmatrix}, \quad \begin{matrix} \updownarrow r = p - q - k \\ \updownarrow k \end{matrix} \tag{5.13}$$

internally stabilizes $P = D^{-1}N$.

Proof. Let us define $T_k = 0$. Then, by hypothesis, the matrix

$$R_k = (D : -N_r) - \Lambda T_k = (D : -N_r)$$

has a left-inverse; i.e., it is unimodular. Therefore, the hypothesis that $P_r = D^{-1} N_r$ admits a left-coprime factorization implies that $R = (D : -N)$ is k-stable. Then, the result directly follows from Theorem 5.1 and $T_k = (T_1 : T_2) = 0$. □

Example 5.2. Let us consider $A = RH_\infty$, $K = Q(A)$, and the transfer matrix

$$P = \begin{pmatrix} \frac{s+1}{s-1} & 0 \\ \frac{1}{(s-1)^2} & \frac{s+1}{s-1} \end{pmatrix} \in K^{2\times 2}.$$

P admits a fractional representation $P = D^{-1} N$, where $R = (D : -N) \in A^{2\times 4}$ is defined by

$$R = \begin{pmatrix} \frac{s-1}{s+1} & 0 & -1 & 0 \\ \frac{1}{(s+1)^2} & -\frac{(s-1)}{(s+1)} & 0 & 1 \end{pmatrix}.$$

The matrix formed by the first two columns of R is not unimodular, but

$$R_1 = \begin{pmatrix} \frac{s-1}{s+1} & 0 & -1 \\ \frac{1}{(s+1)^2} & -\frac{(s-1)}{(s+1)} & 0 \end{pmatrix}$$

is unimodular because we have

$$\begin{pmatrix} \frac{s-1}{s+1} & 0 & -1 \\ \frac{1}{(s+1)^2} & -\frac{(s-1)}{(s+1)} & 0 \end{pmatrix} \begin{pmatrix} \frac{s-1}{s+1} & 4 \\ \frac{1}{(s+1)^2} & -\frac{(s+3)}{(s+1)} \\ -\frac{4\,s}{(s+1)^2} & 4\,\frac{(s-1)}{(s+1)} \end{pmatrix} = I_3.$$

Thus, we can apply Corollary 5.2 to P with $p = 4$, $q = 2$, $k = 1$, $r = 1$ to obtain a stabilizing controller C of P defined by

$$C = \begin{pmatrix} Y\,X^{-1} \\ 0 \end{pmatrix} = -4 \begin{pmatrix} \frac{1}{s+1} & 1 \\ 0 & 0 \end{pmatrix}.$$

Finally, let us notice that P is strongly stabilizable because C is stable.

6. A general structure of the stabilizing controllers based on the stable range. In the rest of the paper, we shall need the following definition.

Definition 6.1 (see [17, 41]). *Let p and q be two positive integers which satisfy $1 \leq q \leq p$. The ring A is said to* satisfy $\mathrm{sr}_k(q, p, A)$ *if every unimodular matrix $R \in A^{q\times p}$ is k-stable. If no confusion arises, we shall write* $\mathrm{sr}_k(q, p)$ *for* $\mathrm{sr}_k(q, p, A)$.

In particular, if A satisfies $\mathrm{sr}(A) = n < +\infty$, then A satisfies $\mathrm{sr}_1(1, n + 1)$.

Theorem 6.2 (see [17, 41]). *We have the following equivalences:*

1. $\mathrm{sr}_1(1, n) \Leftrightarrow \mathrm{sr}_1(1, m)\ \forall\ m \geq n$,
2. $\mathrm{sr}_1(1, n) \Leftrightarrow \mathrm{sr}_k(1, n + k - 1)\ \forall\ k \geq 1$,
3. $\mathrm{sr}_k(1, n) \Leftrightarrow \mathrm{sr}_k(m, n + m - 1)\ \forall\ m \geq 1$.

Corollary 6.3. *Let A be a ring satisfying* $\mathrm{sr}(A) < +\infty$. *Then, for every p, $q \in \mathbb{Z}_+$ which satisfies $p - q \geq \mathrm{sr}(A)$, we have*

$$\mathrm{sr}_{p-q-\mathrm{sr}(A)+1}(q, p);$$

namely, for every unimodular matrix $R = \mathrm{col}(R_1 : \cdots : R_p) \in A^{q\times p}$, there exists a matrix $T_{\mathrm{sr}(A)} \in A^{(p-q-\mathrm{sr}(A)+1)\times(q+\mathrm{sr}(A)-1)}$ such that

$$R_{\mathrm{sr}(A)} = \mathrm{col}(R_1 : \cdots : R_{q+\mathrm{sr}(A)-1}) + \mathrm{col}(R_{q+\mathrm{sr}(A)} : \cdots : R_p)\, T_{\mathrm{sr}(A)} \tag{6.1}$$

is a unimodular matrix.

Proof. Using the fact that we have $\mathrm{sr}(A) = n$, A satisfies $\mathrm{sr}_1(1, n+1)$, and thus, by 1 of Theorem 6.2, we have $\mathrm{sr}_1(1, m)\ \forall\, m \geq n+1$. Then, by 2 of Theorem 6.2, A satisfies $\mathrm{sr}_k(1, m+k-1)$ for $k \geq 1$. Finally, by 3 of Theorem 6.2, A satisfies $\mathrm{sr}_k(l, l+m+k-2)\ \forall\, k,\ l \geq 1$ and $m \geq n+1$.

Now, let $p,\ q \in \mathbb{Z}_+$ such that $p-q \geq \mathrm{sr}(A)$. Let us define $k = p-q-\mathrm{sr}(A)+1 \geq 1$. We have $p = q + (\mathrm{sr}(A)+1) + (p-q-\mathrm{sr}(A)+1) - 2$ and, if we define

$$\begin{cases} l = q \geq 1, \\ n = \mathrm{sr}(A), \\ m = \mathrm{sr}(A)+1, \\ p = l+m+k-2, \end{cases}$$

then A satisfies $\mathrm{sr}_k(l, l+m+k-2)$, i.e., $\mathrm{sr}_{p-q-\mathrm{sr}(A)+1}(q, p)$. Finally, from Lemma 3.4, there exists $T_{\mathrm{sr}(A)} \in A^{(p-q-\mathrm{sr}(A)+1)\times(p+\mathrm{sr}(A)-1)}$ such that the matrix $R_{\mathrm{sr}(A)}$ defined by (6.1) is unimodular. ∎

Now, we are in position to state the second main result of this paper.

Corollary 6.4. *Let $P \in K^{q\times(p-q)}$ be a transfer matrix which admits a left-coprime factorization $P = D^{-1}N$, $R = (D : -N) \in A^{q\times p}$ and satisfies $p-q \geq \mathrm{sr}(A)$. Then, there exist two stable matrices*

$$\begin{cases} T_1 \in A^{(p-q-\mathrm{sr}(A)+1)\times q}, \\ T_2 \in A^{(p-q-\mathrm{sr}(A)+1)\times(\mathrm{sr}(A)-1)} \end{cases} \tag{6.2}$$

such that the matrix $R_{p-q-\mathrm{sr}(A)+1} = (D - \Lambda T_1 : -(N_{\mathrm{sr}(A)-1} + \Lambda T_2)) \in A^{q\times(q+\mathrm{sr}(A)-1)}$ admits a right-inverse, with the notation

$$R = (D : -N) = (\underbrace{D}_{q} : \underbrace{-N_{\mathrm{sr}(A)-1}}_{\mathrm{sr}(A)-1} : \underbrace{-\Lambda}_{p-q-\mathrm{sr}(A)+1}) \in A^{q\times p}. \tag{6.3}$$

Let us denote by $S_{p-q-\mathrm{sr}(A)+1} = (U^T : V^T)^T \in A^{(q+\mathrm{sr}(A)-1)\times q}$ any right-inverse of $R_{p-q-\mathrm{sr}(A)+1}$ such that $\det U \neq 0$. Then, the controller C defined by

$$C = \begin{pmatrix} V\,U^{-1} \\ T_1 + T_2\,(V\,U^{-1}) \end{pmatrix}, \quad \begin{matrix} \updownarrow\, \mathrm{sr}(A)-1 \\ \updownarrow\, p-q-\mathrm{sr}(A)+1 \end{matrix} \tag{6.4}$$

internally stabilizes the plant $P = D^{-1}N$. Moreover, if $\det(D - \Lambda T_1) \neq 0$, then the controller $C_{\mathrm{sr}(A)-1} = V\,U^{-1}$ internally stabilizes the plant

$$P_{\mathrm{sr}(A)-1} = (D - \Lambda T_1)^{-1}\,(N_{\mathrm{sr}(A)-1} + \Lambda T_2).$$

Finally, the unstable part of the controller (6.4) is $C_{\mathrm{sr}(A)-1} = V\,U^{-1}$ and its dimension is equal to $(\mathrm{sr}(A)-1)\times q$.

Proof. By Corollary 6.3, every matrix of $A^{q\times p}$ is $k = (p-q-\mathrm{sr}(A)+1)$-stable. Then, the result directly follows from Theorem 5.1. ∎

Corollary 6.5. *Let us consider $A = RH_\infty$ and $K = Q(A) = \mathbb{R}(s)$. Then, every transfer matrix $P \in \mathbb{R}(s)^{q\times(p-q)}$ admits a stabilizing controller of the form*

$$C = \begin{pmatrix} V\,U^{-1} \\ T_1 + T_2\,(V\,U^{-1}) \end{pmatrix}, \quad \begin{matrix} \updownarrow\, 1 \\ \updownarrow\, p-q-1 \end{matrix},$$

where

$$\left\{ \begin{array}{l} T_1 \in A^{(p-q-1)\times q}, \\ T_2 \in A^{(p-q-1)\times 1}, \end{array} \right.$$

$P = D^{-1}\,N$ *is a left-coprime factorization of* P, $S_{p-q-1} = (U^T : V^T)^T \in A^{(q+1)\times q}$ *is any right-inverse of* $R_{p-q-1} = (D - \Lambda\,T_1 : -(N_1 + \Lambda\,T_2)) \in A^{q\times(q+1)}$ *such that* $\det U \neq 0$, *and*

$$R = (D : -N) = (\underbrace{\quad D \quad}_{q} : \underbrace{-N_1}_{1} : \underbrace{-\Lambda}_{p-q-1}) \in A^{q\times p}.$$

Proof. Every MIMO transfer matrix P with entries in $K = \mathbb{R}(s)$ admits a doubly coprime factorization $P = D^{-1}\,N = \tilde{N}\,\tilde{D}^{-1}$ over A,

$$\left(\begin{array}{cc} D & -N \\ -\tilde{Y} + Q\,D & \tilde{X} - Q\,N \end{array} \right) \left(\begin{array}{cc} X - \tilde{N}\,Q & \tilde{N} \\ Y - \tilde{D}\,Q & \tilde{D} \end{array} \right) = I,$$

where Q is an arbitrary matrix. See [42] for more details. Then, applying Lemma 17 on page 112 of [42], we obtain that there exists $Q^\star$ such that the matrix $\det(X - \tilde{N}\,Q^\star) \neq 0$. Using the facts that $\mathrm{sr}(RH_\infty) = 2$ (see Corollary 2.7) and

$$(U^T : V^T)^T = ((X - \tilde{N}\,Q^\star)^T : (Y - \tilde{D}\,Q^\star)^T)^T,$$

the result follows from Corollary 6.4. □

We have the following straightforward consequence of Corollary 6.4.

COROLLARY 6.6. *If* $\mathrm{sr}(A) = 1$, *then every transfer matrix which admits a left-coprime factorization is strongly stabilizable (i.e., it is internally stabilized by a stable controller). In particular, this result holds for* $A = W_+$ *or* $A(\mathbb{D})$.

Moreover, every internally stabilizable plant, defined by a transfer matrix P *with entries in the quotient field of* $A = H_\infty(\mathbb{D})$ *or* $H_\infty(\mathbb{C}_+)$, *is strongly stabilizable.*

Proof. The first part of the corollary directly follows from Corollary 6.4 and the fact that $\mathrm{sr}(A) = 1$. Moreover, by Theorem 2.8, we know that $\mathrm{sr}(W_+) = 1$ and $\mathrm{sr}(A(\mathbb{D})) = 1$. Finally, if $A = H_\infty(\mathbb{C}_+)$ or $H_\infty(\mathbb{D})$, then it is well known that P is internally stabilizable iff P admits a doubly coprime factorization [25, 26, 36]. The last result directly follows from this fact, Corollary 6.4, Theorem 2.4, and Corollary 2.5. □

Let us notice that the second part of Corollary 6.6 extends Treil's result [38] to MIMO systems. The question of the possibility of having the matrix analogous to Treil's result was asked in [9]. However, the issue consisting in computing effectively the stable stabilizing controllers of a stabilizable plant, defined by a transfer matrix with entries in $K = Q(H_\infty(\mathbb{D}))$ or $K = Q(H_\infty(\mathbb{C}_+))$, is still open.

COROLLARY 6.7. *If* $\mathrm{sr}(A) = 1$, *then every pair of plants, defined by two transfer matrices* P_0 *and* P_1 *with entries in* $K = Q(A)$, *having the same dimensions, and admitting doubly coprime factorizations, is simultaneously stabilized by a controller (simultaneous stabilization). In particular, this result holds for* $A = W_+$ *or* $A(\mathbb{D})$.

Moreover, if $A = H_\infty(\mathbb{D})$ *or* $H_\infty(\mathbb{C}_+)$ *and* P_0, P_1 *are two stabilizable plants with entries in* $K = Q(A)$, *then* P_0 *and* P_1 *are simultaneously stabilized by a controller.*

Proof. Following the proof of Theorem 14 of section 8.3 of [42], there exists a stabilizing controller of P_0 and P_1 iff there exists a matrix T with entries in A such that $U + V\,T$ is a square unimodular matrix, where

$$\left\{ \begin{array}{l} U = D_1\,X_0 - N_1\,Y_0, \\ V = -D_1\,\tilde{N}_0 + N_1\,\tilde{D}_0, \end{array} \right.$$

and $P_i = D_i^{-1} N_i = \tilde{N}_i \tilde{D}_i^{-1}$ is a doubly coprime factorization of P_i, $i = 0, 1$; i.e.,

$$\begin{pmatrix} D_i & -N_i \\ -\tilde{Y}_i & \tilde{X}_i \end{pmatrix} \begin{pmatrix} X_i & \tilde{N}_i \\ Y_i & \tilde{D}_i \end{pmatrix} = I, \quad \begin{pmatrix} X_i & \tilde{N}_i \\ Y_i & \tilde{D}_i \end{pmatrix} \begin{pmatrix} D_i & -N_i \\ -\tilde{Y}_1 & \tilde{X}_i \end{pmatrix} = I.$$

The matrix $(U : V)$ is unimodular because we have $U X - V Y = I$, where

$$\begin{cases} X = D_0 X_1 - N_0 Y_1, \\ Y = \tilde{Y}_0 X_1 + \tilde{X}_0 Y_1. \end{cases}$$

Using the fact that $\mathrm{sr}(A) = 1$, by Corollary 6.3, we obtain that there exists T with entries in A such that $U + V T$ is a square unimodular matrix, and thus every couple of plants is simultaneously stabilized by a controller. Finally, by Theorem 2.8, we know that $\mathrm{sr}(W_+) = 1$ and $\mathrm{sr}(A(\mathbb{D})) = 1$.

Let P_1 and P_2 be two stabilizable transfer matrices with entries in $A = H_\infty(\mathbb{D})$ or $H_\infty(\mathbb{C}_+)$. Then, from [25, 26, 36], we know that P_1 and P_2 admit doubly coprime factorizations. The results directly follow from Theorem 2.4, Corollary 2.5, and the previous point. □

7. Some more results based on stable range.

7.1. Topological stable range. Let us recall the definition of a Banach algebra.

DEFINITION 7.1 (see [13]). *A k-algebra A ($k = \mathbb{R}, \mathbb{C}$) is a* Banach algebra *if A is a Banach k-vector space w.r.t. the norm $\| \cdot \|_A$ and satisfies*

1. $\| 1 \|_A = 1$,
2. $\| a\, b \|_A \leq \| a \|_A \| b \|_A$ *(continuity of the product in each factor).*

Example 7.1. The Hardy space $H_\infty(\mathbb{C}_+)$ of the holomorphic functions in $\mathbb{C}_+$ bounded w.r.t. the norm $\| f \|_\infty = \sup_{s \in \mathbb{C}_+} |f(s)|$ is a Banach algebra [5]. Moreover, the disc algebra $A(\mathbb{D})$ (resp., the Wiener algebra W_+), defined in Theorem 2.8, with the norm $\| f \|_{A(\mathbb{D})} = \sup_{s \in \overline{\mathbb{D}}} |f(s)|$ (resp., $\| f \|_{W_+} = \sum_{n=0}^{+\infty} |a_n|$), are two Banach algebras [13, 42].

DEFINITION 7.2. *If A is a Banach algebra, then the* topological stable range $\mathrm{tsr}(A)$ *of A is the smallest $n \in \mathbb{N} \cup \{+\infty\}$ such that $\mathrm{U}_n(A)$ is dense in A^n for the product topology.*

As for the stable range, the topological stable range $\mathrm{tsr}(A)$ is sometimes called the *topological stable rank* of A.

THEOREM 7.3. *We have the following results:*

- [37] $\mathrm{tsr}(H_\infty(\mathbb{D})) = 2$,
- [31] $\mathrm{tsr}(A(\mathbb{D})) = 2$.

PROPOSITION 7.4. *If A is a Banach algebra such that $\mathrm{tsr}(A) = 2$, then every SISO plant, defined by the transfer function $P = n/d$ ($0 \neq d, n \in A$), satisfies*

$$\forall\, \epsilon > 0,\ \exists\, (d_\epsilon :\ n_\epsilon) \in \mathrm{U}_2(A) : \quad \begin{cases} \| n - n_\epsilon \|_A \leq \epsilon, \\ \| d - d_\epsilon \|_A \leq \epsilon. \end{cases}$$

If $d_\epsilon \neq 0$, then, in the product topology, P is as close as we want to a transfer function $P_\epsilon = n_\epsilon / d_\epsilon$ which admits a coprime factorization. In particular, this result holds for $A = H_\infty(\mathbb{D})$ or $A(\mathbb{D})$.

Proof. Let us consider the vector $(d :\ -n) \in A^{1 \times 2}$. Using the fact that $\mathrm{tsr}(A) = 2$, we obtain

$$\forall\, \epsilon > 0,\ \exists\, (d_\epsilon :\ -n_\epsilon) \in \mathrm{U}_2(A) : \quad \begin{cases} \| d - d_\epsilon \|_A \leq \epsilon, \\ \| n - n_\epsilon \|_A \leq \epsilon. \end{cases}$$

Finally, using the fact that $(d_\epsilon : -n_\epsilon) \in \mathrm{U}_2(A)$, there exist x_ϵ, $y_\epsilon \in A$ such that we have $d_\epsilon\, x_\epsilon - n_\epsilon\, y_\epsilon = 1$, and thus $p_\epsilon = n_\epsilon/d_\epsilon$ admits a coprime factorization. □

In particular, if P is not internally stabilizable, then there exists a stabilizable plant P_ϵ as close as we want to P in the product topology.

7.2. Unit 1-stable range and *n*-fold. Let us introduce a few definitions.

Definition 7.5. *We have the following definitions* [4, 14, 40]*:*

- [14] *A ring A satisfies* unit 1-stable range *if, for every $a = (a_1 : a_2) \in \mathrm{U}_2(A)$, there exists an element $u \in U(A)$ such that $a_1 + a_2\, u \in U(A)$.*
- [39] *A ring A is said to be* n-fold *if, for every n-tuple $a^i = (a_1^i : a_2^i) \in \mathrm{U}_2(A)$, $1 \leq i \leq n$, there exists $b \in A$ such that $a_1^i + a_2^i\, b \in \mathrm{U}(A)$ for $1 \leq i \leq n$.*

Example 7.2. Using a result of Handelman [15], one can easily prove that $\mathrm{sr}(L_\infty(\mathbb{T})) = 1$, where $\mathbb{T} = \{z \in \mathbb{C}\,|\,|z| = 1\}$ is the unit circle, because $L_\infty(\mathbb{T})$ is a *commutative von Neumann algebra* [23], and thus $L_\infty(\mathbb{T})$ has unit 1-stable range (for a $C^\star$-algebra A with a unit [23], it is well known that $\mathrm{sr}(A) = 1$ is equivalent to A has unit 1-stable range [14]). See [18] for the study of stabilization problems over $A = L_\infty(\mathbb{T})$. For the sake of simplicity, in this paper we have studied only the case of integral domains A of SISO stable plants. However, all the results can be easily extended to any ring A with zero divisors.

Proposition 7.6. *We have the following results:*

1. *If A satisfies unit* 1*-stable range, then any SISO plant—defined by the transfer function $P = n/d$ $(d \neq 0, n \in A)$—admitting a coprime factorization is* bistably stabilizable*; namely it is stabilized by a bistable controller (i.e., a stable and inverstable controller)* [2].
2. *If A is an n-fold ring, then every n-tuple of SISO plants—defined by the transfer function $P_i = n_i/d_i$ $(d_i \neq 0, n_i \in A)$ with $1 \leq i \leq n$—having coprime factorizations is stabilized by a stable controller.*

Proof. 1. Let $P = n/d$ be a plant which has a coprime factorization. We may assume that we have $d\, x + n\, y = 1$ with $x, y \in A$. Thus, we have $(d : -n) \in \mathrm{U}_2(A)$. Using the fact that A satisfies unit 1-stable range, there exists $u \in U(A)$ such that $d - n\, u \in \mathrm{U}(A)$, and thus a stabilizing controller is given by $C = u \in \mathrm{U}(A)$; i.e., P is bistably stabilizable.

2. Let $i = 1, \ldots, n$, and let $P_i = n_i/d_i$ be a transfer function admitting a coprime factorization. We may assume that we have $d_i\, x_i + n_i\, y_i = 1$ for certain $x_i, y_i \in A$. Thus, we have $(d_i : -n_i) \in \mathrm{U}_2(A)$. Using the fact that A is n-fold, there exists $y \in A$ such that we have $d_i - n_i\, y \in \mathrm{U}(A)$ for $i = 1, \ldots, n$. Thus, the stable controller defined by $C = y$ simultaneously stabilizes the family of plants $\{P_i\}_{1 \leq i \leq n}$. □

Conclusion. In this paper, we have shown that the concept of stable range was an interesting one in the study of the strong and simultaneous stabilization problems. In particular, we proved that a plant, defined by means of a transfer matrix which admits a left-coprime factorization $P = D^{-1}\, N$, is internally stabilized by a controller, where its unstable and stable parts are separated and the dimension of the unstable part depends only on the k-stability of the matrix $R = (D : -N) \in A^{q \times p}$. Then, using the fact that the stable range of A gives a lower bound of the k-stability of every matrix with entries in A, we proved that, if the stable range of A is 1, then every plant, defined by a transfer matrix admitting a left-coprime factorization, is strongly stabilizable. In particular, using the fact that the stable range of $H_\infty(\mathbb{D})$ is 1 (see [38]), we proved that every stabilizable plant, defined by a transfer matrix with entries in the quotient field of $H_\infty(\mathbb{C}_+)$ or $H_\infty(\mathbb{D})$, is strongly stabilizable. Moreover, we were able to prove that there always exists a stabilizing controller which stabilizes

simultaneously two stabilizable plants defined by a transfer matrix with entries in the quotient field of $H_\infty(\mathbb{C}_+)$ or $H_\infty(\mathbb{D})$. Finally, using the fact that the topological stable range of $H_\infty(\mathbb{D})$ is equal to 2 (see [37]), we proved that every unstabilizable SISO plant, defined by a transfer function with entries in $Q(H_\infty(\mathbb{D}))$, is as close as we want to a stabilizable plant in the product topology.

In this paper, we proved the existence of some particular stabilizing controllers. However, the algorithmical aspects of their constructions were not developed. In forthcoming publications, we shall try to develop this difficult problem.

The concept of a stable range of A was developed by Bass [1] in order to "stabilize" the computation of the group $K_1(A)$ which is the quotient of the group $GL(A)$ of invertible matrices with entries in A by its normal subgroup $EL(A)$ of elementary matrices with entries in A. The connections between the strong stabilization problem and the computation of this group $K_1(A)$ need to be clarified. Moreover, in [35], the obstruction of the simultaneous stabilization of two n-D plants is explicitly expressed in terms of the vanishing of a certain cohomology class. Using the concept of the *Chern character*, it would be interesting to study the links between the results developed in [35] and *topological K-theory*. More generally, it seems that some mathematical tools developed in *algebraic/topological/Hermetian K-theory* are useful for some stabilization problems. Hence, we believe that the study of stabilization problems within a K-theoretical approach should give new interesting results [29].

Finally, a necessary condition for strong stabilizability is the existence of a doubly coprime factorization for the plant (see Proposition 4.5). However, internal stabilizability is generally not equivalent to the existence of doubly coprime factorizations (see [24, 25, 26, 27, 28] and the references therein). Hence, if we do not assume the existence of doubly coprime factorizations for the plants, then the existence of a controller which simultaneously stabilizes two plants P_1 and P_2 is generally not equivalent to the existence of a stable controller for a certain plant P built from P_1 and P_2. For more details, see [30].

Acknowledgments. We would like to thank both Prof. V. Blondel of Louvain-la-Neuve University (Belgium) and Prof. A. Feintuch of the Ben-Gurion University (Israel) for interesting discussions on stable range and strong stabilization. Both of them already knew that the concept of stable range was interesting for the study of strong stabilization [3, 9, 10] (see also the bibliography of [2] and Chapter 6 of [8]). Independently and quite recently, we were led to the concept of stable range while we were developing a K-theoretical approach to stabilization problems (see the forthcoming [29] for more information). Finally, we would like to thank an anonymous referee and an associate editor for all their interesting comments that have improved the quality of the paper.

REFERENCES

[1] H. Bass, *K-theory and stable algebra*, Inst. Hautes Études Sci. Publ. Math., 22 (1964), pp. 5–60.
[2] V. Blondel, *Simultaneous Stabilization of Linear Systems*, Lecture Notes in Control and Inform. Sci. 191, Springer-Verlag, New York, 1994.
[3] V. Blondel, *Private communication*, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, 2001.
[4] H. Chen, *Rings with stable range conditions*, Comm. Algebra, 26 (1998), pp. 3653–3668.
[5] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Texts Appl. Math. 21, Springer-Verlag, New York, 1995.
[6] C. A. Desoer, R.-W. Liu, J. Murray, and R. Saeks, *Feedback system design: The fractional representation approach to analysis and synthesis*, IEEE Trans. Automat. Control, 25 (1980), pp. 399–412.

[7] D. Estes and J. Ohm, *Stable range in commutative rings*, J. Algebra, 7 (1967), pp. 343–362.
[8] A. Feintuch, *Robust Control Theory in Hilbert Space*, Appl. Math. Sci. 130, Springer-Verlag, New York, 1998.
[9] A. Feintuch, *The strong stabilization problem for linear time-varying systems*, Problem 21 in 2002 MTNS Problem Book Open Problems on the Mathematical Theory of Systems, available at http://www.inma.ucl.ac.be/∼blondel/op/.
[10] A. Feintuch, *Private communication*, Ben-Gurion University, Beer-Sheva, Israel, 2003.
[11] M. R. Gabel and A. V. Geramita, *Stable range for matrices*, J. Pure Appl. Algebra, 5 (1974), pp. 97–112; *Erratum*, J. Pure Appl. Algebra, 7 (1976), pp. 239.
[12] M. R. Gabel, *Lower bounds on the stable range of polynomial rings*, Pacific J. Math., 61 (1975), pp. 117–120.
[13] I. Gelfand, D. Raikov, and G. Shilov, *Commutative Normed Rings*, AMS Chelsea Publishing, Providence, RI, 1999.
[14] K. R. Goodearl and P. Menal, *Stable range one for rings with units*, J. Pure Appl. Algebra, 54 (1988), pp. 261–287.
[15] D. Handelman, *Stable range in $AW^\star$-algebras*, Proc. Amer. Math. Soc., 76 (1979), pp. 241–249.
[16] K. Hoffman, *Banach Spaces of Analytic Functions*, Dover, New York, 1962.
[17] Y. Hong, *Remarks on stable range for matrices*, Northeast. Math. J., 11 (1995), pp. 302–306.
[18] B. Jacob, *Stabilizability and Causality of Discrete-time Systems over the Signal Space $l_2(\mathbb{Z})$*, Habilitation thesis, University of Dortmund, Dortmund, Germany, 2001.
[19] C. U. Jensen, *Some curiosities of rings of analytic functions*, J. Pure Appl. Algebra, 38 (1985), pp. 277–283.
[20] P. Jones, D. Marshall, and T. Wolff, *Stable range of the disc algebra*, Proc. Amer. Math. Soc., 96 (1986), pp. 603–604.
[21] J. J. Loiseau, *Algebraic tools for the control and stabilization of time-delay systems*, IFAC Reviews, Annual Reviews in Control, 24 (2000), pp. 135–149.
[22] J. Milnor, *Introduction to Algebraic K-Theory*, Princeton University Press, Princeton, NJ, 1971.
[23] M. A. Naimark, *Normed Algebras*, Wolters-Noordhoff, Groningen, The Netherlands, 1972.
[24] A. Quadrat, *Une approche de la stabilisation par l'analyse algébrique* I. *Factorisations doublement faiblement copremières,* II. *Stabilisation interne,* III. *Sur une structure générale des contrôleurs stabilisants basée sur le rang stable*, in Proceedings of the Conférence Internationale Francophone d'Automatique (CIFA), Nantes, France, 2002.
[25] A. Quadrat, *The fractional representation approach to synthesis problems: An algebraic analysis viewpoint. Part* I: *(Weakly) doubly coprime factorizations*, SIAM J. Control Optim., 42 (2003), pp. 266–299.
[26] A. Quadrat, *The fractional representation approach to synthesis problems: An algebraic analysis viewpoint. Part* II: *Internal stabilization*, SIAM J. Control Optim., 42 (2003), pp. 300–320.
[27] A. Quadrat, *On a generalization of the Youla-Kučera parametrization. Part* I: *The fractional ideal approach to SISO systems*, Systems Control Lett., 50 (2003), pp. 135–148.
[28] A. Quadrat, *A generalization of the Youla-Kučera parametrization for MIMO stabilizable systems*, in Proceedings of the Workshop on Time-Delay Systems (TDS03), INRIA, Rocquencourt, France, 2003.
[29] A. Quadrat, *A systemic K-theory*, in preparation.
[30] A. Quadrat, *An introduction to internal stabilization of infinite-dimensional linear systems*, to appear in the electronic journal e-STA (http://www.e-sta.see.asso.fr/).
[31] M. A. Rieffel, *Dimension and stable rank in the K-theory of $C^\star$-algebras*, Proc. London Math. Soc. (3), 46 (1983), pp. 301–333.
[32] J. Rosenberg, *Algebraic K-Theory and Its Applications*, Grad. Texts in Math. 147, Springer-Verlag, New York, 1996.
[33] J. J. Rotman, *An Introduction to Homological Algebra*, Academic Press, New York, 1979.
[34] R. Rupp, *Stable rank of holomorphic function algebras*, Studia Math., 97 (1990), pp. 85–90.
[35] S. Shankar, *An obstruction to the simultaneous stabilization of two n-D plants*, Acta Appl. Math., 36 (1994), pp. 289–301.
[36] M. C. Smith, *On the stabilization and the existence of coprime factorizations,* IEEE Trans. Automat. Control, 34 (1989), pp. 1005–1007.
[37] D. Suárez, *Trivial Gleason parts and the topological stable rank of H^∞*, Amer. J. Math., 118 (1996), pp. 879–904.
[38] S. Treil, *The stable range of H^∞ equals to* 1, J. Funct. Anal., 109 (1992), pp. 130–154.
[39] W. Van der Kallen, *The K_2 of rings with many units*, Ann. Sci. École Norm. Sup. (4), 10 (1977), pp. 473–515.

[40] W. Van der Kallen, H. Maazen, and J. Stienstra, *A presentation for some $K_2(n, R)$*, Bull. Amer. Math. Soc., 81 (1975), pp. 934–936.
[41] L. N. Vasershtein, *Stable range of rings and the dimension of topological spaces*, Funct. Anal. Appl., 5 (1971), pp. 102–110.
[42] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.

SIAM J. CONTROL OPTIM.
Vol. 42, No. 6, pp. 2286–2304

ORIENTED DISTANCE FUNCTION AND ITS EVOLUTION EQUATION FOR INITIAL SETS WITH THIN BOUNDARY*

MICHEL C. DELFOUR† AND JEAN-PAUL ZOLÉSIO‡

Abstract. The central result of this paper is a *new nonlinear equation* which describes the evolution of the oriented distance function b_Ω of a set Ω with thin boundary under the influence of a velocity field. We relate it to equations and constructions used in the context of *level set methods*. We further introduce a new *moving narrow-band method* which not only can be readily implemented to solve our evolution equation, but could also be used for equations of motion by curvatures. In the process we review and sharpen the characterization of smooth sets and manifolds and sets of positive reach (e.g., local semiconvexity in an extended sense of the oriented distance function of the closure of the set). For $W^{2,p}$-Sobolev domains a new characterization and a compactness theorem are given in terms of the Laplacian of the oriented distance function rather than its whole Hessian matrix.

Key words. oriented distance function, signed distance function, level set method, narrow-band method, extension velocity, nonlinear evolution equation, smooth sets, submanifold, shape, design, image, vision

AMS subject classifications. 49J53, 49Q, 49Q10, 49L, 53A, 58, 35R35

DOI. 10.1137/S0363012902411945

1. Introduction.

In problems where a geometric object is the variable, the object can be identified with a family of functions indexed by the sets, such as the characteristic functions, the distance functions, the oriented distance functions, or the support functions. Such functions play the role of the *state variable* associated with the set. Special metrics can be constructed from such function to measure the distance between two objects and to induce topologies from which existence and characterization of optimal objects can be obtained for design, identification, or control purposes. The choice of the function and the metric is obviously problem dependent and corresponds to pertinent technological, physical, or geometric entities associated with the problem at hand. For instance, distance functions have been used for theoretical and computational purposes in free boundary problems [19, 20, 27], image processing and computer vision [31, 37, 38, 5, 4, 13], [32, 30, 1, 21], [41, 42], and robotics [23, 24, 25, 40].

In this paper we focus on the oriented distance function, its associated metric topologies, and its use in the characterization of special families of sets. The central result is a *new nonlinear equation* which describes the evolution of the oriented distance function of a set with thin boundary under the influence of a velocity field. It can be viewed as the *state equation* of the moving set under the influence of the velocity field which plays the role of a *distributed control function*. This equation and the associated technical results find applications in computing shape derivatives of objective functions involving the oriented distance function, the normal, or the curvatures

*Received by the editors July 23, 2002; accepted for publication (in revised form) October 2, 2003; published electronically May 17, 2004. This research has been supported by National Sciences and Engineering Research Council of Canada discovery grant A–8730 and by a FQRNT grant from the Ministère de l'Éducation du Québec.

http://www.siam.org/journals/sicon/42-6/41194.html

†Centre de recherches mathématiques et Département de mathématiques et de statistique, Université de Montréal, C. P. 6128, succ. Centre-ville, Montréal, QC, Canada H3C 3J7 (delfour@CRM.UMontreal.CA).

‡CNRS and INRIA, INRIA, 2004 route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex, France (Jean-Paul.Zolesio@sophia.inria.fr).

(cf. [12]). In the present paper we relate our new evolution equation to equations and constructions used in the context of *level set methods*. We further introduce a new *moving narrow-band method* which not only can be readily implemented to solve our evolution equation, but could also be used for equations of motion by curvatures.

Section 2 recalls basic definitions and results on the oriented distance function b_Ω of a set Ω. Section 3 reviews the families of sets which will be used in the paper and sharpens their characterizations and properties (sets of positive reach, smooth sets and submanifolds, and $W^{2,p}$-Sobolev domains). The function b_Ω is shown to be locally semiconvex in an extended sense[1] for sets of positive reach. For $W^{2,p}$-Sobolev domains a new characterization and a compactness theorem are given in terms of the Laplacian of b_Ω rather than its whole Hessian matrix. Their boundary is shown to have zero measure and, for $p > N$, the boundary integral to be continuous for special classes of functions. Section 4 reviews the *velocity method* which transforms an initial domain Ω into domains $\Omega_t(V)$ indexed by the real parameter t under the action of a velocity field V. We compute the partial derivative of the oriented distance function of $\Omega_t(V)$ with respect to t. Section 5 is devoted to the new *nonlinear evolution equation* for the oriented distance function for initial sets with thin boundary evolving in a velocity field and its connection with level set methods and the use of the zero-extension introduced in Definition 3.3 to create new *moving narrow-band methods* (cf. Remark 5.5).

2. Oriented distance function and metrics. Given an integer $N \geq 1$, m_N and H_{N-1} will denote the N-dimensional Lebesgue and $(N-1)$-dimensional Hausdorff measures. The inner product and the norm in $\mathbf{R}^{\mathrm{N}}$ will be written $x \cdot y$ and $|x|$. The *complement* $\{x \in \mathbf{R}^{\mathrm{N}} : x \notin \Omega\}$ and the boundary $\overline{\Omega} \cap \overline{\complement\Omega}$ of a subset Ω of $\mathbf{R}^{\mathrm{N}}$ will be, respectively, denoted by $\complement\Omega$ or $\mathbf{R}^{\mathrm{N}} \setminus \Omega$ and by $\partial\Omega$ or Γ. The *distance function* $d_A(x)$ from a point x to a subset $A \neq \varnothing$ of $\mathbf{R}^{\mathrm{N}}$ is defined as $\inf\{|y - x| : y \in A\}$.

Given $\Omega \subset \mathbf{R}^{\mathrm{N}}$, $\Gamma \neq \varnothing$, the *oriented distance function*[2] is defined as

$$b_\Omega(x) \stackrel{\text{def}}{=} d_\Omega(x) - d_{\complement\Omega}(x). \tag{2.1}$$

There is a one-to-one correspondence between b_Ω and the equivalence class[3]

$$[\Omega]_b \stackrel{\text{def}}{=} \left\{\Omega' \subset \mathbf{R}^{\mathrm{N}} \,:\, \Gamma' = \Gamma \text{ and } \overline{\Omega'} = \overline{\Omega}\right\}. \tag{2.2}$$

The function b_Ω is Lipschitz continuous of constant 1, and ∇b_Ω exists and $|\nabla b_\Omega| \leq 1$ almost everywhere in $\mathbf{R}^{\mathrm{N}}$. Thus $b_\Omega \in W^{1,p}_{\mathrm{loc}}(\mathbf{R}^{\mathrm{N}})$ for all p, $1 \leq p \leq \infty$. Definition (2.1) and the associated equivalence classes seem to have been first introduced in 1994 in [9]. This terminology and notation emphasize the fact that ∇b_Ω coincides with the exterior normal to the boundary (when it exists). In the literature, the *signed distance function* is defined for a closed submanifold M under the assumption that there exists an open set Ω such that $\Gamma = M$ as the distance d_Γ to the boundary Γ with a change

[1] Cf. footnote 10 and Theorem 3.1 (ii).

[2] The function b_Ω captures many of the geometric properties of the set Ω. For instance, in 1985 [3] showed that a proper closed domain Ω is convex if and only if $-b_\Omega$ is superharmonic and that, for $N = 2$, the result still holds if $-b_\Omega$ is superharmonic only on Ω. They also show that, for Ω compact, d_Γ *is subharmonic on* $\complement\Omega$ *if and only if* Ω *is convex* (cf. also [11], Chap. 5, Lem. 7.1). This work was pursued in 1987 by [33] and in 1988 by [34]. In 1994 it was shown in [9, 11] that the property that $\overline{\Omega}$ *is convex if and only if* d_Ω *is convex* remains true with b_Ω in place of d_Ω.

[3] In general $d_{\overline{\Omega}} = d_\Omega \leq d_{\operatorname{int}\Omega}$ and $d_{\overline{\complement\Omega}} = d_{\complement\Omega} \leq d_{\complement\overline{\Omega}}$, but we only have $b_{\overline{\Omega}} \leq b_\Omega \leq b_{\operatorname{int}\Omega}$. For convex sets we have $b_{\overline{\Omega}} = b_\Omega$; for sets verifying the segment property we have $b_{\overline{\Omega}} = b_\Omega = b_{\operatorname{int}\Omega}$.

of sign across Γ (cf., for instance, [28]). In some contexts this definition is ambiguous since once a submanifold M has been specified there are several choices of open sets Ω with the *same boundary* $\Gamma = M$ but *completely different* functions b_Ω.[4]

The function b_Ω offers definite advantages over the function d_Ω. It captures the geometric and smoothness properties of the set Ω. For instance, Ω is of class $C^{1,1}$ if and only if b_Ω is locally $C^{1,1}$ in a neighborhood of its boundary as we shall see in Theorem 3.2. This characterization is not possible with d_Ω whose gradient is discontinuous across the boundary Γ. It makes it possible to simultaneously deal with open N-dimensional subsets and embedded submanifolds of $\mathbf{R}^{\mathrm{N}}$ in the same framework. Indeed when Ω is a closed embedded submanifold of $\mathbf{R}^{\mathrm{N}}$ of codimension greater or equal to one, then $\Omega = \overline{\Omega} = \Gamma$ and $b_\Omega = d_\Omega = d_\Gamma$.[5]

DEFINITION 2.1.

(i) *Given a nonempty subset D of $\mathbf{R}^{\mathrm{N}}$, define the families*

(2.3)
$$C_b(D) \overset{\text{def}}{=} \{b_\Omega : \Omega \subset \overline{D} \text{ and } \Gamma \neq \varnothing\}, \quad C_b^0(D) \overset{\text{def}}{=} \{b_\Omega \in C_b(D) : \mathrm{m}_N(\Gamma) = 0\}.$$

(ii) *The boundary Γ of a subset Ω of $\mathbf{R}^{\mathrm{N}}$ is said to be* thin[6] *if* $\mathrm{m}_N(\Gamma) = 0$.

The space $C_b^0(D)$ corresponds to the subfamily of subsets of $\mathbf{R}^{\mathrm{N}}$ with a *thin boundary* that is a more natural family than the family $C_b(D)$ in applications.

In this paper we specialize to the following complete metrics[7] associated with b_Ω over the subsets Ω of a bounded open hold-all D

$$\rho_{C(D)}([\Omega'], [\Omega]) \overset{\text{def}}{=} \max_{x \in \overline{D}} |b_{\Omega'}(x) - b_\Omega(x)|, \tag{2.4}$$

$$\rho_{L^p(D)}([\Omega'], [\Omega]) \overset{\text{def}}{=} \left\{ \int_D |b_{\Omega'} - b_\Omega|^p \, dx \right\}^{1/p}, \tag{2.5}$$

$$\rho_{W^{1,p}(D)}([\Omega'], [\Omega]) \overset{\text{def}}{=} \left\{ \int_D |b_{\Omega'} - b_\Omega|^p + |\nabla b_{\Omega'} - \nabla b_\Omega|^p \, dx \right\}^{1/p}. \tag{2.6}$$

The space $C_b(D)$ is a complete metric space for the metrics (2.4), (2.5), and (2.6), but the space $C_b^0(D)$ is complete only with respect to the metric (2.6) (e.g., [11, Chap. 5]).[8] The metrics (2.6) are all equivalent for $1 \leq p < \infty$. The following theorem is central. It shows that convergence and compactness in the metric $\rho_{W^{1,p}(D)}$ will imply the same properties in all other topologies (cf. [11, Chap. 5, Thm. 5.1]).

THEOREM 2.1. *Let D be a bounded open subset of $\mathbf{R}^{\mathrm{N}}$. The map*

$$b_\Omega \mapsto (b_\Omega^+, b_\Omega^-, |b_\Omega|) = (d_\Omega, d_{\complement\Omega}, d_{\partial\Omega}) : C_b(D) \subset W^{1.p}(D) \to W^{1.p}(D)^3 \tag{2.7}$$

and, for all p, $1 \leq p < \infty$, the map

$$b_\Omega \mapsto (\chi_{\partial\Omega}, \chi_{\operatorname{int}\Omega}, \chi_{\operatorname{int}\complement\Omega}) : W^{1,p}(D) \to L^p(D)^3$$

are continuous.

---

[4]For instance, the unit circle C in $\mathbf{R}^2$ is the boundary of the open unit ball but it is also the boundary of the open set $\mathbf{R}^2 \setminus C$.

[5]Recall that $b_\Omega^+ = d_\Omega$, $b_\Omega^- = d_{\complement\Omega}$, and $|b_\Omega| = d_\Gamma$, and that $\chi_{\operatorname{int}\Omega} = |\nabla d_{\complement\Omega}|$, $\chi_{\operatorname{int}\complement\Omega} = |\nabla d_\Omega|$, and $\chi_\Gamma = 1 - |\nabla d_\Gamma|$ a.e. in $\mathbf{R}^{\mathrm{N}}$.

[6]This terminology is not to be confused with the one of thin set in *Capacity Theory*.

[7]Other complete metrics can be defined with d_Ω, $d_{\complement\Omega}$, d_Γ in place of b_Ω.

[8]The completeness of the metric (2.4) is not a trivial consequence of the proof by Dellacherie [14] in 1972 of the completeness of the Hausdorff metric associated with d_Ω which is different but equivalent to the classical definition of D. Pompéju [36] in 1905 and F. Hausdorff [22] in 1914. To our best knowledge, the metrics (2.4) and (2.6) were first introduced by [9] in 1994.

The weak $W^{1,p}$-topologies are all equivalent on $C_b(D)$ for $1 \le p < \infty$. For the convergence of sets with thin boundary, we have the following equivalence.

LEMMA 2.1. *Given a bounded open subset D of $\mathbf{R}^{\mathrm{N}}$, let $\{\Omega_n\}$ be a sequence of subsets of $\overline{D}$ such that $\Gamma_n \neq \varnothing$ and $\mathrm{m}(\Gamma_n) = 0$. Further assume that there exists $\Omega \subset \bar{D}$ such that $\Gamma \neq \varnothing$ and $\mathrm{m}(\Gamma) = 0$. Then*

$$b_{\Omega_n} \rightharpoonup b_\Omega \text{ in } W^{1,2}(D)\text{-weak} \quad \Rightarrow b_{\Omega_n} \to b_\Omega \text{ in } W^{1,2}(D)\text{-strong},$$

and hence in $W^{1,p}(D)$-strong for all p, $1 \le p < \infty$.

Proof. Same proof as in part (ii) of the proof of Theorem 10.1 in [11]. Since, for all $n \ge 1$, $\mathrm{m}(\Gamma_n) = 0 = \mathrm{m}(\Gamma)$, $|\nabla b_\Omega| = 1 = |\nabla b_{\Omega_n}|$ almost everywhere in D (cf. [11, Thm. 3.2, p. 215]). As a result

$$\int_D |\nabla b_{\Omega_n} - \nabla b_\Omega|^2 \, dx = \int_D |\nabla b_{\Omega_n}|^2 + |\nabla b_\Omega|^2 - 2\nabla b_{\Omega_n} \cdot \nabla b_\Omega \, dx$$
$$= 2\int_D (1 - \nabla b_{\Omega_n} \cdot \nabla b_\Omega) \, dx \to 2\int_D (1 - |\nabla b_\Omega|^2) \, dx = 2\int_D \chi_\Gamma \, dx = 0.$$

Therefore $\nabla b_{\Omega_n} \to \nabla b_\Omega$ in $L^2(D)^N$-strong and $b_{\Omega_n} \to b_\Omega$ in $W^{1,2}(D)$-strong, since the convergence $b_{\Omega_n} \to b_\Omega$ in $L^2(D)$-strong follows from the weak convergence in $W^{1,2}(D)$. The convergence in $W^{1,p}(D)$-strong follows from the equivalence of the topologies on $C_b(D)$ (cf. [11, Chap. 5, Thm. 5.1 (i)]). ∎

The points of $\mathbf{R}^{\mathrm{N}}$ where the gradient of b_Ω does not exist can be divided into two categories: the ones on the boundary Γ and the ones outside of Γ.

DEFINITION 2.2. *The set of* projections *of a point $x \in \mathbf{R}^{\mathrm{N}}$ onto the boundary Γ of a set Ω, $\Gamma \neq \varnothing$,*

$$\Pi_\Gamma(x) \stackrel{\text{def}}{=} \left\{p \in \mathbf{R}^{\mathrm{N}} \,:\, |b_\Omega(x)| = |p - x|\right\}$$

since $|b_\Omega(x)| = d_\Gamma(x)$; the skeleton *of Ω*

$$\mathrm{Sk}(\Omega) \stackrel{\text{def}}{=} \left\{x \in \mathbf{R}^{\mathrm{N}} \,:\, \Pi_\Gamma(x) \text{ is not a singleton}\right\} \tag{2.8}$$

(by definition $\mathrm{Sk}(\Omega) \subset \mathbf{R}^{\mathrm{N}} \setminus \Gamma$); the set of cracks *of Ω*

$$\mathrm{C}(\Omega) \stackrel{\text{def}}{=} \left\{x \in \mathbf{R}^{\mathrm{N}} \,:\, \nabla b_\Omega^2(x) \text{ exists but } \nabla b_\Omega(x) \text{ does not exist}\right\}.$$

The terminology *crack* is used here in a very broad sense and $C(\Omega)$ can contain subsets of arbitrary codimension. In dimension 2 the *corners* along a piecewise smooth boundary belong to the set of cracks.

THEOREM 2.2. *Let Ω be a subset of $\mathbf{R}^{\mathrm{N}}$ with $\Gamma \neq \varnothing$.*

(i) *For all $x \in \Gamma$, $\nabla b_\Omega^2(x)$ exists and $\nabla b_\Omega^2(x) = 0$; for all $x \notin \Gamma$*

$$\nabla b_\Omega^2(x) \text{ exists} \quad \Longleftrightarrow \quad \nabla b_\Omega(x) \text{ exists}.$$

Hence $\nabla b_\Omega(x)$ exists if and only if $x \notin \mathrm{Sk}(\Omega) \cup \mathrm{C}(\Omega)$. Moreover,

$$\mathrm{Sk}(\Omega) = \left\{x \in \mathbf{R}^{\mathrm{N}} \,:\, \nabla b_\Omega^2(x) \text{ does not exist}\right\}$$

and $\mathrm{Sk}(\Omega) \subset \mathbf{R}^{\mathrm{N}} \setminus \Gamma$ and $\mathrm{C}(\Omega) \subset \Gamma$ have zero m_N-measure.

(ii) *The* projection *$p_\Gamma(x)$ of a point $x \notin \mathrm{Sk}(\Omega)$ onto the boundary Γ is given by*

$$p_\Gamma(x) = x - \frac{1}{2}\nabla b_\Omega^2(x) = x - b_\Omega \, \nabla b_\Omega(x). \tag{2.9}$$

(iii) *The Hadamard semiderivative*[9] *of* b_Ω^2 *always exists:*

$$\forall v \in \mathbf{R}^N, \quad d_H b_\Omega^2(x;v) = 2 \min_{p\in\Pi_\Gamma(x)} (x-p)\cdot v. \tag{2.10}$$

(iv) *For all points* $x \notin \Gamma$*, the Hadamard semiderivative of* b_Ω *exists and*

$$\forall v \in \mathbf{R}^N, \quad d_H b_\Omega(x;v) = \frac{1}{b_\Omega(x)} \min_{p\in\Pi_\Gamma(x)} (x-p)\cdot v. \tag{2.11}$$

For all points $x \in \Gamma$*,* $d_H b_\Omega(x;v)$ *exists if and only if*

$$\forall v \in \mathbf{R}^N, \quad \lim_{t\searrow 0} \frac{b_\Omega(x+tv)}{t} \text{ exists.} \tag{2.12}$$

Proof. (i) and (ii) Cf. [11, Chap. 5, Thm. 4.4, and Chap. 8, Sects. 5, 2, 3, and p. 369]). (iii) Cf. [11, Thm. 3.1 (iii), p. 164]. (iv) The proof is obvious. □

3. Some families of sets and their properties. We review the families of sets which will be used in the paper and sharpen their characterization and properties (sets of positive reach, smooth sets and manifolds, and Sobolev domains). For instance, the distance function is shown to be locally semiconvex[10] for sets of positive reach. For $W^{2,p}$-Sobolev domains a new characterization and a compactness theorem are given in term of the Laplacian of b_Ω rather than on the whole Hessian matrix. Their boundary is thin.

Given $h > 0$ the *open and closed tubular neighborhoods* of a set A are defined as

$$U_h(A) \overset{\text{def}}{=} \left\{x \in \mathbf{R}^N \,:\, d_A(x) < h\right\}, \quad A_h \overset{\text{def}}{=} \left\{x \in \mathbf{R}^N \,:\, d_A(x) \le h\right\}. \tag{3.1}$$

Recalling that $d_\Gamma(x) = |b_\Omega(x)|$ we also have $U_h(\Gamma) = \left\{x \in \mathbf{R}^N : |b_\Omega(x)| < h\right\}$.

3.1. Sets and boundaries of positive reach. The sets of positive reach were introduced by Federer [18] in 1959.

Definition 3.1.

(i) $\Omega \subset \mathbf{R}^N$, $\Omega \neq \varnothing$, *is said to have* positive reach *greater or equal to* h *if the projection* $p_\Omega(x)$ *onto* $\overline{\Omega}$ *is unique for all points* x *in the open tubular neighborhood* $U_h(\Omega)$ *of* Ω.

(ii) *The boundary* $\Gamma \neq \varnothing$ *of a set* $\Omega \subset \mathbf{R}^N$ *is said to have* positive reach *greater or equal to* h *if the projection* $p_\Gamma(x)$ *onto* Γ *is unique for all points* x *in the open tubular neighborhood* $U_h(\Gamma)$ *of* Γ.

Theorem 3.1. *Let* Ω *be a nonempty subset of* $\mathbf{R}^N$.

(i) Ω *has positive reach greater or equal to* h *if and only if* $d_\Omega^2 \in C^{1,1}_{\text{loc}}(U_h(\Omega))$.

(ii) *If any one of the conditions of part* (i) *is verified, then for all* $0 < r < h$

$$\forall a \in \overline{\Omega}, \quad b_{\overline{\Omega}}(x) + \frac{1}{2r}\left[|x|^2 - b_{\overline{\Omega}}^2(x)\right] \text{ is convex in } B_h(a), \tag{3.2}$$

where $B_h(a)$ *denotes the open ball of radius* h *and center* a.

---

[9] A function $f : \mathbf{R}^N \to \mathbf{R}$ has a Hadamard semiderivative in x in the direction v if

$$d_H f(x;v) \overset{\text{def}}{=} \lim_{\substack{t\searrow 0 \\ w\to v}} \frac{f(x+tw)-f(x)}{t} \text{ exists}$$

(cf. [11, Chap. 8, Def. 2.1 (ii)]).

[10] Here the terminology *semiconvex* is used in the *extended sense* that there exists a convex function k such that $b_\Omega + k$ is convex as stated in Theorem 3.1 (ii).

Proof. (i) The proof is from Federer [18] or [11, Chap. 4, Thm. 7.1, p. 192]. (ii) For, $\Omega_r = \{x \in \mathbf{R}^N : d_\Omega(x) \le r\}$ the function $\left[|x|^2 - b^2_{\Omega_r}(x)\right]/2$ is convex and continuous. From [11, Chap. 5, Thm. 8.2 (ii), p. 244] for $0 < r < h$ we have $b_{\Omega_r}(x) = b_{\overline{\Omega}}(x) - r$ on $U_h(\Omega)$ and hence for each $a \in \overline{\Omega}$

$$\begin{aligned}\frac{1}{2}\left[|x|^2 - (b_{\overline{\Omega}}(x) - r)^2\right] &= \frac{1}{2}\left[|x|^2 - b^2_{\overline{\Omega}}(x) + 2r b_{\overline{\Omega}}(x) - r^2\right] \\ &= r\left[b_{\overline{\Omega}}(x) + \frac{1}{2r}\left[|x|^2 - b^2_{\overline{\Omega}}(x)\right] - \frac{r}{2}\right] \\ \Rightarrow\ & b_{\overline{\Omega}}(x) + \frac{1}{2r}\left[|x|^2 - b^2_{\overline{\Omega}}(x)\right]\end{aligned}$$

is convex and continuous in $B_h(a)$. □

When the boundary Γ of a set Ω has positive reach, both $\overline{\Omega}$ and $\overline{\complement\Omega}$ have positive reach and we get the following results since $b_{\overline{\complement\Omega}} = -b_{\operatorname{int}\Omega}$.

COROLLARY 3.1.1. *Let Ω be a subset of $\mathbf{R}^N$ such that $\Gamma \neq \varnothing$. Assume that there exists $h > 0$ such that $b^2_\Omega \in C^{1,1}_{\mathrm{loc}}(U_h(\Gamma))$ and $\{x \in \mathbf{R}^N : |b_\Omega(x)| \le h\} \neq \mathbf{R}^N$. For all r, $0 < r < h$, (3.2) is verified and*

$$\forall a \in \overline{\complement\Omega}, \quad -b_{\operatorname{int}\Omega}(x) + \frac{1}{2r}\left[|x|^2 - b^2_{\operatorname{int}\Omega}(x)\right] \text{ is convex in } B_h(a),$$

$$\forall a \in \Gamma, \quad b_\Gamma(x) + \frac{1}{2r}\left[|x|^2 - b^2_\Gamma(x)\right] \text{ is convex in } B_h(a).$$

3.2. Smooth sets and manifolds. First, recall the relation between the smoothness of a set Ω and the smoothness of b_Ω in a neighborhood of its boundary.

THEOREM 3.2. [11] *Let Ω be a subset of $\mathbf{R}^N$ such that $\Gamma \neq \varnothing$.*

(i) *The set Ω is of class $C^{1,1}$* (resp., C^k, $k \ge 2$) *in a neighborhood $U(x)$ of $x \in \Gamma$ if and only if $b_\Omega \in C^{1,1}(W(x))$* (resp., C^k, $k \ge 2$) *in some neighborhood $W(x)$ of x. Moreover, under the $C^{1,1}$ assumption,*

$$C(\Omega) \cap W(x) = \varnothing \text{ and } \nabla b_\Omega = n \circ p_\Gamma \text{ in } W(x),$$

where n is the unit exterior normal to Γ at the point $p_\Gamma(x)$.

(ii) *Let λ, $0 \le \lambda < 1$, be a real number. Ω is of class $C^{1,\lambda}$ and its boundary Γ has positive reach if and only if b_Ω is $C^{1,\lambda}$ in a neighborhood of Γ.*

Proof. (i) Cf. [11]. (ii) ($\Leftarrow$) See Theorem 4.2 in [11, Chap. 5, p. 218]. ($\Rightarrow$) The beginning of the proof of Theorem 4.3 (i) [11, Chap. 5, pp. 219–220] only requires that Ω be of class C^1, and we get for some neighborhood of $U(\Gamma)$ of Γ

$$\forall y \in U(\Gamma),\ \forall p \in \Pi_\Gamma(y), \quad p = y - b_\Omega(y)\, n(p),$$

[11] It is important to note that Theorem 3.2 is not true when b_Ω is replaced by d_Γ since its gradient ∇d_Γ is discontinuous across Γ. Part (i) in the direction ($\Rightarrow$) was asserted by Serrin [39] in 1969 for $N = 3$, proved in 1977 by Gilbarg and Trudinger [19] for $k \ge 2$ (provided that d_Γ is replaced by b_Ω in Lemma 1, p. 382, in [19] and Lemma 14.16, p. 355, in [20]) with a different proof by Krantz and Parks [28] in 1981. Another proof with the function b_Ω was given in 1994 by Delfour and Zolésio [9], who extended the result in the direction ($\Rightarrow$) down to the $C^{1,1}$ case and established the equivalence ($\Leftrightarrow$) in the whole range from C^∞ to $C^{1,1}$. The counterexample for domains of class $C^{1,1-\epsilon}$ given in [11] is the same as the one provided earlier in [28], where they only observe that the domain is $C^{2-\varepsilon}$ leaving the reader under the misleading impression that part (i) of the theorem would not be true for domains ranging from class $C^{1,1}$ to C^2. Part (ii) in the direction ($\Rightarrow$) was proved in the C^1 case by Krantz and Parks [28] in 1981. The equivalence ($\Leftrightarrow$) here for $C^{1,\lambda}$, $0 \le \lambda < 1$, seems to be new.

where $n(p)$ is the unit outward normal to Γ in p. But, by assumption, $\Pi_\Gamma(y) = \{p_\Gamma(y)\}$ is a singleton and

$$\forall y \in U(\Gamma), \quad p_\Gamma(y) = y - b_\Omega(y)\, n(p_\Gamma(y)).$$

Since Γ has positive reach we get from [11, Chap. 5, Lem. 8.2] that

$$\forall y \in U(\Gamma)\backslash\Gamma, \quad p_\Gamma(y) = y - b_\Omega(y)\, \nabla b_\Omega(y)),$$

and hence, for all y in $U(\Gamma)\backslash\Gamma$, $\nabla b_\Omega(y)) = n(p_\Gamma(y))$. But the boundary Γ of a C^1-domain has zero measure, and hence ∇b_Ω is a.e. equal to the $C^{1,\lambda}$ function $n \circ p_\Gamma$ on $U(\Gamma)$ as the composition of $n \in C^{0,\lambda}(U(\Gamma))$ and $p_\Gamma \in C^{0,1}(U(\Gamma))$. Hence $b_\Omega \in C^{1,\lambda}(U(\Gamma))$. □

For a domain Ω of class $C^{1,1}$ the unit exterior normal $n(x)$ to Ω in a point x of its boundary Γ coincides with the gradient of b_Ω, and the Hessian matrix $D^2 b_\Omega(x)$ on Γ coincides with the *second fundamental form* of Γ. The *additive curvature* H of Γ is the sum of the eigenvalues of the second fundamental form and coincides with the *tangential divergence* of n,

$$H \overset{\text{def}}{=} \operatorname{div}_\Gamma n = \Delta b_\Omega|_\Gamma, \quad H_{N-1}\text{-a.e. on } \Gamma.$$

Since 0 is an eigenvalue of $D^2 b_\Omega$ and H is the sum of the eigenvalues of $D^2 b_\Omega$, Δb_Ω is equal to $(N-1)$ times the standard *mean curvature* of Γ (cf. [8, 10]).

Theorem 3.2 covers only sets whose thin boundary is a submanifold of $\mathbf{R}^N$ of codimension one. For closed submanifolds M of $\mathbf{R}^N$ where ∇b_M does not exist on M, the smoothness of M is related to d_M^2 and the existence of ∇d_M^2 in a neighborhood of M implies that M is locally of positive reach. The following analysis of the smoothness of M was given by Poly and Raby [35] in 1984.

THEOREM 3.3. *Let M be a closed nonempty subset of $\mathbf{R}^N$ and $k \geq 2$ be an integer ($k = \infty$ and ω[12] included). Then*

$$\operatorname{sing}_k M = M \cap \operatorname{sing}_k d_M^2,$$

where $\operatorname{sing}_k d_M^2 = \mathbf{R}^N \backslash \operatorname{reg}_k d_M^2$, $\operatorname{sing}_k M = \mathbf{R}^N \backslash \operatorname{reg}_k M$, *and*

$$\operatorname{reg}_k d_M^2 = \left\{x \in \mathbf{R}^N \,:\, d_M^2 \text{ is } C^k \text{ in a neighborhood of } x\right\}$$
$$\operatorname{reg}_k M = \left\{x \in M \,:\, M \text{ is a } C^k\text{-submanifold of } \mathbf{R}^N \text{ in a neighborhood of } x\right\}.$$

For $k = 1$ [35] gives the one-dimensional counterexample $\Omega =]-\infty, 0]$ for which $d_\Omega^2 \in C^{1,1}$, and this counterexample readily extends to a closed half space of $\mathbf{R}^N$. This can be fixed by using b_Ω for which $b_\Omega^2(x) = |x|^2 \in C^\infty$ recalling that $|b_\Omega(x)| = d_\Gamma(x)$. It yields the analogue of Theorem 3.2.

THEOREM 3.4. [13]*Let Ω be a subset of $\mathbf{R}^N$ with nonempty thin boundary Γ and $k \geq 2$ be an integer ($k = \infty$ and ω included), and x be a point of Γ. Then b_Ω^2 is C^k in a neighborhood of a point $x \in \Gamma$ if and only if Γ is a C^k-submanifold in a*

[12] ω indicates the analytical case.

[13] Note that Γ can have several connected components with the same smoothness k but different dimension. When $\Gamma = \mathbf{R}^N$ b_Ω is identically zero. Another set of results for Hölderian sets was obtained by [29] (see also [7]) by introducing a regularized distance function.

neighborhood of x. Moreover, the dimension of Γ in x is equal to the rank of $Dp_\Gamma(x)$ and $Dp_\Gamma(x)$ is the orthogonal projector onto the tangent space in x.

3.3. Sobolev domains. The notion of *Sobolev domain* has been introduced in [10] as an instrument to classify domains which fall in the gaps between Hölderian domains. We further characterize $(2,p)$-Sobolev domains by introducing a special extension of b_Ω by zero outside $U_h(\Gamma)$.

DEFINITION 3.2. *Given $m > 1$ and $p \geq 1$, a subset Ω of $\mathbf{R}^N$ is said to be an* (m,p)-Sobolev domain *if $\Gamma \neq \varnothing$ and*

$$\exists h > 0 \text{ such that } b_\Omega \in W^{m,p}_{\mathrm{loc}}(U_h(\Gamma)).$$

For sets Ω of locally bounded curvature

$$\forall p,\ 1 \leq p < \infty, \quad \forall \eta,\ 0 \leq \eta < 1/p, \quad b_\Omega \in W^{1+\eta,p}_{\mathrm{loc}}(\mathbf{R}^N),$$

and m can range from 1 to 2. It is convenient to introduce a special extension of b_Ω by zero outside of $U_h(\Gamma)$ to work in $\mathbf{R}^N$.

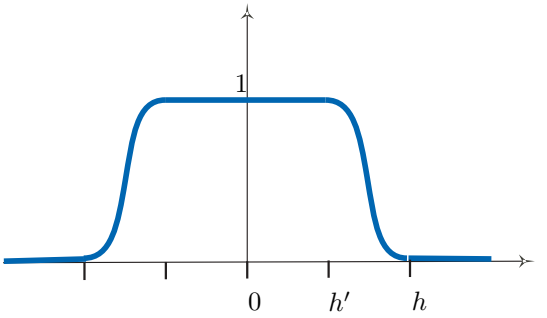


FIG. 3.1. *The function ρ_h with parameters (h',h), $0 < h' < h$.*

DEFINITION 3.3. *Given $h > 0$ and a subset Ω of $\mathbf{R}^N$ with nonempty boundary Γ, let $\rho_h \in \mathcal{D}(\,]-h,h[\,)$ be a nonnegative function which is equal to 1 in a neighborhood $V = \,]-h',h'[$, $0 < h' < h$, of 0. Define the* smooth h-extensions of b_Ω and 1 by zero:

$$b^h_\Omega \overset{\text{def}}{=} \rho_h \circ b_\Omega\, b_\Omega, \quad e^h_\Omega \overset{\text{def}}{=} \rho_h \circ b_\Omega + b_\Omega \rho'_h \circ b_\Omega. \tag{3.3}$$

It is readily seen that $b^h_\Omega = b_\Omega$ and $e^h_\Omega = 1$ in the tubular neighborhood $b_\Omega^{-1}(V) = U_{h'}(\Gamma) \subset U_h(\Gamma)$ of Γ. By construction $e^h_\Omega \in C^{0,1}_0(U_h(\Gamma))$. Moreover, the extension b^h_Ω preserves the smoothness properties of b_Ω in $U_h(\Gamma)$ and e^h_Ω can be viewed as an extension of 1 by zero outside $U_h(\Gamma)$ with the same smoothness as b_Ω in $U_h(\Gamma)$. By construction

$$\nabla b^h_\Omega = [\rho_h \circ b_\Omega + b_\Omega\, \rho'_h \circ b_\Omega]\, \nabla b_\Omega = e^h_\Omega \nabla b_\Omega. \tag{3.4}$$

If there exist $p \geq 1$ and $h > 0$ such that $\Delta b_\Omega \in L^p_{\mathrm{loc}}(U_h(\Gamma))$, then

$$\Delta b^h_\Omega = e^h_\Omega \Delta b_\Omega + \nabla e^h_\Omega \cdot \nabla b_\Omega \in L^p_{\mathrm{loc}}(\mathbf{R}^N) \quad (L^p(\mathbf{R}^N) \text{ if } \Gamma \text{ is bounded}).$$

Therefore by elliptic regularity

$$\begin{gathered} b^h_\Omega \in W^{2,p}_{\mathrm{loc}}(\mathbf{R}^N) \quad (W^{2,p}(\mathbf{R}^N) \text{ if } \Gamma \text{ is bounded}) \\ \Rightarrow\ \forall h',\ 0 < h' < h, \quad b_\Omega \in W^{2,p}_{\mathrm{loc}}(U_{h'}(\Gamma)) \quad (W^{2,p}(U_{h'}(\Gamma)) \text{ if } \Gamma \text{ is bounded}) \\ \Rightarrow\ b_\Omega \in W^{2,p}_{\mathrm{loc}}(U_h(\Gamma)). \end{gathered}$$

THEOREM 3.5. *Given an integer $N \geq 1$, let Ω be a subset of $\mathbf{R}^{\mathrm{N}}$, $\varnothing \neq \Gamma \neq \mathbf{R}^{\mathrm{N}}$.*
(i) *If there exist $p \geq 1$ and $h > 0$ such that $\Delta b_\Omega \in L^p_{\mathrm{loc}}(U_h(\Gamma))$, then*

$$b^h_\Omega \in W^{2,p}_{\mathrm{loc}}(\mathbf{R}^{\mathrm{N}}) \text{ and } b_\Omega \in W^{2,p}_{\mathrm{loc}}(U_h(\Gamma)) \tag{3.5}$$

and $\mathrm{m}_N(\Gamma) = 0$.[14] *The gradient ∇b_Ω exists in all points of $U_h(\Gamma)\backslash\Gamma$ and $|\nabla b_\Omega| = 1$. If Γ is compact*

$$b^h_\Omega \in W^{2,p}_0(\mathbf{R}^{\mathrm{N}}) \text{ and } \quad \forall h',\, 0 < h' < h, \quad b_\Omega \in W^{2,p}(U_{h'}(\Gamma)). \tag{3.6}$$

(ii) *If, in addition to the assumptions of part* (i), *$p > N$, then Ω is a Hölderian set of class $C^{1,1-N/p}$ and $b_\Omega \in C^{1,1-N/p}_{\mathrm{loc}}(U_h(\Gamma))$.*

Proof. For convenience we write b for b_Ω. (i) The proof is from the discussion preceding the theorem and [11, Chap. 5, Thm. 6.5, p. 235]. To show that $\mathrm{m}_N(\Gamma) = 0$, recall that $\nabla b \in [W^{1,p}_{\mathrm{loc}}(U_h(\Gamma)) \cap L^\infty_{\mathrm{loc}}(U_h(\Gamma))]^N$. Therefore, for each $x \in \Gamma$, $|\nabla b|^2 \in W^{1,p}(B(x,h)) \cap L^\infty(B(x,h))$ and $\nabla|\nabla b|^2 = 2\, D^2 b\, \nabla b$. But we have shown that ∇b exists and $|\nabla b|^2 = 1$ in $U_h(\Gamma)\backslash\Gamma$. This implies that $D^2 b \nabla b = 0$ in $U_h(\Gamma)\backslash\Gamma$. Moreover, we know that $\nabla b = 0$ a.e. in Γ. Therefore $\nabla|\nabla b|^2 = 2\, D^2 b\, \nabla b = 0$ a.e. in $U_h(\Gamma)$ and $|\nabla b|^2 \in W^{1,p}(B(x,h)) \cap L^\infty(B(x,h))$. Hence $|\nabla b|^2$ is constant in each connected component of $U_h(\Gamma)$ and the points of Γ can be divided into two categories,

$$\Gamma_0 = \{x \in \Gamma \,:\, B(x,h) \subset \Gamma\} \text{ and } \Gamma_1 = \{x \in \Gamma \,:\, B(x,h)\backslash\Gamma \neq \varnothing\}$$

$$\Rightarrow |\nabla b(x)|^2 = \begin{cases} 0, & x \in U_h(\Gamma_0) \\ 1, & x \in U_h(\Gamma_1) \end{cases} \Rightarrow U_h(\Gamma_0) \cap U_h(\Gamma_1) = \varnothing \text{ and } \Gamma_0 \cap \Gamma_1 = \varnothing,$$

and Γ splits into two closed disjoint subsets. Therefore $U_h(\Gamma_0) \cap \Gamma_1 = \varnothing$ and $U_h(\Gamma_0) \subset \Gamma$ necessarily imply $U_h(\Gamma_0) = \Gamma_0$. Thus Γ_0 is both open and closed. But this can only happen when Γ_0 is $\mathbf{R}^{\mathrm{N}}$ or $\varnothing$. By assumption, the first case cannot happen since $\mathbf{R}^{\mathrm{N}} = \Gamma_0 \subset \Gamma$. Therefore $\Gamma_0 = \varnothing$ and $|\nabla b|^2 = 1$ in $U_h(\Gamma)$. Finally $\chi_\Gamma = 1 - |\nabla b| = 1 - |\nabla b|^2 = 0$ and $\mathrm{m}_N(\Gamma) = 0$.

(ii) The proof is from [11, Chap. 5, Thm. 6.5 (i), p. 235]. □

We have the following new compactness theorem which relaxes the boundedness condition (3.7) from $D^2 b_{\Omega_n}$ in [11] to Δb_{Ω_n}.

THEOREM 3.6. *Let $D \subset \mathbf{R}^{\mathrm{N}}$ be nonempty bounded and open and $\{\Omega_n\}$, $\Gamma_n \neq \varnothing$, be a sequence of subsets of $\overline{D}$. Assume that there exist p, $1 \leq p < \infty$, h, and c such that*

$$\forall n,\ \|\Delta b_{\Omega_n}\|_{L^p(U_h(\Gamma_n))} \leq c. \tag{3.7}$$

Then there exist a subsequence $\{\Omega_{n_k}\}$ and a subset Ω, $\Gamma \neq \varnothing$, of $\overline{D}$ such that for all h', $0 < h' < h$, $b_\Omega \in W^{2,p}(U_{h'}(\Gamma))$ and for all $\bar{p}$, $1 \leq \bar{p} < \infty$,

$$b_{\Omega_{n_k}} \to b_\Omega \text{ in } W^{1,\bar{p}}(U_h(D))\text{-strong}. \tag{3.8}$$

[14] In order to further characterize a bounded $W^{2,p}$-domain, one could try to find the conditions on Γ under which $\nabla b^2_\Omega \in W^{1,p}_0(\Omega)^N$ and $b^{-1}_\Omega \nabla b^2_\Omega \in W^{1,p}(\Omega)^N$. This is the vectorial version of the problem of finding the conditions on Γ such that $W^{1,p}_0(\Omega) = \{f \in W^{1,p}(\Omega) : f\, d^{-1}_\Gamma \in L^p(\Omega)\}$ holds for some p, $1 \leq p < \infty$. The general problem of finding the conditions on Γ under which $W^{k,p}_0(\Omega) = \{f \in W^{k,p}(\Omega) : f\, d^{-k}_\Gamma \in L^p(\Omega)\}$, $1 \leq p < \infty$, $k = 1, 2, \ldots$ (especially $k = 1$), holds is studied in [17].

Moreover,[15] *for* $p > 1$, $\|\Delta b_\Omega\|_{L^p(U_h(\Gamma))} \le c$ *and for all* h', $0 < h' < h$,

$$D^2 b_{\Omega_{n_k}} \chi_{U_{h'}(\Gamma_{n_k})} \rightharpoonup D^2 b_\Omega \chi_{U_{h'}(\Gamma_n)} \text{ in } L^p(U_{h'}(\Gamma))\text{-weak} \tag{3.9}$$

and $\|D^2 b_\Omega\|_{L^p(U_{h'}(\Gamma))} \le c'$ *for some constant* c'.

Proof. (i) Since D is bounded, there exists $b_\Omega \in C_b^0(D)$ and a subsequence, still indexed by n, such that $b_{\Omega_n} \to b_\Omega$ in $C(\overline{U_h(D)})$. For convenience denote b_{Ω_n} and b_Ω by b_n and b. Hence for all p', $1 \le p' < \infty$, $b_n \rightharpoonup b$ in $W^{1,p'}(U_h(D))$-weak and $b_n^h \rightharpoonup b^h$ in $W_0^{1,p'}(\mathbf{R}^N)$-weak. By assumption from Theorem 3.5 (i), $b_n^h \in W_0^{2,p}(U_h(D))$ and

$$\Delta b_n^h = e_n^h \Delta b_n + \nabla e_n^h \cdot \nabla b_n = e_n^h \Delta b_n + (e_n^h)' \quad \Rightarrow \|\Delta b_n^h\|_{L^p(U_h(D))} \le c,$$

where $(e_n^h)' = 2\rho_h' \circ b_n + b_n\, \rho_h'' \circ b_n$. By equivalence of norms in $W_0^{2,p}(U_h(D))$, there exists another constant c' such that

$$\|D^2 b_{\Omega_n}^h\|_{L^p(U_h(D))} \le c' \quad \Rightarrow \forall h',\, 0 < h' < h, \quad \|D^2 b_{\Omega_n}\|_{L^p(U_{h'}(\Gamma_n))} \le c'.$$

From the last condition on the Hessian matrices, the assumptions of Theorem 9.2 [11, Chap. 5, p. 250] for sets of locally bounded curvature are satisfied and we get the compactness in the $W^{1,p}$-topology.

(ii) It remains to check condition (3.9). For all p, $1 \le p < \infty$, $b_{\Omega_n} \to b_\Omega$ in $W^{1,p}(U_h(D))$-strong and for each $\Phi \in \mathcal{D}^1(U_{h'}(\Gamma))^{N\times N}$,

$$\lim_{n\to\infty} \int_{U_{h'}(\Gamma)} \nabla b_{\Omega_n} \cdot \overrightarrow{\operatorname{div}}\, \Phi \, dx = \int_{U_{h'}(\Gamma)} \nabla b_\Omega \cdot \overrightarrow{\operatorname{div}}\, \Phi \, dx.$$

Each such Φ has compact support in $U_{h'}(\Gamma)$, and

$$\exists \varepsilon = \varepsilon(\Phi) > 0,\ 0 < 3\varepsilon < h, \text{ such that } \operatorname{supp}\Phi \subset U_{h'-2\varepsilon}(\Gamma).$$

Moreover, there exists $N(\varepsilon) > 0$ such that

$$\forall n \ge N(\varepsilon), \quad U_{h'-2\,\varepsilon}(\Gamma_n) \subset U_{h'-\varepsilon}(\Gamma) \subset U_{h'}(\Gamma_n)$$

(cf. part (i) of the proof of [11, Chap. 5, Thm. 9.2, pp. 251–252]. For $n \ge N(\varepsilon)$ consider the integral

$$\int_{U_{h'}(\Gamma)} \nabla b_{\Omega_n} \cdot \overrightarrow{\operatorname{div}}\, \Phi \, dx = \int_{U_{h'-2\varepsilon}(\Gamma)} \nabla b_{\Omega_n} \cdot \overrightarrow{\operatorname{div}}\, \Phi \, dx = \int_{U_{h'}(\Gamma_n)} \nabla b_{\Omega_n} \cdot \overrightarrow{\operatorname{div}}\, \Phi \, dx.$$

By assumption $D^2 b_{\Omega_n} \in L^p(U_{h'}(\Gamma))^{N\times N}$ and its norm is bounded by c'

$$\left| \int_{U_{h'}(\Gamma_n)} \nabla b_{\Omega_n} \cdot \overrightarrow{\operatorname{div}}\, \Phi \, dx \right| \le \|D^2 b_{\Omega_n}\|_{L^p(U_{h'}(\Gamma_n))} \|\Phi\|_{L^q(U_{h'}(\Gamma_n))} \le c\, \|\Phi\|_{L^q(U_{h'-2\varepsilon}(\Gamma))}$$

$$\Rightarrow \left| \int_{U_{h'}(\Gamma)} \nabla b_{\Omega_n} \cdot \overrightarrow{\operatorname{div}}\, \Phi \, dx \right| \le c\, \|\Phi\|_{L^q(U_{h'}(\Gamma))},$$

where $q^{-1} + p^{-1} = 1$. By convergence of ∇b_{Ω_n} to ∇b_Ω in the space $L^p(U_{h'}(D))^N$, then for all $\Phi \in \mathcal{D}^1(U_{h'}(\Gamma))^{N\times N}$

$$\left| \int_{U_{h'}(\Gamma)} \nabla b_\Omega \cdot \overrightarrow{\operatorname{div}}\, \Phi \, dx \right| \le c' \|\Phi\|_{L^q(U_{h'}(\Gamma))} \quad \Rightarrow \|D^2 b_\Omega\|_{L^p(U_{h'}(\Gamma))} \le c.$$

For $p > 1$ the weak L^p-convergence now follows by the same argument as in the proof of Theorem 3.5 (ii). This concludes the proof. ∎

[15] For $p = 1$ we get the convergence of $D^2 b_{\Omega_{n_k}}$ as a matrix of bounded measures.

4. Shape semiderivatives and application to b_Ω. In this section the elements of the velocity method and the notion of Eulerian semiderivative are briefly reviewed (cf., for instance, Chapter 8 of [11]) and applied to the computation of the semiderivative of $b_\Omega(x)$ which will be used in section 5.

4.1. Velocity method and shape semiderivative. In shape analysis the derivative of an objective function with respect to a set is obtained by generating perturbations of the set via a nonautonomous velocity field $V : [0,\tau] \times \mathbf{R}^N \to \mathbf{R}^N$, $0 < \tau < \infty$, verifying the conditions

$$\begin{aligned} &\forall x \in \mathbf{R}^N, \quad V(\cdot,x) \in C\big([0,\tau];\mathbf{R}^N\big), \\ &\exists c > 0,\ \forall x,y \in \mathbf{R}^N, \quad \|V(\cdot,y) - V(\cdot,x)\|_{C([0,\tau];\mathbf{R}^N)} \le c|y-x|, \end{aligned} \tag{4.1}$$

where $V(\cdot,x)$ is the function $t \mapsto V(t,x)$. The *parameter* t can be viewed as an artificial time. A point X is moved to the position $x(t) = x(t;X)$ via the differential equation

$$\frac{dx}{dt}(t) = V(t,x(t)), \quad 0 < t < \tau, \quad x(0) = X \in \mathbf{R}^N. \tag{4.2}$$

It will be convenient to define the velocity fields

$$x \mapsto V(t)(x) \overset{\text{def}}{=} V(t,x) : \mathbf{R}^N \to \mathbf{R}^N, \quad 0 \le t \le \tau. \tag{4.3}$$

This yields the families of transformations $\{T_t\}$ and perturbations $\{\Omega_t\}$

$$\forall t,\ 0 < t < \tau, \quad \left| \begin{aligned} &X \mapsto T_t(X) \overset{\text{def}}{=} x(t) = x(t;X) \\ &\forall \Omega \subset \mathbf{R}^N, \quad \Omega_t(V) \overset{\text{def}}{=} T_t(V)(\Omega). \end{aligned} \right. \tag{4.4}$$

THEOREM 4.1. *Given $\tau > 0$, assume that the map $V : [0,\tau] \times \mathbf{R}^N \to \mathbf{R}^N$ satisfy conditions* (4.1).

(i) *The transformation $T(t,x) = T_t(x)$ specified by* (4.2)–(4.4) *has the following properties:*

$$\begin{aligned} &\text{(T1)} && \forall X \in \mathbf{R}^N, \quad T(\cdot,X) \in C^1\big([0,\tau];\mathbf{R}^N\big) \text{ and } \exists c > 0, \\ &&& \forall X,Y \in \mathbf{R}^N, \quad \|T(\cdot,Y) - T(\cdot,X)\|_{C^1([0,\tau];\mathbf{R}^N)} \le c|Y-X|, \\ &\text{(T2)} && \forall t \in [0,\tau],\ X \mapsto T_t(X) : \mathbf{R}^N \to \mathbf{R}^N \text{ is bijective}, \\ &\text{(T3)} && \forall x \in \mathbf{R}^N, \quad T^{-1}(\cdot,x) \in C\big([0,\tau];\mathbf{R}^N\big) \text{ and } \exists c > 0, \\ &&& \forall x,y \in \mathbf{R}^N, \quad \|T^{-1}(\cdot,y) - T^{-1}(\cdot,x)\|_{C([0,\tau];\mathbf{R}^N)} \le c|y-x|, \end{aligned} \tag{4.5}$$

where $T^{-1}(t,y) = T_t^{-1}(y)$.

(ii) *Let Ω be a subset of $\mathbf{R}^N$. Then $\operatorname{int}\Omega_t = T_t(\operatorname{int}\Omega)$, $\Gamma_t = \Gamma_t = T_t(\Gamma)$, Γ_t is thin if and only if Γ is thin, $\overline{\Omega_t} = \Gamma_t$ if $\overline{\Omega} = \Gamma$, and Γ_t is of locally finite H_{N-1}-measure if and only if Γ is of locally finite H_{N-1}-measure.*

(iii) *If, in addition to conditions* (4.1),[16] *$V \in C([0,\tau];C^0(\overline{\mathbf{R}^N},\mathbf{R}^N))$, then for all bounded open subsets D of $\mathbf{R}^N$*

$$t \mapsto b_{\Omega_t} \text{ (resp., } b^2_{\Omega_t}) : [0,\tau] \to C(\overline{D}) \tag{4.6}$$

[16] $C^0(\overline{\mathbf{R}^N},\mathbf{R}^N)$ is the space of bounded uniformly continuous mappings from $\mathbf{R}^N$ to $\mathbf{R}^N$.

is continuous. If, in addition, Γ *is thin, then for all bounded open subsets* D *of* $\mathbf{R}^N$ *and all* p, $1 \le p < \infty$,

$$t \mapsto b_{\Omega_t} \text{ (resp., } b^2_{\Omega_t}) : [0,\tau] \to W^{1,p}(D) \tag{4.7}$$

is continuous.

(iv) *If* Ω *is of class* C^1 *and* $V \in C([0,\tau]; C^1(\overline{\mathbf{R}^N}, \mathbf{R}^N))$, *then* Ω_t *is of class* C^1 *for all* t *in a neighborhood of* $t = 0$. *If* Ω *is of class* $C^{1,1}$, $V \in C([0,\tau]; C^1(\overline{\mathbf{R}^N}, \mathbf{R}^N))$, *and there exists* $c > 0$ *such that for all* $t \in [0,\tau]$,

$$\forall x, y,\ |V(t,y) - V(t,x)| \le c|y-x| \text{ and } |DV(t,y) - DV(t,x)| \le c|y-x|,$$

then Ω_t *is of class* $C^{1,1}$ *for all* t *in a neighborhood of* $t = 0$.

Proof. (i) Cf. [11, Chap. 7, Thm. 4.1, p. 300]. (ii) From (T1)–(T2) the transformation T_t is bi-Lipschitzian. It transports interiors onto interiors, boundaries onto boundaries, and sets of zero measure onto sets of zero measure. Moreover, if $\overline{\Omega} = \Gamma$, then $\overline{\Omega_t} = T_t(\overline{\Omega}) = T_t(\Gamma) = \Gamma_t$.

(iii) Given $\varnothing \ne A \subset \mathbf{R}^N$, for any $t \in [0,\tau]$, $x \in \mathbf{R}^N$ and $y \in A$

$$|T_t(y) - x| \le |T_t(y) - y| + |y - x| \le \|T_t - I\|_{C(\mathbf{R}^N)} + |y - x|,$$
$$|y - x| \le |T_t(y) - y| + |T_t(y) - x| \le \|T_t - I\|_{C(\mathbf{R}^N)} + |T_t(y) - x|,$$

and by taking the infimum with respect to $y \in A$ on both sides

$$\left.\begin{array}{l} d_{T_t(A)}(x) \le \|T_t - I\|_{C(\mathbf{R}^N)} + d_A(x) \\ d_A(x) \le \|T_t - I\|_{C(\mathbf{R}^N)} + d_{T_t(A)}(x) \end{array}\right| \quad \Rightarrow \|d_{T_t(A)} - d_A\|_{C(\overline{D})} \le \|T_t - I\|_{C(\mathbf{R}^N)}.$$

Under conditions (4.1) and $V \in C([0,\tau]; C^0(\overline{\mathbf{R}^N}, \mathbf{R}^N))$, the map $f(t) = T_t - I$ belongs to $C^1([0,\tau]; C^0(\overline{\mathbf{R}^N}, \mathbf{R}^N)) \cap C^0([0,\tau]; C^{0,1}(\overline{\mathbf{R}^N}, \mathbf{R}^N))$ (cf. Theorem 4.3 in [11, p. 305]). Therefore $\|T_t - I\|_{C(\mathbf{R}^N)}$ goes to zero as t goes to 0 and we have the continuity. Of course the proof at $s = 0$ now extends to all $s \in [0,\tau]$. The continuity of b_{Ω_t} for $\Gamma \ne \varnothing$ now follows from the fact that $b_{\Omega_t} = d_{\Omega_t} - d_{\complement\Omega_t}$ and $\overline{T_t(\complement\Omega)} = \complement T_t(\Omega)$ for the homeomorphism T_t. By application of the result with A equal to Ω and $\complement\Omega$

$$\left.\begin{array}{l} d_{\Omega_t}(x) \le \|T_t - I\|_{C(\mathbf{R}^N)} + d_\Omega(x) \\ -d_{\complement\Omega_t}(x) \le \|T_t - I\|_{C(\mathbf{R}^N)} - d_{\complement\Omega}(x) \end{array}\right| \quad \Rightarrow b_{\Omega_t}(x) \le 2\|T_t - I\|_{C(\mathbf{R}^N)} + b_\Omega(x),$$

and we get the same inequalities by interchanging Ω_t and Ω. The continuity in $C(\overline{D})$ implies the continuity in $L^2(D)$ and since $|\nabla b_{\Omega_t}(x)|$ is a.e. bounded by 1, we get the weak continuity in $W^{1,2}(D)$. But we have seen in part (ii) that if $\mathrm{m}_N(\Gamma) = 0$, then $\mathrm{m}_N(\Gamma_t) = 0$. By Lemma 2.1 this implies the continuity in $W^{1,p}(D)$-strong for all finite $p \ge 1$.

(iv) By assumption on V, $t \mapsto T_t - I$ belong to $C([0,\tau]; C^1(\overline{\mathbf{R}^N}, \mathbf{R}^N))$ and $t \mapsto T_t^{-1} - I$ to $C([0,\tau']; C^1(\overline{\mathbf{R}^N}, \mathbf{R}^N))$ for some $0 < \tau' \le \tau$ (cf. [11, Thm. 4.5, p. 312]). Hence $\Omega_t = T_t(\Omega)$ is of class C^1 for a C^1-domain Ω and $t \in [0,\tau']$. Similarly the $C^{1,1}$ case follows from [11, Thm. 4.3, p. 305]. □

When the Ω's are subsets of a smooth open hold-all D, it is sufficient to work with velocity fields $V(t) : D \to \mathbf{R}^N$ such that $V(t) \cdot n_{\partial D} = 0$ on ∂D.

DEFINITION 4.1. *Given a* shape function $f(\Omega)$ *defined on the subsets* Ω *of* $\mathbf{R}^N$ *or* D, *we say that* f *has a* Eulerian semiderivative *at* Ω *in the direction* V *if the following limit exists*

$$df(\Omega; V) \stackrel{\text{def}}{=} \lim_{t \searrow 0} \frac{f(\Omega_t(V)) - f(\Omega))}{t}.$$

4.2. Partial derivative of b_Ω. In this section we compute the partial derivative of b_Ω which will be used, in the next subsection, to construct a new evolution equation for b_{Ω_t} when Ω has a thin boundary. First, introduce the notation

$$b'_\Omega(x) \overset{\text{def}}{=} \frac{\partial}{\partial t} b_{\Omega_t}(x)\bigg|_{t=0^+} \text{ and } (b^2_\Omega)' \overset{\text{def}}{=} \frac{\partial}{\partial t} b^2_{\Omega_t}(x)\bigg|_{t=0^+}.$$

THEOREM 4.2. *Let $\tau > 0$ and $V : [0,\tau] \times \mathbf{R}^N \to \mathbf{R}^N$ satisfying conditions (4.1) be given.*

(i) *Let Ω be a subset of $\mathbf{R}^N$ such that $\Gamma \neq \varnothing$. Then*

$$\begin{aligned} (b^2_\Omega)' &= -\nabla b^2_\Omega \cdot (V(0) \circ p_\Gamma) \text{ in } \mathbf{R}^N \setminus \mathrm{Sk}(\Omega), \\ b'_\Omega &= -\nabla b_\Omega \cdot (V(0) \circ p_\Gamma) \text{ in } \mathbf{R}^N \setminus (\Gamma \cup \mathrm{Sk}(\Omega)), \end{aligned} \tag{4.8}$$

and those identities are, respectively, verified a.e. in $\mathbf{R}^N$ and $\mathbf{R}^N \setminus \Gamma$. If, in addition, Γ has positive reach greater or equal to h, the above identities are, respectively, verified in $U_h(\Gamma)$ and $U_h(\Gamma) \setminus \Gamma$.

(ii) *Let Ω be a subset of $\mathbf{R}^N$ such that $\Gamma \neq \varnothing$ and $\mathrm{m}_N(\Gamma) = 0$. Then*

$$b'_\Omega = -\nabla b_\Omega \cdot (V(0) \circ p_\Gamma) \text{ a.e. in } \mathbf{R}^N. \tag{4.9}$$

If, in addition, $V \in C([0,\tau]; C^0(\overline{\mathbf{R}^N}, \mathbf{R}^N))$, for all bounded open subset D of $\mathbf{R}^N$, $b'_\Omega \in L^\infty(D)$ and

$$\forall \phi \in \mathcal{D}(\mathbf{R}^N), \quad \frac{d}{dt} \int_{\mathbf{R}^N} b_{\Omega_t} \phi \, dx\bigg|_{t=0^+} = \int_{\mathbf{R}^N} b'_\Omega \phi \, dx. \tag{4.10}$$

Remark 4.1. This theorem extends the earlier result of [15] and [16, Lem. 3.3, p. 248] from domains Ω of class C^2 to *arbitrary sets* with a thin boundary.

Proof. (i) First, compute the partial derivative of $b^2_{\Omega_t}(x)$,

$$\frac{\partial}{\partial t} b^2_{\Omega_t}(x)\bigg|_{t=0} = \frac{\partial}{\partial t} d^2_{\Gamma_t}(x)\bigg|_{t=0}, \quad d^2_{\Gamma_t}(x) = \min_{p_t \in \Gamma_t} |p_t - x|^2 = \min_{p \in \Gamma} |T_t(p) - x|^2.$$

Using the theorem on the derivative through a minimum (cf. [11, Chap. 9, Sect. 2.3, and Thm. 3.1, p. 164]) we get

$$\frac{\partial}{\partial t} d^2_{\Gamma_t}(x)\bigg|_{t=0} = \min_{p \in \Pi_\Gamma(x)} 2\,(p - x) \cdot \frac{\partial}{\partial t} T_t(p)\bigg|_{t=0} = \min_{p \in \Pi_\Gamma(x)} 2\,(p - x) \cdot V(0, p).$$

For $x \notin \mathrm{Sk}(\Omega)$ the semiderivative is linear with respect to $V(0)$ and the projection $p(x)$ of x onto Γ is unique and given by (2.9). Hence

$$\frac{\partial}{\partial t} b^2_{\Omega_t}(x)\bigg|_{t=0} = 2\,(p(x) - x) \cdot V(0, p(x)) = -\nabla b^2_\Omega(x) \cdot V(0, p(x)).$$

As for the second identity, for all $x \notin \Gamma \cup \mathrm{Sk}(\Omega)$, $b_\Omega(x) \neq 0$ and

$$\begin{aligned} \frac{\partial}{\partial t} b^2_{\Omega_t}(x)\bigg|_{t=0} &= 2\, b_{\Omega_t}(x)\, \frac{\partial}{\partial t} b_{\Omega_t}(x)\bigg|_{t=0} = 2\, b_\Omega(x)\, b'(x) \\ \Rightarrow\ b'(x) = \frac{\partial}{\partial t} b_{\Omega_t}(x)\bigg|_{t=0} &= \frac{1}{2 b_\Omega(x)}\, \frac{\partial}{\partial t} b^2_{\Omega_t}(x)\bigg|_{t=0} \ \Rightarrow\ b' = -\nabla b_\Omega \cdot (V(0) \circ p). \end{aligned}$$

Finally, since $\mathrm{Sk}(\Omega)$ has zero measure, the identities (4.8) are verified a.e. in $\mathbf{R}^{\mathrm{N}}$ and $\mathbf{R}^{\mathrm{N}}\setminus\Gamma$. If Γ has positive reach greater or equal to h, $U_h(\Gamma)\cap\mathrm{Sk}(\Omega)=\varnothing$ and (4.8) are verified everywhere in $U_h(\Gamma)$ and $U_h(\Gamma)\setminus\Gamma$.

(ii) From part (i) since $\mathrm{m}_N(\mathrm{Sk}(\Omega)\cup\Gamma)=0$. The a.e. derivative b'_Ω is in fact the $L^1(D)$-limit for all bounded open D of the differential quotients

$$\delta_t b(x) \stackrel{\text{def}}{=} \frac{b_{\Omega_t}(x)-b_\Omega(x)}{t}$$

as t goes to zero by Lebesgue dominated convergence theorem. First, the right-hand side of (4.9) is measurable as the inner product of $\nabla b_\Omega\in L^\infty(D)^N$ and $V(0)\circ p_\Gamma$ which is the composition of the Lipschitz continuous mapping $V(0)$ and the projection p_Γ which is itself the gradient of the continuous convex function $f(x)=(|x|^2-b_\Omega^2(x))/2$. Therefore f is uniformly Lipschizian on D, $p_\Gamma\in L^\infty(D)^N$, and $b'_\Omega\in L^\infty(D)$. We have already shown in part (i) that for a set with thin boundary $\delta_t b(x)\to b'_\Omega(x)$ for almost all x. As for the dominance, from Theorem 4.1 (iii)

$$|\delta_t b(x)| = \left|\frac{b_{\Omega_t}(x)-b_\Omega(x)}{t}\right| \le \left\|\frac{T_t-I}{t}\right\|_{C(\mathbf{R}^{\mathrm{N}})}.$$

Under conditions (4.1) and $V\in C([0,\tau];C^0(\overline{\mathbf{R}^{\mathrm{N}}},\mathbf{R}^{\mathrm{N}}))$, the map $f(t)=T_t-I$ belongs to $C^1([0,\tau];C^0(\overline{\mathbf{R}^{\mathrm{N}}},\mathbf{R}^{\mathrm{N}}))\cap C^0([0,\tau];C^{0,1}(\overline{\mathbf{R}^{\mathrm{N}}},\mathbf{R}^{\mathrm{N}}))$ (cf. Theorem 4.3 in [11, p. 305]). Therefore $(T_t-I)/t\to f'(0)$ in $C^0(\overline{\mathbf{R}^{\mathrm{N}}},\mathbf{R}^{\mathrm{N}})$ and the conditions of the Lebesgue dominated convergence theorem are satisfied. In particular, this establishes the property (4.10). □

5. Evolution equations for b_{Ω_t} and p_{Γ_t}. In this section we construct the new nonlinear evolution equation (5.2) of the oriented distance function b_{Ω_t} as a function of t for sets Ω with a thin boundary, along with a nonlinear evolution equation for the projection p_{Γ_t}. It applies not only to the evolution of open domains with thin boundaries, but also to arbitrary sets of zero N-dimensional Lebesgue measure such as a cloud of points, curves, objects of arbitrary codimension larger or equal to one, or a mixture of such objects. Equation (5.2) is to be compared with the weak evolution equation of the characteristic function χ_{Ω_t}: for any $\varphi\in\mathcal{D}(\mathbf{R}^{\mathrm{N}})$,

$$-\frac{d}{dt}\int_{\mathbf{R}^{\mathrm{N}}}\chi_{\Omega_t}\,\varphi\,dx+\int_{\mathbf{R}^{\mathrm{N}}}\chi_{\Omega_t}\,\mathrm{div}(V(t)\,\varphi)\,dx=0,\quad \chi_{\Omega_0}=\chi_\Omega. \tag{5.1}$$

This last equation is useless for sets with zero N-dimensional measure since their characteristic function is zero a.e. In what follows, it will be useful to introduce the notation $b'_{\Omega_t}(x)=\partial b_{\Omega_t}/\partial t(x)$.

THEOREM 5.1. *Let $\tau>0$ and $V:[0,\tau]\times\mathbf{R}^{\mathrm{N}}\to\mathbf{R}^{\mathrm{N}}$ be a map which satisfies the conditions* (4.1) *and, in addition, such that $V\in C([0,\tau];C^0(\overline{\mathbf{R}^{\mathrm{N}}},\mathbf{R}^{\mathrm{N}}))$. Assume that Ω is a subset of $\mathbf{R}^{\mathrm{N}}$ with nonempty thin boundary Γ.*

(i) *For all bounded open subsets D of $\mathbf{R}^{\mathrm{N}}$ and all finite $r\ge 1$, the function $t\mapsto b_{\Omega_t}$ belongs to $C^1([0,\tau];L^r(D))\cap C^0([0,\tau];W^{1,r}(D))$ and it satisfies*

$$\begin{aligned}&\frac{\partial}{\partial t}b_{\Omega_t}+\nabla b_{\Omega_t}\cdot(V(t)\circ p_{\Gamma_t})=0\ \forall t,\ 0\le t\le\tau,\ \text{and a.e. in }\ \mathbf{R}^{\mathrm{N}},\\ &b_{\Omega_0}=b_\Omega,\end{aligned} \tag{5.2}$$

where p_{Γ_t} is the projection onto Γ_t,

$$p_{\Gamma_t}(x)=x-\frac{1}{2}\nabla b^2_{\Omega_t}(x)\ \text{a.e. in }\ \mathbf{R}^{\mathrm{N}}. \tag{5.3}$$

(ii) *For all bounded open subsets D of $\mathbf{R}^N$ and all finite $r \geq 1$, the function $t \mapsto p_{\Gamma_t} : [0,\tau] \to L^r(D)^N$ is continuous and solution of*

$$\begin{aligned} &\frac{d}{dt}\int_{\mathbf{R}^N} p_{\Gamma_t}\cdot\vec{\phi}\,dx + \int_{\mathbf{R}^N}(p_{\Gamma_t}-I)\cdot(V(t)\circ p_{\Gamma_t})\ \mathrm{div}\,\vec{\phi}\,dx = 0, \\ &p_{\Gamma_0} = p_\Gamma \end{aligned} \tag{5.4}$$

for all $\vec{\phi} \in \mathcal{D}(\mathbf{R}^N)^N$ and $t \in [0,\tau]$.

Example 5.1. Let $\Omega = \{X\}$. Clearly $\Omega_t = \{x(t)\}$, where $x(t)$ is the solution of the differential equation $dx(t)/dt = V(t,x(t))$, $x(0) = X$. For all $x \in \mathbf{R}^N$, $p_{\Gamma_t}(x) = x(t)$ is a vector which only depends on t. Hence $V(t, p_{\Gamma_t}(x)) = V(t,x(t))$ is also a vector which only depends on t, and (5.4) reduces to

$$\frac{\partial}{\partial t}p_{\Gamma_t} = V(t)\circ p_{\Gamma_t} \text{ in } \mathbf{R}^N, \quad p_{\Gamma_0} = p_\Gamma.$$

Moreover, $b_{\Omega_t}(x) = |x - x(t)|$, and

$$\frac{\partial}{\partial t}|x-x(t)| + V(t,x(t))\cdot\frac{x-x(t)}{|x-x(t)|} = 0, \quad b_\Omega(x) = |x-X|.$$

For each t the equation is verified everywhere in $\mathbf{R}^N$ except on Γ_t.

Remark 5.1. To our best knowledge, the nonlinear equations (5.2) for b_{Ω_t} and (5.4) for the projection p_{Γ_t}, and the specification of their spaces of solution are new for initial sets with thin boundary. Once the velocity field has been specified it can describe the evolution of submanifolds in $\mathbf{R}^N$ in the spirit of the project of De Giorgi as referred to in [2]. In this paper the existence of solution for all $t \in [0,\tau]$ has been established in a constructive way by studying directly the properties of b_{Ω_t} and p_{Γ_t} and showing that they satisfy the equations. Upon substitution of expression (5.3) for p_{Γ_t}, (5.2) can be viewed as a form of Hamilton–Jacobi equation, but it is not the classical one for motion driven by curvatures since the velocity $V(t)$ on Γ_t is not constrained to be carried by the normal to Γ_t and that normal is not even assumed to exist on Γ_t as illustrated in Example 5.1. Note that the evolution equation for $b^2_{\Omega_t} = d^2_{\Gamma_t}$ is given by

$$\begin{aligned} &\frac{\partial}{\partial t}b^2_{\Omega_t} + \nabla b^2_{\Omega_t}\cdot\left(V(t)\circ\left[I - \frac{1}{2}\nabla b^2_{\Omega_t}\right]\right) = 0\ \forall t,\ 0\leq t\leq\tau,\ \text{a.e. in } \mathbf{R}^N, \\ &b^2_{\Omega_0} = b^2_\Omega. \end{aligned} \tag{5.5}$$

It only involves first order derivatives.

Remark 5.2. The solution of (5.4) determines p_{Γ_t}. Alternatively the level sets Γ_t can be constructively determined, from Γ by solving the equation

$$\frac{dx}{dt}(t) = V(t,x(t)), \quad x(0) = X \in \Gamma.$$

From the knowledge of Γ_t the projections p_{Γ_t} can be determined in a constructive manner a.e. in $\mathbf{R}^N$ and for all $t \in [0,\tau]$. Once p_{Γ_t} is determined, the velocity field $\tilde{V}(t) = V(t)\circ p_{\Gamma_t}$ is completely specified and (5.2) becomes a linear equation in b_{Ω_t}. If Γ and the Γ_t's have positive reach, the projection p_{Γ_t} and a fortiori $\tilde{V}(t)$ are Lipschitzian mappings from $\mathbf{R}^N$ to $\mathbf{R}^N$. Hence we can associate with the solution of the equation

$$\frac{dx}{dt} = \tilde{V}(t,x(t)), \quad x(0) = X,$$

the transformation $\tilde{T}_t$ of $\mathbf{R}^N$, and it is easy to verify by classical arguments that the solution of (5.2) is $b_{\Omega_t} = b_\Omega \circ \tilde{T}_t^{-1}$. Potential extensions of (5.2)–(5.4) to more general velocity fields would probably require techniques from set-valued differential equations and mutational analysis [4, 5].

Remark 5.3. Equation (5.2) describes the evolution of the zero-level sets, $\Gamma_t = \{x \in \mathbf{R}^N : b_{\Omega_t}(x) = 0\}$, of the function $(t,x) \mapsto \phi(t,x) = b_{\Omega_t}(x(t)) : [0,\tau] \times \mathbf{R}^N \to \mathbf{R}^N$ since for all $x(t) \in \Gamma_t$ we have $\phi(t,x(t)) = b_{\Omega_t}(x(t)) = 0$. But the displacement of the points of $\mathbf{R}^N$ is governed by the equation

$$\frac{dx}{dt}(t) = V(t,x(t)), \quad x(0) = X,$$

for some velocity field $V(t)$, and we formally get

$$\begin{aligned} &\frac{\partial \phi}{\partial t}(t,x(t)) + \nabla\phi(t,x(t)) \cdot V(t,x(t)) = 0 \\ &\Rightarrow \frac{\partial \phi}{\partial t}(t) + \nabla\phi(t) \cdot V(t) = 0 \text{ on } \Gamma_t. \end{aligned} \tag{5.6}$$

It would be tempting to say that for each t the above equation is verified not only on Γ_t but in all of $\mathbf{R}^N$. Yet (5.2) says that outside of Γ_t the velocity $V(t)$ has to be modified to the velocity $V(t) \circ p_{\Gamma_t}$. Furthermore, for smooth embedded submanifolds Ω of codimension strictly greater than one, ∇b_{Ω_t} does not exist on Γ_t and, a fortiori, (5.6) has no meaning. So it is really fascinating that, at each t, b_{Ω_t} be effectively determined by (5.2) *outside* of Γ_t.

Remark 5.4. In the context of the *level set method* which corresponds to a velocity constrained to be carried by the normal ∇b_{Ω_t} to the front Γ_t with a scalar speed $F(t)$ which is a function of the curvatures, the concept of *extension velocities* was formalized by Adalsteinsson and Sethian [1] in 1999 to "serve several purposes: (1) to provide a way of building velocities for neighboring level sets in the case where the velocity is defined only on the front itself; (2) to provide a sub-grid resolution in some cases not present in the standard level set approach; (3) to provide a way to update an interface according to a given velocity field prescribed on the front in such a way that the signed distance function is maintained and the front is never re-initialized." Among all the choices of extension velocities, the one constructed by Malladi, Sethian, and Vemuri [30] in 1995 corresponds to our velocity $\tilde{V}(t) = V(t) \circ p_{\Gamma_t}$. Quoting from [1]: "In cases where there is no available choice for an extension velocity, one approach is to simply extrapolate; standing at each grid point, the value of the speed function at the closest point on the front is used as the extension velocity at that point. This is the approach used in [30]." The same construction involving p_{Γ_t} was obtained by Gomes and Faugeras [21] in 2000 with several numerical implementations. Theorem 5.1 establishes, in the context of an unconstrained velocity and by completely independent arguments, that this is indeed the *right theoretical choice*. It is also interesting to observe that when all the functions involved are smooth the Jacobian matrix $D\tilde{V}(t)$ verifies the condition

$$\begin{aligned} D\tilde{V}(t)\,\nabla b_{\Omega_t} &= D\left(V(t) \circ p_{\Gamma_t}\right) \nabla b_{\Omega_t} = DV(t) \circ p_{\Gamma_t} Dp_{\Gamma_t} \nabla b_{\Omega_t} \\ &= DV(t) \circ p_{\Gamma_t} [I - \nabla b_{\Omega_t}\, {}^*\nabla b_{\Omega_t} - b_{\Omega_t} D^2 b_{\Omega_t}] \nabla b_{\Omega_t} \\ &= DV(t) \circ p_{\Gamma_t} [\nabla b_{\Omega_t} - \nabla b_{\Omega_t} - b_{\Omega_t} D^2 b_{\Omega_t} \nabla b_{\Omega_t}] = 0 \end{aligned}$$

since $|\nabla b_{\Omega_t}|^2 = 1$ and $D^2 b_{\Omega_t} \nabla b_{\Omega_t} = 0$. This is the generalization of the equation $\nabla F_{ext} \cdot \nabla\phi = 0$ given in [1, 21] for the scalar speed F, $V = F\nabla\phi/|\nabla\phi|$, to build an extension velocity from an extension F_{ext} of the scalar speed F.

Remark 5.5. Another interesting observation is that the zero-extension $b^h_{\Omega_t} = \rho_h \circ b_{\Omega_t}\, b_{\Omega_t}$ introduced in Definition 3.3 satisfies the evolution equation (5.2),

$$\begin{aligned} &\frac{\partial}{\partial t} b^h_{\Omega_t} + \nabla b^h_{\Omega_t} \cdot (V(t) \circ p_{\Gamma_t}) = 0 \ \forall t,\ 0 \le t \le \tau, \text{ and a.e. in } \mathbf{R}^{\mathrm{N}}, \\ &b^h_{\Omega_0} = b^h_{\Omega}, \end{aligned} \tag{5.7}$$

since $\partial b^h_{\Omega_t}/\partial t = e^h_{\Omega_t} \partial b_{\Omega_t}/\partial t$ and $\nabla b^h_{\Omega_t} = e^h_{\Omega_t} \nabla b_{\Omega_t}$. As the choice of the thickness h and the function ρ^h is independent of Ω and t, this suggests a new type of *narrow-band* method in the level set context where the band $U_h(\Gamma_t)$ naturally evolves with time as the support of $b^h_{\Omega_t}$. By construction, $b^h_{\Omega_t}$ coincides with b_{Ω_t} on the smaller tubular neighborhood $U_{h'}(\Gamma_t) \subset U_h(\Gamma_t)$, $0 < h' < h$. The other part of the solution can be thrown away.

Proof. (i) Equation (5.2) at time t is obtained by the same arguments as the one at $t = 0$ since Γ_t is thin for each t. Note that, in view of the linearity of the expression with respect to V, we get a derivative for all t in $]0.\tau[$ and not just a derivative from the right. Under the additional assumption that $V \in C([0,\tau]; C^0(\overline{\mathbf{R}^{\mathrm{N}}}, \mathbf{R}^{\mathrm{N}}))$, from Theorem 4.1 (iii), b_{Ω_t} is continuous in $W^{1,p}(D)$-strong which means that $t \mapsto \nabla b_{\Omega_t}$ is continuous in $L^p(D)^N$-strong. Moreover,

$$\begin{aligned} &\|V(t) \circ p_{\Gamma_t} - V(0) \circ p_{\Gamma}\|_{L^q(D)} \\ &\le \|(V(t) - V(0)) \circ p_{\Gamma_t}\|_{L^q(D)} + \|V(0) \circ p_{\Gamma_t} - V(0) \circ p_{\Gamma}\|_{L^q(D)} \\ &\le \|(V(t) - V(0))\|_{C(\overline{D})}\, \mathrm{m}_N(D)^{1/q} + c\|p_{\Gamma_t} - p_{\Gamma}\|_{L^q(D)} \\ &\le \|(V(t) - V(0))\|_{C(\overline{D})}\, \mathrm{m}_N(D)^{1/q} + c/2\|\nabla b^2_{\Omega_t} - \nabla b^2_{\Omega}\|_{L^q(D)}. \end{aligned}$$

But the assumption $V \in C([0,\tau]; C^0(\overline{\mathbf{R}^{\mathrm{N}}}, \mathbf{R}^{\mathrm{N}}))$ implies that $\|(V(t) - V(0))\|_{C(\overline{D})}$ goes to zero as t goes to zero and from the continuity of $t \mapsto \nabla b_{\Omega_t}$ we get the continuity of $t \mapsto \nabla b^2_{\Omega_t}$ for $1 \le q < \infty$. Finally since $b'_{\Omega_t} = -\nabla b_{\Omega_t} \cdot (V(t) \circ p_{\Gamma_t})$ we get the continuity of b'_{Ω_t} in $L^r(D)$ for $r \ge 1$ by choosing $p = q = 2r$ in the previous estimates. In view of those properties, (5.2) becomes an equation with values in $L^r_{\mathrm{loc}}(\mathbf{R}^{\mathrm{N}})$ and the solution $t \mapsto b_{\Omega_t}$ belongs to $C^1([0,\tau]; L^r(D)) \cap C^0([0,\tau]; W^{1,p}(D))$ for all bounded open subset D of $\mathbf{R}^{\mathrm{N}}$.

(ii) (5.4) is obtained by similar arguments and the previous properties of $b^2_{\Omega_t}$. ∎

Acknowledgments. The authors would like to thank one of the referees for constructive comments which led to a significant revision and expansion of section 5 and Anne Bourlioux from the Université de Montréal for pointing out the references [30, 1] to extension velocities.

REFERENCES

[1] D. Adalsteinsson and J. A. Sethian, *The fast construction of extension velocities in level set methods*, J. Comput. Phys., 148 (1999), pp. 2–22.

[2] L. Ambrosio and C. Mantegazza, *Curvature and distance function from a manifold*, J. Geom. Anal., 8 (1998), pp. 723–748.

[3] D. H. Armitage and Ü. Kuran, *The convexity of a domain and the superharmonicity of the signed distance function*, Proc. Amer. Math. Soc., 93 (1985), pp. 598–600.

[4] J.-P. Aubin, *Mutational equations in metric spaces*, Set-Valued Anal., 1 (1993), pp. 3–46.

[5] J.-P. Aubin, *Mutational and Morphological Analysis, Tools for Shape Evolution and Morphogenesis*, Birkäuser, Boston, Basel, Berlin, 1999.

[6] D. Bucur and J.-P. Zolésio, *Free boundary problems and density perimeter*, J. Differential Equations, 126 (1996), pp. 224–243.

[7] V. I. Burenkov, *Regularized distance*, Trudy Moskov. Inst. Radiotehn. Élektron. i Avtomat. Vyp. 67 Mat. (1973), pp. 113–117, 152 (in Russian).
[8] M. C. Delfour, *Tangential differential calculus and functional analysis on a $C^{1,1}$ submanifold*, in Differential Geometric Methods in the Control of Partial Differential Equations, Contemp. Math. 268, R. Gulliver, W. Littman, and R. Triggiani, eds., AMS, Providence, RI, 2000, pp. 83–115.
[9] M. C. Delfour and J.-P. Zolésio, *Shape analysis via oriented distance functions*, J. Funct. Anal., 123 (1994), pp. 129–201.
[10] M. C. Delfour and J.-P. Zolésio, *Shape analysis via distance functions: Local theory*, in Boundaries, Interfaces, and Transitions, CRM Proc. Lecture Notes 13, M. C. Delfour, ed., AMS, Providence, RI, 1998, pp. 91–123.
[11] M. C. Delfour and J.-P. Zolésio, *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*, Adv. Des. Control 4, SIAM, Philadelphia, 2001.
[12] M. C. Delfour and J.-P. Zolésio, *Shape Identification via Metrics Constructed from the Oriented Distance Function*, CRM Report, July 2002, first revision May 9, 2003, Université de Montréal.
[13] M. C. Delfour and J.-P. Zolésio, *The New Family of Cracked Sets and the Image Segmentation Problem Revisited*, CRM Report, May 2003, Université de Montréal; Communications in Information and Systems, to appear.
[14] C. Dellacherie, *Ensembles analytiques, capacités, mesures de Hausdorff*, Springer-Verlag, Berlin, Heidelberg, New York, 1972.
[15] F. R. Desaint and J.-P. Zolésio, *Shape boundary derivative of tangential boundary value problems*, Z. Angew. Math. Mech., 76, suppl. 2 (1996), pp. 85–88.
[16] F. R. Desaint and J.-P. Zolésio, *Manifold derivative in the Laplace-Beltrami equation*, J. Funct. Anal., 151 (1997), pp. 234–269.
[17] D. E. Edmunds and R. M. Edmunds, *Distance functions and Orlicz-Sobolev spaces*, Canad. J. Math., 38 (1986), pp. 1181–1198.
[18] H. Federer, *Curvature measures*, Trans. Amer. Math. Soc., 93 (1959), pp. 418–491.
[19] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, New York, 1977.
[20] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, New York, 1983.
[21] J. Gomes and O. Faugeras, *Reconciling distance functions and level sets*, J. Visual Com. and Image Representation, 11 (2000), pp. 209–223.
[22] F. Hausdorff, *Grundzüge der Meugenlehre*, Leipzig, 1914; *Set Theory*, translation from the German 3rd edition (1937) into English by J. R. Aumann et al., Chelsea, New York, 1957.
[23] C. M. Hoffmann, F. P. Preparata, P. C. Kanellakis, and S. Micali, eds., *Advances in Computing Research: Issues in Robotics and Nonlinear Geometry*, JAI Press, Greenwich, CT, 1992.
[24] C. M. Hoffmann, *Algebraic and numerical techniques for offsets and blends*, in Computation of Curves and Surfaces, W. Dahmen et al., eds., Kluwer, Dordrecht, 1990, pp. 499–528.
[25] C. M. Hoffmann, *How to construct the skeleton of CSG objects*, in Computer-Aided Surface Geometry and Design (Bath, 1990), Inst. Math. Appl. Conf. Ser. New Ser. 48, A. Bowyer, ed., Oxford University Press, New York, 1994, pp. 421–437.
[26] R. B. Holmes, *Smoothness of certain metric projections on Hilbert space*, Trans. Amer. Math. Soc., 184 (1973), pp. 87–100.
[27] H. Ishii and P. Souganidis, *Generalized motion of noncompact hypersurfaces with velocity having arbitrary growth on the curvature tensor*, Tohoku Math. J. (2), 47 (1995), pp. 227–250.
[28] S. G. Krantz and H. R. Parks, *Distance to C^k hypersurfaces*, J. Differential Equations, 40 (1981), pp. 116–120.
[29] G. M. Lieberman, *Regularized distance and its applications*, Pacific J. Math., 117 (1985), pp. 329–352.
[30] R. Malladi, J. A. Sethian, and B. C. Vemuri, *Shape modeling with front propagation: A level set approach*, IEEE Trans. on Pattern Analysis and Machine Intelligence, 17 (1995), pp. 158–175.
[31] G. Matheron, *Examples of topological properties of skeletons*, in Image Analysis and Mathematical Morphology, J. Serra, ed., Academic Press, London, 1988, pp. 217–238.
[32] S. Osher and J. A. Sethian, *Front propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulation*, J. Comput. Phys., 79 (1988), pp. 12–49.
[33] M. J. Parker, *The fundamental function of the distance*, Bull. London Math. Soc., 19 (1987), pp. 337–342.

[34] M. J. Parker, *Convex sets and subharmonicity of the distance function*, Proc. Amer. Math. Soc., 103 (1988), pp. 503–506.
[35] J.-B. Poly and G. Raby, *Fonction distance et singularités*, Bull. Sci. Math. (2), 108 (1984), pp. 187–195.
[36] D. Pompéju, Ann. de Toulouse (2), 7 (1905).
[37] J. Serra, *Image Analysis and Mathematical Morphology*, English version revised by N. Cressie, Academic Press, London, 1984.
[38] J. Serra, *Hausdorff distances and interpolations*, in Mathematical Morphology and Its Applications to Image and Signal Processing (Amsterdam, 1998), Comput. Imaging Vision 12, H. J. A. M. Heijmans and J. B. T. M. Roerdink, eds., Kluwer, Dordrecht, 1998, pp. 107–114.
[39] J. Serrin, *The problem of Dirichlet for quasilinear elliptic differential equations with many independent variables*, Philos. Trans. Roy. Soc. London, Ser. A, 264 (1969), pp. 413–496.
[40] S. Stifter, *The Roider method: A method for static and dynamic collision detection*, in Advances in Computing Research: Issues in Robotics and Nonlinear Geometry, C. Hoffmann et al., eds., JAI Press, Greenwich, CT, 1992.
[41] L. Younes, *Computable elastic distances between shapes*, SIAM J. Appl. Math., 58 (1998), pp. 565–586.
[42] L. Younes, *Optimal matching between shapes via elastic deformations*, Image and Vision Computing, 17 (1999), pp. 381–389.